6

The functional architecture of speech perception

DAVID POEPPEL AND MARTIN HACKL

6.1 Introduction

The language system is that aspect of mind/brain function that forms the basis for phonological, morphological, syntactic, and semantic computation. The "currencies" (or the ontology) of this central and abstract computational system are representations that are amodal, for example the concepts "feature" (phonology) or "affix" (morphology) or "phrase" (syntax) or "generalized quantifier" (semantics). Representation and computation with such concepts is typically considered independent of sensory modalities. Of course, the linguistic computational system is not isolated but interacts with other cognitive systems and with sensory–motor interface systems.

With regard to the input and output, the system has at least three modality-specific interfaces: an acoustic-articulatory system (speech perception and production), a visuo-motor system (reading/writing and sign), and a somatosensory interface (Braille). Speech and sign are the canonical interfaces and develop naturally; written language and Braille are explicitly taught: barring gross pathology, every child learns to speak or sign (rapidly, early, without explicit instruction, to a high level of proficiency), whereas learning to read/write Braille requires explicit instruction, is not universal, and occurs later in development.

In this chapter we focus on speech perception, specifically with regard to linguistic constraints and cortical organization. We first outline the key linguistic assumptions, including the concept of "distinctive feature," and then discuss a functional-anatomic model that captures a range of empirical findings.

Topics in Integrative Neuroscience: From Cells to Cognition, ed. James R. Pomerantz. Published by Cambridge University Press.

6.2 The linguistic basis of speech perception

6.2.1 *The central importance of words for language use and understanding*

An essential part of the cognitive ability underlying the linguistic behavior of a competent speaker of a language consists of knowing the words of the language or their constituents (roots). Words cast the two fundamental aspects of language – form and meaning – into single, elementary units. These units are the basic building blocks that are combined in various ways to form larger expressions (pairs of form and meaning) such as phrases, sentences, or texts that are used for communicating information. Models of linguistic competence therefore typically assume two core components: an inventory of building blocks (the set of words stored in the mental lexicon) and a generative engine that manipulates these building blocks to form larger expressions (Figure 6.1).

A central property of this architecture that accounts for the versatility and unparalleled expressive power of natural language is that it is compositional; that is, while at the word level the particular combination of form and meaning is entirely arbitrary, the form and meaning of combinations of words is to a large extent determined by the form and meaning of the words they contain and the particular way these words are put together. To give an example: the English word *cow* is a combination of the phonological form [kau] and the meaning [fully grown female of domestic cattle]. This particular combination of phonological form and meaning into one expression ⟨[kau],[fully grown female of domestic cattle]⟩ is entirely arbitrary. Nothing in the meaning of the word *cow* dictates that its phonological form is [kau]. In fact, the same concept can be described for instance in German with the word *Kuh*, whose phonological form is [ku:]. Vice versa: nothing in the phonological form of *cow* dictates that its (dominant) meaning exponent is [fully grown female of domestic cattle]. In German the meaning associated with the same phonological form [kau] is the root as well as the imperative form of the verb *chew*. Since the particular combination of form and meaning cast into a word is unpredictable, speakers have to learn words one by one and store them in a repository called the mental lexicon. Once words are combined with other words, the resulting expression has predictable form and meaning exponents. For instance, if *cow* is combined with the determiner

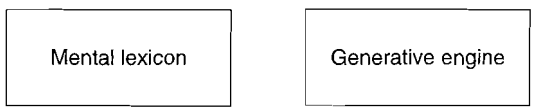

Figure 6.1 The two main components of the language system in the context of contemporary generative linguistic theories include the repository of lexical knowledge as well as the set of elementary operations that generate expressions.

quantifier *every*, the result is the phrase *every cow* whose form is [every cow] and whose meaning is the generalized quantifier [{A: {x: x is a cow} ⊆ A}] – both of which are determined by the properties of the components and the particular way English syntax and semantics demands them to be combined.[1]

A simple illustration of the importance of the compositionality of natural language is provided by the fact that competent speakers understand sentences that they have never encountered before with (roughly) the same ease with which they understand sentences they have encountered many times. To give an example, consider the sentence in (1a), which even though you most likely have never seen before you understand easily to mean the same as the sentence in (1b).

(1) a. John read more books than there are prime numbers smaller than 5.
b. John read more than three books.

This is a remarkable feat that every competent speaker of English is able to accomplish with astonishing ease because she knows all the words in the sentence (1a) and the particular rules that determine how these words are combined to form that sentence. In general, then, understanding an utterance requires of a listener to analyze the signal so that the words that make up the utterance can be identified. The primary cues to achieve this are given by the phonological form of the words. Once the phonological form of a word is recognized it can be used to access the meaning of the word, which in turn is used to build up a representation of the information conveyed by the utterance.

6.2.2 *Identifying words in written language is easy*

The specifics of the task of identifying the words in an utterance depend, of course, on the modality in which the utterance is presented to the recipient. If the utterance is in English and presented in written form, the task is relatively easy because the writing system used to transcribe English typically indicates word boundaries with blank spaces.[2] If that were not the case, understanding written language would be a lot harder. For instance, even a skilled reader will find it much more difficult to read the paragraph below (although it says exactly the same thing as the following paragraph) simply because word boundaries are omitted.

(2) sincetherearenowordboundarysignsinspokenlanguagethedifficulty wefeelinreadingandunderstandingtheaboveparagraphprovidesasimple illustrationofoneofthemaindifficultieswehavetoovercomeinorderto understandspeechratherthananeatlyseparatedsequenceofletterstrings correspondingtothephonologicalformofwordsthespeechsignalisa continuousstreamofsoundsthatrepresentthephonologicalformsof

wordsinadditionthesoundsofneighboringwordsoftenoverlapwhich makestheproblemofidentifyingwordboundariesevenharder

6.2.3 Identifying words in spoken language should be much harder

Since there are no word boundary signs in spoken language, the difficulty we feel in reading and understanding the above paragraph provides a simple illustration of one of the main difficulties we have to overcome when we try to understand speech. Rather than a neatly separated sequence of letter strings corresponding to the phonological form of words, the speech signal is a continuous stream of sounds that represent the phonological forms of words. Worse, not only are the sounds that correspond to the words in an utterance not neatly separated by pauses, they often overlap with sounds of neighboring words (the problem of "linearity"). Additional difficulties arise because actual speech sounds are highly variable across speakers, speech rate and acoustics of the environment (the invariance problem), making the task of speech perception – even if we simplify it in a first approximation as a process of mapping a continuous acoustic signal to a sequence of discrete phonological forms of words – seemingly impossible to master. It is therefore prima facie astonishing how effortless and robust speech perception is for competent speakers of a language.

6.2.4 Speech sounds that correspond to words are highly structured acoustic events

The robustness of speech perception across adverse conditions such as speaker variability, rate of speech, environmental conditions, etc. makes it highly unlikely that all there is to speech perception is a simple, analogue one-to-one mapping between speech sound and the phonological form of a word. (The violation of linearity in the signal is due to factors such as coarticulation.) Instead, it suggests that the speech signal is broken down into more abstract and invariant, linguistically significant components, while many acoustic properties of the signal are filtered out for the purpose of understanding a spoken utterance.[3] But, what are the linguistically relevant components of a speech sound and how is the phonological form of a word represented in an acoustic signal?

We can approach these questions from the other end, so to speak. Minimally, identifying a specific word requires the listener to distinguish it from all other words – in particular from those words that are very similar and differ only minimally from the target. The difference between minimal word pairs can typically be localized to segments of the word. For instance, the difference between the minimal pair *cup* and *cop* is localized in the quality of the vowel. The consonants flanking the vowel are identical. On the other hand, the differences between the minimal pairs *but* and *cut* and *cup* and *cut* are located at the beginning and end of the words, in the identity of the initial and final consonants,

respectively. Observations of this kind lead naturally to the view that the components of the phonological form of a word are segments with distinct melodic identity and the task that speech perception has to accomplish is to identify the segments of the words in the acoustic signal.[4] Segments whose particular melodic identity is exploited by the language to code different words are called *phonemes* and a competent speaker needs to have a representation of the phonemes of her language; that is, she needs to know the ways in which phonological forms of words can minimally differ in her language. On the other hand, the inventory or distinctive segments provide a rough characterization of the space of possible words of a language.[5] Knowing what the possible word forms in your language are is rather useful to solve the problem of speech perception because it constrains the search for the target word given an input signal.

6.2.5 *Sequencing constraints*

Of course, it is not the case that any old combination of phonemes results in a legitimate word. In fact, there are severe restrictions as to what kinds of phoneme sequences are possible. For instance, there are no words in English that contain the phoneme sequence [pf], even though both sounds are phonemes of English. German, on the other hand, allows this sequence. Similarly, certain phoneme sequences are highly restricted in their distribution within a word. For instance, there are no words in English that start with the sequence [rt] although the sequence itself is allowed, as illustrated by the final phoneme sequence of the word *cart*. Constraints of this sort – known as "sonority sequencing constraints" – typically make reference to prosodic units such as syllables, feet, etc. within a word. For instance, the distribution of the sequence [rt] in English is restricted to follow the nucleus of a syllable that is occupied by a vowel while the sequence [tr] is restricted to precede the nucleus as in *track*. Constraints of this sort are highly significant for speech perception because they suggest that the signal is broken down into linguistically significant chunks like syllables and feet[6] within which sequencing constraints provide a powerful filter that constrains the mapping between acoustic signal and phoneme. Although there are a number of universal constraints on syllable structure and sequencing, it is worth pointing out that languages differ as to what kinds of syllables and what kinds of sequencing constraints they employ. If syllable structure and sequencing constraints indeed matter for speech perception, it goes to show that the linguistic competence of speakers – which is thought to be a rather abstract knowledge base of one's language – affects speech perception, often thought to be a lower-level cognitive ability. Compelling results pointing to the relevance of native phonology (syllabic constraints) to perception are shown by, among others, Dupoux *et al.* (1999) and Dehaene-Lambertz *et al.* (2000).

6.2.6 *Allophonic variability*

While this simplifies the perceptual task considerably by temporally narrowing down the problem of identifying which portion of the acoustic signal corresponds to which phoneme, the difficulties mentioned above (variability across conditions, coarticulation) still have to be dealt with. As a simple illustration of the difficulties speakers of English encounter, consider the well-known variation in the realization of the phoneme [t] in American English exemplified in the words listed below (examples from Kenstowicz, 1994).

(3)	a.	**st**em	[t]	
	b.	**t**en	[t^h]	"aspirated t"
	c.	**str**ip	[ṭ]	"retroflexed t"
	d.	a**t**om	[D]	"flapped"
	e.	pan**t**y	[N]	"nasal flap"
	f.	hi**t**	[$t^?$]	"glottalized t"
	g.	bo**tt**le	[$^?$]	"glottal stop"
	h.	pan**ts**		zero

The examples in (3) show that there are at least eight rather different acoustic-phonetic realizations of the same phoneme. Interestingly, speakers of American English report to "hear" the same sound in all these contexts despite the large range of phonetic variability. This suggests that there is a common core to all of these sounds, namely the phoneme /t/, while the various sounds described in (4) are allophonic variations of it.[7] Of course, knowing that English does not make phonemic distinctions between the various realizations listed above simplifies the task of identifying the phoneme in the signal considerably. This suggests once more that linguistic competence matters a great deal for speech perception. Even more, it suggests that the problem of speech perception should be stated in terms of linguistically significant units – especially if they are rather abstract entities such as phonemes that are related to the signal by a many-to-one mapping.

6.2.7 *Phonological processes*

In addition to allophonic variation, there are numerous systematic alternations that phonemes undergo when the words they appear in are combined with other morphemes to form a larger unit. A simple example is given by the alternation the English indefinite article *a* undergoes, depending on whether the following word starts with a vowel or consonant.

(4) a. a book
b. a dog

c. a cat
d. an apple
e. an egg
f. an island

Processes of this kind are very common across languages. Modern Arabic, for instance, displays a slightly more radical version of this phenomenon involving the definite article *al*. Specifically, as the examples listed in (5) and (6) show, the final phoneme of the definite article *al* mimics the melodic identity of the first phoneme of the word following the article (data from Kaye, 1989).

(5) a. al bab "the father"
b. al firaash "the bed"
c. al γurfa "the bedroom"
d. al miftaah "the key"
e. al baab "the door"
f. al qamar "the moon"
g. al kitaab "the book"
h. al yasaar "the left"

(6) a. ad dars "the lesson"
b. ar ruzz "the rice"
c. az zuba "the butter"
d. al turb "the land"
e. as sayyaara "the car"
f. al luγa "the language"
g. an naas "the people"
h. ash shams "the sun"

An inspection of these examples suggests that the particular phonemic make-up of the word-beginning determines whether the preceding definite article changes its appearance or not. In (5), the article keeps its basic form if it combines with words like the ones listed. The data listed in (6) illustrate a robust generalization in Modern Arabic: words that begin with one of the phonemes [d], [r], [z], [t], [s], [l], [n], or [sh] *always* assimilate the preceding determiner, while words that do not begin with one of these phonemes do not (cf. 5). Interestingly, this grouping of phonemes into ones that do and do not affect the preceding article is not random. All of the phonemes listed in the second group that trigger assimilation share an articulatory gesture. Specifically, all of them are produced with the front of the tongue raised toward the top of the mouth, while none of the phonemes listed in the first set of examples (that do not affect the shape of the preceding definite determiner) employs this gesture. The gesture is called *coronal*.

Observations of this kind are abundant across languages and have been taken by linguists to show that phonemes are not atomic units. Instead, they are composites of more elementary entities, so called *distinctive features*. The idea that phonemes are complexes of distinctive features provides a natural and elegant explanation for the fact that phonemes can be grouped into natural classes with respect to phonological processes. In the Arabic example above, the phonemes that trigger regressive assimilation are those that contain the feature [+coronal].[8] A well-known example from the plural morphology of English provides a nice illustration of the explanatory power of the hypothesis that phonological processes are defined over the features that make phonemes rather than the phonemes themselves. Regular English plural formation of nouns employs three distinct suffixes, as the examples in (7) demonstrate (examples from Halle, 1990).

(7) [ɪz] places, mazes, porches, cabbages, ambushes, camouflages
[s] lips, lists, maniacs, telegraphs, hundredths
[z] clubs, herds, colleagues, phonemes, terns, fangs, holes, gears, pies, apostrophes, avenues, cellos, violas

Inspection of these examples shows – quite similar to the assimilation process of Modern Arabic – that the choice of the particular plural suffix is governed by the last phoneme of the word that is pluralized. Specifically, the pattern can be described by the rule in (8):

(8) [ɪz] if the word ends with [s],[z],[č],[š], or [ž], otherwise
[s] if the word ends with [p], [t], [k], [f], or [θ], otherwise
[z]

Even though the rule in (8) is descriptively adequate, it is intrinsically unsatisfactory because it does not explain why the phonemes are grouped in exactly those ways rather than any other combination. The hypothesis that phonemes are feature complexes provides the means to identify the various groups in a principled way: phonemes form a natural class with respect to phonological processes if they share a distinctive feature relevant for the process in question. In the case of regular plural morphology of English nouns, the features in question are [+coronal], [+strident], and [−voice]. The rule that describes the generalization therefore takes on the form in (9).

(9) [ɪz] if the word ends with [+coronal, +strident], otherwise
[s] if the word ends with [−voice], otherwise
[z]

Even though both rules are equally successful in describing the pattern listed in (7), they make different predictions for words that end in phonemes that are not

native to English. One such example is provided by the German name *Bach*. Since the last phoneme of this word is the velar fricative [χ] which contains as one of its components the feature [–voice], the rule schema in (9) predicts correctly that its plural form is realized by [s] and that speakers of English will say [baχs]. The rule schema in (8), on the other hand, incorrectly predicts that the plural of *Bach* is marked by [z] because the phoneme [χ] is not listed in the set that requires the [s]-plural. Clearly, the rule schema that makes reference to the distinctive feature [–voice] offers the better explanation of these facts. Generalizations stated in terms of distinctive features are more powerful in that they do not depend on the specific inventory of phonemes. Instead, they depend on the presence or absence of features, which allows words that employ nonnative phonemes to behave regularly as long as their feature make-up subjects the item to phonological processes native to the language.

Phenomena of this sort are far from being isolated cases. On the contrary, they have been documented in language after language in numerous morphological environments, supporting the same conclusion: phonological processes are defined over units that are smaller than phonemes, that is distinctive features.

The theoretical framework that incorporates these results rejects the significance of the concept phoneme. Rather than being the elementary unit of phonological processes, the phoneme appears to be a mere epiphenomenon that alphabetic writing systems misleadingly present as the fundamental and atomic unit of sound structure. Current phonological theories assume a universal feature inventory of up to 20 distinct features that are used in various combinations by various languages to generate the phoneme inventories of these languages. By the same token, words are no longer viewed as sequences of phonemes. Instead they are sequences of feature complexes. Furthermore, to explain, among other things, the fact that not any combination of features makes a good phoneme, it is typically assumed that the set of features that make up a phoneme is partially hierarchically organized. To illustrate these ideas, consider how the word *cat*, traditionally represented as the phoneme sequence [cæt], is represented in Figure 6.2 in an abbreviated and simplified way as a sequence of feature complexes each associated to a distinct timing slot "x."

6.2.8 *Distinctive features have an articulatory interpretation*

The set of distinctive features is not only motivated by phonological processes; distinctive features have – as pointed out in the example from Modern Arabic – articulatory significance. Recall that the distinctive feature [+coronal] that unifies the phonemes that trigger regressive assimilation of the definite determiner has an interpretation in terms of articulatory gestures. Specifically,

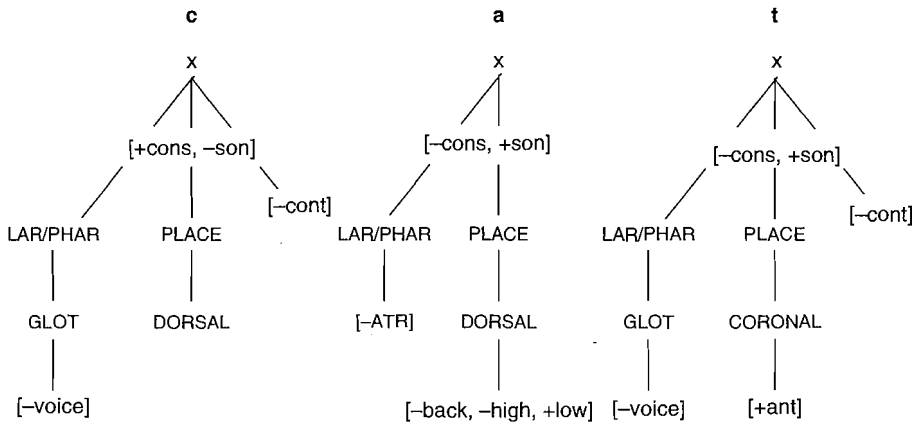

Figure 6.2 The specification of the word "cat" using distinctive features. Each timing slot "x" contains a bundle of features which specify a particular speech sound. This abstract characterization of how words are represented makes explicit the articulatory basis of lexical representation.

[±coronal] represents instructions that the corona of the tongue has to execute in the production of the sounds that are classified as [±coronal]. Similarly, the feature [±voice] represents instructions that the vocal cords execute in the production of sounds that are voiced or voiceless. Quite generally and rather surprisingly, distinctive features, which are the basic units of phonological organization, can be seen as (abstract) instructions of articulatory movements. These articulatory movements have very specific acoustic effects. For instance, the feature [±voice] determines whether the acoustic signal is periodic while the feature [±coronal] has a distinct pattern of formant transitions as an acoustic correlate. In general, the feature complex that constitutes a particular phoneme can be seen as a set of instructions of articulatory movements (quite similar to a ballet score) that the vocal tract has to execute in order to pronounce that phoneme.

6.2.9 *The role of distinctive features in perception*

Given the central importance of distinctive features for the organization of linguistically significant sounds and the fact that their articulatory interpretation results in specific acoustic correlates, it is natural to assume that one of the central aspects of speech perception is the extraction of distinctive features from the signal. In other words, the fact that the basic units of phonological organization can be interpreted as articulatory gestures with distinct acoustic consequences suggests a rather tight and efficient architectural organization of the language system in which speech production and speech perception are intimately connected through the unifying concept of distinctive features.

6.3 The neural basis of speech perception

In the first section we motivated the critical roles that words, syllables, and distinctive features play for the representation of speech. Specifically, we argued that distinctive features play a unifying role in the characterization and explanation of the mental lexicon, speech production, and speech perception. We now turn to a model of the functional anatomy of speech perception. The model builds on the concept of distinctive feature and illustrates how a model for the cortical organization of speech sound processing is natural in the context of the assumptions detailed above.

6.3.1 *The auditory cortex (bilaterally) builds spectro-temporal representations*

The basic challenge for the perceptual system is to transform the incoming signal, a continuously time-varying waveform, into a format that allows the information in the signal to interface with words in the mental lexicon. If words are stored in a format that uses features (Figure 6.2), the goal is thus to extract features (or feature complexes) from the input waveform. The auditory word recognition process thus must minimally include the analysis of the acoustic signal in the ascending auditory pathway and the construction of a spectro-temporal representation of the signal, the extraction of featural information from that representation, and the interface with stored lexical forms. Figure 6.3 illustrates the implicated processes. There is debate about the extent to which these processes are entirely bottom-up or top-down modulated; this debate is not critical to our considerations, although based on present evidence one might favor an "analysis-by-synthesis" view, by which a significant proportion of perceptual analysis involves the (internal) synthesis of potential candidate representations based on sparse data. Recent physiological evidence (van Wassenhove *et al.*, 2005) suggests that such a model is viable.

What brain areas are implicated in this set of processes? Ignoring the (large) contribution of the ascending auditory pathway up to and including the medial geniculate body, the present evidence suggests that primary (core) auditory cortex and adjacent cortical fields (i.e., Brodmann areas 42 and 22) construct

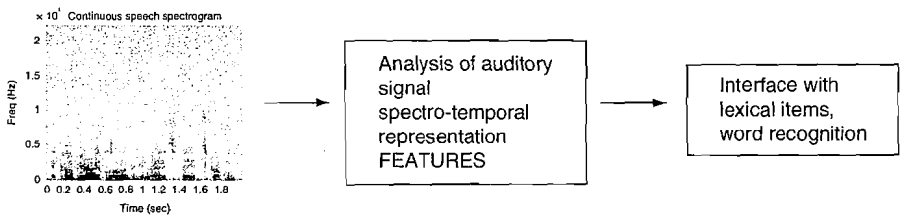

Figure 6.3 The processes implicated in the transformation of the input signal.

spectrotemporal representations of the signal. Neurophysiological data show that there are multi-scale representations that reflect frequency, amplitude, phase, and timing information of acoustic signals. Furthermore, speech perception, when occurring in an ecologically valid way (i.e., "passive" listening without executing laboratory tasks), typically is associated with *bilateral* activity in the auditory cortices, notwithstanding the "left hemisphere imperialism" typical of neurolinguistic research. A variety of findings suggest that the construction of auditory representations of speech is mediated by both hemispheres. *First*, hemodynamic imaging studies show that the activation pattern obtained when subjects listen to speech is always bilateral, including nonprimary areas along Brodmann area 22/STS (Binder *et al.*, 1996, 2000; Mazoyer *et al.*, 1993; Norris & Wise, 2000; Poeppel *et al.*, 1996, 2004; Scott *et al.*, 2000; Zatorre *et al.*, 1992). *Second*, deficit-lesion data suggest that the most selective speech perception deficit, pure word deafness, is a consequence of a lesion pattern that implicates both hemispheres, either directly or by virtue of deafferenting the relevant areas from one another (Buchman *et al.*, 1986; Griffiths *et al.*, 1999; Poeppel, 2001). *Third*, patients in which the dominant (typically left) hemisphere is anaesthetized as part of a presurgical evaluation perform quite well at speech discrimination tasks (Boatman *et al.*, 1998). Overall, the data suggest that (at least one aspect of) speech perception is mediated by the superior temporal lobes of both hemispheres. This general point has been discussed and reviewed in detail by Norris and Wise (2000), Binder *et al.* (2000), and Hickok and Poeppel (2000, 2004).

The hypothesis that the auditory cortex is not just analyzing acoustic structure but is actually sensitive to *featural* information has recently been tested in several MEG studies. Phillips *et al.* (2000) used a mismatch negativity design and manipulated the standard/deviant distributions in a manner such that the analysis of featural information could yield the canonical mismatch responses. These investigators showed that an auditory mismatch field was generated when the only available cue was the featural mismatch, that is the acoustic variation in the test and control conditions could not predict the response. This response localized to auditory cortex. Further experiments by Phillips suggested that the sensitivity to the featural composition is probably left lateralized. To what extent left and right auditory cortices are involved and what their respective contributions might be is discussed below.

6.3.2 *The interface of auditory representations of speech with lexical information*

By conceptual necessity, there must be an interface between sound-based representations of speech and lexical-semantic representations. Where are words represented – or where is the interface? Several kinds of evidence speak to this question. *First*, neuropsychological and neuroimaging data have

implicated left posterior temporo-parietal cortex (but also middle and inferior anterior temporal lobe) in conditions such as semantic dementia, in which the lexical-semantic system appears to be compromised (Damasio, 1992; Price, 2000). *Second*, neuroimaging data often show activation in left posterior temporal cortex for words, including the middle and inferior temporal gyri. *Third*, MEG studies show that the typical "lexical" response (N400/M350) is generated in left posterior temporal cortex, consistent with the position that this part of cortex plays a privileged role in lexical processing (Helenius *et al.*, 1998; Pylkkanen *et al.*, 2002). *Finally*, the large literature on Wernicke's aphasia has as its main generalization that posterior temporal cortex is the neural substrate for the processing of word meaning. Based on such data we hypothesize that the output of the analysis executed in the (bilateral) temporal lobes interfaces with left posterior temporal lobe areas to jointly mediate lexical access.

6.3.3 *Frontal areas are involved in production – but also segmentation*

So far we have been concerned with comprehension and have argued that a feature-based theory is consistent with a view in which auditory speech recognition involves the interplay between auditory cortical areas and posterior cortical areas (retrieve the item that connects auditory form and meaning – i.e., lexical access). We now turn to production. In that domain, the central role of features has been known for a long time. Indeed, the concept of distinctive feature has an articulatory origin and interpretation, as discussed above. When a word is selected for pronunciation, its featural composition must be known in order to provide the correct commands to the articulators. The intuition is illustrated in Figure 6.4, which summarizes the last few steps in the production process, phonological encoding and syllabification, phonetic encoding – a feature-based encoding – and articulation. For a detailed analysis of the steps in the production process, see Levelt (1989) and Indefrey and Levelt (2004). The cortical areas assumed to be involved in these final steps of production are, primarily, Broca's area (for syllabification) and motor cortical areas and other areas known for motor planning (e.g., SMA, cerebellum).

The basic model one might derive is that production is largely a frontal process and perception a purely posterior process. Some recent evidence has complicated this idea. The importance of left inferior frontal cortex in perceptual tasks has been documented in several neuropsychological and imaging studies. For example, Broca's area is reliably activated when subjects are asked to perform sub-lexical tasks involving auditorily presented speech. Recent work has suggested that this anterior activation is driven primarily by processes involved in segmentation (Burton *et al.*, 2000; Zatorre *et al.*, 1992, 1996).

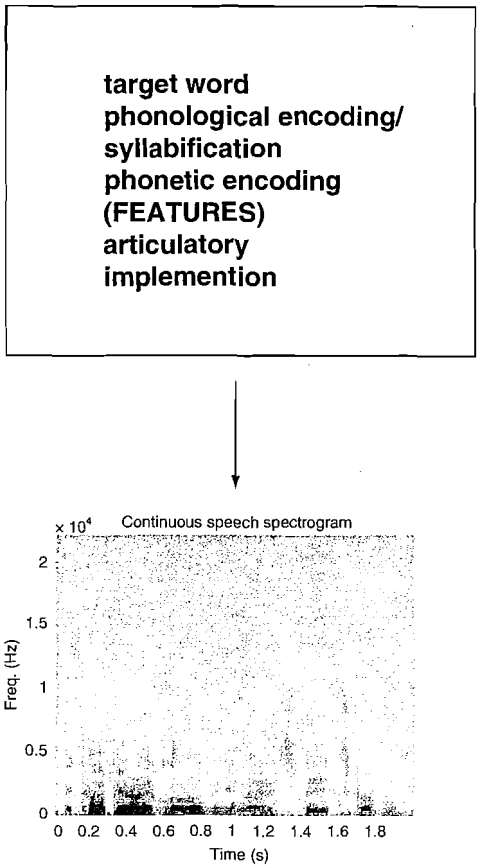

Figure 6.4 Several processes implicated in the planning of speaking a word.

6.3.4 *The coordinate frame problem: how to go from acoustic to articulatory space*

We have satisfactory evidence that frontal areas mediate production (and maybe also aspects of perception) and that temporal areas mediate perception. What, however, about the connection? The challenge is intuitively straightforward: acoustic information is specified in time-frequency coordinates (as shown in the spectrograms in Figures 6.3 and 6.4), but articulatory commands must be specified in motor coordinates, or joint space. It is with respect to this issue that the distinctive feature concept is particularly useful. Because distinctive features have an acoustic and an articulatory interpretation, they may be the currency that can be traded in "brain space" to allow for coordinate transformations. To illustrate why coordinate transformations may be necessary independent operations, consider the task of repeating nonwords. An experimenter provides the auditory stimulus "blicket" or "Krk" and you are asked to repeat it. To execute this trivial task, you cannot turn to lexical information (because there is no lexical entry; in

fact, one can repeat items for which there are no similar items at all). Therefore, to execute the task, you must analyze the signal and turn it into units that can provide instructions for pronunciation. Because the input is in time–frequency coordinates and the output in time–articulator coordinates, there must be a representation that allows you to connect the two representational variants. Features appear to have the right kind of properties. They may be the representational substrate that allows the speaker/listener to transform information in ways to execute both perceptual and motor tasks.

Recent brain imaging data support this hypothesis. Specifically, the role of a temporal/parietal area has been studied. The data show that at least one critical region is deep within the posterior aspect of the Sylvian fissure at the boundary between the parietal and temporal lobes, a region referred to as "area Spt" (Sylvian-parietal-temporal) (Buchsbaum *et al.*, 2001; Hickok & Poeppel, 2004). Area Spt appears to be a crucial part of the network that performs a type of coordinate transformation suggested above, mapping between auditory representations of speech and motor representations of speech. This network could provide a mechanism for the maintenance of parity between auditory and motor representations of speech, as suggested, for example, by the motor theory of speech perception (Liberman & Mattingly, 1985).

6.3.5 *The functional anatomy of the speech processing system*

The functional-anatomic model that emerges has the following properties:

(1) The primary cortical substrate in which sound-based representations of speech are constructed is the bilateral superior temporal cortex (Binder *et al.*, 2000; Hickok & Poeppel, 2000, 2004; Norris & Wise, 2000).
(2) These areas must be organized such that the differentiation between different levels of representation (specifically acoustics, phonetics, and phonology) is maintained (Phillips, 2001; Poeppel, 2001).
(3) Sound-based representations interface (in task-dependent ways) with other systems. An acoustic-phonetic-articulatory "coordinate transformation" occurs in a temporal–parietal–frontal pathway (Buchsbaum *et al.*, 2001; Hickok & Poeppel, 2004) that links auditory representations to motor representations in superior temporal/parietal areas. A second, superior temporal to inferior temporal pathway interfaces speech-derived representations with lexical semantic representations.
(4) Anterior cortical regions play a role in specific perceptual speech segmentation tasks (Burton, 2001). This functional neuroanatomic model is shown in Figure 6.5 and accounts well for activation data as well as the clinical profiles from fluent aphasics (for detailed discussion, see Hickok & Poeppel, 2004).

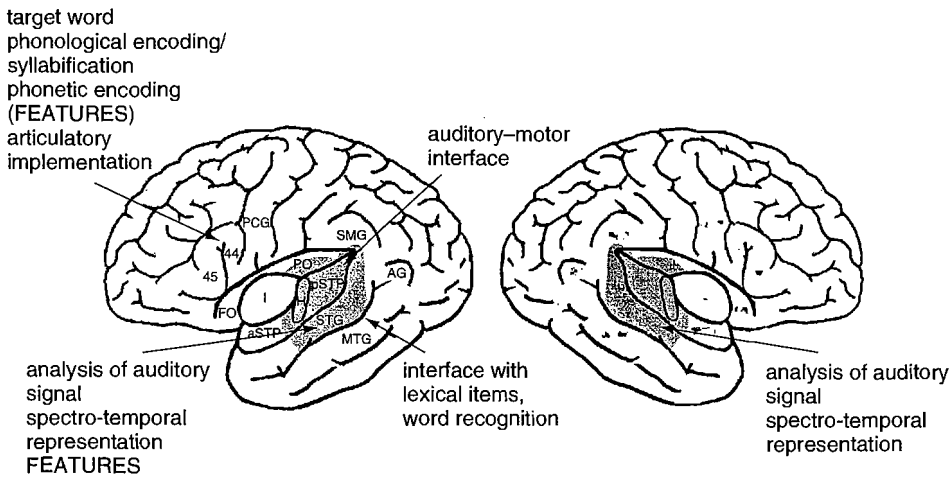

Figure 6.5 The functional anatomy of speech sound processing. The left and right hemispheres are "unfolded" at the Sylvian Fissure to permit visualization of auditory areas. Areas 44/45 are typically taken to be Broca's area. PCG – pre-central gyrus; SMG – supramarginal gyrus; AG – angular gyrus; MTG – middle temporal gyrus; STG – superior temporal gyrus; H – Heschl's gyrus; STP – superior temporal plane; PO – parietal operculum; FO – frontal operculum; I – insula.

6.3.6 *Maintaining functional asymmetry in the auditory areas: the AST model*

Whereas the majority of processes associated with speech and language processing are lateralized, there is an undeniable component to the process that is bilateral. We now turn to the question of what the two hemispheres are doing concurrently in the speech perception process. A growing body of evidence suggests that the right temporal lobe (superior temporal gyrus and superior temporal sulcus, in addition to primary auditory projection areas) plays a role in the analysis of the speech signal (Belin *et al.*, 2000; Binder *et al.*, 2000; Buchman *et al.*, 1986; Burton *et al.*, 2000; Hickok & Poeppel, 2000; Scott *et al.*, 2000) and it is now uncontroversial that an integrated model of the anatomy and physiology of speech perception needs to account for the contribution of both temporal cortices.

It is important to remember, in this context, that the lateralization characteristic of language processing is also well established. The data are consistent with the position that language processing *beyond* the analysis of the input signal is lateralized (Poeppel *et al.*, 2004). The computations that constitute the speech interface are mediated bilaterally, but the "central" computational system (generative engine) that we associate with phonological, morphological, syntactic, and semantic computation is (for the most part) lateralized to the

dominant hemisphere. The bilateral model of speech perception outlined above then brings up an obvious problem: if both hemispheres, specifically both superior temporal gyri, play a role in the analysis of speech, do both areas execute the same computations? The hypothesis proposed here, Asymmetric Sampling in Time (AST), argues that the crucial hemispheric difference derives from the way in which auditory signals are quantized in the time domain. This perspective allows one to maintain the anatomically bilateral nature of processing while preserving functional asymmetry. Moreover, the proposal connects to the question of the primitives argued for in psycholinguistic research.

6.3.7 *Asymmetric sampling in time: the premises*

6.3.7.1 *Temporally evolving information is chunked: integration windows*

We typically think of the passage of time as an arrow, a continuous variable. The central nervous system, on the other hand, takes ongoing events and chunks them in time. Indeed, both psychophysical and electrophysiological data show that perceptual information is analyzed in temporally delimited windows (Näätänen, 1992; Theunissen & Miller, 1995). The importance of the concept of a temporal integration window is that it highlights the discontinuous processing of information in the time domain. The CNS, on this view, treats time not as a continuous variable but as a series of temporal windows, and extracts data from a given window. Recent perspectives on the concept of temporal windows, temporal processing, and temporal integration are provided by Hirsh and Watson (1996), Pöppel (1997), Viemeister and Plack (1993), and Warren (1999).

The link between "temporal integration window" and physiological mechanisms are hypothesized to be oscillatory neuronal activity: the period of an oscillation is assumed to be the duration of the temporal window. Gamma band activity (~40 Hz) is, thus, associated with temporal windows on the order of ~25 ms, theta activity with ~200 ms windows. Neurophysiological data support the idea of "temporal windows" or "sampling". Theunissen and Miller (1995) outline temporal coding in nervous systems and provide physiological definitions of integration windows. Several windows receive support from a neurophysiological perspective, a window associated with a short sampling period (25 ms or 40 Hz) and a window associated with a longer sampling period (200 ms or ~5 Hz); but there are also other integration constants (~2–3 ms and 1000+ ms) that will not be discussed here. Moreover, many electrophysiological recordings in animal preparations have documented sampling at these rates in the form of stimulus-induced or stimulus-related rhythmic brain activity (oscillations) and other physiologic indicators. The short integration window

concept is supported by the data on 40 Hz oscillations in perception. Llinas and colleagues (Joliot *et al.*, 1994) and Singer and colleagues (Singer, 1993) have made arguments for 40 Hz oscillations and synchronization, respectively, as time-based mechanisms to coordinate information. Moreover, high frequency (e.g., gamma) activity has been documented noninvasively in the auditory and visual systems.

Supporting evidence for longer windows comes, for example, from EEG and MEG studies. In particular, Näätänen and colleagues (Näätänen, 1992; Yabe *et al.*, 1997) have argued for long (200 ms/5 Hz) temporal integration windows in auditory cognition. Overall, the notion of temporal integration is motivated by a range of auditory research, both psychophysical and neurophysiological, and very short duration (<5 ms), short-duration (~25 ms), and long-duration (~200 ms) windows have received empirical and theoretical support. Recent psychophysical evidence that shows the relevance of a ~150–300 ms window comes from studies of audiovisual speech desynchronization; it is observed that AV speech tolerates asynchronies within these ranges without serious perceptual degradation (Grant *et al.*, 2004).

One important qualification is that the AST model assumes that there is ongoing gamma band activity, which reflects the cortical "sampling rate." The larger gamma bursts, in this model, occur when the ongoing sampling activity is enhanced during the processing of some stimulus; from the AST perspective, these two aspects of gamma band activity are related but independent.

6.3.7.2 *Sensitivity to time structure*

Numerous hypotheses have been proposed to account for the demonstrable lateralization of function seen across many experimental tasks. For lower-level perceptual processes, there has been some convergence: auditory and visual psychophysical tasks that require fine-grained temporal information for their execution typically implicate the left hemisphere. For example, experiments probing the detection or discrimination of temporal order, temporal sequencing, gap detection, and masking have, on balance, implicated the left hemisphere (for review, see Nicholls, 1996). Recent brain imaging evidence supports the basic notion that there is a leftward bias for the analysis of rapid spectral changes (Zatorre & Belin, 2001; Zatorre *et al.*, 2002).

A frequently articulated view of speech perception argues that the neural mechanisms for speech are lateralized to the left hemisphere. Specifically, it is argued that (1) since the left hemisphere appears to be suited for processing rapid changes and (2) since the speech stream contains many rapid temporal changes there is a natural connection between rapid temporal information and

the left hemisphere (e.g., Tallal *et al.*, 1993). If this hypothesis is on the right track, it is necessary to account for a variety of facts that are problematic on this view; for example: why does the imaging literature on speech perception consistently implicate both hemispheres? Why do neuropsychological data, for example data from pure word deafness, implicate both hemispheres? How are slow spectral changes and small frequency changes analyzed? The model outlined here attempts to capture some of these observations in a unified manner. The model suggests that there may be a bias in left-hemisphere mechanisms for rapidly changing spectral information but (1) there is a stronger bilateral contribution to speech perception than previously assumed and (2) there is a slight bias for spectrally fine-grained and slowly varying information in right-hemisphere mechanisms.

6.3.7.3 Time scales in speech

The critical information contained in speech occurs on multiple time scales. At an intuitive level one can appreciate the temporal (duration) difference between formant transitions, a syllable ("bar"), a multi-syllabic word ("barkeeper"), and a phrase or sentence ("barkeepers listen to drunks"). Rosen (1992) provides a summary of the acoustic and linguistic aspects of the temporal information in speech signals. He shows how the temporal envelope, periodicity, and spectral fine structure are differentially weighted in the encoding of segmental and supra-segmental linguistic contrasts.

Two time scales are relevant to develop the AST hypothesis: the short-duration time constant relevant for encoding formant transitions in stop consonants, approximately 20–40 ms; and the medium-duration time constant relevant for encoding syllables, approximately 150–300 ms. The role of the rapid formant transitions in the encoding of place-of-articulation differences has been appreciated for a long time (Liberman *et al.*, 1967). More recent work has emphasized the importance of syllables. For example, Greenberg has recently argued for the critical importance of syllables in speech recognition (e.g., Greenberg, 1998), and Mehler and colleagues (e.g., Mehler, 1981) have argued for a long time for the primacy of syllables in speech acquisition.

One contrast that is often cited as illustrating time-scale differences is the contrast between consonants (especially stop consonants) and vowels. There exist demonstrable distinctions between vowel and consonant processing. Pisoni (1973), for instance, has shown that short-term memory for vowels is different than short-term memory for consonants in a way that leads to appreciable processing differences. The model we are outlining here is not based on that distinction but on a purely timing-based distinction. The reason the vowel-consonant

distinction is not sufficient to capture the relevant differences is that there is considerable overlap in time/duration between these classes. For example, many consonants can be long (consider /s/ /f/ /sh/ /m/) in the context of short vowels.

6.3.8 *The AST hypothesis and its characteristics*

The AST hypothesis posits that left (nonprimary) auditory areas, perhaps in the superior temporal gyrus, preferentially extract information from short (20–50 ms) temporal integration windows. The right-hemisphere homologues extract information from long (150–250 ms) integration windows. Why these windows? On the one hand, human listeners can resolve the rapid frequency changes typical of formant-transitions. Moreover, listeners have no problem distinguishing temporal order in words (say, e.g., pets vs. pest). This requires a high temporal resolution, at least on the order of 20–50 ms. On the other hand, listeners are able to distinguish among very small frequency changes (say on the order of 5 Hz), for example in the context of prosodic information and music perception. This requires high-frequency resolution. If we assume a frequency resolving power of about 5 Hz, a 200 ms window of analysis is required. By contrast, an analysis window of 25 ms allows a resolution of at best 40 Hz. If we attribute to normal listeners a frequency resolving power of 5 Hz and a temporal resolving power of 25 ms (order threshold), the multiple integration window proposal provides a way to maintain both types of information.

We assume that the initial representation of spectro-temporal receptive fields in primary (core) auditory cortex is bilaterally symmetric. The input signal (heavily preprocessed in the ascending auditory pathway) is analyzed – maybe a multiscale cortical decomposition is performed (Shamma, 2001) – but no strong lateral asymmetry is introduced in core auditory cortex. Subsequently, a "temporally asymmetric" elaboration of the cortical representation occurs in nonprimary areas. The hypothesized mechanism for this is that the proportion of neuronal ensembles with a temporal integration constant of ~25 ms is somewhat larger in left nonprimary areas; in contrast, the proportion of neuronal ensembles with a temporal integration constant of ~200 ms is somewhat larger in the right. As schematized in Figure 6.6, left and right cortical fields contain ensembles with multiple associated scale, but the slight asymmetry in proportion or preference leads to compelling functional asymmetry. Recent work by Zatorre and colleagues (Zatorre & Belin, 2001; Zatorre *et al.*, 2002) as well as by Ivry and colleagues (Ivry & Lebby, 1998; Ivry & Robertson, 1998) addresses similar problems, attempting to account for the lateralization of perceptual phenomena in speech and vision.

The figure also illustrates how to conceptualize the different information types related to the different integration windows. The same input signal will be

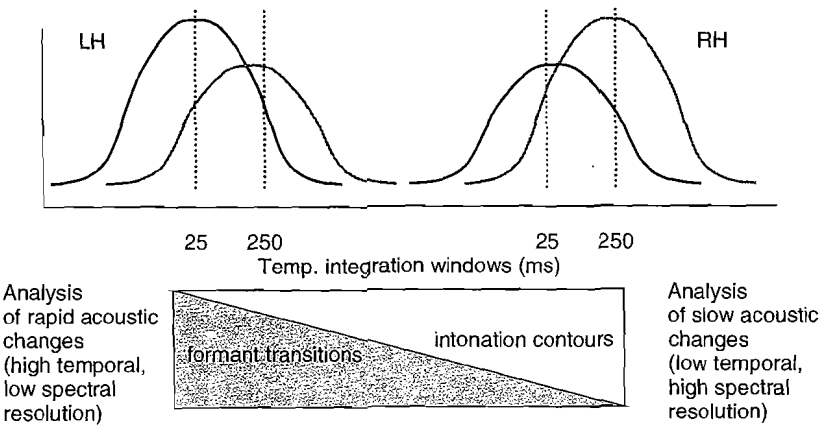

Figure 6.6 Elements of the asymmetric sampling in time (AST) model. The distributions illustrated in the black and gray curves represent the proportion of neuronal circuits with a preferred integration time. The black distributions of neuronal ensembles have a modal integration constant of ~20–50 ms, the gray ensembles a constant of ~150–300ms. Both populations of cells are represented bilaterally in the superior auditory cortex, but by hypothesis, their distribution is asymmetric; the right hemisphere predominately integrates over long-time constants, and the left hemisphere over short-time constants. This asymmetry in temporal integration windows leads to functional asymmetries, as indicated in the bottom panel of the figure.

subjected to two types of analysis that yield complementary information types. If rapidly changing information is relevant, left cortical regions provide the more appropriate neuronal substrate; more gradually changing information or information that requires fine-grained spectral distinctions will be predominantly analyzed by the right auditory cortex. An alternative way to think about this is that there is a "global," lower time-resolution analysis at the syllabic scale and a "local," high temporal resolution analysis at the sub-syllabic scale.

A physiologically motivated way to characterize the AST model is to view it as a sampling issue: the sampling rate of nonprimary auditory areas differs. Left-hemisphere areas sample the spectro-temporal cortical representations built in core auditory cortex at higher frequencies (~40 Hz; gamma band) and right-hemisphere areas at lower frequencies (4–10 Hz; theta and alpha bands).

6.3.9 *Empirical support and challenges*

The model makes a variety of predictions, some of which are unambiguously supported, others of which are problematic. For example, (1) linguistic and affective prosody (at the level of intonation contour) should be associated with right-hemisphere mechanisms. Neuropsychological data investigating the

comprehension of affective prosody support this prediction (Ross *et al.*, 1997). However, experiments on linguistic prosody are problematic. Gandour *et al.* (2000) have shown that at least some aspects of prosody are clearly driven by left-anterior areas. (2) Phonetic phenomena occurring at the level of syllables should be more driven by right-hemisphere mechanisms. This prediction is difficult to examine because syllables by definition contain their phonemic constituents, and the experiments require selective processing of syllables vs. their constituent phonemes. However, there does exist support for the prediction: a recent dichotic listening study. Meinschaefer *et al.* (1999) showed that there was a rightward lateralization when the task demanded a focus on syllabicity rather than the phonemic structure of a given syllable. (3) Music perception should lateralize to the right for most musical attributes (including pitch). Work by Zatorre and colleagues supports this proposal (e.g., Zatorre *et al.*, 1994).

One very specific prediction, the connection between temporal integration windows and oscillatory activity, has been tested. If temporal integration is physiologically reflected as oscillatory activity, shorter time windows associated with the left hemisphere should yield oscillations in the gamma band that have more power in the left. Using whole-head MEG we tested this hypothesis using presentation of auditory stimuli of varying spectral complexity, ripples (dynamic broadband stimuli). High-frequency responses were robustly different for left and right regions, with gamma activity (25–60 Hz) being more pronounced in left temporal cortex (Poeppel *et al.*, 2000). This observation is consistent with the prediction that sensory input is analyzed on different timescales in the left and right.

Zatorre and colleagues have presented some very persuasive work on functional segregation and lateralization in auditory cognition. For example, in the seminal PET study by Zatorre *et al.* (1992) the same consonant-vowel-consonant stimulus set was associated with a strong leftward (frontal) lateralization when subjects made judgments requiring place-of-articulation analysis and a rightward lateralization when subjects judged pitch differences among the stimuli. In work on music perception, Zatorre has shown an association between melodic analysis and rightward lateralization, both using imaging and neuropsychological techniques (Zatorre, 1997; Zatorre *et al.*, 1994). Zatorre discusses the hemispheric differences observed from a perspective that is very comparable to ours: he argues that left-temporal cortex is specialized for temporal analysis and right-auditory cortex for spectral analysis. What we offer in addition is a proposed mechanisms that builds on the time constants of neuronal ensembles. Importantly, recent fMRI evidence supports various aspects of the model. Boemio *et al.* (2005), using nonspeech signals inspired by certain auditory properties of speech, tested the AST hypothesis rather directly and report a timing-

induced asymmetry. An anatomic model is offered to account for the activations and their distributional differences as a function of stimulus timing. Hesling *et al.* (2005) use a speech-derived stimulus and also support the hypothesized generalizations.

6.4 Conclusions

Speech perception is the process of extracting information from an acoustic signal and constructing the appropriate representation that can interface with the stored items in your mental lexicon and the linguistic computational system (Blumstein, 1995; Chomsky, 1995). In the first part of the article we showed why speech perception is hard – for example, because there is no one-to-one mapping from stretches of sound to phonemes and because there are no (obvious) invariant properties in the signal. That these difficulties are not trivial is attested by the fact that automatic speech recognition technology is not particularly far along. Nevertheless, the human brain deals with the problems effectively. We suggest that the efficacy of the system derives from at least three properties of the speech processor. First, a speaker's *knowledge* of phonology significantly helps the process. Second, the problem is broken down in *space*: multiple areas contribute to different aspects of the problem (much like in vision). Third, the problem is broken down in *time* by analyzing signals on different time scales.

A prerequisite for the development of a model of the cognitive neuroscience of speech is theoretical agreement on what the appropriate linguistic units of study are. Here, we built on the assumption that the basic unit of speech that makes sense of neuronal data is the distinctive feature. It is the concept that best connects linguistic theory to biological data.

Notes

1. The predictability claim has to be qualified somewhat. There are cases of larger expressions whose particular form – meaning combination is not (entirely) predictable from their components. Well-known examples are idiomatic expressions like *kick the bucket*, whose meaning [die] is not predictable from the meaning of the components and their combination. Unpredictability is often taken to be a defining property of items that are stored in the lexicon.
2. Of course this is not always true. Morphological derivatives of words such as compounds or inflectional derivatives are typically not signaled by blank spaces. On the other hand, phrasal idioms like the ones mentioned in the previous footnote are often treated as basic lexical units. Nevertheless, the orthographic rules demand the use of blank spaces inside those idioms.
3. A plausibility argument can be given as follows:

imagine that there was no internal structure to speech sounds or to the phonology of a word. Each word in a language would therefore have a unique acoustic exponent. These acoustic signals would be simply listed without any inherent organization expressible for instance through a similarity matrix. Such a system would show an effect of the lexicon size on the efficacy of speech perception. For example, we estimate the average lexicon size of an English speaker at 10 000 to 20 000 words, while speakers of some Southeast Asian languages are estimated to have a vocabulary size of over 100 000 words. Given this difference, it should be much harder for speakers of one of these Southeast Asian languages to identify any word in the signal, no matter what its phonological form or acoustic exponent is simply because the search space is an order of magnitude larger. However, while it is well-known that the size of the lexicon matters locally, that is, if there are many similar sounding words it takes longer to identify one specific word within this set, it has never been reported that the lexicon size has a global effect.

4. Writing systems such as the one used to transcribe English represent relatively closely the intuitions of speakers that words are made of segments.
5. The space is determined by the phoneme inventory and prosodic constraints on words.
6. Feet and higher prosodic units like phonological word and phrase are the relevant unit for assignment of stress and intonation patterns.
7. The term "allophone" describes a particular realization of a phoneme. Since languages have different phoneme inventories, what is an allophone in one language can be a phoneme in another.
8. The term "coronal" appeals to the corona (tip and blade) of the tongue. Distinctive features are typically but not always assumed to be equipollent, that is specified for ±.

References

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, **403**(6767), 309–12.

Binder, J. R., Frost, J. A., Hammeke, T. A., *et al.* (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cerebral Cortex*, **10**, 512–28.

Binder, J. R., Frost, J. A., Hammeke, T. A., Rao, S. M., and Cox, R. W. (1996). Function of the left planum temporale in auditory and linguistic processing. *Brain*, **119**, 1239–47.

Blumstein, S. (1995). The neurobiology of the sound structure of language. In M. Gazzaniga, ed., *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Boatman, D., Hart, J., Lesser, R. P. *et al.* (1998). Right hemisphere speech perception revealed by amobarbital injection and electrical interference. *Neurology*, **51**(2), 458–64.

Boemio, A., Fromm, S., Braun, A., and Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience*, **8**, 389–95.

Buchman, A., Garron, D., Trost-Cardamone, J. E., Wichter, M. D., and Schwartz, M. (1986). Word deafness: one hundred years later. *Journal of Neurol Neurosurgery Psychiatry*, **49**(5), 489–99.

Buchsbaum, B. R., Hickok, G., and Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science*, **25**, 663–78.

Burton, M. W. (2001). The role of inferior frontal cortex in phonological processing. *Cognitive Science*, **25**, 695–709.

Burton, M. W., Small, S. L., and Blumstein, S. E. (2000). The role of segmentation in phonological processing: an fMRI investigation. *Journal of Cognitive Neuroscience*, **12**, 679–90.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Damasio, A. R. (1992). Aphasia. *The New England Journal of Medicine*, **326**(8), 531–9.

Dehaene-Lambertz, G., Dupoux, E., and Gout, A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience*, **12**, 635–47.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J. (1999). Epenthetic vowels in Japanese: a perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, **25**, 1568 –78.

Gandour, J., Wong, D., Hsieh, L., *et al.* (2000). A crosslinguistic PET study of tone perception. *Journal of Cognitive Neuroscience*, **12**(1), 207–22.

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

Greenberg, S. (1998). A syllable-centric framework for the evolution of spoken language. *Brain and Behavioral Sciences*, **21**(4), 518.

Griffiths, T. D., Rees, A., and Green, G. G. R. (1999). Disorders of human complex sound processing. *Neurocase*, **5**, 365–78.

Halle, M. (1990). Phonology. In D. Osherson and H. Lasnik, eds., *Language. An Invitation to Cognitive Science*. Cambridge, MA: MIT Press.

Helenius, P., Salmelin, R., Service, E., and Connolly, J. E. (1998). Distinct time courses of word and context comprehension in the left temporal cortex. *Brain*, **121**, 1133–42.

Hesling, I., Dilharreguy, B., Clement, S., Bordessoules, M., and Allard, M. (2005). Cerebral mechanisms of prosodic sensory integration using low-frequency bands of connected speech. *Human Brain Mapping* **26**(3), 157–69.

Hickok, G. and Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends Cognitive Sciences*, **4**, 131–8.

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

Hirsh, I. and Watson, C. S. (1996). Auditory psychophysics and perception. *Annual Review of Psychology*, **47**, 461–84.

Indefrey, P. and Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, **92**, 101–44.

Ivry, R. and Lebby, P. (1998). The neurology of consonant perception: specialized module or distributed processors? In M. Beeman and C. Chiarello, eds., *Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience*. Mahwah, NJ: Erlbaum, pp. 3–25.

Ivry, R. B. and Robertson, L. C. (1998). *The Two Sides of Perception*. Cambridge, MA: MIT Press.

Joliot, M., Ribary, U., and Llinas, R. (1994). Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. *Proceedings of the National Academy of Sciences USA*, **91**(24), 11748–51.

Kaye, J. (1989). *Phonology: A Cognitive View*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kenstowicz, M. (1994). *Phonology in Generative Grammar*. Cambridge, MA: Blackwell.

Levelt, W. J. M. (1989). *Speaking*. Cambridge, MA: MIT Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431–61.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1–36.

Mazoyer, B. M., Dehaene, S., Tzourio, N., *et al.* (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, **5**(4), 467–79.

Mehler, J. (1981). The role of syllables in speech processing – infant and adult data. *Philosophical Transactions of the Royal Society B* **295**, 333–52.

Meinschaefer, J., Hausmann, M., and Güntürkün, O. (1999). Laterality effects in the processing of syllable structure. *Brain and Language*, **70**, 287–93.

Näätänen, R. (1992). *Attention and Brain function*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Nicholls, M. E. (1996). Temporal processing asymmetries between the cerebral hemispheres: evidence and implications. *Laterality*, **1** (2), 97–137.

Norris, D. and Wise, R. (2000). The study of prelexical and lexical processes in comprehension: psycholinguistics and functional neuroimaging. *The New Cognitive Neurosciences*. G. M. Cambridge: MIT Press.

Phillips, C. (2001). Levels of representation in the electrophysiology of speech perception. *Cognitive Science*, **25**, 711–31.

Phillips, C., Pellathy, T., Marantz, A., *et al.* (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *Journal of Cognitive Neuroscience*, **12**, 1038–55.

Pisoni, D. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, **13**, 253–60.

Poeppel, D. (2001). Pure word deafness and the bilateral processing of the speech code. *Cognitive Science*, **25**, 679–93.

Poeppel, D., Boemio, A., Simon, J., *et al.* (2000). *High-frequency Response Asymmetry to Auditory Stimuli of Varying Spectral Complexity*. New Orleans: Society for Neuroscience.

Poeppel, D., Wharton, C., Fritz, J., *et al.* (2004). FM sweeps, syllables, and word stimuli differentially modulate left and right non-primary auditory areas. *Neuropsychologia*, **42**, 183–200.

Poeppel, D., Yellin, E., Phillips, E., *et al.* (1996). Task-induced asymmetry of the auditory evoked M100 neuromagnetic field elicited by speech sounds. *Cognitive Brain Research*, **4**, 231–42.

Pöppel, E. (1997). A hierarchical model of temporal perception. *Trends Cognitive Sciences*, **1**(2), 56–61.

Price, C. (2000). The anatomy of language: contributions from functional neuroimaging. *Journal of Anatomy*, **197**, 335–59.

Pylkkanen, L., Stringfellow, A., and Marantz, A. (2002). Neuromagnetic evidence for the timing of lexical activation: an MEG component sensitive to phonotactic probability but not to neighborhood density. *Brain and Language*, **81**, 666–78.

Rosen, S. (1992). Temporal information is speech: acoustic, auditory, and linguistic aspects. *Philosophical Transactions of the Royal Society London B*, **336**, 367–73.

Ross, E. D., Thompson, R. D., and Yenkosky, J. (1997). Lateralization of affective prosody in brain and the callosal integration of hemispheric language functions. *Brain and Language*, **56**(1), 27–54.

Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, **123**, 2400–6.

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends Cognitive Sciences*, **5**, 340–8.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review Physiology*, **55**, 349–74.

Tallal, P., Miller, S., and Fitch, R. (1993). Neurobiological basis of speech: a case for the preeminence of temporal processing. *Annals New York Academy of Science*, **682**, 27–47.

Theunissen, F. and Miller, J. P. (1995). Temporal encoding in nervous systems: a rigorous definition. *Journal of Computational Neuroscience*, **2**, 149–62.

van Wassenhove, V., Grant, K., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences USA*, **102**, 1181–6.

Viemeister, N. F. and Plack, C. J. (1993). Time analysis. In W. A. Yost, A. N. Popper, and R. R. Fay, eds., *Human Psychophysics*. New York: Springer.

Warren, R. M. (1999). *Auditory Perception*. Cambridge, UK: Cambridge University Press.

Yabe, H., Tervaniemi, M., Reinikainen, K., and Näätänen, R. (1997). Temporal window of integration revealed by MMN to sound omission. *Neuroreport*, **8**, 1971–4.

Zatorre, R. J. (1997). Cerebral correlates of human auditory processing: perception of speech and musical sounds. In J. Syka, ed., *Acoustical Signal Processing in the Central Auditory System*. New York: Plenum Press, pp. 453–468.

Zatorre, R. and Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, **11**, 946–53.

Zatorre, R., Belin, P., and Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, **6**(1), 37–46.

Zatorre, R. J., Evans, A. C., and Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *Journal of Neuroscience*, **14**, 1908–19.

Zatorre, R. J., Evans, A. C., Meyer, E., and Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science*, **256**, 846–9.

Zatorre, R. J., Meyer, E., Gjedde, A., and Evans, A. C. (1996). PET studies of phonetic processing of speech: review, replication, and reanalysis. *Cerebral Cortex*, **6**(1), 21–30.