Published in final edited form as:

Lang Cogn Process. 2014 January 1; 29(1): 2–20. doi:10.1080/01690965.2013.834370.

The architecture of speech production and the role of the phoneme in speech processing

Gregory Hickok

Department of Cognitive Sciences, University of California, Irvine, California, 92697, USA

Gregory Hickok: gshickok@uci.edu

Abstract

Speech production has been studied within a number of traditions including linguistics, psycholinguistics, motor control, neuropsychology, and neuroscience. These traditions have had limited interaction, ostensibly because they target different levels of speech production or different dimensions such as representation, processing, or implementation. However, closer examination of reveals a substantial convergence of ideas across the traditions and recent proposals have suggested that an integrated approach may help move the field forward. The present article reviews one such attempt at integration, the state feedback control model and its descendent, the hierarchical state feedback control model. Also considered is how phoneme-level representations might fit in the context of the model.

Speech production is a scientific problem that has been studied from a range of theoretical and methodological perspectives including linguistic, psycholinguistic, motor control, neuropsychology, and cognitive neuroscience. It is often assumed, perhaps implicitly, that these approaches tap into only partially overlapping levels or aspects of the process. For example, one typical characterization is that linguistics (phonology in particular) provides information on the nature of the abstract units involved, psycholinguistics and neuropsychology tell us about the higher-level stages of processing, motor control research tells us about the lower-level stages of controlling the vocal tract, and cognitive neuroscience attempts to relate all of this to neural circuits (or in some cases just ignores most prior work and builds new, neurocentric models). In short, each approach claims its own corner of the sandbox and doesn't interact all that productively with others.

A prime example of this is the assumed division of labor between fairly well-developed psycholinguistic and motor control models of speech production. Speech production models as uncovered by decades of research on naturalistic observation of slips of the tongue and laboratory studies using naming, tongue-twister, and other clever paradigms typically restrict themselves to traditional linguistic domains such as semantic, lexical, and phonological levels (Dell, 1986; Fromkin, 1971; Garrett, 1975; Willem J. M. Levelt, Roelofs, & Meyer, 1999). The output of such models is assumed to provide the input to lower-level articulatory processes that are modeled by motor-control architectures that deal with theoretical objects or computations involving controllers, plants, efference copies, internal models, sensory feedback, and Kalman filters (Fairbanks, 1954; F.H. Guenther, Hampson, & Johnson, 1998; J.F. Houde & Jordan, 1998).

A closer look, however, reveals impressive similarities between the two traditions. Both have hit on the notion of feedback monitoring, including internal and external loops (W. J. Levelt, 1983; Nozari, Dell, & Schwartz, 2011); both incorporate a hierarchical organization

(S. T. Grafton & Hamilton, 2007; Haruno, Wolpert, & Kawato, 2003); and each approach has theoretical tendrils that reach into the other's domain. For example, motor planning cannot be limited to single articulatory gestures but must involve planning over sequences of gestures (coarticulation), which corresponds to planning units above the phoneme and extending into the syllable (Bohland & Guenther, 2006; F. H. Guenther, Ghosh, & Tourville, 2006) – the typical domain of higher-level models. And from the reverse perspective, the phoneme is a central unit of representation in psycholinguistic models, one that is defined in vocal tract articulatory space (feature bundles) (Chomsky & Halle, 1968) – the very features that need to be controlled in motor control models. It seems that the theoretical and terminological chasm between psycholinguistic and motor control traditions is more apparent than real (Hickok, 2012a).

In two recent papers, my colleagues and I have proposed an integrated model of speech production that draws on work, to varying degrees, from psycholinguistic, motor control, neuropsychological, and cognitive neuroscience traditions (Hickok, 2012a; Hickok, Houde, & Rong, 2011). My aim here is to describe this integrated approach by outlining two variants, the state feedback control (SFC) model and its descendent, the hierarchical state feedback control (HSFC) model. I then spend some time considering how the theoretical constructs from linguistics, such as the phoneme, fit within the context of the proposed model. To preview my conclusions regarding the latter, I will reject the commonly held assumption that phonemes are amodal knowledge representations used in the service of both speech production and perception. I will argue instead that phoneme representations as identified by generative phonology are objects relevant to sensory-motor interaction, part of a theory of “dorsal stream” function, and are not necessarily relevant to speech recognition. This view is consistent with, but not identical to, previous claims that the perceptual unit of speech perception is the syllable (Greenberg, 1996; Massaro, 1972).

Motor control and internal models

Sensory feedback is a critical component of motor control, yet overt feedback is inefficient for online control because it is delayed, intermittent, and sometimes noisy (Kawato, 1999a; Shadmehr & Krakauer, 2008).¹ To get around this problem, motor control models incorporate an internal model of the motor effector that allows the system to predict the state of the motor effector as well as the sensory consequences of actions, a so-called “forward model” (Figure 1) (Kawato, 1999b; Shadmehr & Krakauer, 2008; Shadmehr, Smith, & Krakauer, 2010; Wolpert, Ghahramani, & Jordan, 1995). The internal forward model is particularly useful for online movement control because the effects of a movement command can be evaluated for accuracy and potentially corrected before external feedback; it also can facilitate movement corrections required by incoming external sensory feedback. By contrast, external feedback is critical for three purposes: to learn the internal model in the first place, i.e., the relationship between motor commands and their sensory consequences; to update the internal model in case of persistent mismatches (errors) between the predicted and measured states; and to detect and correct for sudden perturbations. In many cases, the two sources of feedback work together such as when a perturbation is detected via sensory feedback and a correction signal is generated using internal forward predictions of the state of the effector. Motor control models with these feedback properties are often referred to as state feedback control models because feedback from the predicted (internal) state as well as the measured state of the effector is used as input to the controller.

¹By “sensory feedback” I mean any kind of input to the motor system from sensory systems. By “overt feedback” or “external feedback” I refer to feedback on the consequences of actions that is detected using peripheral sensory receptors. By “internal feedback” or “internal sensory feedback” I refer to sensory-to-motor inputs derived from predicted rather than overt sensory feedback. Thus, sensory feedback can be either internally or externally derived.

The existence of internal models in state feedback control has been supported experimentally (Shadmehr & Mussa-Ivaldi, 1994; Tian & Poeppel, 2010; Wolpert et al., 1995) and has proven highly influential and widely accepted within the visuomotor domain (S. T. Grafton, 2010; Kawato, 1999a; Wolpert, Doya, & Kawato, 2003). Feedback control generally, as opposed to internal or state feedback control specifically, has also been empirically demonstrated in the speech domain using overt sensory feedback alteration paradigms and other approaches (J. Perkell et al., 1997). This work has shown that when speaking, people adjust their speech output to compensate for sensory feedback ‘errors’ (experimentally induced shifts) in both the auditory (Burnett, Freedland, Larson, & Hain, 1998; J.F. Houde & Jordan, 1998; Larson, Burnett, Bauer, Kiran, & Hain, 2001; Tourville, Reilly, & Guenther, 2008) and somatosensory systems (Tremblay, Shiller, & Ostry, 2003). Evidence for *internal* state feedback is not readily found in the motor speech control literature. However, if one looks outside of the motor control tradition, strong evidence can be found for the existence of internal state feedback control in speech production (see below).

With respect to hierarchical organization, it is worth pointing out that in the visuomotor literature state feedback models for motor control are increasingly hypothesized to be hierarchically organized (Diedrichsen, Shadmehr, & Ivry, 2010; S.T. Grafton, Aziz-Zadeh, & Ivry, 2009; S. T. Grafton & Hamilton, 2007; Haruno et al., 2003), a concept that is consistent with long-standing appreciation of the existence of a sensorimotor hierarchy (Jackson, 1887). See Gracco for a speech related discussion of this issue (Gracco, 1994).

The psycholinguistic perspective

Psycholinguistic speech production models typically start with a conceptual or message level representation and end with a phonological representation (that is, the output) that feeds into the motor control system. Thus, phonological representations are considered abstract representations that are distinct from motor control structures in most, but not all (Browman & Goldstein, 1992; Plaut & Kello, 1999), psycholinguistic models of speech production.

Speech production models span the range of hierarchical levels from the phrase down to the phoneme or even feature level (Bock, 1999; Dell, 1995; Fromkin, 1971; Garrett, 1975; W.J.M. Levelt, 1989). Here I will focus on stages involved in the production of single words, which is generally thought to involve two stages of processing: a lexical (or ‘lemma’) level and a phonological level (Dell, 1986; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Willem J. M. Levelt et al., 1999). In such models (Dell, 1986; W.J.M. Levelt, 1999) (Figure 2), input to the system comes from the conceptual system, that is, the particular concept or message that the speaker wishes to express. The concept is mapped onto a corresponding lexical item, often referred to as a lemma representation, which codes abstract word properties such as a word’s grammatical features but does not code a word’s phonological form. Phonological information is coded at the next level of processing. Evidence for such a two-stage model comes from a variety of sources including the distribution of speech error types (Dell, 1995; Fromkin, 1971; Garrett, 1975), chronometric studies of interference in picture naming (W.J.M. Levelt, 1989), tip-of-the-tongue phenomena (Vigliocco, Antonini, & Garrett, 1998) and speech disruption patterns in patients with aphasia (Dell et al., 1997).

As noted above, feedback correction mechanisms — including both internal and external feedback monitoring loops — have been proposed to form part of psycholinguistic models of speech production (W. J. Levelt, 1983). That external feedback is monitored and used for error correction is evident in everyday experience when the occasional misspoken word or

phrase is noticed by a speaker and is corrected. The timing of such error detection in some cases reveals that internal error detection is also operating. For example, it has been pointed out that documented error corrections such as “v-horizontal” (starting to say vertical and correcting to horizontal) occur too rapidly to be carried out by an external feedback mechanism (Hartsuiker, Kolk, & Martensen, 2005; Nozari et al., 2011). Other studies have shown that errors can be detected when external feedback is masked (Lackner & Tuller, 1979), or when speech is not overtly produced at all but only imagined (Oppenheim & Dell, 2008), which provides further evidence for an internal feedback control system. Still other work has shown, using a tongue twister paradigm, that onset exchange errors (e.g., *barn door* → *darn bore*) are infrequent when the resulting error comprises taboo words (e.g., *tool kits* →), suggesting the existence of an internal monitor (Motley, Camden, & Baars, 1982). The symptom complex of conduction aphasia is also indirect evidence for internal feedback control (Hickok, 2012a; Hickok, Houde, et al., 2011) (see also below).

Within the psycholinguistic tradition, the nature of the internal and external feedback correction mechanisms in speech production has received increasing empirical and theoretical attention over the past two decades (Huettig & Hartsuiker, 2010; Nickels & Howard, 1995; Oppenheim & Dell, 2008; Ozdemir, Roelofs, & Levelt, 2007; Postma, 2000), including the suggestion that error detection and correction in speech may not rely on sensory systems (Nozari et al., 2011), a notion that is not consistent with assumptions in the motor control literature.

Towards an integrated model

The state feedback control (SFC) model of speech production was a first attempt at developing a model of speech production that integrates the various traditions (Hickok, Houde, et al., 2011) (Figure 3). The model builds on previous work by Guenther and colleagues (F.H. Guenther et al., 1998), Tian and Poeppel (Tian & Poeppel, 2010) and Houde and Nagarajan (J.F. Houde & Nagarajan, 2011). The essence of the model is the assumption that the feedback control architecture applies not only at low levels of motor control but at higher, “phonological levels”, as well; see the discussion below regarding what “phonological” might mean in this context. And it is at the phonological level (broadly speaking) that the model is primarily cast. In motor control terms, this phonological level is the “internal model”, which is divided into two representational components, a motor-phonological component (~the internal model of the motor effector) and an auditory-phonological component (where the auditory consequences of actions are coded).

Consistent with psycholinguistic models, input to the phonological level of processing comes from a lexical-conceptual level, thus instantiating the familiar two-stage architecture. Different from many psycholinguistic models, but consistent with some neuropsychological models (Howard & Nickels, 2005; Jacquemot, Dupoux, & Bachoud-Levi, 2007; Martin, Lesch, & Bartha, 1999; Shelton & Caramazza, 1999), the phonological processing level is split into two components as noted above, an auditory-based and a motor-based system (input and output lexicons, in other terminology). This division, while not particularly popular in psycholinguistics, is highly consistent with feedback control architectures for the following reason. If the point of state feedback control architectures is to evaluate the sensory consequences of an action coded in the motor system, then there better be both a motor code and a distinct code representing the sensory target, otherwise there is nothing to evaluate. A brief digression is helpful here to underline this point.

There is both behavioral and neurophysiological evidence for the claim that auditory feedback is used by the motor system for speech motor control and that the targets of speech acts are auditory (F.H. Guenther et al., 1998). In terms of the role of auditory feedback, it is well-documented that overt auditory feedback of various sorts affects speech output. This is

seen in altered auditory feedback experiments, which have shown that delayed auditory feedback has a disruptive effect on speech fluency (Stuart, Kalinowski, Rastatter, & Lynch, 2002; Yates, 1963) and that artificially shifted pitch (Burnett et al., 1998; Burnett, Senner, & Larson, 1997; Larson et al., 2001) or formant frequencies (J.F. Houde & Jordan, 1998; Tourville et al., 2008) results in compensatory responses and (if the shift is persistent) adaptation in speech articulation. What these findings show is that motor speech acts lead to predictions about the sensory consequences of those acts, that these consequences are monitored on some level, and that information derived from the error between the predicted and actual consequences are used to adjust ongoing and future speech gestures. This would not be possible if there were not separate motor and auditory codes related to speech. As further confirmation of this view of the motor speech system it has been shown neurophysiologically that speech articulation modulates the auditory response to that speech, presumably reflecting forward sensory prediction (Aliu, Houde, & Nagarajan, 2009; Heinks-Maldonado, Nagarajan, & Houde, 2006; J. F. Houde, Nagarajan, Sekihara, & Merzenich, 2002; Ventura, Nagarajan, & Houde, 2009). Further evidence that the targets of speech acts are indeed auditory include the tendency for talkers to acquire the speech patterns in their local linguistic environment, so-called gestural drift (Sancier & Fowler, 1997), the tendency to unconsciously mimic characteristics of ambient speech with short exposure time (Delvaux & Soquet, 2007; Kappes, Baumgaertner, Peschke, & Ziegler, 2009), and the fact that speech gestures associated with a given sound can be quite variable in terms of the vocal tract shape. American English /r/ is an example of the latter in which very different articulatory gestures are used while the acoustic signal is nearly constant (F. H. Guenther et al., 1999). Facts such as these are neatly explained if motor speech gestures aim toward hitting targets coded in auditory space (F.H. Guenther et al., 1998; J. S. Perkell, 2012).

In the SFC model, the auditory- and motor-phonological systems are wired up via an interface network, which performs a kind of coordinate transform, linking phonological codes in auditory space to corresponding phonological codes in articulatory sequence space and vice versa. The hypothesized existence of sensory-motor interface networks derives from work in neuroscience (Andersen, 1997; Fogassi et al., 2001; Grefkes & Fink, 2005; Hickok, Okada, & Serences, 2009; Jeannerod, Arbib, Rizzolatti, & Sakata, 1995) but is also consistent with computational architectures that are familiar to language scientists such as hidden layers in connectionist networks that map between representational levels (Rumelhart, Hinton, & McClelland, 1986).

As noted, the phonological level receives input from a lexical-conceptual level. One unique aspect of the SFC model is the parallel input from the lexical-conceptual level to both auditory and motor phonological systems. This feature is not characteristic of motor control architectures (which typically start with the execution of a motor act) or two-stage psycholinguistic models, but it does have roots in classical 19th century models of the neural organization of language (Lichtheim, 1885; Wernicke, 1874/1977) and has similarities to more recent dual-route models of speech repetition (Hanley, Dell, Kay, & Baron, 2004). The parallel input assumption is key to explaining conduction aphasia as we will see.

The model does not explicitly deal with what size units are coded in the auditory- and motor-phonological networks. The question of whether phoneme-sized units are coded in this level of the network is taken up below. Here I would like to consider the question of how larger-scale units, such as sequences of phonemes or even sequences of syllables in words or common phrases, might be coded. Previous work has suggested that sequences of phonemes that form the syllables or words of a language might be efficiently coded as motor chunks, a mental syllabary to use Levelt's term (Cholin, Levelt, & Schiller, 2006; W. J. Levelt & Wheeldon, 1994; Willem J. M. Levelt et al., 1999). Adopting this view in the present model, these motor chunks would be associated with similar-sized chunks of

auditory spectrotemporal patterns that would serve as the acoustic targets of motor sequences. Frequency effects in articulating syllables or sequences of syllables suggests that experience with a particular sequence is a factor in the formation of the mental syllabary (Cholin et al., 2006). The implication for (high-level) motor control is that more frequent sequences will be more “strongly chunked” in the motor system and therefore require less segment-by-segment guidance from sensory systems than less frequent sequences. Coding larger sequences, such as syllables, also provides a natural solution to the problem of “knowing where you are in a word” during articulation: if one conceives of a word as a sequence of phonemes, then the representation of that word has to comprise both the individual phonemes and their sequence. But if the word is coded at some level as a holistic chunk – a syllable or sequence of syllables – without internal segmental structure, then the order of the phonemes is built in to the higher-level code. This higher-level code can then serve as a roadmap for lower-levels of motor control that need to hit individual articulatory targets corresponding to the phonemic segments of the word. I’ll return to this point in the context of development.

Now we will consider how the system functions, starting with development. The system must learn two sets of mappings, both anchored by auditory representations. On one hand, the system must learn the relation between sound patterns in the environment (phonological words) and conceptual representations, the “ventral stream” function of dual route models (Hickok & Poeppel, 2000, 2004, 2007; Rauschecker & Scott, 2009; Wernicke, 1874/1969). Acquiring this mapping between sound and meaning – word learning, as it is more commonly called – is a non-trivial process and the topic of much developmental psycholinguistic work (Bloom, 2000), which is beyond the scope of the present discussion. On the other hand, the system must learn the relation between speech gestures and the sounds they make, or to flip the problem around, how to reproduce the sound pattern of the language with vocal tract gestures. This is the “dorsal stream” function of dual route frameworks and the purview of models of speech production and speech motor control models. Thus, I view the present model as an elaboration of the dorsal stream, auditory-motor interface proposed in Hickok and Poeppel (Hickok & Poeppel, 2000, 2004, 2007).

The mapping between sound and meaning on the one hand and sound and speech articulation on the other are computationally distinguishable tasks in the sense that one mapping can be achieved independently of the other. For example, the ability to articulate a sequence of syllables is not dependent on knowledge of the semantic referent(s) of that sequence (although semantic information certainly aids long-term memory for word forms). Conversely, expert word learning (sound to meaning association) can be achieved without involvement of the motor speech system, as cases of developmental anarthria and cerebral palsy have shown convincingly (Bishop, Brown, & Robson, 1990; Lenneberg, 1962). This is not to say that the two mappings are noninteracting. Certainly they do interact, particularly in that they both involve representations of the sound patterns of words. The point is that the two mappings are computationally distinguishable due to the representations that the sound patterns are mapped onto in the two cases, conceptual versus motor, thus requiring different sorts of transformations. The distinction in terms of computational mappings associated with the two tasks is inescapable and likely the reason why dual stream models of sensory processing have surfaced repeatedly in the last century or so (Hickok & Poeppel, 2000; Milner & Goodale, 1995; Poljak, 1926; Rauschecker, 1998; Rauschecker & Scott, 2009; Schneider, 1969; Trevarthen, 1968; Ungerleider & Mishkin, 1982; Wernicke, 1874/1977).

I suggest that these two mappings in speech are learned somewhat independently at first and linked later. As proposed in the DIVA model, auditory-motor learning likely starts with babbling, which associates motor gestures with their auditory consequences thus training the internal model (the mapping relation between action plans and vocal tract-generated sounds)

(F. H. Guenther, 2006). In the mean time, the sound pattern of words in the ambient language are acquired via sensory learning and via whatever processes link phonological words to meanings; the sound patterns are presumably stored in the form of high-level auditory representations, although the precise level (or levels) is debatable. DIVA, for example, assumes that the representations are categorical and correspond to phoneme representations (F. H. Guenther, 1995), but recent evidence suggests that talkers will compensate for within category acoustic shifts, although to a lesser extent than cross category shifts (Niziolek & Guenther, 2013). Now, given that the system has acquired some auditory targets (the stored sound patterns) and given that the system has learned a basic mapping between motor gestures and the sounds they make, activation of an auditory target (via imitation of a heard word or just via internal volitional speech) should be able to trigger a motor attempt to “hit” that target, and in fact infants appear to attempt to hit speech targets provided by ambient speech samples (P. K. Kuhl & Meltzoff, 1996). The success of these attempts can then be evaluated against the stored auditory targets using overt acoustic feedback. This process further tunes the internal model (the auditory-motor associations) and in addition establishes the double-association between auditory sound patterns and motor systems on one hand and auditory sound patterns and conceptual systems on the other. See (P. K. Kuhl, 2010) for a more detailed discussion of sensory-motor processes in the early stages of speech acquisition and see (F. H. Guenther, 1995) for a more detailed discussion of the computational issues.

The auditory sound patterns, *auditory phonological representations* as I will call them, play a linking or hub-like role in the network in that they serve both as targets for speech gestures and as access codes to the lexical-conceptual network. However, focusing on the production side, this does not imply that speech output *must* involve an auditory-phonological code in every case and to an equal degree for successful articulation. In reaching for a cup, it is critical to have a sensory target that provides information about the location, size, and orientation of the object. However, it is also the case that if you reach for the same cup in the same location repeatedly, you will need to rely less and less on sensory information for achieving a successful reach. In general, there is an inverse relation between familiarity with the action-object pairing and the need for sensory involvement in the action: the more familiar the situation, the less you need sensory guidance (Halsband & Lange, 2006; Preilowski, 1977). The same is true in speech, I suggest. Articulating less familiar words will require more input from the auditory-phonological component of the network than articulating highly familiar words. In the limit, repeating novel non-words is absolutely dependent on auditory-phonological targets because this would be akin to reaching for a completely novel object in a novel location; there is no motor preset to rely on, so you can’t possibly do it without reference to the sensory target. Frequency effects in naming (faster, more accurate responses for more familiar items) are broadly consistent with this view (Almeida, Knobel, Finkbeiner, & Caramazza, 2007; Jescheniak & Levelt, 1994; Willem J. M. Levelt et al., 1999; Nozari, Kittredge, Dell, & Schwartz, 2010; Oldfield & Wingfield, 1965).

In the SFC framework, the “bypass route” for highly familiar words is the parallel, direct link between lexical-conceptual systems (or the word/lemma network in the more detailed HSFC model) and motor-phonological codes, consistent with recent neuropsychological models (Jacquemot et al., 2007; Nozari et al., 2010). I hypothesize that this link is established via Hebbian learning: at first, lexical-concepts (words/lemmas) activate their associated auditory-phonological units, which in turn activate their associated motor-phonological units. This results in co-activation of lexical and motor-phonological units, which, when firing together, end up wiring together. The end result in the mature network is that when a lexical node is activated, it activates both the auditory- and motor-phonological units with which it is associated and according to the strength of the associations. A word

that is spoken more often will have stronger associations between lexical and motor units and therefore will rely less on the activation of the sensory target; again, this notion is borrowed wholesale from the motor learning literature (Halsband & Lange, 2006; Preilowski, 1977).

The production network is therefore dynamic in the following sense. If one has learned a word that involves a syllable sequence that has not been spoken previously (e.g., as when learning technical terms), activation will primarily involve a lexical-auditory-motor route because few or no associations between conceptual and motor systems have been built up. The idea is that each syllable will have to be guided individually by previously learned auditory-motor associations. Conversely, a highly learned word (on the production side) will primarily rely on a direct lexical-motor route, where the motor sequence is chunked and with little required input from the auditory system. This is different than saying the auditory units are not activated; they are, I assume, it is just that their input is not needed as much to achieve success. A mid-frequency word (or a low-frequency word composed of high-frequency subcomponents, say *tempest*) would fall somewhere in between with direct activation of motor units requiring more input from the auditory target units. Consistent with these assumptions, error rates in aphasia are strongly correlated with word familiarity (Nozari et al., 2010).

What does it mean computationally to require input from the auditory target units? The proposal put forward in the SFC and HSFC is inspired by state feedback control models in motor research (J.F. Houde & Nagarajan, 2011; Shadmehr & Krakauer, 2008). In typical visuomotor experiments, targets are defined by a sensory stimulus, such as an object in a spatial location, and the action is directed at that external target. Natural speech production (c.f., shadowed or repeated speech) is different in the sense that the target is not defined in the immediate sensory environment but is an *internal* representation of previously heard auditory stimuli. But in both cases it is still the sensory target that grounds the action. Motor control models often assume a kind of cancelling operation as the basis for comparing motor predictions with overt sensory feedback, with the motor prediction implemented as an inhibitory signal. If the prediction (–) and the overt sensory input (+) cancel, then this indicates an accurate prediction. Prediction error, then, is sensory activation that is not cancelled. Consistent with this general approach, the present model assumes excitatory inputs from the lexical-conceptual level to both the auditory- and motor-phonological units, excitatory inputs from auditory- to motor-phonological units, and inhibitory inputs from motor- to auditory-phonological units (see Figure 4). The excitatory connections between auditory- and motor-phonological units can be conceptualized as an “error signal” in the sense described above, although in this case input to sensory units are coming not from the external environment but from an internal top-down source. But the prediction error signal computation is identical: if an auditory target is activated and there is no activity in the motor units, then the excitatory inputs from auditory to motor will correct this non-activation “error” by activating the corresponding motor units. If an auditory target is activated and the corresponding motor units are also activated then inhibitory motor-to-auditory inputs will cancel the auditory-to-motor excitation, thus squashing the error signal (i.e., the motor network is on the right track and doesn’t need any “correction” from the auditory network). If an auditory target is activated and the wrong (non-corresponding) motor units are activated, the motor-to-sensory feedback will inhibit non-target auditory units and therefore “allow” the actual auditory target to continue sending excitatory inputs to the corresponding motor units thus correcting the error. In this way, the same prediction/error detection calculation that is used for external feedback is used for internal error detection and correction. Thus, all of this is going on, I hypothesize, prior to speech articulation *as part of the motor programming process*. So while the auditory targets are always activated, they only play a major role in the process when something goes wrong,

either because the wrong motor units were activated or because no motor units were activated. Given that the whole computation is anchored by the auditory target, it is worth noting that if the system happened to activate the wrong auditory target, then the result will be an error because the system is built to hit whatever target is activated. Such errors should go undetected by the talker (at least at the phonological level) because the incorrect auditory target will be correctly hit by the motor component of the system.

The present proposal incorporates *internal* error detection and correction via sensory target activation and forward sensory prediction *as part of the initial motor planning process*. This differs from typical conceptualizations of internal models, which come into play only after a motor command has been initiated and an efference copy is received by the internal model, although the internal model can influence an ongoing motor plan. In fact, in the SFC and HSFC models, true efference copies (copies of executed motor commands) do not do any computational work. In fact, it is better to conceptualize the internal model in the present approach simply as *the* motor planning mechanism, not something that is separate from it. That said, the present framework still implements the predictive mechanisms that make more traditional internal models (with efference copies) useful, it just builds the internal model into the motor planning circuit itself and generates forward predictions prior to motor execution, effectively taking advantage of these predictive mechanisms at all stages of motor planning. Direct empirical evidence for auditory system involvement during the motor planning process is limited but there are some suggestive findings. Levelt and colleagues report MEG-localized activity in the left posterior Sylvian region (approximately at the location of Spt) prior to motor cortex activity in a naming task (W.J.M Levelt, Praamstra, Meyer, Helenius, & Salmelin, 1998) and a recent intracranial electrocortical recording study found high-gamma activity in this same region coincident with motor cortex activation just prior to speech articulation (Edwards et al., 2010).

One consequence of the model architecture is that any time an auditory phonological unit is activated it will propagate its activity to motor units, even during perception. This is an unwanted feature of the model in the sense that most heard speech is not immediately shadowed by the listener. Some inhibitory inputs to motor units, say from prefrontal or basal ganglia circuits, must be added to the model to gate their output (Shadmehr & Krakauer, 2008). However, the automatic activation of motor units is consistent with the observation that perceived speech often activates motor speech regions even when no output is required (Wilson, Saygin, Sereno, & Iacoboni, 2004) and can provide an explanation of conditions such as echolalia, which appear to involve a release of inhibition causing inappropriate repetition of heard speech (Christman, Boutsen, & Buckingham, 2004; Duffy, 1995).

These ideas have yet to be tested in a large-scale implemented model, but a small-scale connectionist simulation demonstrated the feasibility of its architectural and computational assumptions: the network can indeed correct errors in motor unit activation (Hickok, 2012a).

Suppression of sensory target activity in the speech domain makes sense computationally for two reasons. One is to prevent interference with the next sensory target. In the context of connected speech, auditory phonological targets (syllables) need to be activated in a rapid series. Residual activation of a preceding phonological target may interfere with activation of a subsequent target if the former is not quickly suppressed. An inhibitory motor-to-sensory input provides a mechanism for achieving this. The second benefit of target suppression is that it can enhance detection of off-target sensory feedback. Detection of deviation from the predicted sensory consequence of an action is a critical function of forward prediction mechanisms, as it allows the system to update the internal model. Recent work on selective attention has suggested that attentional gain signals that are applied to flanking or 'off-target' sensory features comprise a computationally effective and

empirically supported mechanism to detect differences between targets and non-targets (Jazayeri & Movshon, 2006, 2007; Regan & Beverley, 1985; Scolari & Serences, 2009). In the present context, target suppression would have the same functional consequence on detection as increasing the gain on flanking non-targets, namely, to increase the detectability of deviations from expectation (Hickok, Houde, et al., 2011).

The target suppression mechanism also resolves a noted problem in psycholinguistics concerning simultaneously monitoring both inner and external feedback by the same system given the time delay between the two (Nozari et al., 2011; Vigliocco & Hartsuiker, 2002). In the SFC model, internal and external monitoring are just early and later phases, respectively, of the same mechanism. In the early, internal phase, errors in motor activation fail to inhibit the driving activation of the sensory representation, which acts as a ‘correction’ signal; in the later, external monitoring phase, the sensory representation is suppressed, consistent with some models of top-down sensory prediction (Friston, 2010; Summerfield & Egner, 2009), which enhances detection of deviation from expectation; that is, the detection of errors (Hickok, Houde, et al., 2011).

Motor influences on perception

There is much fuss currently about the possible involvement of motor speech systems in perception (D’Ausilio et al., 2009a, 2009b; Fadiga, Craighero, & D’Ausilio, 2009; Meister, Wilson, Deblieck, Wu, & Iacoboni, 2007; Wilson, 2009). Despite early claims for a necessary role, it is clear that the motor system’s participation is quite limited given how well speech perception can be accomplished with a damaged, deactivated, undeveloped, or genetically lacking motor speech system (Bishop et al., 1990; Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Hickok, Costanzo, Capasso, & Miceli, 2011; Hickok et al., 2008; P.K. Kuhl & Miller, 1975; Rogalsky, Love, Driscoll, Anderson, & Hickok, 2011; Rogalsky, Pitz, Hillis, & Hickok, 2008). However, a small modulatory contribution of the motor system to speech perception remains a possibility and the SFC model architecture provides a mechanism for such influence in that activation of motor speech units generate a forward prediction for their corresponding auditory speech units (Hickok, Houde, et al., 2011). In the context of a state feedback control model, though, forward predictions are instantiated as inhibitory signals, which would be expected to *decrease* sensitivity to the predicted features, contrary to the idea that motor involvement enhances perception (Hickok, 2012b). For this reason, it is not obvious that motor-based prediction will be all that useful for speech perception, in contrast to well-known and robust lexical, semantic, or syntactic (i.e., ventral stream) top-down influences (Marslen-Wilson & Tyler, 1980; Miller, Heise, & Lichten, 1951; Obleser, Wise, Alex Dresner, & Scott, 2007). A recent discussion of predictive coding in the context of a dual stream model of sensory processing can be found in (Hickok, 2012b).

Conduction aphasia

Conduction aphasia is one of the strongest sources of empirical validation for the proposed model. People with such aphasia have fluent speech yet produce relatively frequent and predominantly phonemic speech errors (paraphasias) that they often detect and attempt to correct, mostly unsuccessfully (Baldo, Klosternann, & Dronkers, 2008; Damasio & Damasio, 1980; Goodglass, 1992). Although speech perception and auditory comprehension at the word and conversational level are well preserved in such individuals, verbatim repetition is impaired, particularly for complex phonological forms and non-words (Goodglass, 1992). Explaining of the co-occurrence of these features —that is, generally fluent output, impaired phonemic planning, and preserved speech perception — has proven difficult. In models that assume one phonological level network for perception and

production, a central phonological deficit could yield phonemic output problems but would also be expected to affect perception. Alternatively, assuming that separate phonological input and output systems exist, impairment to a phonological output system could explain the paraphasias but should also cause dysfluency. Furthermore, the lesions in conduction aphasia are in auditory-related temporal–parietal cortex (Baldo et al., 2008; B.R. Buchsbaum et al., e-pub 2011; Damasio & Damasio, 1980), not in frontal cortex where one would expect to find motor-related systems. Damage to a phonological input system is more consistent with the lesion location, explains the preserved fluency because the motor phonological system is still intact, and could explain paraphasias if one assumes a role for the input system in speech production. However, again there is no explanation for why the system can easily recognize errors perceptually that it fails to prevent in production. For this reason, some authors have attempted to explain conduction aphasia as a deficit to a buffer in phonological working memory, separate from the phonological input and output systems (Baldo et al., 2008), but then one loses the explanation of paraphasic errors (B. R. Buchsbaum et al., 2011).

Wernicke's original hypothesis that conduction aphasia is caused by a disconnection between sensory and motor speech systems is a viable solution (Hickok et al., 2000): fluency is preserved because the motor system is intact, perception is preserved because the sensory system is intact, and paraphasias occur because the sensory system can no longer play its role in speech production once the systems are disconnected (see also (B. R. Buchsbaum et al., 2011; Jacquemot et al., 2007) for similar arguments). What was lacking from Wernicke's account, though, was a principled explanation for why the sensory system plays a role in production. Internal state feedback control (as included in the SFC model) provides such a principled explanation: the sensory speech system is involved in production because the sensory system defines the targets of speech actions, and without access to information about the targets, actions will sometimes miss their mark. The only other modern adjustment that is needed to Wernicke's account is the anatomy. He proposed a white matter tract as the source of the disconnection, for which there is little evidence (Anderson et al., 1999; Dronkers & Baldo, 2009; Hickok, 2000). Modern findings instead implicate a cortical system that computes a sensorimotor coordinate transformation, which we have identified as area Spt (B. Buchsbaum, Hickok, & Humphries, 2001; B.R. Buchsbaum et al., e-pub 2011; Hickok, Buchsbaum, Humphries, & Muftuler, 2003; Hickok et al., 2009), which lies in the lesion distribution of conduction aphasia (B. R. Buchsbaum et al., 2011).

Extending the model

So far the discussion has been fairly vague about what kind of phonological information is coded in the auditory and motor networks and it has neglected a large literature showing the involvement of the somatosensory system in speech production. Here, I will outline an extension of the integrated SFC model, the *hierarchical* state feedback control (HSFC) model (Figure 4) that partially remedies this situation. In the HSFC model there are two hierarchically organized levels of state feedback control, which are similar to the levels proposed by Gracco and Lofqvist (Gracco, 1994; Gracco & Lofqvist, 1994b). The higher level codes speech information roughly at the syllable level (that is, vocal tract opening and closing cycles) and involves a sensory–motor loop that includes sensory targets in auditory cortex, motor programs coded in the Brodmann area (BA) 44 portion of Broca's area and/or lower BA6, and area Spt, which computes a coordinate transform between the sensory and motor areas (Hickok, 2012a; Hickok et al., 2003; Hickok et al., 2009). This is the loop described in the earlier SFC model (Hickok, Houde, et al., 2011). The lower level of feedback control codes speech information at the level of articulatory feature clusters, that is, the collection of feature values that are associated with the targets of a vocal tract opening

or closing gesture. Given that phonemes are defined in generative linguistics in terms of articulatory feature clusters (Chomsky & Halle, 1968), one can conceptualize this level of the speech motor control circuit as the most closely aligned to the theoretical construct of the phoneme (Gracco, 1994; Gracco & Lofqvist, 1994b). The circuit is hypothesized to include sensory targets coded primarily in somatosensory cortex (as suggested by V. Gracco, personal communication), motor programs coded in lower primary motor cortex (M1), and a cerebellar circuit mediating the relation between the two.

To get an intuitive sense of what the two hierarchical levels of motor speech control are doing, it is helpful to consider an analogy with a visuomotor grasping task. The target, say a cup, is provided in terms of a visual representation and the overall structure of the action is constrained by the ultimate goal of positioning the hand such that it matches the shape of the cup and is in the same location as the cup. But to actually effect the reach, the system needs information about the current position of the hand, what kind of loads might be on the limb, and so on; for this, one needs (or at least predominantly relies on) somatosensory input, as sensory deafferentation research has shown (Sanes, Mauritz, Evarts, Dalakas, & Chu, 1984). So in a sense, and oversimplifying just a bit, visual input defines the overall goal and may be associated with a coarse motor plan, while somatosensory input fine-tunes a particular implementation of the movement (there are many degrees of freedom) on a more local level given the current context (Desmurget & Grafton, 2000), which can vary with limb position, loads, fatigue and other factors. And at the very end stage of the action, somatosensory input takes primacy in determining how much force to exert as the hand closes on the cup. Speech motor control works similarly. The higher-level goal is defined by the auditory sequence that corresponds to the intended word and a coarsely coded. But in order to execute the motor commands to reproduce those sounds, the system needs to know where the speech motor effectors are in articulatory space, whether there are any unusual loads (e.g. food in the mouth), and other factors that vary depending on context. Moreover, speech acts are cyclic, alternating between open and closed vocal tract states, which sets up a motor planning situation that oscillates from one target (e.g., a particular closed position corresponding to a consonant) to another (e.g., an open position corresponding to a vowel) through a trajectory that depends on the endpoint states, resulting in coarticulation effects. The somatosensory system appears to be critical for defining the endpoints of these open-closed targets (e.g., the feel of the lips coming together) that comprise sub-goals in reproducing a particular sound sequence (Gracco & Lofqvist, 1994a; Tremblay et al., 2003). The inclusion of both auditory and somatosensory targets, as well as a cerebellar loop is not unique to the HSFC model: DIVA also includes these components but does not organize them hierarchically (Golfopoulos, Tourville, & Guenther, 2010; F.H. Guenther et al., 1998).

I will not rehash the arguments for somatosensory involvement in speech production (see, (Hickok, 2012a; Tremblay et al., 2003)), nor will I spend time on the arguments for the role of the cerebellum in the lower-level circuit (see (Hickok, 2012a)) or its possible role in auditory-motor circuits (an open question (Knolle, Schroger, Baess, & Kotz, 2012)). But I will spend some time on the observation that the theoretical definition of phonemes line up more neatly with somatosensory codes than auditory codes. Voiceless stop consonants are an interesting case-in-point because the vocal tract state, the feature bundle that define stops correspond to breaks in acoustic energy; by definition a stop blocks airflow and therefore acoustic transmission. If a phoneme has no acoustic realization, then an acoustic representation cannot define the target for such a phoneme. Therefore, phonemes – some of them at least – can't be coded in auditory space. A reviewer pointed out that while this point is amusing, it is completely incorrect because immediately following the silent period is an unvoiced release that has acoustic characteristics that very depending on the place of constriction. This is true, of course, but misses the point. The linguistic representation of the

phoneme /t/, for example, has features such as [–continuant, –voiced, +coronal/anterior], but it doesn't have the feature [+release]. So the representation that defines the phoneme linguistically, does not by itself have an auditory consequence, making it hard to think about auditory codes for stop consonants. However, we can easily think about these linguistic representations in somatosensory space because they correspond to a vocal tract state that has somatosensory consequences (relaxed larynx, pressure on tongue and alveolar ridge). Now, perhaps this characterization is extreme to the point of absurdity, but it illustrates the point that as a class phonemes are more naturally amenable to being grounded in the somatosensory than the auditory system, precisely because they are defined not by the way they sound but by vocal tract states.

This is not to say, of course, that phoneme sized units cannot have auditory representations or that these representations can't serve as auditory targets. After all, vowels and most consonants have an acoustic realization, and some words are phoneme size. The real question, from a motor control standpoint, is what a phoneme size unit is useful for controlling? Because phonemes define individual articulatory states and not sequences of states, they are best suited to control movements targeting those articulatory states, which implicates somatosensory systems. The auditory system, on the other hand, is well-suited to coding sequences of sounds, or longer time scale spectrotemporal patterns (syllables and sequences of syllables) and therefore will be better suited to controlling motor programming at the sequence level (syllable+).

Stepping back a bit, a recent large-scale lesion study provides some support for these claims. Schwartz et al. studied 106 unselected aphasic patients and documented 1718 phonological errors on a naming task. Error rates in these patients were correlated with damaged regions in the auditory dorsal stream including prominent involvement of somatosensory cortex (post-central gyrus) and a more posterior supramarginal gyrus region that included area Spt and surrounds (the authors suggested that Spt wasn't implicated but it clearly is in there analysis that excludes cases with apraxia of speech; see their figure 4). This pattern of findings is consistent with the hypothesis that auditory-motor as well as somatosensory-motor loops play a substantial in the control of speech production at the phonological level(s).

The role of the phoneme in speech perception

Phonemes are often assumed to be abstract, amodal creatures that play a role both in speech production and speech perception (Figure 5). For example, dominant models of speech recognition include phonemes as representational units (Luce, Goldinger, Auer, & Vitevitch, 2000; Marslen-Wilson & Tyler, 1980; McClelland & Elman, 1986; Norris, 1994; Stevens, 2002). The discussion above already hints at the possibility that this may not be the case. Here I will address this issue more directly by discussing three observations that have led me to question the standard view of the phoneme as a core unit in speech recognition.

One is the fact that phonemes are defined in terms of speech articulator space. This is an odd state of affairs for a unit of representation that is supposed to be amodal with equal involvement in production and perception. It makes perfect sense, however, if the phoneme is a unit of representation in a computational stream controlling speech production, i.e., the dorsal stream.

The second observation, virtually axiomatic, is that the behavioral task one chooses for studying a given system is a major factor in determining which computational networks get involved. David Poeppel and I have argued at length that even two nominally receptive speech tasks, syllable discrimination and word recognition, recruit rather different computational streams, the dorsal and ventral streams respectively (Hickok & Poeppel,

2004, 2007). Research in phonology is to a very large extent concerned with building a theory of how people speak (e.g., explaining why the same phoneme, /t/, is aspirated [t^h] in one context like *table* and unaspirated [t] in another context like *stable*), not how people listen. Given that the task employed by phonologists is effectively speech production, we can assume that the resulting theories and representational units will apply to that task. It is an empirical question whether the theories and units will apply to other tasks, such as speech perception, which brings us to observation three.

While the existence of the phoneme is well accepted in theoretical phonology, its role in speech perception has been questioned for decades. Early work questioned the existence of the phoneme and pointed to the syllable as a relevant unit instead (Massaro, 1972; Savin & Bever, 1970), an idea that has been equally challenged (Foss & Swinney, 1973; Nusbaum & DeGroot, 1991) but has persisted into more recent discussions centered on low-frequency modulation rates being a critical feature in the acoustic signal (Giraud & Poeppel, 2012; Greenberg & Arai, 2004; Poeppel, 2003). Two recent imaging studies have attempted to investigate whether subsyllabic information is represented in auditory regions, one using a phonotactic manipulation (Vaden, Piquado, & Hickok, 2011) and the other using consonant clusters to separately manipulate the number of syllables versus the number of segments (McGettigan et al., 2011); both reported subsyllabic effects in motor-related areas but not in auditory related cortex. Other authors have questioned whether there is a fundamental unit of speech perception at all, proposing instead more dynamic and context-dependent accounts of the process (Goldinger & Azuma, 2003; Grossberg, 2003; Holt, Lotto, & Kluender, 2000; Nusbaum & DeGroot, 1991). The point is that research on the perceptual side has turned up a set of empirical results and processing models that does not necessarily implicate the phoneme.

My working hypothesis regarding the place of the phoneme in speech processing is as follows. Phonemes (articulatory feature-defined segments) constitute a level of representation inherent to the computations of the sensorimotor, dorsal stream. Specifically, phonemes are defined by somatosensory-motor circuits: the articulatory targets (open or closed vocal tract positions coded in somatosensory space) for individual speech gestures and the motor routines tasked with “hitting” those targets. These representations are abstract in the sense that movement trajectories into and out of these somatosensory states will vary with the context in which they are executed, but they are still tied to the motor system. Because phonemes are defined by *somatosensory-motor* circuits for controlling speech gestures, this is not an inherent level of representation in the ventral computational stream that maps sound onto meaning.

Do we lose “parity” between auditory and motor speech systems by abandoning the phoneme in models of speech recognition? Not at all. It is true that the system must maintain a systematic mapping between auditory and motor representations of speech, but there is no reason why the phoneme has to do all the work. I suggest that the common currency is whatever units the auditory system is coding, likely something closer in level to the syllable or syllable sequence (Figure 5).

Of course, my reference to phonemes and syllables is an oversimplification, a convenient shorthand for referring to levels of representational granularity. So let me summarize and clarify my claims. For speech motor planning, I suggest that there are two kinds of sensory targets, auditory targets (the ultimate goal of a speech act) and somatosensory targets (which are useful for planning individual movement trajectories in the local context). With respect to auditory targets, I think the representational granularity can vary from individual speech sounds (defined in acoustic space), ranging from vowels and consonants that make noise to syllables to sequences of syllables. The granularity of the code that serves as an action target

likely varies with the familiarity of the sequence and/or may code targets on multiple scales. These scales may relate to inherent temporal windows of integration (Poeppel, 2003) or periodotopic maps (Barton, Venezia, Saberi, Hickok, & Brewer, 2012) in the auditory system.

Animal models and the evolution of speech motor control

One significant advantage of integrating traditional linguistics-based models of speech production with speech motor control models that are built on more general sensorimotor neural architectures is that it bridges research on speech with work on sensorimotor circuits in animal models. Research on birdsong, for example, has long been touted as a viable model of speech processing in that it involves auditory learning, auditory-motor integration and motor skill learning, and also exhibits a critical period for acquisition (Brainard & Doupe, 2013; Doupe & Kuhl, 1999). Much progress has been made in mapping the neural circuits involved in birdsong, the role of different components of the circuit for different aspects and developmental phases of the behavior, the role of social factors, and its genetic basis (Brainard & Doupe, 2013). Although there are obvious parallels with human speech, there has been a disconnect between traditional models of speech production (e.g., the two stage framework) and birdsong. However, integrating traditional psycholinguistic models with motor control architectures may facilitate comparative research by showing how a basic sensorimotor architecture common to both bird and human can be extended into high-levels of linguistic representation and processing.

Although I will not attempt a detailed comparative analysis here, it is worth highlighting one observation in the birdsong model that has not yet been developed in the (H)SFC. A cortical-basal ganglia circuit has been found to be particularly important for reinforcement learning in the birdsong model (Brainard & Doupe, 2013). Given that the basal ganglia has a similar role in learning more broadly in humans (Graybiel, 2008), that it has been incorporated into feedback control models in the manual action domain (Shadmehr & Krakauer, 2008), and that it has been included in some speech and language models (Golfenopoulos et al., 2010; Ullman, 2004), it is likely that this circuit plays an important role in the presently proposed circuits. I will leave this issue to future work.

Although birdsong is clearly a useful model for understanding aspects of the speech circuit, there are significant limitations. A major one is relatively limited repertoire of birdsong, which contrasts with the highly productive speech system. This may have consequences for the organization or the plasticity of the systems. But if the architecture of motor control circuits is similar across speech and manual control domains, then we are not restricted to animal models of vocal behaviors and can look to both human and non-human primate research on manual- or oculo-motor control as a constraining source of information on the development of speech production models.

With these links between speech production and sensorimotor control in animal work, we are in a position to carry out fairly standard comparative research that could inform theories of the evolution of speech. And, to end on a speculation, if it is possible to build sensorimotor control models beyond the word level and incorporate combinatorial processes – and I don't see a principled reason why this shouldn't be possible -- we may be well on our way to understanding the evolutionary building blocks of language. Similar suggestions have been put forward in regards to the sensorimotor foundation of speech and language from the perspective of mirror neurons (Rizzolatti & Arbib, 1998), although I would argue that placing these cells and their proposed “action understanding” function at the core of the mechanism is not empirically defensible (Hickok, 2009).

Conclusions

My aim in this work is to develop a model of speech production that integrates research in linguistics, psycholinguistics, computational motor control, neuropsychology, and neuroscience. All of these subfields represent different approaches to understanding the same fundamental problem and therefore should provide mutual constraint on theory development. The models outlined here provide some indication that this research program is fruitful and at the very least deserves further study. More broadly I think that language science could benefit from this sort of integrated approach and hope to see more attempts to bridge research approaches in the future and in computationally explicit ways.

Acknowledgments

This work was supported by a grant (DC009659) from the National Institutes of Health.

References

- Aliu SO, Houde JF, Nagarajan SS. Motor-induced suppression of the auditory cortex. *Journal of Cognitive Neuroscience*. 2009; 21(4):791–802.10.1162/jocn.2009.21055 [PubMed: 18593265]
- Almeida J, Knobel M, Finkbeiner M, Caramazza A. The locus of the frequency effect in picture naming: when recognizing is not enough. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Psychon Bull Rev*. 2007; 14(6):1177–1182. [PubMed: 18229493]
- Andersen R. Multimodal integration for the representation of space in the posterior parietal cortex. *Philosophical Transactions of the Royal Society of London B Biological Sciences*. 1997; 352:1421–1428.
- Anderson JM, Gilmore R, Roper S, Crosson B, Bauer RM, Nadeau S, Heilman KM. Conduction aphasia and the arcuate fasciculus: A reexamination of the Wernicke-Geschwind model. *Brain and Language*. 1999; 70:1–12. [PubMed: 10534369]
- Baldo JV, Klostermann EC, Dronkers NF. It's either a cook or a baker: patients with conduction aphasia get the gist but lose the trace. *Brain and Language*. 2008; 105(2):134–140. S0093-934X(07)00301-X [pii]. 10.1016/j.bandl.2007.12.007 [PubMed: 18243294]
- Barton B, Venezia JH, Saberi K, Hickok G, Brewer AA. Orthogonal acoustic dimensions define auditory field maps in human cortex. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Proc Natl Acad Sci U S A*. 2012; 109(50):20738–20743.10.1073/pnas.1213381109 [PubMed: 23188798]
- Bishop DV, Brown BB, Robson J. The relationship between phoneme discrimination, speech production, and language comprehension in cerebral palsied individuals. [Research Support, Non-U.S. Gov't]. *Journal of Speech and Hearing Research*. 1990; 33(2):210–219. [PubMed: 2359262]
- Bloom, P. How children learn the meanings of words. Cambridge, MA: MIT Press; 2000.
- Bock, K. Language production. In: Wilson, RA.; Keil, FC., editors. *The MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press; 1999. p. 453–456.
- Bohland JW, Guenther FH. An fMRI investigation of syllable sequence production. *Neuroimage*. 2006
- Brainard MS, Doupe AJ. Translating birdsong: songbirds as a model for basic and applied medical research. *Annual Review of Neuroscience*. 2013; 36:489–517.10.1146/annurev-neuro-060909-152826
- Browman CP, Goldstein L. Articulatory phonology: an overview. [Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. Review]. *Phonetica*. 1992; 49(3–4):155–180. [PubMed: 1488456]
- Buchsbaum B, Hickok G, Humphries C. Role of Left Posterior Superior Temporal Gyrus in Phonological Processing for Speech Perception and Production. *Cognitive Science*. 2001; 25:663–678.
- Buchsbaum BR, Baldo J, D'Esposito M, Dronkers N, Okada K, Hickok G. Conduction Aphasia and Phonological Short-term Memory: A Meta-Analysis of Lesion and fMRI data. *Brain and Language*. epub 2011. 10.1016/j.bandl.2010.12.001

- Buchsbaum BR, Baldo J, Okada K, Berman KF, Dronkers N, D'Esposito M, Hickok G. Conduction aphasia, sensory-motor integration, and phonological short-term memory - an aggregate analysis of lesion and fMRI data. [Research Support, N.I.H., Extramural Review]. *Brain and Language*. 2011; 119(3):119–128.10.1016/j.bandl.2010.12.001 [PubMed: 21256582]
- Burnett TA, Freedland MB, Larson CR, Hain TC. Voice F0 responses to manipulations in pitch feedback. *J Acoust Soc Am*. 1998; 103(6):3153–3161. [PubMed: 9637026]
- Burnett TA, Senner JE, Larson CR. Voice F0 responses to pitchshifted auditory feedback: a preliminary study. *J Voice*. 1997; 11(2):202–211. [PubMed: 9181544]
- Cholin J, Levelt WJ, Schiller NO. Effects of syllable frequency in speech production. *Cognition*. 2006; 99(2):205–235.10.1016/j.cognition.2005.01.009 [PubMed: 15939415]
- Chomsky, N.; Halle, M. *The sound pattern of English*. New York: Harper & Row; 1968.
- Christman SS, Boutsen FR, Buckingham HW. Perseveration and other repetitive verbal behaviors: functional dissociations. *Seminars in Speech and Language*. 2004; 25(4):295–307.10.1055/s-2004-837243 [PubMed: 15599820]
- D'Ausilio A, Pulvermuller F, Salmas P, Bufalari I, Begliomini C, Fadiga L. The motor somatotopy of speech perception. *Current Biology*. 2009a; 19(5):381–385. S0960-9822(09)00556-9 [pii]. 10.1016/j.cub.2009.01.017 [PubMed: 19217297]
- D'Ausilio A, Pulvermuller F, Salmas P, Bufalari I, Begliomini C, Fadiga L. Speech Perception May Causally Depend from the Activity of Motor Centers: Reply to Hickok. 2009b
- Damasio H, Damasio AR. The anatomical basis of conduction aphasia. *Brain*. 1980; 103:337–350. [PubMed: 7397481]
- Dell GS. A spreading activation theory of retrieval in language production. *Psychological Review*. 1986; 93:283–321. [PubMed: 3749399]
- Dell, GS. Speaking and misspeaking. In: Glietman, LR.; Liberman, M., editors. *An invitation to cognitive science: Language*. 2. Vol. 1. Cambridge, MA: MIT Press; 1995. p. 183-208.
- Dell GS, Schwartz MF, Martin N, Saffran EM, Gagnon DA. Lexical access in aphasic and nonaphasic speakers. *Psychological Review*. 1997; 104:801–838. [PubMed: 9337631]
- Delvaux V, Soquet A. The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*. 2007; 64:145–173. [PubMed: 17914281]
- Desmurget M, Grafton S. Forward modeling allows feedback control for fast reaching movements. *Trends Cogn Sci*. 2000; 4(11):423–431. [PubMed: 11058820]
- Diedrichsen J, Shadmehr R, Ivry RB. The coordination of movement: optimal feedback control and beyond. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Review]. *Trends Cogn Sci*. 2010; 14(1):31–39.10.1016/j.tics.2009.11.004 [PubMed: 20005767]
- Doupe AJ, Kuhl PK. Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*. 1999; 22:567–631.
- Dronkers, N.; Baldo, J. Language: Aphasia. In: Squire, LR., editor. *Encyclopedia of Neuroscience*. Vol. 5. Oxford: Academic Press; 2009. p. 343-348.
- Duffy, JR. *Motor speech disorders: Substrates, Differential diagnosis, and Management*. St Louis: Mosby; 1995.
- Edwards E, Nagarajan SS, Dalal SS, Canolty RT, Kirsch HE, Barbaro NM, Knight RT. Spatiotemporal imaging of cortical activation during verb generation and picture naming. [Research Support, N.I.H., Extramural]. *Neuroimage*. 2010; 50(1):291–301.10.1016/j.neuroimage.2009.12.035 [PubMed: 20026224]
- Eimas PD, Siqueland ER, Jusczyk P, Vigorito J. Speech perception in infants. *Science*. 1971; 171(968):303–306. [PubMed: 5538846]
- Fadiga L, Craighero L, D'Ausilio A. Broca's area in language, action, and music. [Research Support, Non-U.S. Gov't Review]. *Ann N Y Acad Sci*. 2009; 1169:448–458.10.1111/j.1749-6632.2009.04582.x [PubMed: 19673823]
- Fairbanks G. Systematic research in experimental phonetics: 1. A theory of the speech mechanism as a servosystem. *Journal of Speech and Hearing Disorders*. 1954; 19:133–139.

- Fogassi L, Gallese V, Buccino G, Craighero L, Fadiga L, Rizzolatti G. Cortical mechanism for the visual guidance of hand grasping movements in the monkey: A reversible inactivation study. *Brain*. 2001; 124(Pt 3):571–586. [PubMed: 11222457]
- Foss DJ, Swinney DA. On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*. 1973; 12:246–257.
- Friston K. The free-energy principle: a unified brain theory? [Research Support, Non-U.S. Gov't Review]. *Nature Reviews Neuroscience*. 2010; 11(2):127–138.10.1038/nrn2787
- Fromkin V. The non-anomalous nature of anomalous utterances. *Language*. 1971; 47:27–52.
- Garrett, MF. The analysis of sentence production. In: Bower, GH., editor. *The psychology of learning and motivation*. Volume 9: advances in research and theory. New York: Academic Press; 1975. p. 133-177.
- Giraud AL, Poeppel D. Cortical oscillations and speech processing: emerging computational principles and operations. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Nat Neurosci*. 2012; 15(4):511–517.10.1038/nn.3063 [PubMed: 22426255]
- Goldinger SD, Azuma T. Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics*. 2003; 31:305–320.
- Golfinopoulos E, Tourville JA, Guenther FH. The integration of largescale neural network modeling and functional brain imaging in speech motor control. *Neuroimage*. 2010; 52(3):862–874. S1053-8119(09)01094-5 [pii]. 10.1016/j.neuroimage.2009.10.023 [PubMed: 19837177]
- Goodglass, H. Diagnosis of conduction aphasia. In: Kohn, SE., editor. *Conduction aphasia*. Hillsdale, N.J: Lawrence Erlbaum Associates; 1992. p. 39-49.
- Gracco VL. Some organizational characteristics of speech movement control. *J Speech Hear Res*. 1994; 37(1):4–27. [PubMed: 8170129]
- Gracco VL, Lofqvist A. Speech motor coordination and control: evidence from lip, jaw, and laryngeal movements. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]. *J Neurosci*. 1994a; 14(11 Pt 1):6585–6597. [PubMed: 7965062]
- Gracco VL, Lofqvist A. Speech motor coordination and control: evidence from lip, jaw, and laryngeal movements. *Journal of Neuroscience*. 1994b; 14(11 Pt 1):6585–6597. [PubMed: 7965062]
- Grafton ST. The cognitive neuroscience of prehension: recent developments. [Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. Review]. *Exp Brain Res*. 2010; 204(4): 475–491.10.1007/s00221-010-2315-2 [PubMed: 20532487]
- Grafton, ST.; Aziz-Zadeh, L.; Ivry, RB. Relative hierarchies and the representation of action. In: Gazzaniga, MS., editor. *The cognitive neurosciences*. 4. Cambridge, MA: MIT Press; 2009. p. 641-652.
- Grafton ST, Hamilton AF. Evidence for a distributed hierarchy of action representation in the brain. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Review]. *Hum Mov Sci*. 2007; 26(4):590–616.10.1016/j.humov.2007.05.009 [PubMed: 17706312]
- Graybiel AM. Habits, rituals, and the evaluative brain. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Review]. *Annual Review of Neuroscience*. 2008; 31:359–387.10.1146/annurev.neuro.29.051605.112851
- Greenberg, S. Understanding speech understanding: towards a unified theory of speech perception. Paper presented at the Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception; Keele, England. 1996.
- Greenberg S, Arai T. What are the essential cues for understanding spoken language? *IEICE Transactions on Information and Systems*. 2004; E87-D:1059–1070.
- Grefkes C, Fink GR. The functional organization of the intraparietal sulcus in humans and monkeys. *J Anat*. 2005; 207(1):3–17. JOA426 [pii]. 10.1111/j.1469-7580.2005.00426.x [PubMed: 16011542]
- Grossberg S. Resonant neural dynamics of speech perception. *Journal of Phonetics*. 2003; 31:423–445.
- Guenther FH. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. [Research Support, U.S. Gov't, Non-P.H.S. Review]. *Psychol Rev*. 1995; 102(3):594–621. [PubMed: 7624456]
- Guenther FH. Cortical interactions underlying the production of speech sounds. [Research Support, N.I.H., Extramural Review]. *J Commun Disord*. 2006; 39(5):350–365.10.1016/j.jcomdis.2006.06.013 [PubMed: 16887139]

- Guenther FH, Espy-Wilson CY, Boyce SE, Matthies ML, Zandipour M, Perkell JS. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. [Comparative Study Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.]. *J Acoust Soc Am*. 1999; 105(5):2854–2865. [PubMed: 10335635]
- Guenther FH, Ghosh SS, Tourville JA. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang*. 2006; 96(3):280–301. [PubMed: 16040108]
- Guenther FH, Hampson M, Johnson D. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*. 1998; 105:611–633. [PubMed: 9830375]
- Halsband U, Lange RK. Motor learning in man: a review of functional and clinical studies. [Review]. *J Physiol Paris*. 2006; 99(4–6):414–424.10.1016/j.jphysparis.2006.03.007 [PubMed: 16730432]
- Hanley JR, Dell GS, Kay J, Baron R. Evidence for the involvement of a nonlexical route in the repetition of familiar words: A comparison of single and dual route models of auditory repetition. *Cogn Neuropsychol*. 2004; 21(2):147–158.10.1080/02643290342000339 [PubMed: 21038197]
- Hartsuiker, RJ.; Kolk, HHJ.; Martensen, H. The division of labor between internal and external speech monitoring. In: Hartsuiker, RJ.; Bastiaanse, R.; Postma, A.; Wijnen, F., editors. *Phonological encoding and monitoring in normal and pathological speech*. New York: Psychology Press; 2005. p. 187-205.
- Haruno M, Wolpert DM, Kawato M. Hierarchical MOSAIC for movement generation. *International Congress Series*. 2003; 1250:575–590.
- Heinks-Maldonado TH, Nagarajan SS, Houde JF. Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport*. 2006; 17(13):1375–1379. [pii]. 10.1097/01.wnr.0000233102.43526.e900001756-200609180-00002 [PubMed: 16932142]
- Hickok, G. Speech perception, conduction aphasia, and the functional neuroanatomy of language. In: Grodzinsky, Y.; Shapiro, L.; Swinney, D., editors. *Language and the brain*. San Diego: Academic Press; 2000. p. 87-104.
- Hickok G. Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *J Cogn Neurosci*. 2009; 21(7):1229–1243. [pii]. 10.1162/jocn.2009.2118910.1162/jocn.2009.21189 [PubMed: 19199415]
- Hickok G. Computational neuroanatomy of speech production. [Research Support, N.I.H., Extramural]. *Nature Reviews Neuroscience*. 2012a; 13(2):135–145.10.1038/nrn3158
- Hickok G. The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *J Commun Disord*. 2012b10.1016/j.jcomdis.2012.06.004
- Hickok G, Buchsbaum B, Humphries C, Muftuler T. Auditory-motor interaction revealed by fMRI: Speech, music, and working memory in area Spt. *Journal of Cognitive Neuroscience*. 2003; 15:673–682. [PubMed: 12965041]
- Hickok G, Costanzo M, Capasso R, Miceli G. The role of Broca's area in speech perception: Evidence from aphasia revisited. *Brain and Language*. 2011; 119(3):214–220.10.1016/j.bandl.2011.08.001 [PubMed: 21920592]
- Hickok G, Erhard P, Kassubek J, Helms-Tillery AK, Naeve-Velguth S, Strupp JP, Ugurbil K. A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: implications for the explanation of conduction aphasia. *Neuroscience Letters*. 2000; 287:156–160. [PubMed: 10854735]
- Hickok G, Houde J, Rong F. Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*. 2011; 69(3):407–422. S0896-6273(11)00067-5 [pii]. 10.1016/j.neuron.2011.01.019 [PubMed: 21315253]
- Hickok G, Okada K, Barr W, Pa J, Rogalsky C, Donnelly K, Grant A. Bilateral capacity for speech sound processing in auditory comprehension: evidence from Wada procedures. *Brain and Language*. 2008; 107(3):179–184. S0093-934X(08)00126-0 [pii]. 10.1016/j.bandl.2008.09.006 [PubMed: 18976806]
- Hickok G, Okada K, Serences JT. Area Spt in the human planum temporale supports sensory-motor integration for speech processing. *Journal of Neurophysiology*. 2009; 101(5):2725–2732. 91099.2008 [pii]. 10.1152/jn.91099.2008 [PubMed: 19225172]
- Hickok G, Poeppel D. Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*. 2000; 4:131–138. [PubMed: 10740277]

- Hickok G, Poeppel D. Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*. 2004; 92:67–99. [PubMed: 15037127]
- Hickok G, Poeppel D. The cortical organization of speech processing. *Nature Reviews Neuroscience*. 2007; 8(5):393–402.
- Holt LL, Lotto AJ, Kluender KR. Neighboring spectral content influences vowel identification. [Research Support, U.S. Gov't, Non-P.H.S.]. *Journal of the Acoustical Society of America*. 2000; 108(2):710–722. [PubMed: 10955638]
- Houde JF, Jordan MI. Sensorimotor adaptation in speech production. *Science*. 1998; 279:1213–1216. [PubMed: 9469813]
- Houde JF, Nagarajan SS. Speech production as state feedback control. [Review]. *Frontiers in Human Neuroscience*. 2011; 510.3389/fnhum.2011.00082
- Houde JF, Nagarajan SS, Sekihara K, Merzenich MM. Modulation of the auditory cortex during speech: an MEG study. *J Cogn Neurosci*. 2002; 14(8):1125–1138.10.1162/089892902760807140 [PubMed: 12495520]
- Howard D, Nickels L. Separating input and output phonology: Semantic, phonological, and orthographic effects in short-term memory impairment. *Cognitive Neuropsychology*. 2005; 22:42–77. [PubMed: 21038240]
- Huetting F, Hartsuiker RJ. Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes*. 2010; 25:347–374.
- Jackson JH. Remarks on Evolution and Dissolution of the Nervous System. *Journal of Mental Science*. 1887; 33:25–48.
- Jacquemot C, Dupoux E, Bachoud-Levi AC. Breaking the mirror: Asymmetrical disconnection between the phonological input and output codes. *Cognitive Neuropsychology*. 2007; 24(1):3–22. 769783066 [pii]. 10.1080/02643290600683342 [PubMed: 18416481]
- Jazayeri M, Movshon JA. Optimal representation of sensory information by neural populations. *Nat Neurosci*. 2006; 9(5):690–696. nn1691 [pii]. 10.1038/nn1691 [PubMed: 16617339]
- Jazayeri M, Movshon JA. A new perceptual illusion reveals mechanisms of sensory decoding. [Research Support, N.I.H., Extramural]. *Nature*. 2007; 446(7138):912–915.10.1038/nature05739 [PubMed: 17410125]
- Jeannerod M, Arbib MA, Rizzolatti G, Sakata H. Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends Neurosci*. 1995; 18(7):314–320. 016622369593921J [pii]. [PubMed: 7571012]
- Jescheniak JD, Levelt WJM. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1994; 20:824–843.
- Kappes J, Baumgaertner A, Peschke C, Ziegler W. Unintended imitation in nonword repetition. [Research Support, Non-U.S. Gov't]. *Brain Lang*. 2009; 111(3):140–151.10.1016/j.bandl.2009.08.008 [PubMed: 19811813]
- Kawato M. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*. 1999a; 9(6):718–727. S0959-4388(99)00028-8 [pii]. [PubMed: 10607637]
- Kawato M. Internal models for motor control and trajectory planning. [Research Support, Non-U.S. Gov't Review]. *Curr Opin Neurobiol*. 1999b; 9(6):718–727. [PubMed: 10607637]
- Knolle F, Schroger E, Baess P, Kotz SA. The cerebellum generates motor-to-auditory predictions: ERP lesion evidence. [Research Support, Non-U.S. Gov't]. *Journal of Cognitive Neuroscience*. 2012; 24(3):698–706.10.1162/jocn_a_00167 [PubMed: 22098261]
- Kuhl PK. Brain mechanisms in early language acquisition. [Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. Review]. *Neuron*. 2010; 67(5):713–727.10.1016/j.neuron.2010.08.038 [PubMed: 20826304]
- Kuhl PK, Meltzoff AN. Infant vocalizations in response to speech: vocal imitation and developmental change. [Research Support, U.S. Gov't, P.H.S.]. *J Acoust Soc Am*. 1996; 100(4 Pt 1):2425–2438. [PubMed: 8865648]
- Kuhl PK, Miller JD. Speech perception by the chinchilla: Voicedvoiceless distinction in alveolar plosive consonants. *Science*. 1975; 190:69–72. [PubMed: 1166301]

- Lackner, JR.; Tuller, BH. Role of efference monitoring in the detection of self-produced speech errors. In: Cooper, WE.; Walker, ECT., editors. Sentence processing. Hillsdale, N.J: Lawrence Earlbaum Associates, Inc; 1979. p. 281-294.
- Larson CR, Burnett TA, Bauer JJ, Kiran S, Hain TC. Comparison of voice F0 responses to pitch-shift onset and offset conditions. *Journal of the Acoustical Society of America*. 2001; 110(6):2845–2848. [PubMed: 11785786]
- Lenneberg EH. Understanding language without ability to speak: a case report. *Journal of Abnormal and Social Psychology*. 1962; 65:419–425. [PubMed: 13929636]
- Levelt WJ. Monitoring and self-repair in speech. *Cognition*. 1983; 14(1):41–104. [PubMed: 6685011]
- Levelt WJ, Wheeldon L. Do speakers have access to a mental syllabary? *Cognition*. 1994; 50(1–3): 239–269. [PubMed: 8039363]
- Levelt, WJM. Speaking: From intention to articulation. Cambridge, MA: MIT Press; 1989.
- Levelt WJM. Models of word production. *Trends in Cognitive Sciences*. 1999; 3:223–232. [PubMed: 10354575]
- Levelt WJM, Praamstra P, Meyer AS, Helenius P, Salmelin R. An MEG study of picture naming. *Journal of Cognitive Neuroscience*. 1998; 10:553–567. [PubMed: 9802989]
- Levelt WJM, Roelofs A, Meyer AS. A theory of lexical access in speech production. *Behavioral & Brain Sciences*. 1999; 22(1):1–75. [PubMed: 11301520]
- Lichtheim L. On aphasia. *Brain*. 1885; 7:433–484.
- Luce PA, Goldinger SD, Auer ET Jr, Vitevitch MS. Phonetic priming, neighborhood activation, and PARSYN. *Percept Psychophys*. 2000; 62(3):615–625. [PubMed: 10909252]
- Marslen-Wilson W, Tyler LK. The temporal structure of spoken language understanding. *Cognition*. 1980; 8(1):1–71. [PubMed: 7363578]
- Martin RC, Lesch MF, Bartha MC. Independence of input and output phonology in word processing and short-term memory. *Journal of Memory and Language*. 1999; 41:3–29.
- Massaro DW. Preperceptual images, processing time, and perceptual units in auditory perception. *Psychol Rev*. 1972; 79(2):124–145. [PubMed: 5024158]
- McClelland JL, Elman JL. The TRACE model of speech perception. *Cognitive Psychology*. 1986; 18:1–86. [PubMed: 3753912]
- McGettigan C, Warren JE, Eisner F, Marshall CR, Shanmugalingam P, Scott SK. Neural correlates of sublexical processing in phonological working memory. [Research Support, Non-U.S. Gov't]. *J Cogn Neurosci*. 2011; 23(4):961–977.10.1162/jocn.2010.21491 [PubMed: 20350182]
- Meister IG, Wilson SM, Deblieck C, Wu AD, Iacoboni M. The essential role of premotor cortex in speech perception. *Current Biology*. 2007; 17(19):1692–1696. S0960-9822(07)01969-0 [pii]. 10.1016/j.cub.2007.08.064 [PubMed: 17900904]
- Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*. 1951; 41:329–335. [PubMed: 14861384]
- Milner, AD.; Goodale, MA. The visual brain in action. Oxford: Oxford University Press; 1995.
- Motley MT, Camden CT, Baars BJ. Covert formulation and editing of anomalies in speech production: Evidence from experimentally elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*. 1982; 21:578–594.
- Nickels L, Howard D. Phonological errors in aphasic naming: comprehension, monitoring and lexicality. *Cortex*. 1995; 31(2):209–237. [PubMed: 7555004]
- Niziolek CA, Guenther FH. Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *J Neurosci*. 2013; 33(29):12090–12098.10.1523/JNEUROSCI.1008-13.2013 [PubMed: 23864694]
- Norris D. Shortlist: A connectionist model of continuous speech recognition. *Cognition*. 1994; 52:189–234.
- Nozari N, Dell GS, Schwartz MF. Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. [Research Support, N.I.H., Extramural]. *Cognitive Psychology*. 2011; 63(1):1–33.10.1016/j.cogpsych.2011.05.001 [PubMed: 21652015]

- Nozari N, Kittredge AK, Dell GS, Schwartz MF. Naming and repetition in aphasia: Steps, routes, and frequency effects. *J Mem Lang*. 2010; 63(4):541–559.10.1016/j.jml.2010.08.001 [PubMed: 21076661]
- Nusbaum, HC.; DeGroot, J. The role of syllables in speech perception. In: Ziolkowski, MS.; Noske, M.; Deaton, K., editors. *Papers from the parasession on the syllable in phonetics and phonology*. Chicago: Chicago Linguistic Society; 1991.
- Obleser J, Wise RJ, Alex Dresner M, Scott SK. Functional integration across brain regions improves speech perception under adverse listening conditions. [Research Support, Non-U.S. Gov't]. *J Neurosci*. 2007; 27(9):2283–2289.10.1523/JNEUROSCI.4663-06.2007 [PubMed: 17329425]
- Oldfield RC, Wingfield A. Response latencies in naming objects. *Q J Exp Psychol*. 1965; 17(4):273–281. [PubMed: 5852918]
- Oppenheim GM, Dell GS. Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. [Research Support, N.I.H., Extramural]. *Cognition*. 2008; 106(1):528–537.10.1016/j.cognition.2007.02.006 [PubMed: 17407776]
- Ozdemir R, Roelofs A, Levelt WJ. Perceptual uniqueness point effects in monitoring internal speech. *Cognition*. 2007; 105(2):457–465. S0010-0277(06)00214-9 [pii]. 10.1016/j.cognition.2006.10.006 [PubMed: 17156770]
- Perkell J, Matthies M, Lane H, Guenther F, Wilhelms-Tricarico R, Wozniak J, Guind P. Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Communication*. 1997; 22:227–250.
- Perkell JS. Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics*. 2012; 25(5):382–407.10.1016/j.jneuroling.2010.02.011 [PubMed: 22661828]
- Plaut, DC.; Kello, CT. The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In: MacWhinney, B., editor. *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates; 1999. p. 381–416.
- Poeppel D. The analysis of speech in different temporal integration windows: cerebral lateralization as “asymmetric sampling in time”. *Speech Communication*. 2003; 41:245–255.
- Poljak S. The connections of the acoustic nerve. *Journal of Anatomy*. 1926; 60:465–469.
- Postma A. Detection of errors during speech production: a review of speech monitoring models. *Cognition*. 2000; 77(2):97–132. S0010-0277(00)00090-1 [pii]. [PubMed: 10986364]
- Preilowski B. Phases of motor skills acquisition: a neuropsychological approach. *Human Movement Studies*. 1977; 3:169–181.
- Rauschecker JP. Cortical processing of complex sounds. *Current Opinion in Neurobiology*. 1998; 8(4): 516–521. [PubMed: 9751652]
- Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*. 2009; 12(6):718–724. nn.2331 [pii]. 10.1038/nn.2331
- Regan D, Beverley KI. Postadaptation orientation discrimination. *J Opt Soc Am A*. 1985; 2(2):147–155. [PubMed: 3973752]
- Rizzolatti G, Arbib M. Language within our grasp. *Trends in Neurosciences*. 1998; 21:188–194. [PubMed: 9610880]
- Rogalsky C, Love T, Driscoll D, Anderson SW, Hickok G. Are mirror neurons the basis of speech perception? Evidence from five cases with damage to the purported human mirror system. *Neurocase*. 2011; 17(2):178–187. 931806807 [pii]. 10.1080/13554794.2010.509318 [PubMed: 21207313]
- Rogalsky C, Pitz E, Hillis AE, Hickok G. Auditory word comprehension impairment in acute stroke: relative contribution of phonemic versus semantic factors. *Brain and Language*. 2008; 107(2): 167–169. S0093-934X(08)00111-9 [pii]. 10.1016/j.bandl.2008.08.003 [PubMed: 18823655]
- Rumelhart, DE.; Hinton, GE.; McClelland, JL. A general framework for parallel distributed processing. In: Rumelhart, DE.; McClelland, JL., editors. *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press; 1986. p. 45–76.
- Sancier ML, Fowler CA. Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*. 1997; 25:421–436.

- Sanes JN, Mauritz KH, Evarts EV, Dalakas MC, Chu A. Motor deficits in patients with large-fiber sensory neuropathy. [Research Support, Non-U.S. Gov't]. *Proc Natl Acad Sci U S A*. 1984; 81(3):979–982. [PubMed: 6322181]
- Savin HB, Bever TG. The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*. 1970; 9:295–302.
- Schneider GE. Two visual systems. *Science*. 1969; 163(3870):895–902. [PubMed: 5763873]
- Scolari M, Serences JT. Adaptive allocation of attentional gain. *J Neurosci*. 2009; 29(38):11933–11942. 29/38/11933 [pii]. 10.1523/JNEUROSCI.5642-08.2009 [PubMed: 19776279]
- Shadmehr R, Krakauer JW. A computational neuroanatomy for motor control. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Review]. *Exp Brain Res*. 2008; 185(3):359–381.10.1007/s00221-008-1280-5 [PubMed: 18251019]
- Shadmehr R, Mussa-Ivaldi FA. Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience*. 1994; 14(5 Pt 2):3208–3224. [PubMed: 8182467]
- Shadmehr R, Smith MA, Krakauer JW. Error correction, sensory prediction, and adaptation in motor control. [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Review]. *Annual Review of Neuroscience*. 2010; 33:89–108.10.1146/annurev-neuro-060909-153135
- Shelton JR, Caramazza A. Deficits in lexical and semantic processing: Implications for models of normal language. *Psychonomic Bulletin & Review*. 1999; 6:5–27. [PubMed: 12199314]
- Stevens KN. Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustic Society of America*. 2002; 111:1872–1891.
- Stuart A, Kalinowski J, Rastatter MP, Lynch K. Effect of delayed auditory feedback on normal speakers at two speech rates. *J Acoust Soc Am*. 2002; 111(5 Pt 1):2237–2241. [PubMed: 12051443]
- Summerfield C, Egner T. Expectation (and attention) in visual cognition. [Review]. *Trends Cogn Sci*. 2009; 13(9):403–409.10.1016/j.tics.2009.06.003 [PubMed: 19716752]
- Tian X, Poeppel D. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*. 2010; 1:166.10.3389/fpsyg.2010.00166 [PubMed: 21897822]
- Tourville JA, Reilly KJ, Guenther FH. Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*. 2008; 39(3):1429–1443. S1053-8119(07)00883-X [pii]. 10.1016/j.neuroimage.2007.09.054 [PubMed: 18035557]
- Tremblay S, Shiller DM, Ostry DJ. Somatosensory basis of speech production. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]. *Nature*. 2003; 423(6942):866–869.10.1038/nature01710 [PubMed: 12815431]
- Trevarthen CB. Two mechanisms of vision in primates. *Psychol Forsch*. 1968; 31(4):299–348. [PubMed: 4973634]
- Ullman MT. Contributions of memory circuits to language: the declarative/procedural model. *Cognition*. 2004; 92(1–2):231–270. [pii]. 10.1016/j.cognition.2003.10.008S0010027703002324 [PubMed: 15037131]
- Ungerleider, LG.; Mishkin, M. Two cortical visual systems. In: Ingle, DJ.; Goodale, MA.; Mansfield, RJW., editors. *Analysis of visual behavior*. Cambridge, MA: MIT Press; 1982. p. 549–586.
- Vaden KI Jr, Piquado T, Hickok G. Sublexical properties of spoken words modulate activity in Broca's area but not superior temporal cortex: implications for models of speech recognition. [Research Support, N.I.H., Extramural]. *J Cogn Neurosci*. 2011; 23(10):2665–2674.10.1162/jocn.2011.21620 [PubMed: 21261450]
- Ventura MI, Nagarajan SS, Houde JF. Speech target modulates speaking induced suppression in auditory cortex. *BMC Neuroscience*. 2009; 10:58. 1471-2202-10-58 [pii]. 10.1186/1471-2202-10-58 [PubMed: 19523234]
- Vigliocco G, Antonini T, Garrett MF. Grammatical gender is on the tip of Italian tongues. *Psychological Science*. 1998; 8:314–317.
- Vigliocco G, Hartsuiker RJ. The interplay of meaning, sound, and syntax in sentence production. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Review]. *Psychological Bulletin*. 2002; 128(3):442–472. [PubMed: 12002697]

- Wernicke, C. The symptom complex of aphasia: A psychological study on an anatomical basis. In: Cohen, RS.; Wartofsky, MW., editors. Boston studies in the philosophy of science. Dordrecht: D. Reidel Publishing Company; 1874/1969. p. 34-97.
- Wernicke, C. Der aphasische symptom-complex: Eine psychologische studie auf anatomischer basis. In: Eggert, GH., editor. Wernicke's works on aphasia: A sourcebook and review. The Hague: Mouton; 1874/1977. p. 91-145.
- Wilson SM. Speech perception when the motor system is compromised. Trends Cogn Sci. 2009; 13:329–330. [PubMed: 19646917]
- Wilson SM, Saygin AP, Sereno MI, Iacoboni M. Listening to speech activates motor areas involved in speech production. Nature Neuroscience. 2004; 7:701–702.
- Wolpert DM, Doya K, Kawato M. A unifying computational framework for motor control and social interaction. Philos Trans R Soc Lond B Biol Sci. 2003; 358(1431):593–602.10.1098/rstb.2002.1238 [PubMed: 12689384]
- Wolpert DM, Ghahramani Z, Jordan MI. An internal model for sensorimotor integration. [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.]. Science. 1995; 269(5232): 1880–1882. [PubMed: 7569931]
- Yates AJ. Delayed auditory feedback. Psychological Bulletin. 1963; 60:213–251. [PubMed: 14002534]

Figure 1. State feedback control

State feedback control models typically include a motor controller that sends commands to a motor effector, which in turn result in a change of state, such as a change in the position of an arm. State changes are detected by sensory systems. Most models also include an internal forward model that receives a copy of the motor command issued by the controller and generates a prediction of the sensory consequences of the command that can be compared against the measured sensory consequences. The error between the predicted and measured sensory consequences is used as a correction signal to correct such error. Image adapted from Shadmehr and Krakauer 2008.

Figure 2. Two-stage psycholinguistic model of speech production

Psycholinguistic models of speech production typically identify two major linguistic stages of processing, the word (or lemma) stage in which an abstract word form without phonological specification is coded and the phonological stage in which the phonological form of the word is coded. The distinction between these stages can be intuitively understood when considering tip-of-the-tongue states in which we know the word we want to use (that is, we have accessed the lemma) but we cannot retrieve the phonological form. These linguistic stages of processing receive input from the conceptual system and send output to the motor articulatory system. Conceptual and articulatory processes are typically considered outside the domain of linguistic analysis of speech production.

Communication between the auditory and motor systems is achieved by an auditory-motor translation system. The two stages of the psycholinguistic model are evident in the lexical-conceptual system, which is intended to represent, in part, the lemma level, and the motor-auditory phonological systems, which correspond to the phonological level. Reprinted with permission.

Figure 4. The hierarchical state feedback control (HSFC) model

The HSFC model includes two hierarchical levels of feedback control, each with its own internal and external sensory feedback loops. As in psycholinguistic models, the input to the HSFC model starts with the activation of a conceptual representation that in turn excites a corresponding word (lemma) representation. The word level projects in parallel to sensory and motor sides of the highest, fully cortical level of feedback control, the auditory–Spt–BA44 loop. This higher-level loop in turn projects, also in parallel, to the lower-level somatosensory–cerebellum–motor cortex loop. Direct connections between the word level and the lower-level circuit may also exist, although they are not depicted here. The HSFC model differs from the state feedback control (SFC) model in two main respects. First, ‘phonological’ processing is distributed over two hierarchically organized levels implicating a higher-level cortical auditory–motor circuit and a lower-level somatosensory–motor circuit, which roughly map onto syllabic and phonemic levels of analysis, respectively. Second, a true efference copy signal is not a component of the model. Instead, the function served by an efference copy is integrated into the motor planning process. BA, Brodmann area; M1, primary motor cortex; S1, primary somatosensory area; aSMG, anterior

supramarginal gyrus; STG, superior temporal gyrus; STS, superior temporal sulcus; vBA6, ventral BA6.

Figure 5. Schematic models of speech processing

There is agreement that the speech network must interface auditory, motor, and conceptual systems, and there is reasonable convergence on the existence of several intermediate levels of representation including (but not limited to) phonemes, syllables, syllable sequences (such as words), and lemmas (syntactic levels are ignored here). (A). A standard model of how these intermediate levels are arranged with phoneme level representations functioning as the hub for all three “spokes” of the network. (B). An alternative that situations phoneme level representations within the motor spoke, or processing stream, and with a higher level phonological representation serves as the hub of the network.