PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 1071–1086
doi:10.1098/rstb.2007.2160
Published online 21 September 2007

Speech perception at the interface of neurobiology and linguistics

David Poeppel[1,2,*], William J. Idsardi[1] and Virginie van Wassenhove[3]

[1]*Department of Linguistics, and* [2]*Department of Biology, University of Maryland, College Park, MD 20742, USA*

[3]*Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA*

Speech perception consists of a set of computations that take continuously varying acoustic waveforms as input and generate discrete representations that make contact with the lexical representations stored in long-term memory as output. Because the perceptual objects that are recognized by the speech perception enter into subsequent linguistic computation, the format that is used for lexical representation and processing fundamentally constrains the speech perceptual processes. Consequently, theories of speech perception must, at some level, be tightly linked to theories of lexical representation. Minimally, speech perception must yield representations that smoothly and rapidly interface with stored lexical items. Adopting the perspective of Marr, we argue and provide neurobiological and psychophysical evidence for the following research programme. First, at the implementational level, speech perception is a multi-time resolution process, with perceptual analyses occurring concurrently on at least two time scales (approx. 20–80 ms, approx. 150–300 ms), commensurate with (sub)segmental and syllabic analyses, respectively. Second, at the algorithmic level, we suggest that perception proceeds on the basis of internal forward models, or uses an 'analysis-by-synthesis' approach. Third, at the computational level (in the sense of Marr), the theory of lexical representation that we adopt is principally informed by phonological research and assumes that words are represented in the mental lexicon in terms of sequences of discrete segments composed of distinctive features. One important goal of the research programme is to develop linking hypotheses between putative neurobiological primitives (e.g. temporal primitives) and those primitives derived from linguistic inquiry, to arrive ultimately at a biologically sensible and theoretically satisfying model of representation and computation in speech.

Keywords: multi-time resolution; temporal coding; analysis-by-synthesis; predictive coding; forward model; distinctive features

1. INTRODUCTION

We take speech perception to be the set of computations that entail as their 'endgame' and optimal result the identification of words, either presented in isolation or in spoken discourse. This—almost banal—presupposition, that speech perception is primarily about finding words in ecologically natural contexts (and not, say, about spotting phonemes or indicating intelligibility in experimental contexts; see Cleary & Pisoni (2001) for a related perspective), provides an important boundary condition on a programme of research; because words (or syllables or morphemes), once identified, must enter into *subsequent* linguistic computation (phonological, morphological, syntactic) to permit successful language comprehension, the internal representation of words generated by the speech perception processes must be suitable for the range of linguistic operations performed with these words. In short, it is a critical requirement that the output of the processes that constitute speech perception are representations that permit *using* and *manipulating* these representations in specific ways. Such a requirement implies that research on speech perception must interface closely with theories of lexical representation. It is, in our view, not a sufficient answer to state that a word has been recognized without specifying rather explicitly what the format of the representation is. More colloquially, if the neural code for lexical representation is written in, say, Brain++, speech perception must transform the input signal, a continuously varying waveform, into Brain++ objects. On this view, any theory of speech perception thus requires making commitments to theories of lexical representation.

Based on this perspective, we outline a research programme on speech perception that is strongly influenced by Marr's (1982) approach to understanding visual perception. Marr's suggestion to distinguish between computational, algorithmic and implementational levels of description when investigating computational systems in cognitive neuroscience seems to be very helpful to us in fractionating the problem and organizing the set of questions one faces in the study of speech perception. We adopt the taxonomic organizational principles outlined by Marr and discuss from that perspective three major properties of speech perception

* Author and address for correspondence: Department of Linguistics, 1401 Marie Mount Hall, University of Maryland, College Park, MD 20742, USA (dpoeppel@umd.edu).

One contribution of 13 to a Theme Issue 'The perception of speech: from sound to meaning'.

that we take to require a principled explanation. *First*, at the *implementational* level of description, speech perception is a *multi-time resolution process*, with signal analysis occurring concurrently on (at least) two time scales relevant to speech, syllabic-level (approx. 5 Hz) and segmental-level (approx. 20 Hz) temporal analyses. Naturally, multiresolution processing is but one of many relevant implementational issues, but it has received recent empirical support in both human and animal studies (Boemio *et al.* 2005; Narayan *et al.* 2006) and has interesting consequences for the architecture of the system; consequently we focus on that issue here. Multiresolution processing is widely observed in other systems (e.g. vision) and can, we suggest, be used profitably in engineering approaches to speech recognition. *Second*, at the *algorithmic* level of description, the central algorithm we invoke is *analysis-by-synthesis*. This constitutes a set of operations first discussed in the 1950s and 1960s (and specifically for the speech case by Halle & Stevens (1959, 1962) and Stevens & Halle (1967)) that provide an approach to bottom-up processing challenges in perception by using 'hypothesize-and-test' methods. Based on minimal sensory information, the perceptual system generates knowledge-based 'guesses' (hypotheses) about possible targets and internally synthesizes these targets. Matching procedures between the synthesized candidate targets and the input signal ultimately select the best match; in other words, the analysis is guided by internally synthesized candidate representations. In the terminology of contemporary cognitive neuroscience, analysis-by-synthesis as we develop it here is closely related to the concept of internal forward models. In the terminology of automatic speech recognition and statistics, analysis-by-synthesis is also conceptually related to Bayesian classification approaches. *Third*, at the *computational* level of description, we commit to a specific representational theory, that of *distinctive features* as the primitives for lexical representation and phonological computation. Our proposal contrasts with views that argue for *strictly* episodic (acoustic) representations—although we are sympathetic to the fact that the rich evidence for episodic effects must be accommodated, and we articulate a proposal in §5. In our view, words are represented in the mind/brain as a series of segments each of which is a bundle of distinctive features that indicate the articulatory configuration underlying the phonological segment. As decades of research show, phonological generalizations are stated over features (neither holistic phonemes nor *a fortiori* 'epiphones'), reflecting their epistemological primacy. Given the importance of features for the organization of linguistically significant sounds and given the fact that their articulatory implementation results in specific acoustic correlates (Stevens 1998, 2002), we assume that one of the central aspects of speech perception is the extraction of distinctive features from the signal. The fact that the elements of phonological organization can be interpreted as articulatory gestures with distinct acoustic consequences suggests a tight and efficient architectural organization of the speech system in which speech production and perception are intimately connected through the unifying concept of distinctive features.

The ideas we raised earlier provide a new perspective on some challenges in speech recognition that we take to be fundamental: the problem of linearity (segmentation); the problem of invariance; and the problem of perceptual constancy. These three problems are, of course, closely related and constitute irritating stumbling blocks for automatic speech recognition research as well as accounts of human speech perception. We are in no position to provide answers to these foundational challenges, and the paper is not focused on segmentation and invariance. However, the three properties of speech perception that we argue for here may provide a wedge into dealing with these challenges to the recognition process. For example, we argue that multi-time resolution processing—in the context of which segmentation occurs concurrently on segmental and syllabic time scales—relates closely to the 'landmarks' approach advocated by Stevens (2002). In particular, our approach allows for a 'quick and coarse' sample of the input that can subsequently be refined by the further analysis in a parallel stream. This concept is very similar to Stevens' notion of looking for informative landmarks and then verifying and testing the information around these landmarks to specify the speech information at that time point in the waveform. If this model is on the right track, and if Stevens and we are on the right track in hypothesizing that distinctive features are both the basis for speech representation and have acoustic realizations, one can begin to formulate models that try to link *acoustic information on multiple time scales* to featural information. Secondly, once one has such featural hypotheses, one can generate internal guesses that can then guide further perceptual processing. That is, guesses based on coarsely represented spectro-temporal representations would constitute a way to ignite the analysis-by-synthesis algorithm that we take to be particularly useful to rapidly recognize incoming speech based on predictions that are conditioned by both the prior speech context and higher-order linguistic knowledge. It is conceivable that the linearity and invariance problems are not the principled limitations they are now if one adopts a multi-time resolution perspective, because it is possible that when one looks at the information on multiple scales, there is more robustness in the acoustic-to-feature mapping than when looking only at processing on one time scale (as is typical in hidden Markov models for automatic speech recognition systems). Certainly this is no solution to invariance, but it does provide one new perspective on how to approach this highly important and vexing issue in the perception of spoken language.

Figure 1 schematizes the operative representations in speech perception in the context of the present proposal. Based on a continuously varying signal at the periphery (figure 1*a*), the afferent auditory pathway constructs a detailed spectro-temporal representation (figure 1*b*), by hypothesis based on operations well described by the Fourier transform (at the periphery) and the Hilbert transform (in cortex, to extract envelope information). These assumptions are not particularly controversial and follow from extensive research in auditory theory and neurophysiology. Assuming that the representation of lexical items is a discrete series of segments composed

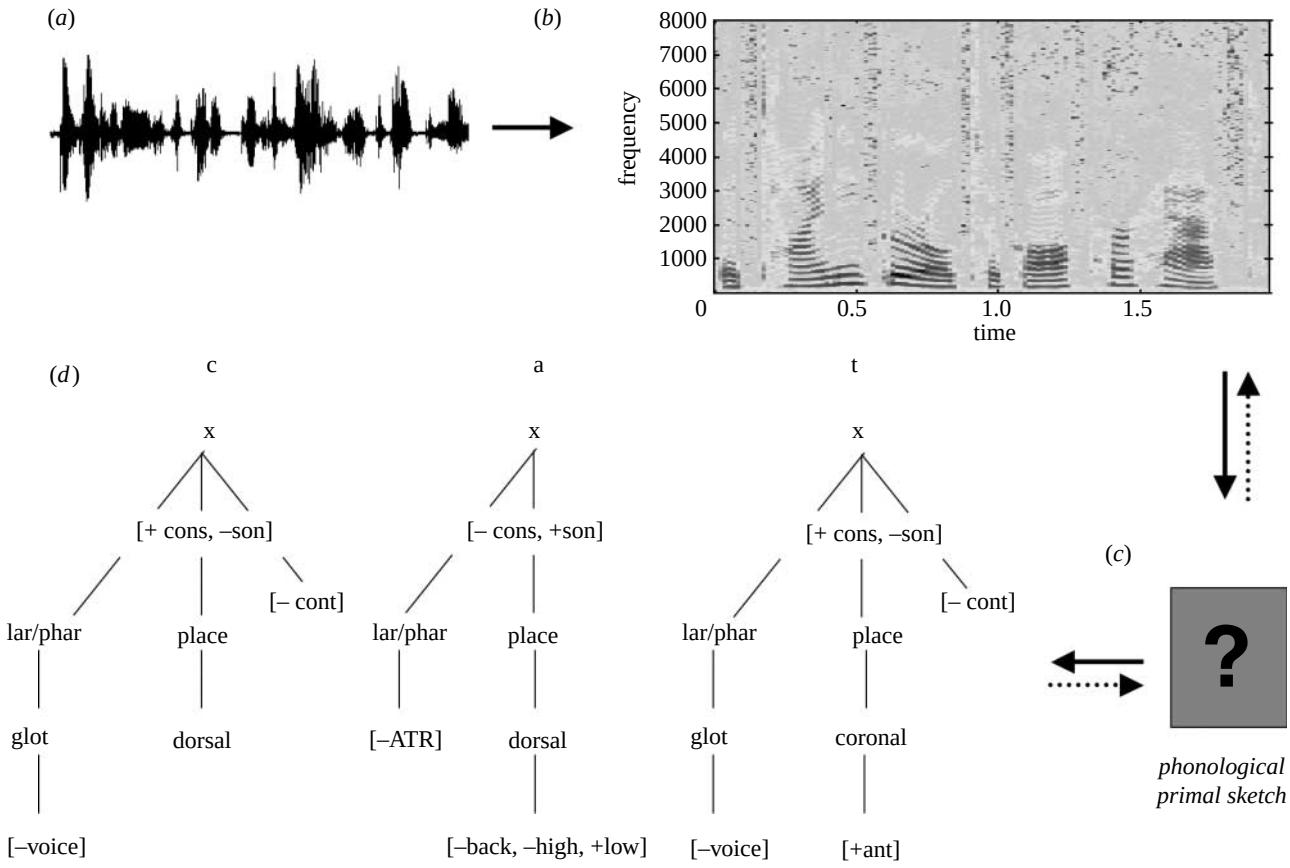

Figure 1. Representations and transformations from input signal to lexical representation. Solid arrows represent logically required steps and dotted arrows reflect hypothesized top-down mappings. (*a*) At the auditory periphery, the listener has to encode a continuously varying acoustic waveform (*x*-axis, time; *y*-axis, amplitude). (*b*) The afferent auditory pathway analyses the input signal in time and frequency. A neural 'analogue' of the spectrogram is generated to highlight both spectral and temporal variations in the signal. (cf. STRFs in auditory cortex.) (*c*) An intermediate representation may be necessary to map from a spectro-temporal representation of the acoustic input signal to the putative abstract representation of the word. The intermediate representation may be a PPS, built on temporal primitives (temporal windows of specific sizes) and spectral primitives. (*d*) The hypothesized representation of the word *cat* in the mind/brain of the speaker/listener. Each of the three segments of this consonant-vowel-consonant word is built from distinctive features that as a bundle are definitional of the segment.

of distinctive features (figure 1*d*)—a view that is also uncontroversial insofar as one accepts the last few decades of phonological research—a central question is how to accomplish the mapping from a spectro-temporal, acoustic, representation to the lexical–phonological one. This mapping may—as suggested in figure 1*c*—or may not involve further intermediate representations. Note that it *is* controversial as to what extent the processing steps can feedback to previous stages (cf. Norris *et al.* 2000). From our perspective that incorporates both multi-time resolution processing and analysis-by-synthesis, it follows that there is an intermediate representation (figure 1*c*; say, the auditory equivalent of Marr's 2 1/2-dimensional sketch), namely (minimally) temporal windows of different sizes that represent different attributes of the signal, the *phonological primal sketch* (PPS). The properties of this putative intermediate representation are largely unknown, for the moment. (For a discussion of the related concept *auditory primal sketch*, see Todd (1994).) For example, calling it 'phonological' implies a categorical representation, but the extent to which the information in each of the two windows is categorical is unclear because it is untested. One could call the representation 'phonetic primal sketch', as well, if the information remains graded. Crucial is that *something* must mediate the mapping from spectro-temporal signal configurations to lexical entries (on our view of lexical representation). Such an intermediate (and fleeting) multi-time resolution representation will retain acoustic properties, but they will differ depending on whether one is looking at the shorter (segmental) or longer (syllabic) temporal primitive. We see the primal sketch as related to Stevens' (2002) notion of landmarks. It is not yet worked out in what way a PPS relates to the representations stipulated by models, such as TRACE, NAM, shortlist or distributed cohort. Because such models do not make explicit reference to multi-time resolution processing, it is not obvious whether short, segmental or long, syllabic temporal primitives can be accommodated best within such theories. Finally, it stands to reason that the locus of computation is the superior temporal cortex, but any claim beyond that must remain speculative. If permitted such speculation in the context of our proposed functional anatomy (see figure 2), one might argue that posterior superior temporal gyrus (STG) and superior temporal sulcus (STS) in the ventral pathway (Hickok & Poeppel 2004) are the relevant part of cortex to construct an interface representation, given that middle temporal gyrus (MTG) and STG are argued to be the substrates for lexical representation and auditory analysis, respectively.

The approach to the speech perception that we advocate here differs from many current proposals in explicitly (re)incorporating linguistic and psycholinguistic considerations, particularly considerations of lexical and phonological representations. Much research makes the implicit (and sometimes explicit) presupposition that the best route for understanding the major challenges to speech recognition comes from trying to bridge auditory theory with auditory neuroscience (e.g. the papers in Greenberg & Ainsworth (2006)), while dismissing the (often largely representational) issues raised by phonological theory. While we are sympathetic to what can be learned from such a research programme, we are convinced that one cannot do without the constraints derived from linguistics, and particularly phonology. Obviously, auditory theory (say, with regard to the importance of critical bands, modulation transfer functions, masking, pitch extraction, stream segregation, the modulation spectrum and so on) is crucial to an understanding of how the incoming signal is analysed and transformed into *representations that form the basis for speech recognition*. Similarly, (i) cellular and systems neuroscience teaches us essential facts about how acoustic signals are analysed in the afferent auditory pathway and (ii) the distributed cortical functional anatomy associated with speech recognition suggests that various dimensions are processed in a segregated manner. (Note that our own work often focuses on these issues, i.e. we are not just sympathetic to auditory neuroscience and auditory theory, we are also practitioners; e.g. Hickok & Poeppel 2000, 2004). These domains of investigation provide critical knowledge about the *construction* of the representations that constitute speech.

However, the nature of the speech representations *as they enter speech-related computation* is rarely, if ever, spelt out. For example, the field is very comfortable talking about how acoustic signals can be characterized by spectro-temporal receptive fields (STRFs) of auditory cortical neurons (e.g. Shamma 2001). This is a terrific set of results—but such a characterization tells us nothing about how such a (neuronal) representation allows for *further* computation with that token. Suppose one has recognized the acoustic realization of the word '*caterpillar*' using only the machinery of auditory theory and the neuronal concept of STRFs. What is now owed is a set of linking hypotheses from auditory-based representations of that type to whatever machinery or representational structure underlies further, language-based processing. Why? Because the recognized item typically enters into phonological and morphological operations (say, pluralization) as well as syntactic ones (say, subject–predicate agreement, *viz*. '*caterpillar-s change-Ø into butterflies*'). To connect with that aspect of processing, the representations in play must be in the same 'code', a rather straightforward conjecture. Now, if we assume (for us, uncontroversially; for some, shockingly) that there are abstract internal representations that form the basis for linguistic representation and processing, there must be some stage at which auditory signals are translated into such representations. If one is disinclined to invoke linguistically motivated representations early in the processing stream, then one owes a statement of linking hypotheses that connect the different formats (unless one does not, categorically, believe in any internal abstract representations for language processing). Alternatively, perhaps the representations of speech that are motivated by linguistic considerations are in fact active in the analysis process itself and therefore active throughout the subroutines that make up the speech perception process. Unsurprisingly, we adopt the latter view.

A slightly different way to characterize the programme of research is to ask: what are the representational and computational primitives in auditory cortex; what are the primitives for speech; and how can we build defensible linking hypotheses that bridge these domains (cf. Poeppel & Embick 2005)? Here, we will discuss three steps that we take to be essential in the process of transforming signals to interpretable internal representations: (i) multi-time resolution processing in auditory cortex as a computational strategy to fractionate the signal into appropriate 'temporal primitives' commensurate with processing the auditory input concurrently on a segmental and a syllabic scale; (ii) analysis-by-synthesis as a computational strategy linking top-down and bottom-up operations in auditory cortex; and (iii) the construction of abstract representations (distinctive features) that form the computational basis for both lexical representation and transforming between sensory and motor coordinates in speech processing.

In our view, the three attributes of speech representation (features) and processing (multi-time resolution, analysis-by-synthesis) that we raised provide a way to (begin to) think about how one might more explicitly link the acoustic signal to the internal abstractions that are words. Building on the intuitions of Marr (1982), we see the perception and recognition processes as having a number of bottom-up and top-down steps. It is not a 'subtle interplay' of feed-forward and feedback steps that we have in mind, though, but a rather unromantic, mechanical (forward) calculation of perceptual candidates based on very precisely guided synthesis steps. In a first pass, the system attempts a quick reduction (primal sketch) of the total search space for lexical access by finding the—somewhat coarsely specified—*landmarks* (Stevens 2002) through the *articulator-free* (major-class, place-less) features (Halle 2002). That is, the initial pass defines a neighbourhood on broad-class and manner features (e.g. stop-fricative-nasal-approximant; the term 'approximant' covers both glides and vowels). These initial guesses are based on minimal spectro-temporal information (say, two or three analysis windows) and can be stepwise refined in small time increments (approx. 30 ms or so) owing to the multi-time resolution nature of the process. Subsequent to the initial hypotheses triggered by the construction of the PPS, a cohort-type selection is elicited from the articulator-bound (place) features. In this way, we try to have our cake and eat it too—trying to capture both (gross) neighbourhood and (gross) cohort-model effects. Overall, our proposal is similar in spirit (if not in details) to the featurally underspecified lexicon (FUL) model of speech recognition (Lahiri & Reetz 2002).

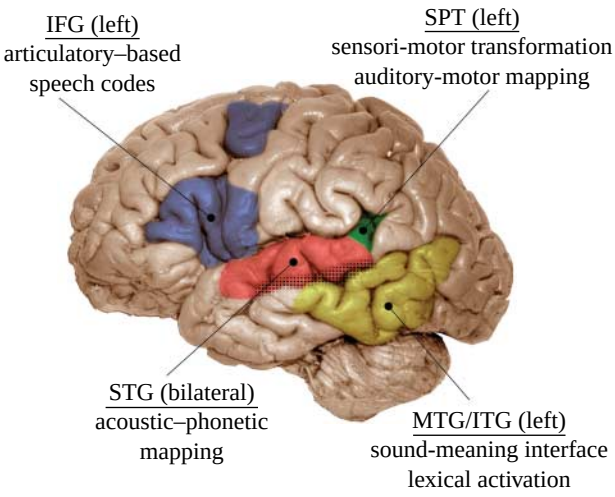

Figure 2. Functional anatomy of speech-sound processing. In the mapping from input to lexical representation, the initial steps are bilateral, mediated by various cortical fields on the STG; subsequent computation is typically left lateralized and extends over many left per-Sylvian areas. IFG, inferior frontal gyrus; SPT, Sylvian parieto-temporal area; MTG, middle temporal gyrus; ITG, inferior temporal gyrus; STG, superior temporal gyrus. (Adapted from Hickok & Poeppel (2004).)

2. FUNCTIONAL ANATOMIC BACKGROUND

Importantly, once again from the perspective of the implementational level of Marr (1982), the research we outline is consistent with (and in part explicitly motivated by) cognitive neuroscience approaches to speech perception, specifically considerations of the cortical functional architecture of speech perception (figure 2). Several recent reviews develop large-scale models of the cortical basis of speech perception and, despite some disagreement on various details, there is also considerable convergence among these proposals (Binder *et al.* 2000; Hickok & Poeppel 2000, 2004; Scott & Johnsrude 2003; Boatman 2004; Indefrey & Levelt 2004; Poeppel & Hackl 2007). We briefly outline the cortical architecture here.

The initial cortical analysis of speech occurs bilaterally in core and surrounding superior auditory areas (see Hackett *et al.* (2001) for relevant human auditory cortex anatomy). Subsequent computations (typically involving lexical-level processing) are largely left lateralized (with the exception of the analysis of pitch change; the analysis of voice; and the analysis of syllable-length signals), encompassing the STG, anterior and posterior aspects of the STS as well as inferior frontal, temporo-parietal and inferior temporal structures (see Poeppel *et al.* (2004) for arguments and imaging evidence that speech is bilaterally mediated). This listing shows that practically all classical, peri-Sylvian language areas are implicated in some aspects of the perception of speech. Therefore, one goal has to be to begin to specify what the computational contribution of each cortical field might be.

With regard to the input signal travelling up the afferent pathway, there are notable asymmetries in the brain stem and even at the cochlear level of sound analysis (e.g. Sininger & Cone-Wesson 2004), but it is not well understood whether these subcortical asymmetries condition the processing in a way that is sufficiently rich to account for the compelling asymmetries that emerge at the cortical level. Imaging studies show very convincingly that the processing of speech at the initial stages is robustly bilateral, at least at the level of core and surrounding STG (Mummery *et al.* 1999; Binder *et al.* 2000; Norris & Wise 2000; Poeppel *et al.* 2004). The fact that imaging studies show bilateral activation does, of course, not imply that the computations executed in left and right core auditory cortices are identical—there are, presumably, important differences in local computation. Nevertheless the processing is bilateral as assessed by haemodynamic and electrophysiological methods. We hypothesize that the STRFs of neurons in bilateral core auditory cortex generate high-resolution neuronal representations of the input signal (which of course is already highly pre-processed in subcortical areas, say, the inferior colliculus).

A growing body of neuroimaging research deals with the question of what exactly is computed in left and right auditory areas during speech and non-speech processing. Zatorre and colleagues have argued on the basis of neuropsychological and neuroimaging data that right hemisphere superior temporal areas are specialized for the analysis of spectral properties of signals, in particular spectral change, and the analysis of pitch, specifically pitch change. In contrast, they argue that left hemisphere areas are better suited to the processing of rapid temporal modulation (Zatorre *et al.* 2002). Their view converges with that of Poeppel (2001, 2003), where it is suggested that the spectral versus temporal right–left asymmetry is a consequence of the size of the temporal integration windows of the neuronal ensembles in these areas. Neuronal ensembles in left (non-primary) temporal cortex are associated with somewhat shorter integration constants (say, 20–50 ms) and therefore left hemisphere cortical fields preferentially reflect temporal properties of acoustic signals. Right hemisphere (non-primary) cortex houses neuronal ensembles, a large proportion of which have longer (150–300 ms) integration windows, and therefore are better suited to analyse spectral change. These ideas are discussed in more detail below, but they build on a long history and literature that investigates hemispheric asymmetry in the auditory cortex related to spectral versus temporal processing (Schwartz & Tallal 1980; Robin *et al.* 1990). In summary, we hypothesize that primary (core) auditory cortex builds high-fidelity representations of the signal, and surrounding non-primary areas differentially 'elaborate' this signal by analysing it on different time scales.

Beyond this initial analysis of sounds that is robustly bilateral and may involve all the steps involved in the acoustic-to-phonetic mapping, there is wide agreement that speech perception is lateralized. The right STG and STS have been shown to play a critical role in the analysis of voice information (Belin *et al.* 2004) and dynamic pitch. The analysis of prosodic features of speech has also been suggested to be lateralized to right STG. The processing of speech *per se*, i.e. that aspect of processing that permits lexical access and further speech-based computation, however, is lateralized to left temporal, parietal and frontal cortices (Binder *et al.* 2000; Hickok & Poeppel 2000; Boatman 2004; Indefrey & Levelt 2004; Scott & Wise 2004).

Beginning in the STG, research in the last few years has identified the emergence of two processing streams. The idea of segregated and parallel pathways is closely related to vision research, where the concept of a 'what' versus a 'where'/'how' pathway is very firmly established (Ungerleider & Mishkin 1982). In the auditory domain, one can also think of a what (ventral) pathway (the pathway responsible for the 'sound-to-meaning mapping'; Hickok & Poeppel 2004) that involves various aspects of the temporal lobe that are apparently dedicated to sound identification. Both more anterior parts of the STS (Scott *et al.* 2000) and more posterior parts of the STG/STS (Binder *et al.* 2000) as well as MTG (Indefrey & Levelt 2004) have been implicated in speech-sound processing. Scott *et al.* (2000) were the first to show that anterior STS plays a crucial role in speech intelligibility. Binder and colleagues have suggested that posterior STG and STS are critical for the transformation from acoustic-to-phonetic information; and, based on a large meta-analysis, Indefrey & Levelt (2004) suggest that the interface of phonetic and lexical information is at least in part mediated by posterior MTG. Note that it is not at all clear which aspect of the so-called what pathway in auditory processing is responsible for lexical access. There are some suggestions that middle and inferior temporal gyri and basal temporal cortex reflect lexical processing, but the fractionation of speech processing and lexical processing is, perhaps unsurprisingly, not straightforwardly reflected in imaging studies. Nevertheless, there is consensus that the STG from rostral to caudal fields and the STS constitute the neural tissue in which many of the critical computations for speech recognition are executed. It is worth bearing in mind that the range of areas implicated in speech processing go well beyond the classical language areas typically mentioned for speech; the vast majority of textbooks still state that this aspect of perception and language processing occurs in Wernicke's area (the posterior third of the STG).

In analogy to the visual what/where distinction, evidence from auditory anatomy and neurophysiology (e.g. Romanski *et al.* 1999) as well as imaging suggests that there is a dorsal pathway that plays a role—not just in where-type computations but also in speech processing (the pathway responsible for the 'sound-to-articulation mapping'; Hickok & Poeppel 2004). The dorsal pathway implicated in auditory tasks includes temporo-parietal, parietal and frontal areas. The specific computational contribution of each area is not yet understood for either where/how tasks in hearing or speech perception tasks. However, there is evidence, from the domain of speech processing, that a temporo-parietal area plays an important role in the (hypothesized) coordinate transformation from auditory to motor coordinates. This Sylvian parieto-temporal area has been studied by Hickok and colleagues (Hickok *et al.* 2003) and is argued to be necessary to maintain parity between input- and output-based speech tasks. Furthermore, aspects of Broca's area (Brodmann areas 44 and 45) are also regularly implicated in speech processing (see Burton (2001) for review). It is of considerable interest that frontal cortical areas are involved in perceptual tasks. Such findings have (i) challenged the view that Broca's area is principally responsible for production tasks or syntactic tasks and (ii) reinvigorated the discussion of a 'motor' contribution to speech perception. In part, these discussions are reflected in the debates surrounding mirror neurons and the renewed interest in the motor theory of speech perception. Figure 2 shows the functional anatomy of speech perception derived from neuropsychological and neuroimaging data. A central challenge to the field is to begin to formulate much more detailed hypotheses about what computations are executed in each of these areas, first for speech perception and second for other linguistic and non-linguistic operations in which the computations mediated by these areas participate in causal ways. We now turn to the three hypothesized attributes of speech perception introduced above and, when possible, relate them to the sketch of the anatomy outlined here.

3. MULTI-TIME RESOLUTION PROCESSING

It is an intuitively straightforward observation that visual signals are processed on multiple spatial scales. For example, faces can be analysed at a detailed, featural level, but also at a coarser, configural level, and these correspond to different spatial frequencies in the visual image. The information carried in these different channels is not identical—different spatial frequencies are associated with differential abilities to convey emotional information (low spatial frequencies) versus image details (high spatial frequencies; e.g. Vuilleumier *et al.* 2003). An alternative way to conceptualize this distinction is to think of it as the tension between global versus local information. Processing on multiple spatial scales in the visual domain has been studied psychophysically and harnessed for provocative analyses of contemporary art (Pelli 1999). Whereas in the visual case the image can be fractionated into different spatial scales, in the auditory case both frequency and time can be thought of as dimensions along which one could fractionate the signal. We pursue the hypothesis that auditory signals are processed in time windows of different sizes, or durations. (For data supporting this conjecture from the domain of neural coding in bird song, see Narayan *et al.* (2006).) The idea that time windows of different sizes are relevant for speech analysis and perception derives from several phenomena. In particular, acoustic as well as articulatory-phonetic phenomena occur on different time scales. Investigation of a waveform and spectrogram of a spoken sentence reveal that at the scale of roughly 20–80 ms, segmental and subsegmental cues are reflected, as well as local segmental *order* (i.e. the difference between 'pe*st*' and 'pe*ts*'). In contrast, at the scale of roughly 150–300 ms (corresponding acoustically to the envelope of the waveform), suprasegmental and syllabic phenomena are reflected. One way to reconcile the tension between local (fast modulation frequency) and global (slower modulation frequency) information is to assume hierarchical processing such that higher-order, longer representations are constructed on the basis of smaller units. Alternatively, perhaps information on multiple time scales is processed concurrently. We explore the

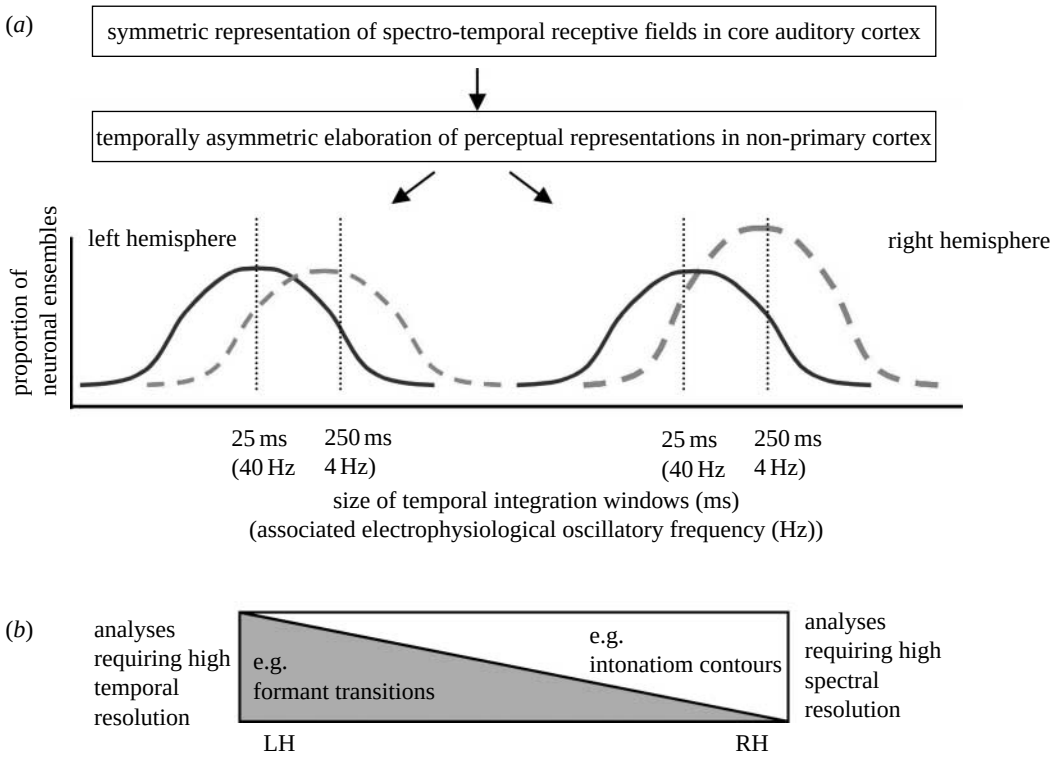

Figure 3. Temporal integration in auditory and speech analysis. (*a*) Temporal integration and multi-time resolution analysis: quantization and lateralization. Both left and right auditory cortices have populations of neurons (wired as neuronal ensembles) that have preferred integration constants of two types. By hypothesis, one set of neurons prefers approximately 25 ms integration, another 250 ms. In electrophysiological studies, such integration windows may be reflected as activity in the gamma and theta bands, respectively. The evidence for a rightward asymmetry of slow integration is growing and the evidence for a leftward asymmetry of rapid integration is unsettled. Minimally, both hemispheres are equipped to deal with subtle temporal variation (Boemio *et al.* 2005). (*b*) Functional lateralization as a consequence of temporal integration. From asymmetric temporal integration of this type, it follows that different auditory tasks will recruit the two populations differentially owing to sensitivity differences and lead to hemispheric asymmetry.

latter possibility here. In particular, we discuss the hypothesis that there are two principal time windows within which a given auditory signal (speech or non-speech) is processed, with the mean durations as above. Although we are not 'two time window imperialists' and recognize the importance of processing on the (sub-)millisecond scale in the brainstem and the 1000+ millisecond scale for phrases, we argue that the two windows we identify play a privileged role in the analysis and perceptual interpretation of auditory signals, and that these two time windows have special consequences for speech perception.

Multi-time resolution processing, as we develop the hypothesis here, is built on the concept of temporal integration windows (figure 3). Both psychophysical and neurobiological evidence suggest that physically continuous information is broken apart and processed in temporal windows. The claim that there is temporal integration is rather uncontroversial. More controversial is the hypothesis that there is not just integration but discretization (Saberi & Perrott 1999; VanRullen & Koch 2003). Either way, it is clear that signals are analysed in a discontinuous fashion.

We hypothesize that there are two integration windows and that their implementation occurs in *non-primary* auditory cortex. As stated above, the auditory signal up to primary auditory cortex is processed in a predominantly symmetric manner (although there are notable asymmetries at the subcortical level). The STRFs in core auditory cortex permit the construction of a relatively high-fidelity representation of the signal. Based on this initial representation, there is a temporally asymmetric elaboration of the signal by 'sampling' the output of core auditory cortex using two window sizes. One window is of the order of 20–50 ms, another of the order of 200 ms. One way to develop a visual intuition for this idea is to imagine two spectrograms of the same signal, one highlighting the rapid temporal changes and representing the glottal pulse, the other representing the narrowband frequency variation (say, formants). What is the purpose of such a proposed temporal quantization of the input waveform? We hypothesize that this sampling serves as a logistical or administrative device to generate auditory representations of the appropriate granularity to interface with higher-order, abstract representations. The model is outlined in more detail in Poeppel (2001, 2003), and is consistent with functional magnetic resonance imaging (fMRI) data in Boemio *et al.* (2005); Schonwiesner *et al.* (2005) and others.

If this type of multi-time resolution model is on the right track, evidence is owed for the model's constituent claims. In particular, one needs to show that there is (i) integration on the 25–80 ms time scale, (ii) integration on the 150–300 ms time scale, (iii) a perceptually

relevant interaction between representations constructed on these two time scales, and (iv) lateralization of function associated with processing on these time scales.

Evidence for temporal integration on the short time scale is relatively abundant and will not be discussed further here (see Poeppel 2003; Wang *et al.* 2003). Both for speech and non-speech signals, it has been shown psychophysically and electrophysiologically that integration on a 20–50 ms time scale has compelling perceptual consequences. In the non-speech case, three clear examples of processing on this time scale are provided by the psychophysical order threshold (Hirsh & Sherrick 1961), by frequency modulation (FM) direction discrimination (Gordon & Poeppel 2002; Luo *et al.* 2007) and click-train integration studies (Lu *et al.* 2001; A. Boemio 2002, unpublished dissertation). Experiments testing the minimum stimulus onset asynchrony (SOA) at which the *order* of two concurrently presented signals can be reliably indicated show that, across sensory modalities, approximately 20–30 ms are the relevant time scale (Hirsh & Sherrick 1961). Experiments testing the minimum signal duration at which one can reliably discriminate between upward and downward moving FM tones (glides) consistently show that the stimulus duration at which one performs robustly is 20 ms (Gordon & Poeppel 2002; Luo *et al.* 2007). Psychophysical and electrophysiological experiments on the processing of click trains also highlight the relevance of processing on this time scale, in both humans and non-human primates (Lu *et al.* 2001; A. Boemio 2002, unpublished dissertation). As one varies the SOA between clicks from 1 to 1000 ms, subjects experience categorically different auditory percepts: at SOAs above 50 ms, each click is perceived as an individuated event; at SOAs below 10 ms, the click train is perceived as a tone with a well-defined pitch. However, there is a sharp perceptual transition between roughly 10 and roughly 50 ms in which subjects have neither clearly discrete nor clearly continuous judgements of click trains; and, importantly, electrophysiological recordings show that this perceptual transition region (associated with the sensation of roughness) is associated with a particular neurophysiological response profile that reflects the transitory nature of this temporal regime at which one begins to construct perceptual pitch. Animal (Lu *et al.* 2001) and human (A. Boemio 2002, unpublished dissertation) studies show that response properties change at precisely that perceptual boundary, possibly associated with a transition from temporal to rate coding schemes.

A compelling example from speech perception comes from the work of Saberi & Perrott (1999), who took spoken sentences, cut them into time slices and locally reversed the direction of each time window. The intelligibility function they reported was strongly conditioned by the size of the window. For segment durations of up to 50 ms, intelligibility was not significantly affected, supporting the notion of the special nature of integration, and perhaps even discretization, on this time scale.

Evidence for temporal integration on a longer, 150–300 ms time scale also comes from physiological and psychophysical studies. For non-speech signals, loudness integration has been shown to require roughly 200 ms of signal to reach asymptotic psychophysical loudness judgement (Moore 1989; Green 1993). Electrophysiological mismatch negativity studies testing temporal integration of non-speech signals also point to 150–300 ms as the relevant time scale (Yabe *et al.* 1997, 2001*a*,*b*). There are two observations from the domain of speech that we consider critical in this context. First, cross-linguistic measurements of mean syllable duration have revealed that although there is tremendous variability in syllable structure across languages, mean acoustic duration is remarkably stable, peaking between 100–300 ms. These values are commensurate with measurements of the modulation spectrum of spoken language (Greenberg 2005). Peaks in the modulation spectrum occur between 2 and 6 Hz and are argued to reflect the underlying syllabic structure of the spoken utterance, which determines its envelope.

A second source of evidence from the speech domain comes from studies on audiovisual (AV) integration. Several studies have replicated the following surprising finding (Massaro *et al.* 1996; Munhall *et al.* 1996; Grant *et al.* 2004; van Wassenhove *et al.* 2007). When one presents listeners with AV syllables and desynchronizes the audio and video tracks, one might expect severe perceptual disturbances given that one has disrupted the temporal alignment of the auditory and the visual information. Surprisingly, listeners tolerate enormous temporal asynchronies when viewing asynchronous AV speech. For example, it has been shown (Grant *et al.* 2004; van Wassenhove *et al.* 2007) that both McGurk auditory–visual stimuli and congruent auditory–visual stimuli are judged to be interpretable (with little performance degradation) despite AV asynchronies of up to 200 ms. In other words, the perceptual system interprets as simultaneous AV speech signals that are within a 200 ms window, the mean duration of a syllable.

Does the information carried on these two time scales interact in perceptually relevant ways? If an auditory signal is indeed analysed concurrently on two time scales, is there a binding of information that modulates speech perception? This question has, to our knowledge, not previously been addressed experimentally. Whereas there are studies exploiting signal-processing techniques to highlight the contributions of higher- or lower-modulation frequencies (Drullman *et al.* 1994*a*,*b*), the interaction between temporal information on different scales has been taken for granted (i.e. it is an implicit presupposition that segmental and syllabic information are 'congruent' in some sense, permitting successful comprehension). A new study by Chait *et al.* (submitted) tests the idea directly. They created signals in which either low-modulation frequency information was maintained across the spectrum (0–4 Hz, corresponding to long temporal window analysis) or higher-modulation frequency information was selectively retained (22–40 Hz, corresponding to the shorter integration constants). Subjects heard sentences (binaurally) that had only low- or high-modulation frequency and were

tested for intelligibility. Consistent with previous work (Drullman *et al.* 1994*a*), low-modulation frequency signals, despite being extremely impoverished relative to a normal speech signal, allow for surprisingly good comprehension, with subjects showing intelligibility scores of well over 40% despite a severely degraded signal. In contrast, signals that retain only higher-modulation frequencies (above 20 Hz) generated low-intelligibility scores (below 20%, which is still remarkable given the restricted nature of the signal). But, crucially, what happens when both signals are presented at the same time (dichotically)? Many patterns *could* be obtained. The two signals could destructively interfere with each other, yielding low-intelligibility scores; the signals could be processed independently, yielding no net gain overall; the two signals could interact and yield an additive or even a supra-additive effect. Interestingly, it is the last possibility that is obtained: dichotically presented signals presented concurrently were apparently bound to generate representations that allowed for intelligibility scores that were significantly larger (68%) than a predicted linear additive effect (approx. 50%). This observation argues for the view that information extracted and analysed on two time scales interacts synergistically and in a perceptually relevant manner, supporting the hypothesis of multi-time resolution processing.

A final point concerns the hypothesis that processing on different time scales is actually associated with different cortical areas and possibly lateralized. In particular, it has been proposed that more slowly modulated signals—on the 150–300 ms scale—preferentially drive right hemisphere (non-primary) auditory areas whereas rapidly modulated signals—say, 20–80 ms—drive left cortical areas. If distinct auditory cortical fields are found to be differentially sensitive to temporal information, such data would support the model that different time scales are processed in parallel. In a recent fMRI study, Boemio *et al.* (2005) tested this hypothesis using non-speech signals that were constructed to closely match certain properties of speech signals. They observed that STG, bilaterally, is exquisitely sensitive to rapid temporal signals; however, right STS was preferentially driven by longer-duration signals, supporting the hypothesis that signals are not only processed on multiple time scales but also that the processing is partially lateralized, with slower signals differentially associated with right hemisphere mechanisms in higher-order auditory (and the canonical multisensory) area, STS. Lateralization of auditory analysis as a function of temporal signal properties has been observed in a number of studies now and can be viewed as a well-established finding (Hesling *et al.* 2005; Meyer *et al.* 2005; Schonwiesner *et al.* 2005). Cumulatively, the psychophysical and neurobiological data are most consistent with the conjecture of multi-time resolution processing, with the two critical time scales being a short (perhaps segmental) temporal integration window of 20–80 ms and a longer (syllabic) window of 150–300 ms. Regardless of the specifics of the values, the multi-time resolution nature of the processing seems like a very solid hypothesis based on a large body of convergent evidence.

4. ANALYSIS-BY-SYNTHESIS—INTERNAL FORWARD MODELS

Models of speech perception and lexical access tend to come in two types. Either the processing is rather strictly bottom-up (e.g. Norris *et al.* 2000) or there is feed-forward and feedback processing during perceptual analysis and lexical access, as is typical in most connectionist-style models. An attribute that tends to be common among such models is that the analysis is relatively 'passive'. That is to say, features percolate up the processing hierarchy in bottom-up models, or activation spreads in interactive models. The approach to recognition that we advocate here differs from such proposals. In particular, analysis-by-synthesis, or perception driven by predictive coding based on internal forward models, is a decidedly active stance towards perception that has been characterized as a 'hypothesize-and-test' approach. A minimal amount of signal triggers internal guesses about the perceptual target representation; the guesses (hypotheses) are recoded, or synthesized, into a format that permits comparison with the input signal. It is this 'forward' synthesis of candidate representations that is the central property of the approach and makes it a completely active process. Hypothesize-and-test models for perception were discussed in the 1950s and 1960s, for example, by Miller *et al.* (1960). Halle & Stevens (1959, 1962) and Stevens & Halle (1967) first developed the idea of analysis-by-synthesis for speech perception and an updated model, very much in line with our thinking, is provided in Stevens (2002).

Anticipating somewhat the discussion of distinctive features from §5, we see analysis-by-synthesis employed as a general computational architecture common to the whole recognition process. Acoustic measurements yield guesses about distinctive feature values in the string. 'Mini-lexicons' of valid syllable types are consulted, and a space of possible parses is constructed as a first-pass analysis. Frequency information is encoded throughout the system, so that the search can proceed on a 'best-first' basis, with more probable parses assigned greater weight in the system. More specifically, we see the landmarks of Stevens (2002), which correspond to the articulator-free features of Halle (2002), and which define the 'major' classes of phonemes (stop, fricative, nasal and approximant), as defining a PPS of the segmental time scale. This primal sketch gives a neighbourhood of words matching the detected landmark sequence. (Note that the speech perceptual model we outline is very sympathetic to lexical access models that emphasize the role of lexical neighbourhoods in processing.) The primal sketch includes enough information to broadly classify certain prosodic characteristics as well (such as the approximate number of moras or syllables in the word).

Of course, the various feature detectors are fallible and return probabilistic information, which can then be inverted using Bayes's rule.

$$\text{Bayes's rule}: \quad p(H|E) = p(E|H) \times p(H)/p(E)$$

(here 'H' can be read as 'hypothesis' and 'E' as 'evidence').

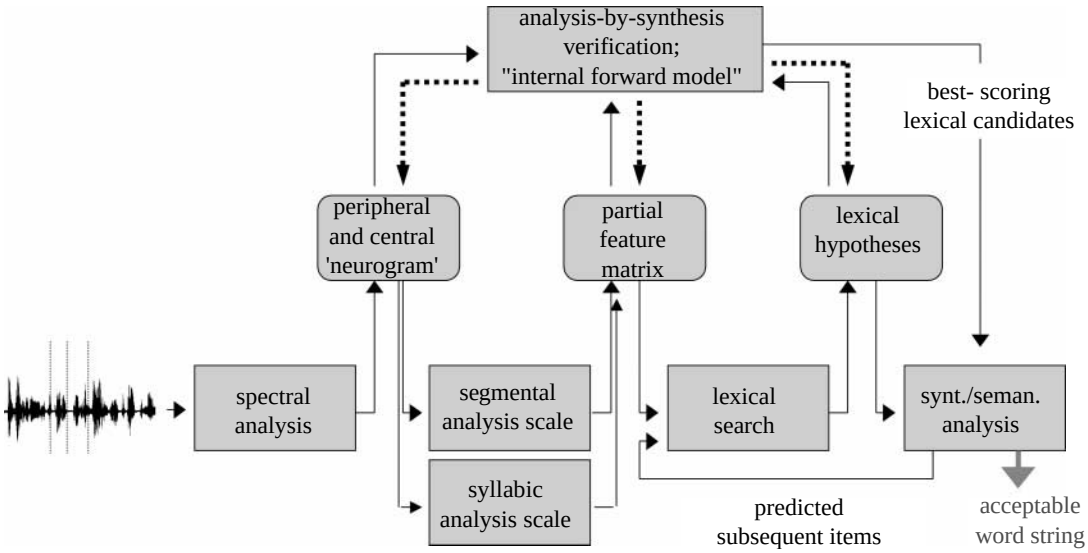

Figure 4. Possible processing steps in an analysis-by-synthesis model. The bottom tier incorporates distinct levels of representation in the mapping from sound to word (spectral analysis–segmental analysis–lexical hypotheses). The intermediate tier shows possible representations and computations that interact with the bottom and top (analysis-by-synthesis) levels to generate the correct mappings. The internal forward model can synthesize the candidates for matching at each level (neuronal, featural decomposition, lexical hypotheses) depending on how much information the forward model has to guide the internal synthesis. We hypothesize that the internal model is updated approximately every 30 ms, i.e. with each new sample that is available. Segmental and syllabic-level analyses of the signal are concurrent (multi-time resolution). Spectro-temporal analysis and the construction of a high-resolution auditory representation that is performed in the afferent pathway and core auditory cortex. Segmental- and syllabic-size analyses are hypothesized to occur in STG and STS (bilaterally), respectively; the mapping from hypothesized featural information to lexical entries may be mediated in STS, the lexical processes (search, activation) in middle temporal gyrus (while the conceptual information associated with lexical entries is likely to be much more distributed). The syntactic and compositional semantic representations further constraining lexical hypotheses are, perhaps, executed in frontal areas. The top-down forward model signals feed to temporal lobe from all connected areas, with a strong contribution from frontal articulatory cortical fields. (Adapted and extended from Klatt (1979).)

The quantity p(H|E) represents the likelihood of the analysis, p(E|H) is the likelihood of the synthesis of the data given the analysis. It is rather astonishing how seldom this connection between Bayesian methodology and analysis-by-synthesis has been drawn in the literature. Note that there are a small number of precedents linking analysis-by-synthesis with a Bayesian perspective, especially Hinton & Nair (in press) on handwriting recognition, and Bayesian perspectives have become much more common in perception (e.g. Knill & Richards 1996) and signal analysis (e.g. Bretthorst 1988) as well as in functional neuroimaging analysis (e.g. Friston *et al.* 2002). Moreover, much work in automatic speech recognition has a Bayesian orientation through the use of inverse-probability techniques such as hidden Markov models. However, much of the work in perception, signal analysis and speech recognition is of the empirical Bayes variety, using non-informative priors. In contrast, our view is that much of the interest is in discovering informative priors, that is, part of the content of universal grammar, in addition to incorporating the priors derived from the online perceptual processing.

The PPS is then specified by identifying the articulator-bound features within the detected landmarks. That is, the 2–1/two-dimensional analogue is constructed using probabilistic information about features such as [labial], [coronal], etc. (the *articulator-bound* features, see below) within the major class defined by the landmark primal sketch. For example, the detection of [labial] or [coronal] place is different within nasals in English (which lacks a velar nasal phoneme) than it is for stops (which contrast all three categories). That is, we see this layer of analysis as evaluating conditional probabilities of the sort p([labial]|[+nasal]). The hypothesized temporally synchronized feature sequences are then matched against the main lexicon and a list of candidates is generated, and then the rules of the phonology of the language are used to resynthesize the predictable features, again using Bayes's rule, and thus allowing a straightforward inclusion of variable rules (Labov 1972, 2001). Figure 4 shows some of the hypothesized set of computations, adapted and extended from Klatt (1979), to make clear where the synthesis process sits within the larger architecture.

One source of evidence for internal forward models of this type comes from studies on AV integration in speech. Several investigators have tested temporal constraints on AV speech integration and, as mentioned above, it is now reasonably well established that AV syllables tolerate signal desynchronization of up to 250 ms (Massaro *et al.* 1996; Munhall *et al.* 1996; Grant *et al.* 2004; van Wassenhove *et al.* 2007). When subjects are presented with AV syllables in which either the audio or the video signals lead or lag by up to 200 ms, subjects apparently integrate the desynchronized signal successfully and interpret the AV signal as coherent and bound. One way to verify this is to use McGurk tokens in which listeners are presented, for example, with an auditory /pa/ and a visual /ka/. In synchronous presentation, subjects typically report perceiving the

syllable /ta/. However, what happens during signal desynchronization is not clear—subjects could report perceiving either the audio or the video or the fused representation. As it turns out, over a time interval of 200 ms or more of desynchronization, listeners reliably perceive the fused syllable /ta/. This evidence is of course consistent with the claim that there is a temporal integration window of roughly 250 ms in AV speech (van Wassenhove *et al.* 2006).

Interestingly, the audio versus video lead or lag is not symmetric. Whereas visual leads are tolerated very well in perceptual experiments, auditory leads are more detrimental. Is such a result plausible from a more ecological perspective? We believe such a psychophysical result follows from the fact that movement of the articulators naturally precedes auditory speech output. In spoken language, auditory and visual onsets are not actually simultaneous in the physical sense and therefore incorporate a natural SOA. Such observations suggest that a tolerance for visual leads is not only preferable but also natural in AV speech perception. Unclear, however, is the issue of whether the information associated with the visual signal plays any specific role in auditory speech perception.

van Wassenhove *et al.* (2005) conducted a combined psychophysical and ERP study to investigate this issue. Subjects listened to and viewed congruent (audio /pa/ and video /pa/ or audio /ta/ and video /ta/ or audio /ka/ and video /ka/) syllables or incongruent (audio /pa/ and video /ka/) McGurk stimuli. In a three-alternative forced choice test, subjects had to categorize what they perceived; concurrently, ERPs were recorded. Based on the multisensory literature, which derives primarily from single-unit studies (Stein & Meredith 1993) or haemodynamic imaging studies (Calvert 2001), the most straightforward prediction was to observe supra-additivity on one of the major auditory evoked responses. In particular, because previous imaging work had observed that responses to AV (non-speech) signals could be supra-additive, it was predicted that the auditory N1 or P2 responses should reflect this multisensory interaction.

In contrast to the supra-additivity prediction, the experiment showed that the auditory-evoked N1 and P2 responses were actually reduced in amplitude in the AV case. The *amplitude reduction* of both the N1 and the P2 were independent of the stimulus. A very interesting pattern was obtained when looking at the response latency. The *peak latency* of the N1 and the P2 responses varied systematically as a function of the correctly identified viseme. In this study, subjects were asked to identify, based only on the visual signal, spoken syllables. For bilabials, correct identification was, predictably, very high (more than 90%); for alveolars, identification was at an intermediate level (approx. 70–85% correct); for velars, the correct identification was typically approximately 60–65%. When plotting electrophysiological response peak latency as a function of correct identification of the visual signal, significant temporal facilitation of the auditory evoked responses (N1 and P2) was observed as a function of how informative the face was for the listeners. In particular, the visual /ka/ was associated with a temporal facilitation of 5–10 ms. In comparison, the visual /pa/, which is almost always correctly identified, was associated with much greater temporal facilitation (up to 25 ms at the P2). In other words, the rate of correct identification of the visual-alone signal predicted the degree of temporal savings of the auditory N1 and P2 components. van Wassenhove *et al.* (2005) interpret these findings in the following way: the visual speech input, which typically precedes the auditory signal, elicits a (broad class of) internal abstract representations (the hypothesis space). These internal representations predict the possible audio targets. The internally synthesized and predicted targets are compared against the auditory speech input and the residual error is calculated and fed back for correction. The more informative the facial information is, the more specific the prediction can be, and the more temporal savings is observed. The (abstract) internal representation that is rapidly elicited by the leading visual signal (which may, of course, be somewhat coarse) elicits a candidate set of possible targets; the fewer targets there are, as in the case of bilabials, the more rapid and precise the synthesis is, and the more temporal facilitation is observed. Cumulatively, these data are most consistent with an analysis-by-synthesis model, an internal forward model in which perceptual analysis is guided by the predictions made based on the internally synthesized candidates that are compared against input signals. The multimodality sensory-motor integration is thus in the articulation that underlies the various outputs, not in a sensory-to-sensory mapping between, say, vision and audition. That is, the integration is in abstract, amodal articulation space.

5. LEXICAL REPRESENTATION AND DISTINCTIVE FEATURES

We adopt the idea (probably first expressed in Bell (1867); cited in Halle 2002, pp. 3–4, see also pp. 97–100) that the mental representation of speech sounds is not as segment-sized (alphabetic) units, but is decomposed into *distinctive features*. Following Jakobson *et al.* (1952; see also Halle 2002, pp. 108–110), the features have dual definitions and provide the fundamental connection between action (articulation) and perception (audition). Each feature is defined by its effective motoric gesture(s), and also by the auditory patterns that trigger its detection. Though the search for phonetic invariants for features has a long and controversial history (see Stevens & Blumstein (1981), in particular, and Perkell & Klatt (1986), in general), we are not ready to abandon the search just yet. As a fairly clear example, the feature [+round] defines the connection between the motor gesture of lip rounding (the enervation of the *orbicularis oris* muscle) and the perceptual pattern of a down sweep in frequencies across the whole spectral range. For the derivation of the acoustic effect from the motor gesture of lip protrusion through the physics of resonant tube models, see Stevens (1998). Thus, in our view, distinctive features are the sort of representational primitives that allow us to talk both about action and perception and about the connection between action and perception in a principled manner. We can do this because distinctive features are stated in both

articulatory (i.e. as gestures performed in a motor coordinate system) and acoustic/auditory terms (i.e. as events statable in an acoustic coordinate system). From this it follows that one must be able to translate between acoustic and motor representations of words: there must be some type of coordinate transformations, and, indeed, there is evidence for thinking about the problem in that way (cf. Hickok & Poeppel 2000, 2004; Hickok *et al.* 2003).

One reason we believe in features as primitives rather than segments is that generalizations in phonology typically affect or involve more than one segment of the language. That is, rules of pronunciation traffic in *natural classes* of sounds rather than individual sounds. To give just one example, the rule of coronal palatalization in Tohono O'odham (formerly known as Papago) given in many introductory phonology texts (e.g. Halle & Clements 1990) changes the *set* of alveolar stops /t d/ to the set of alveopalatal affricates /c j/ when they occur before any vowel in the set of [+high] vowels /i u _/. Likewise, in word-final position Korean neutralizes all coronal obstruents /t t^h c c^h s s'/ to the plain coronal stop /t/ (Martin 1951). The sets of sounds triggering and undergoing changes receive perspicuous description with distinctive features; in an alphabetic–phoneme world they are just arbitrary subsets of the language's sounds. In addition, the changes the sets undergo are also typically simple in terms of features—in Tohono O'odham adding [−back] and [+strident]; in Korean removing [+spread glottis] and [+strident]. Psycholinguistic evidence for features in speech perception has been available since the pioneering study by Miller & Nicely (1955, p. 338) who found evidence for distinctive features in that 'the perception of any one of these five features (which are [±voice], [±nasal], [±strident], duration and place of articulation DP/WJI) is relatively independent of the perception of the others, so that it is as if five separate, simple channels were involved rather than one complex channel.' Neuroimaging evidence for the psychological reality of these sets has been provided by Phillips *et al.* (2000), who found mismatch negativity responses in English subjects between the set of voiced stops /b d g/ and the set of voiceless stops /p t k/.

The rate of change of feature values in running speech varies, but often reaches the level of one feature per segment. That is, therefore, some features must be detected within the segmental time frame (approx. 20–80 ms). However, in addition, there is also abundant phonological evidence for syllable-level generalizations in phonology (a recent survey of such effects is Féry & van de Vijver (2004)). Almost all languages display various (approx. one-per-syllable) phenomena, such as stress, tone and vowel harmony (Archangeli & Pulleyblank 1994). The importance of the syllable in speech perception has also been emphasized by many researchers (e.g. Greenberg (2005) for an important perspective). For example, one study (Kabak & Idsardi submitted) shows differential sensitivity in cases of perceptual vowel epenthesis (the illusion of hearing vowels that are not present in the signal) to the syllabic status of consonants. They found that the onset-only set of consonants in Korean (e.g. [+strident] ones such as /c/ in 'pachma') induce perceptual epenthesis (i.e. they are indistinguishable from 'pachima' by Korean listeners), whereas other syllable contact violations of Korean (such as the /k.m/ in the impossible Korean form 'pakma') are ignored and do not result in the percept of an illusory vowel. More intriguingly, however, languages seem to organize their features so as to minimize the number of features used in the language to distinguish among both consonants and vowels (see especially the discussion of combinatorial specification in Archangeli & Pulleyblank (1994)). That is, for example [+round] is typically a feature only for vowels; much less often is [+round] used contrastively to distinguish different consonants (though one such example is Ponapean and others are documented in Ladefoged & Maddieson (1996)). In the case of canonical CV syllables, this allows the features such as [+round] to 'flow' at the syllable resolution rate (integration time scale) as well as at the segmental resolution rate. Furthermore, processes of vowel–consonant assimilation (such as nasalization of vowels and rounding or palatalization of consonants) serve to further 'smear' featural information into the syllabic time scale.

Thus for speakers to use segments to produce speech requires the segments to be decomposed into features in order to adequately account for rules of pronunciation, and listeners also construct representations using the same features, as shown by various psycholinguistic and neurolinguistic tests. Moreover, the features seem to organize along both a slower (a syllabic-level analysis) and a faster rate of change (a segment-level analysis). We find the convergence to a multiresolution analysis on two time scales of approximately the same size particularly intriguing and very much worth pursuing as one of the fundamental principles for speech recognition. For a related multitier framework on speech that also engages the multiple time scale, spectral integration and representational challenges to recognition, see Greenberg (2005).

6. DISCUSSION

The research programme we have outlined has implications for some of issues in speech perception research that tend to elicit high blood pressure. We briefly mention some of the consequences of our proposal for two major issues here. First, there have been many discussions on the question of whether speech is 'special'. In its most pointed form, the concept that speech is special presumably means that the cerebral machinery we have to analyse speech is specialized for speech signals in many or most parts of the auditory pathway. It is not obvious whether very much can be learned by focusing on whether or not there is this kind of extreme specialization. Presumably, what everybody is actually interested in is to try to understand how speech recognition works. However, some things do need to be said, since we advocate such a strong linguistically oriented position. As has been argued recently, for example by Price *et al.* (2005), all the cortical machinery that is used for speech is also used for other tasks. It seems like a reasonable proposition that in the difficult case of having to analyse complex signals extremely rapidly, you use

whatever is available to you. However, there is a point at which there must be specialization, and that is the point at which the auditory representation interfaces with lexical representation. Lexical representations are *sui generis*: they *may* share properties with other cognitive representations, but they have a number of extremely specialized properties that seem to be restricted to the representation of lexical items in the human brain. So, for example, lexical items do not, at least to our knowledge, look like the internal representations of jingling keys, faces, melodies or odours. Therefore, there is a stage in speech perception at which this format must be constructed, and if that format is of a particular type, there is necessarily specialization. As mentioned above, a critical requirement of lexical representation is that the representations enter into subsequent computation, for example of the morphological or syntactic flavour. One could imagine that lexical roots share properties with the mental/neural representations of non-linguistic sounds, but *some* formal attribute of the representations must be such that they can participate in formal operations ranging from pluralization to compound generation to phrase structure construction. Second, rather than adopt episodic models, we propose that speech is processed in featural and syllabic (categorical) terms. But listeners also pay attention to the location of individual tokens in the acoustic 'clouds' defined by the categories (or types). That is, they detect, know and remember something about the speaker's speech, as compared with the listener's statistical summary of the acoustic variation in the categories that they have already encountered (Goldinger *et al.* 1992). Parallel to Labov (1972) who found *accommodation* by speakers in a variable speech community to the traits of other conversational partners (dropping more /r/'s in 'fourth floor' when their conversational partner dropped their /r/'s), we see the episodicists' findings as showing the willingness of speakers to track and accommodate their low-level speech traits to those of their conversational partners, presumably for sociological reasons (as argued by Labov (1972, 2001)). Strong confirmation of the sociolinguistic mediation of speech accommodation is given by the work of Howard Giles (e.g. Giles 1973) and colleagues. In particular, Bourhis & Giles (1977) were able to produce accent divergence by Welsh speakers to an English-speaking authority figure by having the authority figure profess derogatory attitudes towards the Welsh language and culture. The speech of the Welsh speakers showed more Welsh characteristics after demeaning questions than after neutral questions. Similar results were also found by Bourhis *et al.* (1979) in a study on trilingual Flemish students. Thus, accommodation is not a mechanical exemplar-driven process but is rather mediated by the attitude of the listener to the speaker. That is, we believe that the listener constructs a (statistical) model of the speaker and then decides whether to (temporarily) move the speech targets towards the speakers when a sociological message of convergence is desired, or to move them away when wanting to convey an attitude of divergence with the speaker. That is, the speaker's *knowledge OF language* (in the sense of Chomsky (1986)) serves as the basis for the collection of *knowledge ABOUT language*. We believe that the episodic evidence is best understood as the statistical collection of knowledge *about* language, the sort of knowledge we draw on to complete crossword puzzles and knowledge similar to the statistics collected by all animals in various domains (Gallistel 1990). These concepts are not mutually exclusive—rather, knowledge about language is built upon knowledge of language.

A related way to address this tension comes from the cognitive psychology of concepts. In particular, the concepts literature has struggled with the tension between a range of well-documented surface effects (typically perceptual similarity effects) and the necessary and sufficient conditions that are definitional of the 'classical' accounts of concepts (see for review Murphy 2002). The disagreement has been, principally, about how to account for categorical versus gradient phenomena in conceptual processing. For many inferential psychological processes (say, deductive reasoning), a categorical perspective on concepts has been more successful; in contrast, many other processing effects have been best described by gradient, non-categorical representations of some type. How has this conflict been addressed? One approach has been to appeal to a 'theory' view of concepts in which concepts have, constitutively, a 'core' and a 'periphery', where the core corresponds, roughly, to the necessary and sufficient conditions for category membership and the periphery corresponds to the representational architecture that permits gradient characterization. (Naturally, it can be weighted to what extent core versus periphery are more important for various tasks using that concept.) The kind of phenomena that motivated this relatively complex view of concepts include experiments in which it must be the case that both kinds of information are consulted. For example, Armstrong *et al.* (1983) showed that a concept such as 'odd number', surely a classical concept given its formal definition for category membership, nevertheless is also subject to interesting non-categorical effects; i.e. subjects reliably judge '7' to be a better odd number than '237' despite the fact that both are, for computational inferential purposes, totally non-gradient.

This debate on concepts in cognitive psychology seems to us quite analogous to the conflict between abstractionist versus episodic models of speech perception and lexical representation. We advocated an abstractionist model (distinctive features as the representational primitives for phonology and lexicon) that, we argued, can be linked in principled ways to acoustic implementation and also holds hope for developing spectro-temporal primitives (Stevens 2002). But we are, of course, appreciative that there are gradient effects in speech recognition that require explanation. We are, therefore, not at all hostile to all episodic effects in models of speech perception. However, we are against episodic models insofar as they are not just episodic but also explicitly anti-abstractionist. It seems to us an unnecessary consequence to discard abstraction because there is evidence for episodic encoding. From the important demonstration of gradient episodic effects in recognition

(Goldinger *et al.* 1992), it does not follow that categorical-type abstract representations do not exist. Instead, we believe that we can learn from the concepts literature and accommodate both types of effects. How can the disagreement be resolved? Perhaps lexical representation is like conceptual representation of the type discussed above: the mind/brain representation of lexical items is made up of a core (abstract, categorical, symbolic) and a periphery (close to the signal, gradient, statistical), both of which are essential for successful representation and are responsible for different aspect of lexical processing. Some speech or language tasks can be (or must be) driven by the *type*—say, morphological computation—and some tasks can be or must be conditioned by the *token* of the type. Either way, we see no logical reason why episodic and abstractionist models are mutually exclusive since they, for the most part, are designed to account for very different sets of phenomena.

Supported by NIH DC 05660 to D.P. and a Canada–US Fulbright Program award to W.J.I. We thank Norbert Hornstein for numerous insightful and provocative comments on these issues.

REFERENCES

Archangeli, D. & Pulleyblank, D. 1994 *Grounded phonology*. Cambridge, MA: MIT Press.

Armstrong, S. L., Gleitman, H. & Gleitman, L. 1983 What some concepts might not be. *Cognition* **13**, 263–308. (doi:10.1016/0010-0277(83)90012-4)

Belin, P., Fecteau, S. & Bédard, C. 2004 Thinking the voice: neural correlates of voice perception. *Trends Cogn. Sci.* **8**, 129–135. (doi:10.1016/j.tics.2004.01.008)

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N. & Possing, E. T. 2000 Human temporal lobe activation by speech and nonspeech sounds. *Cereb. Cortex* **10**, 512–528. (doi:10.1093/cercor/10.5.512)

Boatman, D. 2004 Cortical bases of speech perception: evidence from functional lesion studies. *Cognition* **92**, 47–65. (doi:10.1016/j.cognition.2003.09.010)

Boemio, A., Fromm, S., Braun, A. & Poeppel, D. 2005 Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat. Neurosci.* **8**, 389–395. (doi:10.1038/nn1409)

Bourhis, R. Y. & Giles, H. 1977 The language of intergroup distinctiveness. In *Language, ethnicity and intergroup relations* (ed. H. Giles), pp. 119–135. London, UK: Academic Press.

Bourhis, R. Y., Giles, H., Leyens, J. & Tajfel, H. 1979 Psycholinguistic distinctiveness: language divergence in Belgium. In *Language and social psychology* (eds H. Giles & R. St Clair), pp. 158–185. Oxford, UK: Blackwell.

Bretthorst, G. L. 1988 *Bayesian spectrum analysis and parameter estimation*. Berlin, Germany: Springer.

Burton, M. W. 2001 The role of inferior frontal cortex in phonological processing. *Cogn. Sci.* **25**, 695–709. (doi:10.1016/S0364-0213(01)00051-9)

Calvert, G. A. 2001 Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb. Cortex* **11**, 1110–1123. (doi:10.1093/cercor/11.12.1110)

Chait, M., Greenberg, S., Arai, T., Simon, J. Z. & Poeppel, D. Submitted. Multi-time resolution analysis of speech.

Chomsky, N. 1986 *Knowledge of language: its nature, origin, and use*. New York, NY: Praeger.

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

Drullman, R., Festen, J. M. & Plomp, R. 1994*a* Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.* **95**, 1053–1064. (doi:10.1121/1.408467)

Drullman, R., Festen, J. M. & Plomp, R. 1994*b* Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* **95**, 2670–2680. (doi:10.1121/1.409836)

Féry, C. & van de Vijver, R. 2004 *The syllable in optimality theory*. Cambridge, MA: Cambridge University Press.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G. & Ashburner, J. 2002 Classical and Bayesian inference in neuroimaging: theory. *Neuroimage* **16**, 465–483. (doi:10.1006/nimg.2002.1090)

Gallistel, C. R. 1990 *The organization of learning*. Cambridge, MA: MIT Press.

Giles, H. 1973 Accent mobility: a model and some data. *Anthropol. Linguist.* **15**, 87–105.

Goldinger, S. D., Luce, P. A., Pisoni, D. B. & Marcario, J. K. 1992 Form-based priming in spoken word recognition: the roles of competition and bias. *J. Exp. Psychol. Learn. Mem. Cogn.* **18**, 1210–1238.

Gordon, M. & Poeppel, D. 2002 Inequality in identification of direction of frequency change (up vs. down) for rapid frequency modulated sweeps. *ARLO J. Acoust. Soc. Am.* **3**, 29–34.

Grant, K. W., van Wassenhove, V. & Poeppel, D. 2004 Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony. *Speech Commun.* **44**, 43–53. (doi:10.1016/j.specom.2004.06.004)

Green, D. M. 1993 Auditory intensity discrimination. In *Human psychophysics* (eds W. A. Yost, A. N. Popper & R. R. Fay), pp. 13–55. New York, NY: Springer.

Greenberg, S. 2005 A multi-tier theoretical framework for understanding spoken language. In *Listening to speech: an auditory perspective* (eds S. Greenberg & W. A. Ainsworth), pp. 411–433. Mahwah, NJ: Erlbaum.

Greenberg, S. & Ainsworth, W. A. 2006 *Listening to speech: an auditory perspective*. Mahwah, NJ: Erlbaum.

Hackett, T. A., Preuss, T. M. & Kaas, J. H. 2001 Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. *J. Comp. Neurol.* **441**, 197–222. (doi:10.1002/cne.1407)

Halle, M. 2002 *From memory to speech and back: papers on phonetics and phonology 1954–2002*. Berlin, Germany: Mouton de Gruyter.

Halle, M. & Clements, G. N. 1990 *Problem book in phonology: a workbook for courses in introductory linguistics and modern phonology*. Cambridge, MA: MIT Press.

Halle, M. & Stevens, K. N. 1959 Analysis by synthesis. In *Proc. Seminar on Speech Compression and Processing*, vol. 2 (eds W. Wathen-Dunn & L. E. Woods), paper D7.

Halle, M. & Stevens, K. N. 1962 Speech recognition: a model and program for research, Reprinted in Halle. 2002.

Hesling, I., Dilharreguy, B., Clement, S., Bordessoules, M. & Allard, M. 2005 Cerebral mechanisms of prosodic sensory integration using low-frequency bands of connected speech. *Hum. Brain Mapp.* **26**, 157–169. (doi:10.1002/hbm.20147)

Hickok, G. & Poeppel, D. 2000 Towards a functional neuroanatomy of speech perception. *Trends Cogn. Sci.* **4**, 131–138. (doi:10.1016/S1364-6613(00)01463-7)

Hickok, G. & Poeppel, D. 2004 Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* **92**, 67–99. (doi:10.1016/j.cognition.2003.10.011)

Hickok, G., Buchsbaum, B., Humphries, C. & Muftuler, T. 2003 Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. *J. Cogn. Neurosci.* **15**, 673–682.

Hinton, G. & Nair, V. In press. Inferring motor programs from images of handwritten digits. In *Proc. 2005 Neural Information Processing Systems*.

Hirsh, I. J. & Sherrick Jr, C. E. 1961 Perceived order in different sense modalities. *J. Exp. Psychol.* **62**, 423–432. (doi:10.1037/h0045283)

Indefrey, P. & Levelt, W. J. 2004 The spatial and temporal signatures of word production components. *Cognition* **92**, 101–144. (doi:10.1016/j.cognition.2002.06.001)

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

Kabak, B. & Idsardi, W. J. Submitted. Speech perception is not isomorphic to phonology: the case of perceptual epenthesis.

Klatt, D. H. 1979 Speech perception: a model of acoustic–phonetic analysis and lexical access. In *Perception and production of fluent speech* (ed. R. A. Cole), pp. 243–288. Hillsdale, NJ: Erlbaum.

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

Labov, W. 1972 *Sociolinguistic patterns*. Oxford, UK: Basil Blackwell.

Labov, W. 2001 *Principles of linguistic change. Social factors*, vol. 2. Cambridge, MA: Blackwell.

Ladefoged, P. & Maddieson, I. 1996 *The sounds of the World's languages*. Cambridge, MA: Blackwell.

Lahiri, A. & Reetz, H. 2002 *Laboratory phonology. Phonology and phonetics*, vol. 7. Berlin, Germany: Mouton de Gruyter.

Lu, T., Liang, L. & Wang, X. 2001 Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat. Neurosci.* **4**, 1131–1138. (doi:10.1038/nn737)

Luo, H., Boemio, A., Gordon, M. & Poeppel, D. 2007 The perception of FM sweeps by Chinese and English listeners. *Hearing Res.* **224**, 75–83. (doi:10.1016/j.heares.2006.11.007)

Marr, D. 1982 *Vision*. San Francisco, CA: Freeman.

Martin, S. E. 1951 Korean phonemics. *Language* **27**, 519–533. (doi:10.2307/410039)

Massaro, D. W., Cohen, M. M. & Smeele, P. M. 1996 Perception of asynchronous and conflicting visual and auditory speech. *J. Acoust. Soc. Am.* **100**, 1777–1786. (doi:10.1121/1.417342)

Meyer, M., Zaehle, T., Gountouna, V. E., Barron, A., Jancke, L. & Turk, A. 2005 Spectro-temporal processing during speech perception involves left posterior auditory cortex. *NeuroReport* **16**, 1985–1989. (doi:10.1097/00001756-200512190-00003)

Miller, G. A. & Nicely, P. E. 1955 An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* **27**, 338–352. (doi:10.1121/1.1907526)

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

Moore, B. C. J. 1989 *An introduction to the psychology of hearing*. San Diego, CA: Academic Press.

Mummery, C. J., Ashburner, J., Scott, S. K. & Wise, R. J. 1999 Functional neuroimaging of speech perception in six normal and two aphasic subjects. *J. Acoust. Soc. Am.* **106**, 449–457. (doi:10.1121/1.427068)

Munhall, K., Gribble, P., Sacco, L. & Ward, M. 1996 Temporal constraints on the McGurk effect. *Percept. Psychophys.* **58**, 351–362.

Murphy, G. 2002 *Big book of concepts*. Cambridge, MA: MIT Press.

Narayan, R., Grana, G. & Sen, K. 2006 Distinct time scales in cortical discrimination of natural sounds in songbirds. *J Neurophysiol.* **96**, 252–258. (doi:10.1152/jn.01257.2005)

Norris, D. & Wise, R. 2000 The study of prelexical and lexical processes in comprehension: psycholinguistics and functional neuroimaging. In *The new cognitive neurosciences* (ed. M. Gazzaniga), pp. 867–880. Cambridge, MA: MIT Press.

Norris, D., McQueen, J. M. & Cutler, A. 2000 Merging information in speech recognition: feedback is never necessary. *Behav. Brain Sci.* **23**, 299–325. (doi:10.1017/S0140525X00003241)

Pelli, D. G. 1999 Close encounters: an artist shows that size affects shape. *Science* **285**, 844–846. (doi:10.1126/science.285.5429.844)

Perkell, J. S. & Klatt, D. 1986 *Invariance and variability of speech processes*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M. & Roberts, T. P. L. 2000 Auditory cortex accesses phonological categories: an MEG mismatch study. *J. Cogn. Neurosci.* **12**, 1038–1055. (doi:10.1162/08989290051137567)

Poeppel, D. 2001 Pure word deafness and the bilateral processing of the speech code. *Cogn. Sci.* **21**, 679–693. (doi:10.1016/S0364-0213(01)00050-7)

Poeppel, D. 2003 The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Commun.* **41**, 245–255. (doi:10.1016/S0167-6393(02)00107-3)

Poeppel, D. & Embick, D. 2005 The relation between linguistics and neuroscience. In *Twenty-first century psycholinguistics: four cornerstones* (ed. A. Cutler), pp. 103–120. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Poeppel, D. & Hackl, M. 2007 The architecture of speech perception. In *Topics in integrative neuroscience: from cells to cognition* (ed. J. Pomerantz). Cambridge, UK: Cambridge University.

Poeppel, D., Wharton, C., Fritz, J., Guillemin, A., San Jose, L., Thompson, J., Bavelier, D. & Braun, A. 2004 FM sweeps, syllables, and word stimuli differentially modulate left and right non-primary auditory areas. *Neuropsychologia* **42**, 183–200. (doi:10.1016/j.neuropsychologia.2003.07.010)

Price, C. J., Thierry, G. & Griffiths, T. 2005 Speech specific neuronal processing: where is it? *Trends Cogn. Sci.* **9**, 271–276. (doi:10.1016/j.tics.2005.03.009)

Robin, D. A., Tranel, D. & Damasio, H. 1990 Auditory perception of temporal and spectral events in patients with focal left and right cerebral lesions. *Brain Lang.* **39**, 539–555. (doi:10.1016/0093-934X(90)90161-9)

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S. & Rauschecker, J. P. 1999 Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci.* **2**, 1131–1136. (doi:10.1038/16056)

Saberi, K. & Perrott, D. R. 1999 Cognitive restoration of reversed speech. *Nature* **398**, 760. (doi:10.1038/19652)

Schonwiesner, M., Rubsamen, R. & von Cramon, D. Y. 2005 Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *Eur. J. Neurosci.* **22**, 1521–1528. (doi:10.1111/j.1460-9568.2005.04315.x)

Schwartz, J. & Tallal, P. 1980 Rate of acoustic change may underlie hemispheric specialization for speech perception. *Science* **207**, 1380–1381. (doi:10.1126/science.207.4431.665)

Scott, S. K. & Johnsrude, I. S. 2003 The neuroanatomical and functional organization of speech perception.

Trends Neurosci. **26**, 100–107. (doi:10.1016/S0166-2236(02)00037-1)

Scott, S. K. & Wise, R. J. 2004 The functional neuroanatomy of prelexical processing in speech perception. *Cognition* **92**, 13–45. (doi:10.1016/j.cognition.2002.12.002)

Scott, S. K., Blank, C. C., Rosen, S. & Wise, R. J. S. 2000 Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* **123**, 2400–2406. (doi:10.1093/brain/123.12.2400)

Shamma, S. 2001 On the role of space and time in auditory processing. *Trends Cogn. Sci.* **5**, 340–348. (doi:10.1016/S1364-6613(00)01704-6)

Sininger, Y. S. & Cone-Wesson, B. 2004 Asymmetric cochlear processing mimics hemispheric specialization. *Science* **305**, 1581. (doi:10.1126/science.1100646)

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

Stevens, K. N. 1998 *Acoustic phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N. 2002 Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* **111**, 1872–1891. (doi:10.1121/1.1458026)

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

Stevens, K. N. & Blumstein, S. E. 1981 The search for invariant acoustic correlates of phonetic features. In *Perspectives in the study of speech* (eds P. D. Eimas & J. L. Miller), pp. 1–39. Hillsdale: Lawrence Erlbaum.

Todd, N. P. 1994 The auditory primal sketch: a multi-scale model of rhythmic grouping. *J. New Music Res.* **23**, 25–70.

Ungerleider, L. G. & Mishkin, M. 1982 Two cortical visual systems. In *Analysis of visual behavior* (eds D. J. Ingle, M. A. Goodale & R. J. W. Mansfield), pp. 549–586. Cambridge, MA: MIT Press.

Van Rullen, R. & Koch, C. 2003 Is perception discrete or continuous? *Trends Cogn. Sci.* **7**, 207–213. (doi:10.1016/S1364-6613(03)00095-0)

van Wassenhove, V., Grant, K. & Poeppel, D. 2005 Visual speech speeds up the neural processing of auditory speech. *Proc. Natl Acad. Sci. USA* **102**, 1181–1186. (doi:10.1073/pnas.0408949102)

van Wassenhove, V., Grant, K. W. & Poeppel, D. 2007 Temporal window of integration in auditory–visual speech perception. *Neuropsychologia* **45**, 598–607. (doi:10.1016/j.neuropsychologia.2006.01.001)

Vuilleumier, P., Armony, J. L., Driver, J. & Dolan, R. J. 2003 Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nat. Neurosci.* **6**, 624–631. (doi:10.1038/nn1057)

Wang, X., Lu, T. & Liang, L. 2003 Cortical processing of temporal modulations. *Speech Commun.* **41**, 107–121. (doi:10.1016/S0167-6393(02)00097-3)

Yabe, H., Tervaniemi, M., Reinikainen, K. & Näätänen, R. 1997 Temporal window of integration revealed by MMN to sound omission. *NeuroReport* **8**, 1971–1974. (doi:10.1097/00001756-199705260-00035)

Yabe, H., Koyama, S., Kakigi, R., Gunji, A., Tervaniemi, M., Sato, Y. & Kaneko, S. 2001*a* Automatic discriminative sensitivity inside temporal window of sensory memory as a function of time. *Cogn. Brain Res.* **12**, 39–48. (doi:10.1016/S0926-6410(01)00027-1)

Yabe, H., Winkler, I., Czigler, I., Koyama, S., Kakigi, R., Sutoh, T., Hiruma, T. & Kaneko, S. 2001*b* Organizing sound sequences in the human brain: the interplay of auditory streaming and temporal integration. *Brain Res.* **897**, 222–227. (doi:10.1016/S0006-8993(01)02224-7)

Zatorre, R. J., Belin, P. & Penhune, V. B. 2002 Structure and function of auditory cortex: music and speech. *Trends Cogn. Sci.* **6**, 37–46. (doi:10.1016/S1364-6613(00)01816-7)