LANGUAGE AND COGNITIVE PROCESSES
0000, 00 (00), 1–17

Psychology Press
Taylor & Francis Group

Feedforward and feedback in speech perception: Revisiting analysis by synthesis

David Poeppel[1] and Philip J. Monahan[2]

[1]*Department of Psychology, New York University, NY, USA,* [2]*Basque Center on Cognition, Brain, and Language, San Sebastian, Spain*

We revisit the analysis by synthesis (A × S) approach to speech recognition. In the late 1950s and 1960s, Stevens and Halle proposed a model of spoken word recognition in which candidate word representations were synthesised from brief cues in the auditory signal and analysed against the input signal in tightly linked bottom-up/top-down fashion. While this approach failed to garner much support at the time, recent years have brought a surge of interest in Bayesian approaches to perception, and the idea of A × S has consequently gained attention, particularly in the domain of visual perception. We review the model and illustrate some data from speech perception that are well-accounted for in the context of such an architecture. We focus on prediction in speech perception, an operation at the centre of the A × S algorithm. The data reviewed here and the current possibilities to study online measures of speech processing using cognitive neuroscience methods, in our view, add to a provocative series of arguments why A × S should be reconsidered as a contender in speech recognition research, complementing currently more dominant models.

Keywords: Perception; Bayesian; Cognitive neuroscience.

INTRODUCTION

The extent to which knowledge of language, a high-order property of the linguistic cognitive system, influences lower levels of perceptual analysis of the speech signal has been an energetically debated topic in the efforts to understand the psychological and biological computations that underlie speech recognition (McClelland & Elman, 1986; Norris, McQueen, & Cutler, 2000). It is a commonplace—and intuitively straightforward—assumption

Correspondence should be addressed to David Poeppel, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA. E-mail: david.poeppel@nyu.edu

http://www.psypress.com/lcp DOI: 10.1080/01690965.2010.493301

that representations constructed at earlier stages of processing feed immediately higher levels in a *feedforward* manner and that this process proceeds incrementally until access to a "lexical–conceptual" representation has been achieved. In speech recognition (which we take here to be comprised of the set of algorithms responsible for mapping the time-varying acoustic waveform onto lexical representations), this involves a conversion from acoustic features onto phonetic representations, phonetic representations onto phonological representations, and finally access of the lexical item based on its phonological structure. One might think of this, schematically, as *the transformation from vibrations in the ear to abstractions in the brain*. In an alternative architecture, higher order knowledge (in the case of speech recognition, phonological, and lexical structure) constrains and facilitates earlier stages of processing through *feedback* mechanisms. That is, our knowledge of phonological sound patterns and lexical items aids in the perceptual mapping from acoustic to linguistic information. The feedforward vs. feedback debate—and the different hypothesised mechanisms—plays an important role in theories of cognitive architectures more generally, and a comprehensive theory of the ubiquitous process of speech recognition requires a well-motivated answer to the various issues arising in this context. Here we discuss a surprisingly old proposal, one that has received little attention both in the psychological and engineering literatures, until recently, where the idea has been resuscitated in other areas of perception, notably high-level vision.

In the late 1950s and 1960s, Morris Halle and Ken Stevens from MIT published a series of short articles arguing for a model of speech recognition that, at its core, incorporated a procedure called "analysis by synthesis (A × S)" (e.g., Halle & Stevens, 1959; Halle & Stevens, 1962; Stevens, 1960; Stevens & Halle, 1967). While related to concurrently developed motor theories, the A × S approach was distinct in positing an *active*, *top-down* process in which potential signal patterns were internally generated (synthesis) and compared to the incoming signal. Thus, perceptual analysis crucially contained a step of synthetically generating candidate representations (a form of a hypothesis-and-test model). Whereas, motor theories were concerned with identifying the underlying articulatory gestures related to the signal pattern (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), the A × S model proceeded from the assumption that cues from the input signal triggered guesses about the identity of phonemes, and subsequently, the internal synthesis of *potential* phonemes was then *compared* to the input sequence. In light of recent interest in the role of feedback in various domains in cognitive science and cognitive neuroscience (e.g., visual object recognition) and the popularity of Bayesian approaches to perception (Geisler & Diehl, 2003; Geisler & Kersten, 2002; Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996; Mamassian, Landy, &

Maloney, 2002), we revisit the A × S model of speech perception, a model that has garnered little attention in the past 40 years, but includes *feedforward* (in the form of cue extraction and hypothesis generation) *and feedback* (in the form of synthesis and comparison) as primary mechanisms for recognition and can be modelled using Bayesian approaches (Poeppel, Idsardi, & van Wassenhove, 2008; Yuille & Kersten, 2006). (See Bever & Poeppel, in press, for a recent more general perspective on A × S approaches in psycholinguistics.)

We see a strength of the A × S model in the fact that—both in spirit and in terms of its constituent operations—the idea lies at the nexus of a group of concepts that are currently widely discussed in perception, engineering, and computational neuroscience. Specifically, the notion of *predictive coding* has received considerable attention in systems neuroscience (Bar, 2009; Schultz & Dickinson, 2000). Similarly, the use of *internal forward models* plays a central role in the study of motor control and motor planning (Wolpert & Ghahramani, 2004), and a unifying framework for these families of ideas has been put forth in the form of *Bayesian modelling of perceptual processing* (Knill & Richards, 1996). A × S is an internal forward model that capitalises on predictive coding (the synthesis stage), and is arguably well-captured by the formalism provided by Bayesian approaches (Poeppel et al., 2008). For example, the manner in which "guesses"/ hypotheses are generated in the A × S loop (described below) is conditioned by the current state and probabilistic estimates of the next processing step. A Bayesian approach to such problems could be effectively used to characterise some of the core subroutines of the A × S algorithm. In short, the model we describe may constitute a bridge between areas of research that have proceeded in a somewhat isolated fashion in the investigation of speech recognition. A × S thus provides a potential framework to link established concepts from engineering, systems neuroscience, and speech research.

BACKGROUND

As Halle and Stevens (1962, p. 155) stated the issue: "the fundamental problem in pattern recognition is the search for a *recognition function* that will appropriately pair *signals* and *messages*". For them, an *active feedback* process mediates this pairing, and the *analysis* of the speech signal is largely accomplished by *synthesising* candidate representations. Lexical items are composed of a string of discretised segments. These segments are abstract in nature (i.e., they need not bear any direct relationship to actual speech events) and serve to relate the acoustic and articulatory properties of sounds. Moreover, the segments themselves are not monolithic units but

instead composed of a set of "distinctive features" that provide instructions about articulator movements to the production system (e.g., position of the tongue in the oral cavity; whether or not the vocal folds vibrate in the laryngeal cavity, etc.). In addition to the extensive evidence from linguistic research that distinctive features (rather than segments or phonemes) comprise the smallest building blocks of speech sounds, such features have the appealing property of providing links between articulatory specifications and auditory patterns, thereby allowing for a discrete and modality-neutral representation suitable for memory storage of words (see Jakobson, Fant, & Halle, 1952; Ladefoged, 1997; Poeppel et al., 2008; Stevens, 2002; for discussion and illustration). To exemplify, the morpheme/word "cat" contains the segments /kæt/. In turn, the /k/ is specified as [-voice], [-continuant], and has the place value [dorsal]. These feature values characterise this segment as having no vocal chord vibration during the production (compared to a /g/), being "brief" (compared to, say, a vowel, which would be [+ continuant]), and being articulated with the tongue body in a dorsal position. Similar featural decompositions apply to all segments, and an inventory of distinctive features underlies the specification of all human speech sounds.

The goal of speech perception, on this view, then, is to map the acoustic waveform onto these discretised segments and their constitutive features. Crucially, for Halle and Stevens, the listener must possess the knowledge of the set of generative phonological rules shared with the talker that serve to relate these abstract segments and features to their articulatory actualisations, and moreover, this knowledge is exploited in the process of the *active* generation of candidate hypotheses (converting the trial phoneme string into the comparison spectrum) to be matched against the incoming auditory stimulus.

The model proposed by Halle and Stevens (1962) is reproduced in Figure 1. The input speech signal undergoes preliminary analysis in the afferent auditory pathway, perhaps up to and including core and belt auditory areas. It can be placed in a temporary storage and subsequently into the *comparator* component (the anatomic specifications of these operations are not understood, although, presumably executed in auditory areas). A minimal amount of signal is required (what constitutes a minimal signal sample to elicit hypotheses needs further specification), however, to begin the generation of candidate representations, effectively guesses. Once this "minimal" amount of input (perhaps no less than 20–30 ms of signal) has undergone preliminary processing (say Fourier analysis at the periphery and Hilbert transform more centrally), it is placed in a *control* component that serves to order the candidate representations that will be compared to the preprocessed input spectrum. The *control* is not only informed by the preliminary analysis but also by the results of previous A × S loops. From

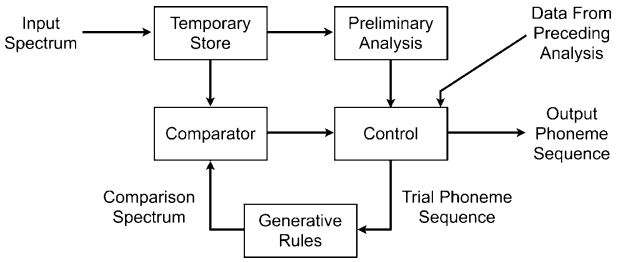

Figure 1. Diagram of the analysis by synthesis model as it appeared in Halle and Stevens (1962). The input spectrum triggers a preliminary analysis based on a minimal amount of the incoming signal. Given this small amount of input, hypothesised representations are generated via stored generative rules, which serve to relate the acoustic and articulatory properties of speech sounds. These candidate representations are subsequently fed back and compared with the input spectrum. This computation is performed until a best match with the input is determined and an output phoneme sequence can be read out.

there, a trial phoneme sequence is produced and fed into the generative rules, which apply and consequently result in a comparison spectrum, a format of representation that can be compared with the input spectrum. The co-ordinate system that permits the comparison operation is likely to be "auditory", that is the representation that forms the basis of the comparison includes information about the spectro-temporal content (real or hypothesised) of the signal. The amount of error between the bottom-up input spectrum and the top-down generated predicted spectrum is calculated, fed back into the *control* component, and this procedure is repeated until a best match is obtained (details regarding what constitutes a "best match" were never discussed at length in the original formulations). The crucial property of this model is that given some small amount of input, candidate representations, i.e., hypotheses, are *actively* generated and their putative acoustic realisations subsequently *fed back* to be compared with the input spectrum.

In order to arrive at an output phoneme sequence, however, two stages of A × S are required (independently of the number of loops necessary to arrive at the best match candidate representation in each stage). The first stage is responsible for translating the input speech spectrum into a set of "quasi-continuous" phonetic parameters, as well as eliminating speaker-dependent sources of variation (not particularly trivial operations in their own right). An A × S procedure similar to that pictured in Figure 1 is responsible for this translation. The output of the first stage is subsequently translated into an output phoneme sequence via the computations performed in the second stage. These computations again are similar to those involved in Figure 1. Critically, the generative rules called on for the synthesis component in the second stage are the same as those involved in the *production* of speech.

Though not directly relevant to the discussion at hand, that presupposition suggests that the same system implicated here for perception underlies speech production, though information flows in the opposite direction. This, then, is the conceptual nexus between the A × S proposal and the motor theories entertained at the same time. In the Liberman motor theory, the listener was aiming to identify the intended articulatory gestures; in the Halle and Stevens' A × S theory, the listener was using knowledge of phonology to internally generate the acoustic consequences of articulatory gestures. In the latter, the auditory signal analysis played a central part because it alone provided the data to generate hypotheses about the phoneme sequences. Moreover, phonological knowledge played a distinct part in the A × S proposal.

The literature on speech recognition is dominated by a few models, the challenges for which are ably reviewed by both Cleary and Pisoni (2001) and Pardo and Remez (2006). To provide some context for our current discussion, the A × S model shares important components with both cohort theory (Marslen-Wilson, 1987) and the TRACEmodel (McClelland & Elman, 1986), because both models explicitly incorporate top-down information and deal with the difficult problem of how to converge on a stored lexical representation based on small chunks of input. There are some important differences between these models and the A × S approach. Three are briefly discussed here. First, unlike cohort theory and TRACE, the A × S model capitalises on the existence of the *phonological rules* internal to a speaker/listener. Second, the stored phonological knowledge that each listener holds is *actively* deployed in the construction of candidate representations. This contrasts with the "passive" percolation (both bottom-up and top-down) of activation throughout a network, which is free of instantiated rules. In other words, both the existence of rules per se, and the active deployment of these to create predicted signals, constitutes a major difference between the three models. Third, the synthesis stage draws on a specific operation that is potentially unique to the A × S approach. In particular, the activated rules generate a candidate sequence that is realised in a co-ordinate system that is appropriate for the comparison operation. A rule that points to, say, an /a/ as a potential target creates an internally predicted representation of a spectrum that can be then compared with the input spectrum created by the auditory periphery. This tight link between predictive operation and the comparison to an actual input is the hallmark of the model and is what sets it apart from the alternative approaches. Activated representations are not passively passed onto higher order operations, but rather the hypothesise-and-test model of the circuit generates target representations that underlie the construction of the perceptual representation.

The A × S model was conceived as an architecture for handling the mapping from acoustic input to lexical representation, i.e., its domain of action is spoken word recognition. In our view, however, there is no compelling reason to limit the A × S algorithm to this level of processing. Indeed, the notion of predicting potential completions (or next steps) based on the current state generalises very naturally to processing connected speech. For example, once a candidate set of lexical representations is entertained by the processor, their associated grammatical information is also immediately available—and, therefore, plays a critical constraining role for the next processing step (recent models of parsing capitalise on similar algorithms; see, for example, Phillips & Wagers, 2007; Townsend & Bever, 2001). To give a most banal example, for speakers of English, a phrase beginning with the word "the" has a series of likely and legitimate next constituents, about which we know at least one absolutely robust attribute: the next word will be a noun (the caterpillar), an adjective (the hungry caterpillar), or an adverb (the very hungry caterpillar). However, it can be nothing else. This point simply serves to illustrate that much like in the perceptual analysis of speech, once we are dealing with the currency of lexical items and their interaction, the same operations can apply fruitfully. How A × S works at the level of sentence processing has been discussed extensively by Townsend and colleagues (2001), and is recently reviewed by Bever and Poeppel (in press). An illustration that tries to schematise how the A × S algorithm scales up to spoken language recognition was published in Poeppel et al. (2008, Figure 4). Just as we hypothesise that the rules underlying speech recognition are deployed in speech production (i.e., there is a single system of phonological rules at the basis of lexical storage, production, and comprehension), we surmise that the grammatical knowledge that forms the basis for comprehension is the same as that for production. A different architecture, in fact, would be highly complex, requiring an additional layer of linking operations that bridge the systems of knowledge mediating comprehension/recognition and production.

In summary, A × S uses (1) the extraction of (necessarily brief and coarse) cues in the input signal to elicit hypotheses, that while coarse, are sufficient to generate plausible guesses about classes of sounds (for example, plosives, fricatives, nasals, and approximants), and that permit subsequent refinement; (2) the actual synthesis of potential sequences consistent with the cues; and (3) a comparison operation between synthesised targets and the input signal delivered from the auditory analysis of the speech. (Some of the linguistically motivated properties of the approach outlined are discussed in more detail in Poeppel & Idsardi, in press). Importantly, in order to generate the appropriate sequences, the synthesis apparatus must be closely linked to the "generative rules" that underlie the mapping from abstract phoneme sequence to the associated auditory pattern. It is worth bearing in mind how

this rather explicit proposal from the 1950s incorporates feedforward and feedback mechanisms, as well access to abstract linguistic knowledge—a prescient stance, to say the least. From a conceptual perspective, the A × S approach incorporates and bridges crucial aspects of both auditory theories and motor theories. Indeed, we argue that linking the ideas put forth in these older treatments with current computational neuroscience and cognitive neuroscience makes for a thoroughly contemporary proposal.

INTERDISCIPLINARY APPROACH

The manner in which information from engineering, linguistics, and speech science was thoughtfully and creatively combined and incorporated in the A × S model seems very modern and provides an example of interdisciplinary research at its best. Similar ideas were discussed in other domains of perception, as well (e.g., MacKay, 1951; and later Miller, Galanter, & Pribram, 1960; Neisser, 1967). The discussions at this time, 40–50 years ago, made an explicit and robust effort to link high-level abstract knowledge, properties of the signal, and the possibilities and limitations associated both with feedforward and feedback processing. Why was this particular approach not pursued more systematically in speech recognition research? It appears that such a knowledge-driven algorithm was not popular at that time. More narrowly signal-oriented approaches, as well as statistical approaches, have dominated the discourse. In part, this may be a sociological artefact of the predominance of research in a more empiricist tradition at that time in psychology. The impact of the work of Neisser and others in psychology and Chomsky and colleagues in linguistics was not yet as widely felt, as the cognitive revolution was in its infancy. Nevertheless, given the sophistication of the ideas developed in this research, it is quite surprising that the subsequent decades of research on speech recognition were dominated by aggressively statistical approaches (e.g., hidden Markov models). While that research has yielded important progress from an engineering perspective and reasonable performance of artificial systems, it is uncontroversial that the performance of these engineering-centred methods pale in comparison to the abilities of humans, and moreover, interface very poorly with perceptual and neurobiological research on speech recognition. The present remarks therefore constitute an attempt to revisit and resuscitate the A × S model for speech recognition. Given that such procedures are now explicitly discussed in the literature on visual perception (Bar, 2004, 2007, 2009; Bar et al., 2006; Hochstein & Ahissar, 2002; Yuille & Kersten, 2006), it is ironic that research on speech recognition needs a reminder that these ideas were developed and discussed many years ago in speech research. Conceptually, the A × S framework must be considered a

precursor to current research on predictive coding and Bayesian approaches to perception. Anticipating our conclusion, A × S must be considered an idea whose time has come—or rather, whose time has come back.

Here we highlight some data and ideas that might contribute to a renewed appreciation of the A × S algorithm. We review data that are well-accounted for in the context of the framework. These data provide further motivation for considering A × S in the context of current cognitive neuroscience research. We also outline recent data from imaging studies that are suggestive with regard to the time course over which A × S may be operating. The perceptual and neurobiological data reviewed, in our view, add to a provocative series of arguments that A × S should be (re)considered as a serious contender in speech recognition research, complementing other, currently more dominant models. In particular, while parallel distributed (McClelland), dynamic cohort (Gaskell & Marslen-Wilson, 1997), or feedforward (Norris et al., 2000) models are widely discussed, A × S adds a valuable perspective on some of the computational subroutines that make the cognitive neuroscience of speech recognition such a challenging problem.

Prediction in audiovisual speech perception

In the last few years, there has been much interest in multisensory processing, in general, and audiovisual speech processing, in particular. Illustrative of this trend, and important in the context of a discussion on analysis and synthesis, is an EEG experiment by van Wassenhove, Grant, and Poeppel (2005). (See also a thoughtful replication and extension by Stekelenburg & Vroomen, 2007 as well as recent work by Arnal, Morillon, Kell, & Giraud, 2009.) The authors presented participants with (video and audio) spoken syllables—of the canonical type, i.e., pa/ta/ka, and so on, while recording EEG. The materials included congruent video and audio signals (for example, a speaker articulating a syllable /pa/—i.e., bilabial viseme—and the audio track playing /pa/) as well as incongruent stimuli of the McGurk and McDonald (1976) type, in which the audio and video signals are mismatched, and speakers identify a spoken syllable that is neither in the audio nor in the video tracks (e.g., visemic stimulus "ka" plus audio stimulus /pa/ yield percept /ta/ in a large proportion of subjects, trial by trial). Of particular interest in this experiment were the neurophysiological responses elicited by the spoken syllables. On a—perhaps relatively naive—reading of the neurophysiological literature, the response elicited by a multimodal, A + V signal was hypothesised to be significantly larger (and perhaps even supra-additive) compared to the response elicited by a single modality, A or V, stimulus. Subcortical nuclei, specifically the colliculus, have shown a significant proportion of cells that yield a supra-additive response to visual plus auditory stimulation compared to unimodal responses (see, e.g., Stein &

Meredith, 1993). This notion was wholeheartedly imported as an explanation into cognitive neuroscience research using imaging; the supra-additivity view was propagated through the literature, a straightforward hypothesis for an EEG study was, therefore, that unimodal electrophysiological responses should be smaller than (congruent) multimodal responses.

Here the plot thickens. Some fMRI data in fact revealed multimodal responses—particularly in the superior temporal sulcus (STS)—whose response amplitude was significantly larger than unimodal hemodymanic activity (e.g., Calvert, Campbell, & Brammer, 2000). Moreover, EEG data associated with nonspeech stimulation (i.e., lights and tones) also yielded larger responses to multisensory stimulation (e.g., Giard & Peronnet, 1999), suggesting that the processing model was plausible, in particular, in that it posited multiple modalities converging on multisensory neurons (perhaps in STS), and that the larger activity associated with those neurons reflects multisensory input and *integration*. However, the data reported by van Wassenhove et al. (2005) did not fit this pattern: the major response components evoked by auditory signals, the N1 and P2, did not show amplitude increases, as predicted, but rather had consistent and significant amplitude decreases at both the N1 and P2 peaks.

A particularly interesting and subtle effect was observed when considering the response latencies. Participants, in a separate session, were asked to identify the visemic information—i.e., the articulated sounds without the benefit of the audio signals. Subjects were essentially perfect (and at ceiling performance) in identifying bilabial visemes, associated with the articulation of /ba/, /pa/, or /ma/. When confronted with alveolar/dental visemes, identification performance was intermediate (just under 80% correct), and upon viewing velar-associated visemes (e.g., saying /ka/), performance dropped sharply, to under 65%. This is hardly surprising: viewing a more or less open mouth configuration is only minimally informative about the intended articulation, whereas viewing a bilabial configuration implicates no more than three segments: /b/, /p/, and /m/. Now, when plotting electrophysiological peak response latencies of the N1 and P2 as a function of visemic identification performance, an interesting pattern emerged. There was significant and differential response facilitation, expressed as temporal "savings"—that is, shorter response latencies for bimodal compared with the unimodal conditions. In particular, for the syllables tested ([ka], [ta], and [pa]), the peak latencies for [ka] were facilitated by 5 and 10 ms (at the N1 and P2, peaks, respectively); the [ta] responses were up to 15 ms faster (at the P2), and the [pa] responses were 10 and 25 ms faster (N1 and P2, respectively). In other words, the rate of correct identification in the "visual-alone" condition predicts the degree of temporal savings at the N1/P2 complex. More predictable facial articulatory configurations yield more auditory temporal savings.

van Wassenhove et al. (2005) interpreted these data along the lines of an A × S model. The facial articulatory information had to be a sufficient cue to generate hypotheses about the possible audio signals about to occur. In natural (audiovisual) speech, the movement of the face (articulators) always precedes the audio signal. This has been documented extensively by Grant, van Wassenhove, and Poeppel (2004), and is, of course, plausible: for a target sound to be produced, articulators have to be moved. The visual information, which tends to precede auditory information by approximately 60–200 ms, therefore, can elicit hypotheses about the associated speech events (and, of course, about articulator configuration). These hypotheses (or guesses) initiate a *synthesis* stage in which a putative speech sound is transformed into the predicted audio pattern (a concept borrowed from the literature on efference copies and internal forward models). The predicted signal—transformed into some form of auditory co-ordinate space—is then compared to the actual input, and this procedure is iterated until error is minimised. With regard to the ERP data in van Wassenhove et al. (2005), the hypotheses differ in precision as a function of how informative the facial cues are. For example, seeing a bilabial configuration will lead to very narrow predictions (b, p, and m) and more temporal savings, whereas a more open-mouthed configuration will lead to more predictions and less temporal savings. In short, the audiovisual speech perception task is cashed out as an example of A × S.

In considering the various component operations, the A × S algorithm includes (1) a stage of generating predictions that are local in time and specific to a listener's grammatical knowledge; (2) a stage of making comparisons between a predicted and an actual input; and (3) a stage of computing (and feeding back) the residual error from the comparison stage. The latter two operations are widely studied in motor control (for example, in the context of the literature on internal forward models; see, e.g., Wolpert & Ghahramani, 2004), and there exist well-supported neurobiological mechanisms that underlie the execution of these stages. The neuronal mechanisms implicated in some of these operations have been reviewed in a number of domains of motor control (see Kawato, 1999; Sommer & Wurtz, 2008), including speech (Guenther, Hampson, & Johnson, 1998; Guenther, Ghosh, & Tourville, 2006). The neuronal circuitry for planning a movement and for evaluating a prediction and monitoring the outcome has been especially well-characterised for eye movements. While the interface with the cognitive systems underlying speech obviously changes the details, we have no reason to believe that executing such operations requires a fundamentally different architecture when dealing with speech or other forms of motor output. We surmise that the neuronal architecture that facilitates these operations in the context of various motor systems is successfully co-opted, both in speech production, and—as we argue here—in speech perception.

By and large, the prediction stage is not so widely investigated in systems neuroscience; in what follows we provide some examples of psycholinguistic speech research that speaks to that issue. In terms of the figure above, the prediction stage encompasses the control (trial sequence), generative rules, and comparator stage.

PREDICTIONS

In A × S models, listeners actively generate hypothesised candidate representations on the basis of some amount of auditory speech signal. These internally generated hypotheses should serve as the basis for actively formulating predictions regarding the nature of the upcoming speech stream (i.e., which sound should I hear next?). In order to demonstrate the plausibility of such models then, it is necessary to show that listeners do, in fact, construct predictions regarding the upcoming signal and that these predictions are rooted in their knowledge of the sound pattern of their native language. Recent work supports such a conclusion. In particular, violations of a prediction elicit processing difficulty, manifested either as slower behavioural reaction times or delayed latencies of evoked electrophysiological components in response to unpredicted segments.

Systematic alternations exist in languages such that we expect listeners to be able to use their stored phonological knowledge about the sound patterns of their language to generate predictions about which sound they are likely to hear next. For example, in English, vowels are almost always produced only with airflow through the oral cavity. There is one particular exception, however: vowels that precede nasal consonants (e.g., [m] and [n]) are produced with additional airflow through the nasal cavity. Given this systematic relationship between nasality on vowels and subsequent nasal consonants, it is plausible to conjecture that listeners should be able to use the phonetic cues contained within the nasalised vowel to predict that the next segment they hear will be a nasal consonant. The fact that listeners are sensitive to violations of these predictions has been demonstrated behaviourally (Fowler & Brown, 2000; Lahiri & Marslen-Wilson, 1991). Using a segment identification task, Fowler and Brown (2000) showed that English listeners are able to use the information contained with the vowel (i.e., whether it is oral or nasal) to create expectations regarding whether the next consonant is oral or nasal. In an MEG experiment, Flagg, Cardy, and Roberts (2005) asked about the time course of these effects: how early in the evoked neuromagnetic response might one see effects of violations of these predictions. They spliced and cross-spliced VCV tokens such that the first VC was either congruent in their specification for nasal (i.e., both oral: [aba]; both nasal: [ãma]) or incongruent (i.e., [ãba] and [ama]). They report no

reliable difference at the onset of the vowel. In other words, whether the vowel was oral or nasal had no influence on the latency of the M100, and early automatic evoked neuromagnetic response. Approximately 50–75 ms postonset of the consonant, however, Flagg et al. (2005) found a reliable difference in the latency of an evoked component for the cases where the consonant was oral (i.e., [b]). The latency of the evoked component in cases where the preceding vowel was oral ([aba]) was reliably shorter as opposed to when it was nasal ([ãba]). They reported no differences when the medial consonant was nasal. Results of this type suggest that listeners are sensitive to violations of expectation, and these violations manifest themselves not only in behavioural reaction times, but also in very early, automatic evoked electrophysiological components, suggesting that these predictions exert their influence on particularly early stages of perception.

In another series of experiments, using both behavioural measures and MEG, we have investigated a different, and more universally attested constraint on sound sequences (Hwang, Monahan, & Idsardi, in press; Monahan, Hwang, & Idsardi, submitted). In English, word-final obstruent consonant clusters (stops, fricatives, and affricates) must agree in their specification for voicing. Cases of mixed voicing (e.g., [kz] or [sd]) never occur. For a listener of English that can exploit this knowledge, we expect the listener to be able to predict that if he/she hears a voiced obstruent followed by a second obstruent, this second obstruent should be voiced in its production. In a segment identification task (listeners were asked to respond whether they heard a [z] or [s]; Hwang et al., in press), we tested English listeners' responses to congruent (e.g., [udz] and [uts]) and incongruent cases (e.g., [uds] and [utz]). Only when the first consonant was voiced (i.e., [d]) did we find the predicted effects of violation of expectation. The congruent case [udz] elicited a significantly higher accuracy and faster reaction times than the incongruent case [uds]. No difference between [uts] and [utz] were found, and this pair elicited significantly slower reaction times than [udz] and significantly faster reaction times than [uds]. This particular pattern of results suggests that only some sounds can serve as the basis for predictions (Lahiri & Reetz, 2002). In this particular case, it has been proposed that the feature [voice] is specified only for voiced sounds (Lombardi, 1995, 1999), and that voiceless sounds are underspecified in their representation of voicing. The logic of this proposal suggests that we see effects of prediction only with [d], because [d] is specified for [voice] and the presence of this feature allows listeners to generate predictions about the upcoming speech signal. We followed up on these results using MEG to further address the time course of these effects and to determine how early in the processing stream we can find elicited differences between congruent and incongruent cases (Monahan et al., submitted). Given that we were only interested in the response to the congruent or incongruent segment (i.e., the final fricative [s]

or [z]), we only compared conditions in which the final consonant was the same. By 150 ms postonset of the violating segment, we observed a reliable difference in the amplitude of the root mean square of the grand average waveform across participants between the congruent [uts] and incongruent [uds]. These results are consistent with Flagg et al. (2005) suggesting that the influence of phonological processes is exerted early in auditory processing.

There are other ways to document predictive processing in speech, and one prominent approach has been to use the electrophysiological (EEG, MEG) mismatch response. Given a sequence of stimuli that repeat—setting up a "standard" sequence—a mismatching, "deviant" stimulus will elicit a distinct response that peaks roughly 150–250 ms postdeviant onset. The stimulus /pa/ in the context /ba ba ba ba ba ba ba pa/ will generate this response. The mismatch paradigm has been applied very widely, to both speech and nonspeech materials, and the fact that a mismatch is generated at all can be taken as evidence that a prediction was formulated. This response has been used profitably to investigate more abstract aspects of phonology, as well (Kazanina, Phillips, & Idsardi, 2006; Phillips et al., 2000). While the data derived from mismatch experiments are certainly consistent with arguments for online and local predictions, in those paradigms prediction can be driven by local, momentary models, i.e., short-term memory, whereas the new results discussed above require prediction and violation of expectation on the basis of linguistic knowledge that is deployed online, at the time scale of tens of milliseconds. In short, the mismatch-based literature broadly supports the A × S conjectures, but the studies, by and large, do not probe the online knowledge-driven predictions for which we are seeking evidence.

Cumulatively, the data provide support for the idea that listeners are able to utilise their knowledge of the sound processes of their language to generate hypotheses regarding the nature of the upcoming speech signal. Moreover, violations of these predictions seem to influence early auditory processing. The fact that listeners are able to generate hypotheses about the nature of the upcoming speech signal is one of the core predictions of a forward model like A × S. While additional research is required to determine more precisely the nature of these predictions, the evidence thus far seems to suggest that listeners are able to generate predictions and anticipate the nature of the upcoming signal.

It has not escaped us that the functional architecture described here and for which we provide some tentative support based on recent experimental data has significant potential implications for automatic speech recognition (ASR). The ASR literature has not succeeded yet in identifying systems that successfully build on human performance to motivate engineering applications. We suggest that a closer second look at A × S as an algorithm, its functional components, and its increasingly rich link to supporting

behavioural and neurological data, can provide a fresh new perspective for ASR research.

Manuscript received September 2009
Revised manuscript received May 2010
First published online month/year

REFERENCES

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A-L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, *29*, 13445–13453.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*, 617–629.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289.

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*, 1235–1243.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmidt, A. M., Dale, A. M., . . . Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences of the USA*, *103*(2), 449–454.

Bever, T., & Poeppel, D. (in press). Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics*.

Calvert, G. A., Campbell, R., & Brammer, M. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*, 649–657.

Cleary, M., & Pisoni, D. B. (2001). Speech perception and spoken word recognition: Research and theory. In B. Goldstein (Ed.), *Handbook of perception* (pp. 499–534). Cambridge: Blackwell.

Flagg, E. J., Cardy, J. E. O., & Roberts, T. P. L. (2005). MEG detects neural consequences of anomalous nasalization in vowel-consonant pairs. *Neuroscience Letters*, *397*(3), 263–269.

Fowler, C. A., & Brown, J. M. (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception & Psychophysics*, *62*, 21–32.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language & Cognitive Processes*, *12*(5–6), 613–656.

Geisler, W. S., & Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, *27*(3), 379–402.

Geisler, W. S., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, *5*(6), 508–510.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473–490.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Communication*, *44*(1–4), 43–53.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*(3), 280–301.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, *105*, 611–633.

Halle, M., & Stevens, K. N. (1959). Analysis by synthesis. In W. Wathen-Dunn & L. E. Woods (Eds.), *Proceedings of the seminar on speech compression and processing*. USAF Camb. Res. Ctr. 2: Paper D7.

Halle, M., & Stevens, K. N. (1962). Speech recognition: A model and a program for research. *IRE Transactions of the PGIT, IT-8*, 155–159.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, *36*, 791–804.

Hwang, S-O., Monahan, P. J., & Idsardi, W. J. (in press). Underspecification and asymmetries in voicing perception. *Phonology*.

Jakobson, R., Fant, G., & Halle, M. (1952). *Preliminaries to speech analysis*. Cambridge, MA: MIT Press.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, *9*(6), 718–727.

Kazanina, N., Phillips, C., & Idsardi, W. (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences of the USA*, *103*(30), 11381–11386.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Ladefoged, P. (1997). Linguistic phonetic descriptions. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 589–618). Oxford: Blackwell.

Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition*, *38*, 245–294.

Lahiri, A., & Reetz, H. (2002). Underspecified recognition. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology* (7th ed., pp. 637–675). Berlin: Mouton de Gruyter.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Lombardi, L. (1995). Dahl's Law and privative voice. *Linguistic Inquiry*, *26*, 356–372.

Lombardi, L. (1999). Positional faithfulness and voicing assimilation in optimality theory. *Natural Language and Linguistic Theory*, *17*, 267–302.

MacKay, D. M. (1951). Mindlike behaviour in artefacts. *British Journal for the Philosophy of Science*, *2*(6), 105–121.

Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian modelling of visual perception. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain* (pp. 13–36). Cambrdige, MA: MIT Press.

Marslen-Wilson, W. G. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*, 71–102.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt.

Monahan, P. J., Hwang, S-O., & Idsardi, W. J. (under revision). Predicting speech: Neural correlates of voicing mismatch using MEG.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–370.

Pardo, J. S., & Remez, R. E. (2006). The perception of speech. In M. Traxler & M. A. Gernsbacher (Eds.), *The handbook of psycholinguistics* (pp. 201–248). New York: Academic Press.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., et al. (2000). Auditory cortex accesses phonological categories: An MEG mismatch study. *Journal of Cognitive Neuroscience*, *12*(6), 1038–1055.

Phillips, C., & Wagers, M. (2007). Relating structure and time in linguistics and psycholinguistics. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 739–756). Oxford, UK: Oxford University Press.

Poeppel, D., & Idsardi, W. J. (in press). Recognizing words from speech: The perception-action-memory loop.

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics: Prospects and problems. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*, 1071–1086.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*, 473–500.

Sommer, M. A., & Wurtz, R. H. (2008). Brain circuits for the internal monitoring of movements. *Annual Review of Neuroscience*, *31*, 317–338.

Stein, B., & Meredith, A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973.

Stevens, K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America*, *32*(1), 47–55.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*(4), 1872–1891.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form* (pp. 88–102). Cambridge, MA: MIT Press.

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Cambridge, MA: MIT Press.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the USA*, *102*(4), 1181–1186.

Wolpert, D. M., & Ghahramani, Z. (2004). Computational motor control. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences III* (pp. 485–494). Cambridge, MA: MIT Press.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.