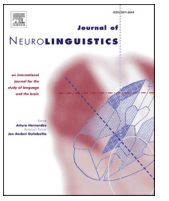
Journal of Neurolinguistics 73 (2025) 101229

Contents lists available at ScienceDirect

Journal of Neurolinguistics

journal homepage: www.elsevier.com/locate/jneuroling



Engineering conversation: Understanding the control requirements of language production in monologue and dialogue

Chiara Gambi [a,b,*], Fan Zhang [c], Martin J. Pickering [d]

[a] Department of Psychology, University of Warwick, CV4 7AL, UK
[b] School of Psychology, Cardiff University, CF10 3AT, UK
[c] Department of Engineering & Design, School of Engineering & Informatics, University of Sussex, BN1 9QJ, UK
[d] Department of Psychology, University of Edinburgh, EH8 9JZ, UK

ARTICLE INFO

Keywords:
Control theory
Forward models
Language production
Planning
Dialogue

ABSTRACT

Both artificial and biological systems are faced with the challenge of noisy and uncertain estimation of the state of the world, in contexts where feedback is often delayed. This challenge also applies to the processes of language production and comprehension, both when they take place in isolation (e.g., in monologue or solo reading) and when they are combined as is the case in dialogue. Crucially, we argue, dialogue brings with it some unique challenges. In this paper, we describe three such challenges within the general framework of control theory, drawing analogies to mechanical and biological systems where possible: (1) the need to distinguish between self- and other-generated utterances; (2) the need to adjust the amount of advance planning (i.e., the degree to which planning precedes articulation) flexibly to achieve timely turn-taking; (3) the need to track changing conversational goals. We show that message-to-sound models of language production (i.e., those that cover the whole process from message generation to articulation) tend to implement fairly simple control architectures. However, we argue that more sophisticated control architectures are necessary to build language production models that can account for both monologue and dialogue.

1. Why control theory?

Speaking and writing are motor actions and, like other motor actions (e.g., arm reaching movements), they can be fruitfully studied within the framework of control theory. In engineering, control theory is widely used to manipulate the operation of (artificial) dynamic systems to achieve a desired purpose. Importantly, control theory provides a framework for describing (in a general way) how these different systems behave and proposing strategies that are most suited to solve the identified challenges. In this paper, we apply this framework to models of language production that deal with stages prior to articulation, and argue that the control architectures instantiated by these models are not sophisticated enough to deal with dialogue. We therefore suggest alternative types of control architectures that could solve the unique challenges of language production in dialogue.

One type of control architecture that we refer to below is a closed-loop feedback system, where the output of the system is used to modify its operation (i.e., it is fed back into the system as a part of the input). A simple example is the domestic oven temperature control. In this example, the output (the oven's internal temperature) is measured and these measurements are fed back to a controller

* Corresponding author. Department of Psychology University Road University of Warwick CV4 7AL, Coventry, UK.
E-mail addresses: chiara.gambi@warwick.ac.uk (C. Gambi), fan.zhang@sussex.ac.uk (F. Zhang).

https://doi.org/10.1016/j.jneuroling.2024.101229
Received 30 August 2023; Received in revised form 25 July 2024; Accepted 16 August 2024
Available online 27 August 2024
0911-6044/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

that decides when and what corrective actions should be issued (i.e., switching the heating element on and off as required). The application of control theory principles ensures that such a system behaves appropriately, avoiding instability (i.e., unbounded changes in the system output over time), fast transient responses (the ability to respond quickly to a change), and accurate steady-state tracking (the ability to maintain low and near-constant deviations from the intended output once the system has reached a stable state).[1] To illustrate with the oven example, the average temperature of the oven can be maintained at the level set by the user (stability), in the event of a sudden heat loss (e.g., because the user has opened the door), the controller can heat up the oven quickly to maintain the temperature (fast transient response) and deviations from the intended temperature are close to zero (accurate steady state-tracking). The controller also handles the impact of external influences (as in the example above, a sudden temperature decrease due to the user opening the oven door), but importantly, in this simple closed-loop feedback control system, external influences are not directly modelled because it is sufficient to measure their effects on the measured output. This system can deal with small feedback delays while maintaining stability, but becomes unstable if the delay increases beyond a threshold.

However, much more sophisticated control systems are needed to deal with non-linearity, noise and external disturbances, longer delays, and so on. In fact, control theory provides a very broad framework within which to understand how any system should optimally use information about itself and its environment to guide the selection of future actions, given the inherent noise and time delays that characterise any physical system-environment interaction. Within this broad framework, control systems can implement a variety of different architectures and use a range of different algorithms to adjust the control strategy in order to achieve the optimal output.

Control theory principles do not just apply to mechanical systems, but have also been fruitfully applied to explain how living organisms sense and move in their environment (i.e., biological control; see McNamee & Wolpert, 2019 for a review). For example, Wolpert and colleagues showed that the human motor system maintains accurate estimates of the position of the arm using a Kalman filter, which optimally combines sensory feedback and sensory predictions based on an internal model of the arm's dynamics (Wolpert, Ghahramani, & Jordan, 1995). More recently, a type of architecture known as adaptive intermittent control was proposed to understand how people perform complex target-tracking tasks in real-time, despite their slow sensorimotor system (Sakaguchi, Tanaka, & Inoue, 2015). Finally, related control mechanisms have also been proposed as a framework for understanding how networks of neurons in the brain can learn any complex function reliably, efficiently, and robustly (Denève, Alemi, & Bourdoukan, 2017).

Importantly for our purposes, there are well-developed models of speech production and perception that incorporate control theory principles. In Section 2, we briefly review how these existing models deal with noise and delay. Our aim is not to provide a comprehensive review, but just to highlight key ideas that will be important in later sections of this paper. In particular, we will focus on one solution to the problem of recovering phonological representations when the auditory signal is noisy or degraded (Bayesian inference) and one solution to the problem of overcoming delays inherent in executing speech motor commands and monitoring one's own speech via the auditory system (coupled inverse and forward models).

In Section 3, we then review which control solutions have been incorporated so far in models of language production that deal with higher linguistic levels (i.e., from message to phonology). We show that the control architectures embedded in these models are generally much less sophisticated than the ones incorporated in models of speech production, and we review the theoretical arguments in support of such simpler architectures and against the inclusion of coupled inverse-forward models of the language production system.

Finally, in Section 4 we argue that current models can account well for language production in monologic contexts, but that more sophisticated control architectures are likely to be needed to account for language use in dialogic contexts. Specifically, we identify three challenges for models of language production that arise from the nature of dialogue: (1) the need to distinguish between self- and other-generated utterances; (2) the need to adjust the amount of advance planning (i.e., the degree to which planning precedes articulation) flexibly to achieve timely turn-taking; (3) the need to track changing conversational goals. For each of these challenges, we discuss which type of control architecture could be implemented, drawing on analogies to mechanical or biological systems whenever possible. We hope that drawing these analogies will foster the development of a novel class of language production models which can address such challenges.

2. Dealing with noisy and delayed sensory signals

When a living organism interacts with its environment, its estimation of the current state of the environment is inherently noisy. Moreover, any information it receives from the senses is somewhat delayed and so it refers to the state of the environment at some time in the past, rather at the present moment.

Bayesian inference provides an optimal solution to the problem of noisy input (McNamee & Wolpert, 2019): Bayes' rule specifies how to optimally combine beliefs about the current state of the environment (i.e., priors) with incoming sensory information (i.e., the data) to derive the posterior probability distribution over possible states of the environment. Decades of research on human vision (Kersten, Mamassian, & Yuille, 2004; Rao & Ballard, 1997; Summerfield & De Lange, 2014) and reasoning (Griffiths & Tenenbaum, 2009) show how this simple principle accounts for many aspects of our perception and decision-making; see Fig. 1 for an example from the visual modality.

Importantly, applications of Bayesian inference have been successful in explaining auditory perception as well. Indeed, an auditory

[1] Note that an unstable system (by definition) does not display the property of steady-state tracking; but not every stable system displays this property, as a system can be stable but fluctuate considerably above/below the intended output value over time.

Fig. 1. A (left): two-tone (or Mooney) image. B (right): original template for the two-tone image. When we observe a two-tone image such as the one in Fig. 1A (Teufel, Dakin, & Fletcher, 2018), the impoverished sensory information typically prevents us from recognising what it depicts (i.e., compare it to the original template image in Fig. 1B). But prior exposure to the template image results in a perceptual "pop-out" effect, allowing the observer to disambiguate the otherwise meaningless patches of black and white (Ludmer, Dudai, & Rubin, 2011). In Bayesian terms, this occurs because exposure to the template generates prior beliefs that "fill in" the missing sensory information in the two-tone image to bias perceptual inference towards the template (Chang, Baria, Flounders, & He, 2016).

equivalent to the visual "pop-out" effect (illustrated in Fig. 1) has been demonstrated for speech. When speech is degraded (e.g., through noise-vocoding), comprehensibility diminishes (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Mattys, Davis, Bradlow, & Scott, 2012). But first playing the original, non-degraded speech to listeners restores perception, despite the sensory information in the signal being the same (Sohoglu, Peelle, Carlyon, & Davis, 2014; see also Corps & Rabagliati, 2020).

Accordingly, Bayesian inference has been implemented in several models of speech perception (e.g., Feldman, Griffiths, & Morgan, 2009; Gagnepain, Henson, & Davis, 2012; Kleinschmidt & Jaeger, 2015; Kronrod, Coppess, & Feldman, 2016; Norris & McQueen, 2008; Skipper, 2014). Discussing the differences between these implementations (see e.g., Davis & Sohoglu, 2020; Mattys et al., 2012; Norris, McQueen, & Cutler, 2016 for discussion) is beyond the scope of this paper; here it is sufficient to note that they all use Bayesian principles to explain how listeners infer speech categories from noisy and variable sensory signals.

Bayesian inference has also been used to model speech production as part of a unified model of speech perception and production (COSMO; Moulin-Frier, Diard, Schwartz, & Bessière, 2015), but here we turn to a different (though related) approach to modelling production which focuses on solving the problem of delays in motor execution and sensory transmission: This approach involves pairs of coupled inverse and forward models (McNamee & Wolpert, 2019). In the context of biological control (e.g., of hand grasping movements), a forward model is an internal model of the motor system, which maps from motor commands to the expected sensory consequences of executing such commands (e.g., a prediction of the hand's maximal degree of aperture). Conversely, an inverse model instantiates the reverse mapping, from sensory consequences to the underlying motor commands (e.g., issued to the hand muscles). The inverse model can be used to select appropriate corrective motor commands if a deviation between intended and observed sensory consequences is detected, as well as more generally to select motor commands that will generate an intended sensory target.

How do coupled inverse-forward model pairs represent a solution to the problem of delays? Forward models allow an organism acting on its environment to predict the sensory consequences of its own movements; to do so, a copy of the motor command is sent as input to the forward model, which maps it onto expected sensory consequences. This is important because motor execution and sensory transmission introduce delays between motor commands being issued and their sensory consequences being observed. But forward model predictions are fast and therefore, when the movement is eventually executed and the actual sensory consequences observed, the latter can be immediately compared to the predicted sensory consequences – for example, the actual aperture of the hand can be compared to the predicted aperture. This comparison generates an error signal which can be fed through the inverse model to issue a prompt correction (Miall & Wolpert, 1996). Forward model predictions also allow the organism to distinguish between self-generated and externally-generated sensory consequences (essentially, because self-generated sensory consequences are linked to a motor command, and can be more accurately predicted).

Coupled inverse-forward model pairs are a key component of the architecture of models of speech motor control (Bohland, Bullock, & Guenther, 2010; Guenther, Ghosh, & Tourville, 2006; Hickok, 2012a; Kello & Plaut, 2004; Tourville & Guenther, 2011) where they deal with the complexity of learning the one-to-many mappings between movements of the articulators and speech sounds. In the Directions into Velocities of Articulators (DIVA) model (Guenther et al., 2006), for example, coupled inverse-forward mappings are learnt by the model during an initial period of vocal exploration (akin to the babbling phase in human infants). Once learnt, these mappings represent an internal model of the speech motor system, which can be used to maintain an estimate of the current state of the system and detect and respond to external perturbations. Indeed, if the auditory feedback to the speaker is systematically manipulated in real-time (e.g., consistently shifted in frequency), the speaker adapts by shifting their vocal production in the opposite direction to compensate for the disturbance (e.g., Houde & Jordan, 1998; Lametti, Nasir, & Ostry, 2012).

In sum, by modelling how sensory data are generated, the organism can identify and respond to perturbations in very specific ways and very quickly, despite significant and variable delays in motor execution and sensory transmission. Importantly, the control system necessary for online movement control is much more sophisticated than the simple oven temperature control example described in Section 1. This closed-loop control system is an example of state feedback control (Hickok, 2012b; Houde & Nagarajan, 2016), where an estimate of the current state of the system is fed back to the controller and guides the selection of future actions. Crucially, the current state of the speech motor system is not directly observable, and thus a model of this system (an "observer") is used to generate

predictions that can be compared to sensory feedback to derive an optimal estimate of the state of the system (Houde & Nagarajan, 2016).

3. Current models of language production, and their limitations

The previous section made brief reference to accounts that assume Bayesian inference explains key aspects of speech perception and accounts that assume coupled inverse-forward model pairs are implicated in speech motor control. But what about aspects of language production and comprehension beyond the level of sounds and syllables?

Language production begins with the selection of a message, and then the speaker chooses a structure to convey that message alongside lexical items corresponding to entities and events; event roles must be bound to specific lexical items and these must then be sequenced in the appropriate linear order given the chosen structure (Slevc, 2023). All of these stages typically occur before phonological encoding and phonetic realisation of individual lexical items (Bock & Levelt, 1994; Levelt, 1989), which are the stages that the models reviewed in section 2 deal with. Conversely, language comprehension involves much more than simply segmenting continuous speech into words – it involves recovering the underlying message using many sources of linguistic information (e.g., the preceding words and sentences, the intonation) and extra-linguistic information (e.g., the speaker's social characteristics, the visual context), and then updating a mental model of the situation under discussion in light of this inferred message (Glenberg, Meyer, & Lindem, 1987; Sanford & Garrod, 1998). Bayesian inference and coupled inverse-forward models have been applied to the understanding of these higher-level processes in language comprehension and production, respectively.

In comprehension, Bayesian inference approaches have been applied to the problem of recovering meaning from a noisy input (e.g., Gibson, Bergen, & Piantadosi, 2013) and also to recovering referential intentions (Frank, Goodman, & Tenenbaum, 2009; Goodman & Frank, 2016). In addition, there is a long tradition of information-theoretic approaches to language comprehension, which link processing time to surprisal – that is the negative log probability of encountering a word or structure in a given context (Hale, 2001; Levy, 2008). These accounts are related to Bayesian inference because surprisal is equivalent to the comparison between prior and posterior probability distributions (Jaeger & Snider, 2013). Finally, recent models of the N400 (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Fitz & Chang, 2019; Rabovsky, Hansen, & McClelland, 2018) propose that the amplitude of this fundamental event-related potential (ERP) component – which peaks around 400 ms after the onset of a word and is inversely proportional to the word's predictability in context (Kutas & Hillyard, 1984) – indexes the extent to which current comprehension representations are updated in light of new bottom-up information (though the models disagree on the nature of this updating process). In sum, models of language comprehension that account for how we build meaning incrementally have largely embraced the idea that comprehension representations are inferred probabilistically by combining prior knowledge with bottom-up noisy input. Thus, Bayesian inference can function as a unifying computational principle across linguistic levels in comprehension.

In contrast, in language production, models that deal with the stages from message formulation to lexical retrieval have traditionally been based on very different computational principles than models of the production of isolated sounds or syllables. Specifically, these models have implemented different control architectures. Control systems in language production have been discussed under the umbrella term of "monitoring" (see Gauvin & Hartsuiker, 2020 for a recent review). One long-standing account of monitoring in language production (Levelt, 1983, 1989) posits that the comprehension system is used to monitor the output of the production system. Another influential account – conflict-based monitoring – assumes that monitoring is performed within the language production system itself (Nozari, Dell, & Schwartz, 2011). Below, we briefly describe these accounts focusing on the key differences in terms of how control is implemented, and point out that both comprehension-based and production-based monitoring implement simpler control architectures compared to the more recent forward modelling account (Garrod & Pickering, 2015; Pickering & Garrod, 2013a, 2014, 2021), which proposes that monitoring is performed via inverse-forward model pairs at all levels of the linguistic hierarchy. We then discuss arguments for and against inverse-forward model pairs beyond the phonological level.

Comprehension-based monitoring (Levelt, 1983; Levelt, Roelofs, & Meyer, 1999) assumes that the comprehension system has access to the acoustic-phonetic output (external monitoring) and also to an internal representation of the phonological output (internal monitoring). In control-theoretic terms, this architecture corresponds to closed-loop feedback; specifically, since the monitor has access to information about the state of the system (i.e., production-based phonological representations), it is a form of state feedback control where an estimate of the current state of the system is fed back to the controller. Note that this state estimate is computed by a highly sophisticated system – the comprehension system, which functions as an "observer" of the production system.

In fact, this approach to monitoring effectively offloads the heavy computational work onto the comprehension system. The acoustic-phonetic signal may be noisy (e.g., due to environmental noise that needs to be filtered out) or incomplete (e.g., due to a degraded signal, or to lapses in attention). Furthermore, processing the acoustic-phonetic input takes time, leading to a delay between the generation of the intended message and the recovery of the underlying message from the signal. The internal monitor, which has direct access to phonological representations (i.e., to an earlier stage of the production process) was actually proposed to address these issues but recovering the underlying message from these phonological representations still requires some processing time which, in addition to the time required to generate the phonological representations themselves (up to 400 ms from the intention to speak; Indefrey, 2011; Indefrey & Levelt, 2004) introduces delay in the monitoring process.

In conflict-based monitoring, the limitations of comprehension-based monitoring are addressed by taking a completely different approach. The level of interference (or conflict) within the production system itself is used as a control signal, as it indexes the likelihood that an error will be produced (Nozari et al., 2011). For example, in a recent model (Gauvin & Hartsuiker, 2020), this control signal is used to trigger a domain-general conflict resolution response (arousal) that boosts the activation of all active production representations to facilitate selection of the correct response. Importantly, this type of control architecture makes use of information

about the state of the production system (as in state feedback control), but it assumes this information is immediately accessible to the monitor (i.e., there is no need for state estimation). Furthermore, the level of conflict acts as a "switch" that determines whether or not to trigger a generic arousal response. The latter is reminiscent of so-called "gain-scheduling", in which the value of a system variable is used to switch between different controller settings (e.g., aircraft altitude is used to select a different set of parameters for an automatic pilot, tailored to the different air density at high vs. low altitude). Note that gain-scheduling is a form of open-loop control because whether or not the "switch" occurs is entirely determined by a measured variable (i.e., the level of conflict within the production system) and there is no additional check on the outcome (e.g., whether the arousal response did actually result in selection of the correct response).

Note that in Gauvin and Hartsuiker's model, conflict can be detected both within the production system and within the perceptual system (accounting for both internal and external monitoring, respectively). Importantly, production and perception representations are separate but tightly linked: Whenever a production representation is activated, activation flows to the corresponding perception representation and vice versa, at every linguistic level. Thus, the monitor can avoid delays, because of the direct mapping between perception and production representations.

Finally, in the forward-modelling account of monitoring, the coupled inverse-forward model architecture of models of speech motor control is extended to all linguistic levels (Pickering & Garrod, 2013a). To avoid delays, the monitor learns to predict the sensory consequences of executing production commands (i.e., it learns forward models of the production and comprehension systems) and to recover the underlying production command from observed sensory effects (i.e., it learns corresponding inverse models). Importantly, forward models are computed more quickly than the corresponding production representations are activated within the production system. Moreover, the forward models predict comprehension representations, and these predicted representations can be directly compared to the observed comprehension representations at all levels (during self- or other-monitoring). An error can therefore be flagged up very quickly, because there is no need to wait for the underlying message to be recovered and the comparison to take place at the message level. If a mismatch is detected, the inverse models can be used to modify the production command appropriately and issue a correction.

However, several authors have argued against the need for coupled inverse-forward models at all levels of the linguistic hierarchy (Dell & Chang, 2014; Hartsuiker, 2013; Meyer & Hagoort, 2013) as proposed by Pickering and Garrod (2013a). There are two types of arguments. The first is that inverse-forward model pairs are not necessary at these levels. Some authors suggested that prediction could be performed by the production system itself, without the need for a separate forward model, which is seen as an unnecessary duplication of the production system. The two functions of forward and inverse models performed at the phonetic level are to reduce feedback delays and to map between motor-based and perceptual-based representations that are quite distinct. These functions may be redundant at higher linguistic levels, and particularly at the level of lexico-syntactic representations.

The rationale behind this first argument is as follows: (1) estimates suggest that production process up to lexical selection are comparatively fast (~275ms; Indefrey & Levelt, 2004) and (2) representations are generally assumed to be fully shared between production and comprehension at levels beyond phonology (Gambi & Pickering, 2017). To elaborate, conceptualization takes around 150–200ms, and lexical selection takes 70–90ms, which is considerably shorter than phonological and phonetic encoding (290–400ms). In addition, the activation flow within the language production system might not be serial as implied above, but rather proceed to a large extent in parallel, with phonological processing taking place at the same time as lexico-semantic processing (Strijkers, Costa, & Pulvermüller, 2017). Finally, as soon as the production representations become activated, they would activate corresponding representations within the comprehension system. In sum, there may not be a need for a forward prediction mechanism that can perform an even faster mapping between production and comprehension (Dell & Chang, 2014; Meyer & Hagoort, 2013). However, such a mechanism could still be useful if it could be deployed as soon as high-level conceptualization of the message to be conveyed begins (which could be in the order of seconds before retrieval of specific lexicalised concepts starts).

The second type of argument is that forward models cannot serve the function for which they have been proposed (Hartsuiker, 2013). For forward models to be faster than the flow of activation within the production system, they must encode some sort of approximate or impoverished representations (Gauvin & Hartsuiker, 2020). For example, Pickering and Garrod (2013a) suggested that a forward model at the lexico-syntactic level might correspond to the prediction that a noun will be selected for production, but without the detail about the full lexical entry. But others have noted that it is unclear how such impoverished representations could be useful in monitoring (Hartsuiker, 2013; Strijkers, Runnqvist, Costa, & Holcomb, 2013). At the phonetic level, it has been proposed that forward model predictions may represent average phoneme realisation so that deviations can be easily detected (Niziolek, Nagarajan, & Houde, 2013), but it is an open question how this idea of representing "average behaviour" could be applied to lexical entries. However, the specific information that is encoded in the forward models may depend on what is useful in the current context, for example the aspects of semantics or phonetics about which an error is most likely to occur (Pickering & Garrod, 2013b). To illustrate, it may be that speakers are sensitive to the fact that a semantic substitution error is very likely to occur when talking about sheep and goats within the same discourse (because of the similarity between these concepts) and the forward models are tuned to predict a specific lexical concept in this situation to catch such errors.

In reviewing these opposing arguments on the role of inverse-forward model pairs in language monitoring, our aim is not to adjudicate between them, but rather to show that the need for this kind of complex control architecture in models of language production is far from agreed upon. In contrast, many scholars have suggested that simpler control architectures are sufficient, at least at higher linguistic levels because of the nature of processes and representations that are hypothesized at those levels (i.e., timing, parity between output and input representations). But crucially all of these arguments relate to language production in a monologic context. In the next section, we argue that more complex control architectures are likely to be needed when dealing with language production in dialogue.

4. Why is monitoring more complex in dialogue contexts?

We argue that the requirements of control for language production in a dialogue context are more complex than in a monologic context because of three unique challenges: (1) distinguishing between self- and other-generated utterances; (2) the (comparative) slowness of language production in the context of fast conversational turn-taking (3) the changing nature of conversational goals. See Fig. 2 for a schematic representation of the three challenges and potential control solutions.

4.1. Distinguishing between self- and other-generated utterances

In section 2, we noted that one of the functions that forward models of motor commands serve in biological control is to differentiate between the sensory effects of internally-generated and externally-generated events. In the control of speech, the operation of this control mechanism is evidenced by larger auditory suppression during overt speech than during passive listening (Houde, Nagarajan, Sekihara, & Merzenich, 2002): The sensory consequences of moving our own articulators can be more accurately predicted and the auditory cortex's responses to them can be dampened, whereas other, less predictable sensory information activates the cortex more strongly. This mechanism can thus be used to detect any deviations from the expected acoustics of one's own speech movements which are due to external influences (e.g., a source of background noise, or experimental manipulation of the participant's jaw in some experiments). Thus, even when producing language in a monologue context, it is useful for the monitor to distinguish between self-produced speech sounds and sounds that have external causes.

In a dialogue context, the speaker must monitor not only their own utterances but also those produced by the interlocutor, either in order to prepare their next turn or to check the interlocutor's understanding of their current turn (e.g., are they signalling understanding via backchannels, such as *mmh* or *yes*?; Schegloff, 1982). This means that the comprehension system is used both to process what the other is saying and to check that the current utterance is being produced as intended. These two functions must often be performed concurrently, because of overlap between backchannels and the current speaker turn, and during instances of interruption of the current utterance by the interlocutor (Heldner & Edlund, 2010). It thus becomes crucial to be able to distinguish between comprehension representations that are internally-generated and those that are generated by the interlocutor's utterance, in order to select an appropriate response.

As an example, let us take a speaker who performs conflict-based monitoring as in Gauvin and Hartsuiker's (2020) model (see section 3). If this speaker is interrupted by an interlocutor producing overlapping speech, some comprehension representations will be activated bottom up while the speaker is activating production representations underlying her utterance. Because activation flows immediately from comprehension to production representations in the model, the speaker may detect high levels of conflict between lexical representations that are concurrently activated in this scenario.

However, this conflict would not be an indication that the speaker's own production system is about to generate an error in this instance. In this situation, the appropriate response may be to stop language production altogether in order to attend to the interlocutor's speech. In contrast, when conflict arises from multiple representations active in the production system, the appropriate response is to increase arousal to boost activation of all active representations, so that eventually the correct word will be produced.

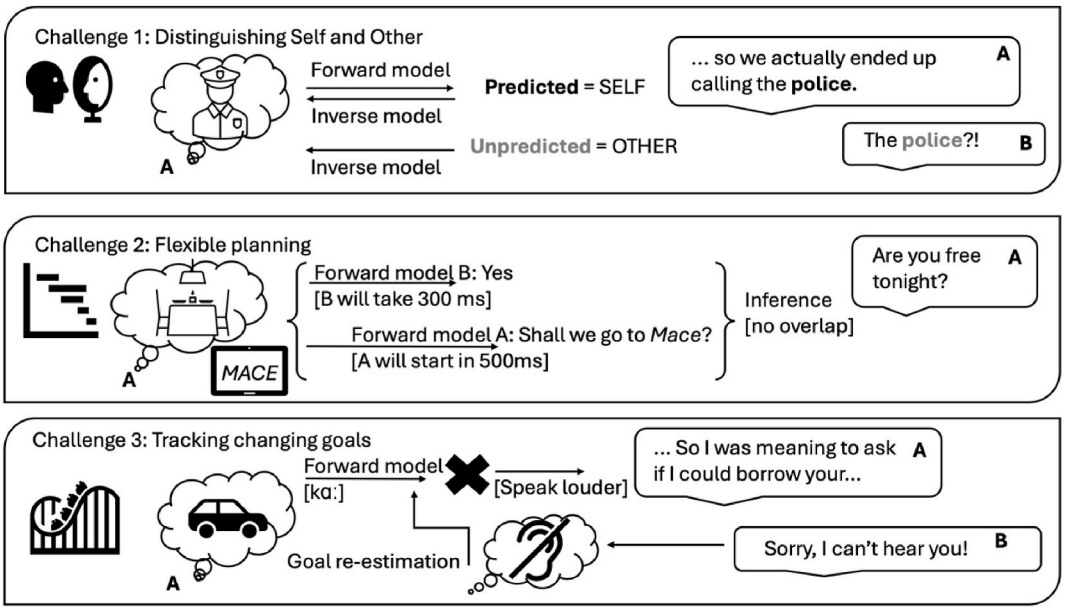


Fig. 2. The three control challenges of language production in dialogue, and three potential solutions. Note that for Challenge 2, only two of many possible forward model predictions are depicted for readability (but see Section 4.2 for alternative examples).

But in order to select the appropriate response, the system must be able to distinguish between self- and other-generated utterances.

As outlined above, forward models represent one type of control architecture that could allow the speaker to distinguish between self- and other-generated utterances. One could argue that a simpler solution to this problem exists: For example, it could be sufficient to "tag" any phonetic representations in the comprehension system as self- or other-generated (perhaps as a result of comparisons to forward model predictions at the phonetic level), and then simply pass this information upwards along the comprehension hierarchy, *without* extending the inverse-forward model architecture to higher linguistic levels. This is possible, though we note that one advantage of extending the inverse-forward model architecture to higher levels of linguistic representations is that it provides a straightforward mechanism by which information at higher levels could affect decisions about whether an utterance is self- or other-produced.

4.2. *Language production is slow but turn-taking is fast*

Chronometric estimates for language production have been reviewed in Section 3 (Indefrey & Levelt, 2004). Here we argue that, even if lexical selection might occur in parallel with phonological encoding to some extent (Strijkers et al., 2017), the delays involved in language production are still potentially problematic when the speaker is under time pressure, such is often the case in dialogue. Most turn-transitions are short (with a median of 200ms according to Stivers et al., 2009) and delayed responses are pragmatically dispreferred (Kendrick & Torreira, 2015) and carry the risk of losing the floor (Wilson & Wilson, 2005). Even if production is facilitated in dialogue because of extensive repetition and formulaic sequences (Pickering & Garrod, 2004), which make the estimates derived from picture naming experiments (Indefrey & Levelt, 2004) likely too long, it is still the case that speakers are under much more pressure to produce quickly in dialogue than in monologue.

It is likely that speakers have adapted to producing language under these challenging circumstances by flexibly distributing planning over time, both before and after they begin articulation. Specifically, there is good evidence that high-level planning at the message/conceptual level often begins well before the end of one's partner's turn (Bögels, 2020; Bögels, Magyari, & Levinson, 2015; Magyari, Bastiaansen, De Ruiter, & Levinson, 2014) and in fact even phonological planning may start a second or two before turn-end and be more protracted in time during dialogue than in monologue (Barthel & Levinson, 2020; Bögels & Levinson, 2023). But note also that much planning (of later portions of the utterance) also takes place while speaking (e.g., Smith & Wheeldon, 1999, 2004), even though no study has yet directly measured the extent to which this happens in dialogue.

Most language researchers agree that speakers can vary the amount of advance planning (Brown-Schmidt & Konopka, 2015; Konopka, 2012; Sjerps, Decuyper, & Meyer, 2020), and such variability can lead to a trade-off between the speed with which next speakers initiate their turn and their ability to comprehend the current speaker's turn (Bögels, Casillas, & Levinson, 2018) – that is, next speakers who opt to start planning earlier (and are thus able to begin speaking faster) also show more shallow processing of the current speaker's turn. Turns vary greatly in duration and complexity, which means that the optimal degree of advance planning is likely to be determined on the fly. Moreover, this computation is dependent on one's partner's choices and behaviours: Will they say something unexpected and important right at the end of their turn? Or is their turn so predictable that cognitive resources can be safely

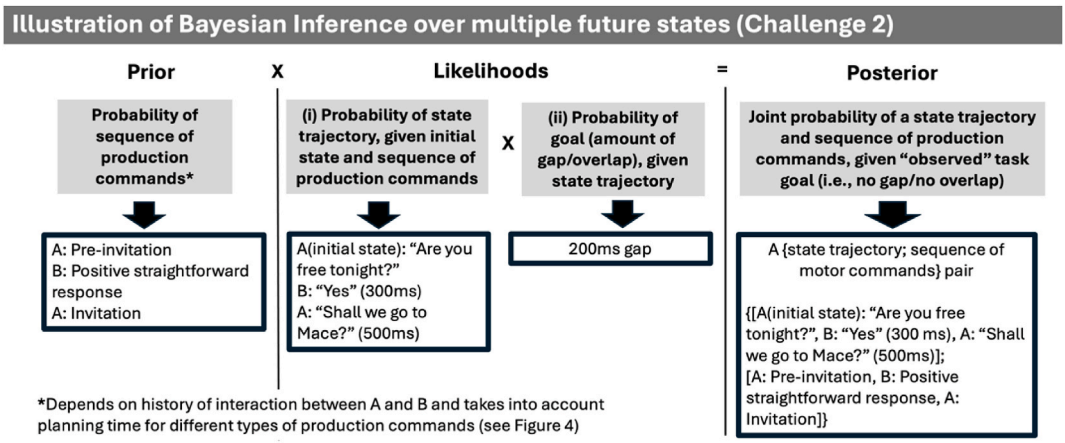


Fig. 3. Illustration of Bayesian inference over multiple future states, based on the example in Fig. 2, Challenge 2 and modelled after McNamee and Wolpert (2019), Equation (5). In this example, speaker A represents priors over possible sequences of production commands (broadly corresponding to communicative intentions or speech acts), which are combined with two likelihood functions, (i) and (ii), to compute the joint probability distribution over states trajectories (i.e., sequences of utterances) and sequences of production commands, given the goal of minimizing gaps and overlaps between successive utterances by A and B. See Fig. 4 for a more detailed illustration of the priors. The white text boxes with a black outline illustrate the most frequent value (mode) for each probability distribution in this example. The first likelihood function (i) represents the probability of a state trajectory (sequence of utterances, including estimates of their planning time) given the initial state (first utterance, produced by A) and a sequence of production commands. The second likelihood function (ii) represents the probability of achieving a particular goal (in this case the amount of gap/overlap between A's and B's turns) given a particular state trajectory (i.e., sequence of utterances). Note that both A's and B's production commands and utterances are represented by A and thus contribute to A's computations.

taken away from understanding it and directed towards advance planning instead?

We argue it is unclear how this level of flexibility can be achieved within existing models of language production. Such models represent the system as a network of connected nodes organised in different representational levels (Dell, 1986; Levelt et al., 1999). They share the assumption that representations are selected as activation spreads through this network and exceeds some threshold. Different models make different assumptions about the way in which activation spreads between representational levels – with some assuming fully interactive architectures (Dell, 1986) and others implementing sequential architectures (Levelt et al., 1999). But these models do not include control mechanisms that can flexibly decide when processing at one representational level should start relative to other levels, or indeed whether resources should be preferentially devoted to comprehension or production at any given point in time (though some models do account for flexibility in the degree to which planning at a given representational level is forward-looking; see Dell, Burger, & Svec, 1997).

What kind of control architecture could allow speakers to assess the future consequences of committing to a particular amount of advanced planning? Above, we presented Bayesian inference as a solution to noisy estimation problems in comprehension and coupled inverse-forward models as a solution to time delays in production. Interestingly, these two approaches have also been combined in theories of biological control. As argued by McNamee and Wolpert (2019), inference can be performed not only with respect to the current state but also over sequences of future states, and can be used to guide the selection of motor commands that will achieve a desired task goal in the future. In other words, Bayesian inference over multiple future states can be used to train a forward model that predicts the future consequences of different courses of action, without the need for the organisms to actually execute them. This can allow the system to choose an optimal sequence of actions when multiple such sequences are possible – that is, it can guide planning (Botvinick & Toussaint, 2012). In sum, this type of control architecture allows an organism interacting with a noisy environment to select an optimal control strategy in the face of uncertainty about the future state of the environment and about the effect that the organism's own future actions will have on the environment.

While Bayesian inference over multiple future states is not the only type of control architecture that could be implemented to achieve flexible planning in language production (Challenge 2, Fig. 2), here we propose it could explain how speakers cope with the time pressures of real-time turn-taking (see Figs. 3 and 4). For example, listeners may entertain multiple predictions about when the current speaker will finish their turn: Imagine A has just asked B a yes/no question (e.g., Are you free tonight?); A might predict that B will produce their answer in about 300ms if their answer is straightforward (e.g., B: Yes), but might take several seconds if they need to explain their answer (e.g., B: Yes, but I am really tired and need to do some cleaning on top of that.). At the same time, A will generate multiple predictions about when they themselves will be ready to start speaking. For example, if A plans to invite B for dinner, they might predict they will be ready to start in 500ms if they have just read the restaurant name on Instagram, but if they are trying to recall it from a conversation they had the previous day they might predict it will take them up to a couple of seconds instead. These predictions may be compared to infer how likely it is that the turn transition will be smooth, and perhaps the inference may be used to decide whether planning needs to start earlier or can be delayed.

Moreover, there is another useful comparison one can draw to motor control architectures specifically designed to deal with coordination problems in the context of joint actions. One common assumption in these architectures is that each agent builds and updates a model of their partner's behaviour in order to predict that behaviour and integrate these predictions into the planning of their own actions (e.g., De Vicaris, Pusceddu, Chackochan, & Sanguineti, 2022; Donnarumma, Dindo, Iodice, & Pezzulo, 2017; Friston & Frith, 2015a, 2015b). This control strategy has specifically been used to model turn-taking behaviour (Donnarumma et al., 2017;

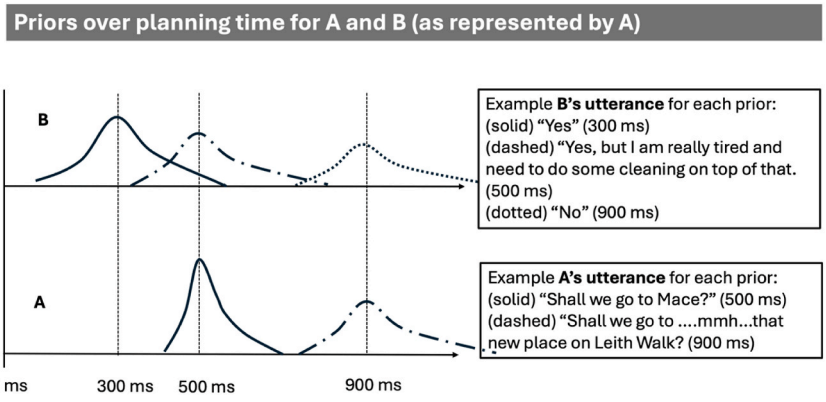


Fig. 4. Schematic illustration of possible priors over planning time for different types of production commands (speech acts) selected by A and B (as represented by A). This refers to the same example (Fig. 2, Challenge 2) as Fig. 3, but note that here we depict priors over single production commands, separately for A and B (rather than sequences of production commands) for ease. Top (B): Priors over B's planning time are generally less precise (representing the fact that A is less certain about B's behaviour than about their own behaviour). Distributions with higher peaks represent more likely priors: based on the previous interaction with B, A expects a positive straightforward response (solid line) to be most likely and be produced on average in 300 ms; positive responses with explanation (dashed lines) are less likely and produced more slowly; negative responses (refusing the invitation outright; dotted line) are least likely and slowest. Bottom (A): Speaker A may be fairly certain to retrieve the restaurant name in about 500ms (solid line, narrower prior); or they may be relatively uncertain about a longer planning time (dashed line, wider prior).

Friston & Frith, 2015a, 2015b), and in this context the use of the same control architecture to model oneself and others' behaviour has been proposed as a control-inspired justification for the existence of turn-taking behaviour in the first place (i.e., speaking and listening at the same time is effortful).

4.3. Tracking changing conversational goals

Our discussion of the challenges of dialogue so far has acknowledged that speakers must monitor their interlocutors' utterances and that they must compute decisions about what to say and how to say it under the strict time constraints imposed by turn-taking. We have also considered how one potential solution to the latter problem may involve modelling the interlocutor herself and her contributions to the conversations. In this section, we further explore the challenges involved in modelling the other: The point is that decisions about what to say and how to say it should take into account a representation of the interlocutor's current understanding and goals, which are however subject to change over the course of a conversation (Pickering & Garrod, 2021).

Dialogue is made up of contingent responses (Schegloff, 2007) and interlocutors typically align on an understanding of the situation under discussion (Pickering & Garrod, 2004), often sharing a joint communicative goal (Clark, 1996). While this helps reduce the uncertainty about the interlocutor's past and future utterances, and the underlying communicative intentions (i.e., it makes dialogue more predictable), it does not eliminate uncertainty altogether. In fact, even fairly simple exchanges are characterised by a high degree of uncertainty about how the conversation will develop, such that it would not be possible to predict utterances at the end of the exchange from utterances at the start of the exchange. This is because, during spontaneous conversations, goals and topics change continuously, and are partly dependent on one's partner (i.e., they are not fully under one's control).

But despite the fast-changing nature of conversational goals, interlocutors do not usually incur major disruption. How is this possible? Pickering and Garrod (2021) suggested that speakers represent the interlocutor's goal and their joint goal in the same way as they represent their own communicative goal. These representations can be used to generate (forward model) predictions about how the conversation will unfold, and comparisons between these predictions and the input can be fed through corresponding inverse models to update the goal representations themselves.

If this is the case, a key question is how these high-level representations and predictions about the state of the conversation interact with lower-level representations and predictions about individual utterances. A useful analogy here is with adaptive intermittent control. Intermittent control was initially developed for complex systems composed of a large set of interconnected nonlinear dynamic subsystems, with the goal of achieving stabilisation and synchronisation of the dynamic components (e.g., Song & Huang, 2015), but here we refer to its application to human behaviour (Nomura, Oshikawa, Suzuki, Kiyono, & Morasso, 2013; Sakaguchi et al., 2015). The key feature of this type of control architecture is that control is applied intermittently rather than continuously. In particular, Sakaguchi et al. (2015) describe a control architecture that is capable of adaptive tracking of changing goals in a manual target-tracking task. This architecture includes a task segmentation component, which dynamically segments time into intervals during which planning and control via coupled inverse-forward model pairs are implemented. At the start of each segment, the forward model which predicts the position of the cursor is re-estimated, and the reliability of this estimation is used to set the segment length, reflecting the fact that high levels of uncertainty and variability mean that planning over shorter time intervals may be less risky. Importantly, a new planning segment can start earlier than the planned duration of the previous segment if a large prediction error is detected (i.e., large distance between cursor and target), allowing the system to cope with large and sudden changes in target velocity and direction. In sum, this architecture allows the system to represent a changing goal by re-estimating the forward models when needed, but not more often than needed (Sakaguchi et al., 2015).

How does this relate to conversation? We suggest that planning segments correspond to stretches of conversation during which the underlying goal/topic is relatively stable. The goal/topic is re-assessed at regular intervals, rather than continuously, and any time there is a breakdown in communication (e.g., because of a misunderstanding, or an interruption) or the speaker detects waning interest in the listener (e.g., because the listener looks away). Thus, unlike the inverse-forward model pairs that underlie production of individual utterances, which are updated at every word to perform monitoring, the inverse-forward models that underlie management of the conversation as a whole can be updated much less frequently. This reflects the nature of the computations involved in the updating. While monitoring typically involves correcting discrepancies between the predicted and observed linguistic representations, which can be achieved by issuing an adjusted production command, managing the conversation instead requires updating and re-estimating the goal itself. The latter is much more computationally intensive, so it adaptive to perform goal re-estimation less often.

5. Conclusion

We have argued that the control architectures instantiated by most current models of language production are comparatively unsophisticated from a control-theoretic perspective. Crucially, we have suggested that more sophisticated control architectures are likely necessary to deal with language use in dialogue. We have identified three challenges for models of language generation that arise from the nature of dialogue and proposed control architectures that might be adopted to deal with these challenges. Specifically, we have argued that (1) coupled inverse-forward model pairs can distinguish between self- and other-generated utterances; (2) the kind of flexible planning needed to achieve timely turn-taking can be implemented as inference over sequences of future states; and (3) changing conversational goals can be tracked using intermittent adaptive control.

CRediT authorship contribution statement

Chiara Gambi: Writing – original draft, Visualization, Funding acquisition, Conceptualization. **Fan Zhang:** Writing – review & editing, Funding acquisition, Conceptualization. **Martin J. Pickering:** Writing – review & editing.

Declaration of competing interest

The authors have no interests to declare.

Data availability

No data was used for the research described in the article.

Acknowledgements

CG and FZ were supported by a grant from the Welsh Crucible (https://welshcrucible.org.uk/). The authors would like to thank all speakers and participants of the Engineering Conversation Workshop, organised by CG and FZ, which took place online on May 21, 2021 (a full recording is available here: https://youtu.be/cUpqx555iZc). CG and FZ would also like to thank Franklin Chang (Kobe City University for Foreign Studies), Severin Lemaignan (Pal Robotics), and Julie Weeds (University of Sussex) for useful discussions.

References

Barthel, M., & Levinson, S. C. (2020). Next speakers plan word forms in overlap with the incoming turn: Evidence from gaze-contingent switch task performance. *Language, Cognition and Neuroscience, 35*(9), 1183–1202.
Bock, K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego: Academic Press.
Bögels, S. (2020). Neural correlates of turn-taking in the wild: Response planning starts early in free interviews. *Cognition, 203*, 104347.
Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia, 109*, 295–310.
Bögels, S., & Levinson, S. C. (2023). Ultrasound measurements of interactive turn-taking in question-answer sequences: Articulatory preparation is delayed but not tied to the response. *PLoS One, 18*(7), Article e0276470.
Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5*(1), Article 12881.
Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). Neural representations and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience, 22*(7), 1504–1529.
Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences, 16*(10), 485–488.
Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science, 41*, 1318–1352.
Brown-Schmidt, S., & Konopka, A. E. (2015). Processes of incremental message planning during conversation. *Psychonomic Bulletin & Review, 22*(3), 833–843.
Chang, R., Baria, A. T., Flounders, M. W., & He, B. J. (2016). Unconsciously elicited perceptual prior. *Neuroscience of consciousness, 2016*(1), Article niw008.
Clark, H. H. (1996). *Using language*. Cambridge, U.K.: Cambridge University Press.
Corps, R. E., & Rabagliati, H. (2020). How top-down processing enhances comprehension of noise-vocoded speech: Predictions about meaning are more important than predictions about form. *Journal of Memory and Language, 113*, Article 104114.
Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*(2), 222–241.
Davis, M. H., & Sohoglu, E. (2020). Three functions of prediction error for Bayesian inference in speech perception. In M. Gazzaniga, R. Mangun, & P. D (Eds.), *The cognitive neurosciences* (6th ed., pp. 177–189). Camb, MA, USA: MIT Press.
De Vicaris, C., Pusceddu, G., Chackochan, V. T., & Sanguineti, V. (2022). Artificial partners to understand joint action: Representing others to develop effective coordination. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30*, 1473–1482.
Dell, G. S. (1986). A spreading-activaction theory of retrieval in sentence production. *Psychological Review, 93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283
Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104*(1), 123–147.
Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1634), Article 20120394.
Denève, S., Alemi, A., & Bourdoukan, R. (2017). The brain as an efficient and robust adaptive learner. *Neuron, 94*(5), 969–977.
Donnarumma, F., Dindo, H., Iodice, P., & Pezzulo, G. (2017). You cannot speak and listen at the same time: A probabilistic model of turn-taking. *Biological Cybernetics, 111*, 165–183.
Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*(4), 752–782.
Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology, 111*, 15–52.
Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*(5), 578–585.
Friston, K., & Frith, C. (2015a). A duet for one. *Consciousness and Cognition, 36*, 390–405.
Friston, K. J., & Frith, C. (2015b). Active inference, communication and hermeneutics. *Cortex, 68*, 129–143.
Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology, 22*(7), 615–621.
Gambi, C., & Pickering, M. J. (2017). Psycholinguistic models linking production and comprehension. In H. Cairns, & E. Fernández (Eds.), *Handbook of psycholinguistics* (pp. 157–181). Wiley/Blackwell. https://doi.org/10.1002/9781118829516.ch7.
Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology, 6*, 751.
Gauvin, H. S., & Hartsuiker, R. J. (2020). Towards a new model of verbal monitoring. *Journal of Cognition, 3*(1). https://doi.org/10.5334/joc.81
Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, 110*(20), 8051–8056.
Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language, 26*(1), 69–83.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829.
Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116*(4), 661–716.
Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*(3), 280–301. https://doi.org/10.1016/j.bandl.2005.06.001
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Paper presented at the proceedings of the second meeting of the north American chapter of the association for computational linguistics on language technologies*.
Hartsuiker, R. J. (2013). Are forward models enough to explain self-monitoring? Insights from patients and eye movements. *Behavioral and Brain Sciences, 36*(4), 357–358.
Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555–568.
Hickok, G. (2012a). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135–145. https://doi.org/10.1038/nrn3158
Hickok, G. (2012b). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual stream model. *Journal of Communication Disorders, 45*(6), 393–402.
Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*, 1213–1216.
Houde, J. F., & Nagarajan, S. S. (2016). Speech motor control from a modern control theory perspective. In G. Hickok, & S. L. Small (Eds.), *Neurobiology of language* (pp. 221–238). Elsevier.
Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech: An MEG study. *Journal of Cognitive Neuroscience, 14*(8), 1125–1138.
Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology, 2*, 255.
Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*(1), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001
Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition, 127*(1), 57–83.
Kello, C. T., & Plaut, D. C. (2004). A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *Journal of the Acoustical Society of America, 116*(4), 2354–2364.
Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes, 52*(4), 255–289.
Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304.
Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203.
Konopka, A. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language, 66*(1), 143–162.
Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review, 23*(6), 1681–1712.
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161–163.
Lametti, D. R., Nasir, S. M., & Ostry, D. J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *Journal of Neuroscience, 32*(27), 9351–9358.
Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*(1), 41–104.
Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1–75.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.
Ludmer, R., Dudai, Y., & Rubin, N. (2011). Uncovering camouflage: Amygdala activation predicts long-term memory of induced perceptual insight. *Neuron, 69*(5), 1002–1014.
Magyari, L., Bastiaansen, M. C., De Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience, 26*(11), 2530–2539.
Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language & Cognitive Processes, 27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006
McNamee, D., & Wolpert, D. M. (2019). Internal models in biological control. *Annual Review of Control, Robotics, and Autonomous Systems, 2*, 339–364.
Meyer, A. S., & Hagoort, P. (2013). What does it mean to predict one's own utterances?[Commentary on Pickering & Garrod]. *Behavioral and Brain Sciences, 36*, 367–368.
Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks, 9*(8), 165–1279.
Moulin-Frier, C., Diard, J., Schwartz, J. L., & Bessière, P. (2015). COSMO ("Communicating about objects using sensory–motor operations"): A bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics, 53*, 5–41.
Niziolek, C. A., Nagarajan, S. S., & Houde, J. F. (2013). What does motor efference copy represent? Evidence from speech production. *Journal of Neuroscience, 33*(41), 16110–16116.
Nomura, T., Oshikawa, S., Suzuki, Y., Kiyono, K., & Morasso, P. (2013). Modeling human postural sway using an intermittent control and hemodynamic perturbations. *Mathematical Biosciences, 245*(1), 86–95.
Norris, D., & McQueen, J. M. (2008). Shortlist B: A bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357–395.
Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience, 31*(1), 4–18.
Nozari, N., Dell, G. S., & Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology, 63*(1), 1–33.
Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–226.
Pickering, M. J., & Garrod, S. (2013a). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–392. https://doi.org/10.1017/S0140525X12001495
Pickering, M. J., & Garrod, S. (2013b). Forward models and their implications for production, comprehension, and dialogue. *Behavioral and Brain Sciences, 36*(4), 377–392. https://doi.org/10.1017/S0140525X12003238
Pickering, M. J., & Garrod, S. (2014). Self-, other-, and joint monitoring using forward models. *Frontiers in Human Neuroscience, 8*, 132.
Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.
Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour, 2*(9), 693–705.
Rao, R. P., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation, 9*(4), 721–763.
Sakaguchi, Y., Tanaka, M., & Inoue, Y. (2015). Adaptive intermittent control: A computational model explaining motor intermittency observed in human behavior. *Neural Networks, 67*, 92–109.
Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes, 26*(2–3), 159–190.
Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse: Text and talk* (pp. 71–93). Washington, DC: Georgetown University Press.
Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I* (Vol. 1). Cambridge university press.
Sjerps, M. J., Decuyper, C., & Meyer, A. S. (2020). Initiation of utterance planning in response to pre-recorded and "live" utterances. *Quarterly Journal of Experimental Psychology, 73*(3), 357–374.
Skipper, J. I. (2014). Echoes of the spoken past: How auditory cortex hears context during speech perception. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), Article 20130297.
Slevc, L. R. (2023). Grammatical encoding. In R. J. Hartsuiker, & K. Strijkers (Eds.), *Language production*. London, UK: Routledge (Chapter 1)).

Smith, M., & Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition, 73*(3), 205–246.
Smith, M., & Wheeldon, L. (2004). Horizontal information flow in spoken language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(3), 675–686.
Sohoglu, E., Peelle, J., Carlyon, R. P., & Davis, M. H. (2014). Top-down influences of written text on perceived clarity of degraded speech. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 186–199.
Song, Q., & Huang, T. (2015). Stabilization and synchronization of chaotic systems with mixed time-varying delays via intermittent control with non-fixed both control period and control width. *Neurocomputing, 154*, 61–69.
Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences, 106*(26), 10587–10592.
Strijkers, K., Costa, A., & Pulvermüller, F. (2017). The cortical dynamics of speaking: Lexical and phonological knowledge simultaneously recruit the frontal and temporal cortex within 200 ms. *NeuroImage, 163*, 206–219.
Strijkers, K., Runnqvist, E., Costa, A., & Holcomb, P. (2013). The poor helping the rich: How can incomplete representations monitor complete ones? *Behavioral and Brain Sciences, 36*(4), 374–375.
Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience, 15*(11), 745–756.
Teufel, C., Dakin, S. C., & Fletcher, P. C. (2018). Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific Reports, 8*(1), Article 10853.
Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language & Cognitive Processes, 26*(7), 952–981.
Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review, 12*(6), 957–968.
Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science, 269*(5232), 1880–1882.