Phonetic variation as communicative system: Perception of the particular and the abstract

Sarah Hawkins
Department of Linguistics, University of Cambridge, UK
sh110@cam.ac.uk

This paper addresses three issues raised by observations of the role of systematically-varying fine phonetic detail in speech: the representation of linguistic knowledge so that it simultaneously accommodates detail and abstraction; the nature of phonetic categories; and the type of processes that must underlie speech perception. Evidence is adduced from the anatomy and physiology of the brain to support arguments for a polysystemic view of speech perception.

1. The particular

1.1 History and definition of the term “fine phonetic detail” (FPD):

When John Local talked in the 1980s about fine phonetic detail, he was describing phonetic phenomena such as resonances associated with liquid consonants in English that were systematically distributed but not systematically treated in conventional phonetic description. They tended to be hard to notice too. Since then, the term FPD has come to be applied to anything that is not considered a major, usually local, perceptual cue for phonemic contrasts in the citation forms of lexical items. In this broader usage, some FPD is indeed “fine”, and subtle, but other types are perfectly audible; they have just not been factored into the prevailing view that perceptual processing of phonetic information is largely aimed at identifying strings of features or phonemes that allow words to be distinguished.

FPD cuts across traditional subdisciplines of enquiry. It does not just distinguish words, but also the wider phonological and grammatical structure of the message. For example, grammatical function words have a narrower range of sound patterns than content words, and undergo different connected speech processes; and each type of function word (e.g. auxiliary verbs, articles) has its own distinct system of contrasts (e.g. Ogden 1999; Local 2003). FPD also reflects function and structure of the smaller units that comprise words, and of larger groupings, influencing everything necessary for successful communication: phonological, morphological, grammatical, pragmatic, interactional (see *Phonetica* Volume 61, especially papers by Local, Ogden, and Plug). FPD indicating a single linguistic distinction can involve many acoustic properties distributed over long stretches of speech (Hawkins and Smith 2001; Local 2003) e.g. traces of English /r/ can occur several syllables before the main /r/ segment and influence perception (Hawkins and Slater 1994; West 1999a, b; Heid and Hawkins 2000; Coleman 2003). Even well-researched distinctions like coda voicing involve multiple distinctions, some of which are less local than was until recently assumed (Hawkins and Nguyen 2004). Thus, much FPD—the sort discarded by traditional abstractionist phonological and perceptual models as uninteresting or due to random effects—in fact systematically reflects many different aspects of meaning that are crucial to the maintenance of normal conversation: lexical, grammatical, and interactional differences.

As Nguyen notes “The goal of current research on FPD is to show that FPD is important in speech [processing], and, therefore, that a change of theoretical perspective is called for.” (in press :8). If such a changed perspective proves valuable, the term fine phonetic detail can largely be replaced by “phonetic information”.

1.2 Perception of FPD

Perceptual experiments show that FPD is often perceptually salient, perhaps especially in adverse listening conditions (reviews: Hawkins and Smith 2001; Hawkins 2003; Nguyen in press). Listeners even know about, and in the right circumstances use, spectral and/or temporal distinctions whose mean distributions in natural speech corpora differ by only a few milliseconds or tens of Hz. However, the right circumstances are crucial: words and phrases can be pronounced in very different ways, especially in the more casual or faster styles typical of most natural speech, and the changes are often systematic, yet the resulting stretch of sound may only be understandable in the context in which it was spoken. Listeners attend to different types of phonetic information, and use the same
information in different ways, depending on the entire phonetic and situational context (cf. Mattys et al. 2005).

Some FPD is always important e.g. grammatically-distinct variants of had: [aɪ hæd sʌŋ] is less usual than [aɪd sʌŋ] and differs in meaning, whereas [aɪ hæd tɛn] is normal and [aɪd tɛn], with no schwa offglide to the diphthong, impossible in most varieties. Conversely, some FPD is hardly noticeable. Yet FPD that is hard to detect in quiet, e.g. /r/-resonances, can increase intelligibility in noise. While not all observed FPD is always perceptually salient, ignorance of FPD may explain the disproportionate difficulty of understanding a foreign language in noise.

These observations are important because standard phonology and most perceptual models assume that lexical form is the most important thing to be distinguished, and that the citation form of the word is the most important lexical form. Yet citation forms are heard least often; and, consistent with their restricted range of linguistic functions, the FPD that they exhibit is relatively restricted. In natural, and perhaps especially conversational speech, different forms of FPD are widespread and enormously informative about the function of the utterance in discourse.

In sum, we know that listeners know about FPD, because they use it when appropriate, and can learn about new instances. So it must be monitored during listening, and remembered. Experiments demonstrate that speech which lacks natural detail is remembered less effectively, and that learning new sound categories is more effective when the stimuli include the variation typical of natural speech (Duffy and Pisoni 1992; Pisoni et al. 1994). We understand little about what governs when we need the FPD, for this is less researched than influences on spoken word recognition like word frequency, ambiguity, and types of competition.

These observations should be accounted for in a unified theory of speech communication, but no current theory approaches this ideal. A step towards it may be provided by Polysp, a perceptual framework whose linguistic system emphasises context and communicative function (including grammar) over phonological form. Influenced by statistical pattern matching and declarative phonology, Polysp simultaneously accommodates detail and abstraction in representing linguistic knowledge. Consequences for the conceptualisation of perceptual processes, including interplay between knowledge and physical signal, reflect behavioural evidence and current understanding of the anatomy and physiology of the brain, as well as being congruent with perceptual processes in other modalities.

2. “Using” the detail: exemplar and abstractionist approaches

If listeners know about, and in the right circumstances use, FPD to understand the meaning of speech, then FPD needs to be represented in our models of speech perception. The obvious class of models to do this are those that assume exemplar representation of speech. These have been discussed in the literature for some years and with some passion. Computational models (e.g. Hintzman 1988; Nosofsky 1988) showed that classification into categories from groups of instances is possible, and the principles have been applied to speech (e.g. Johnson 1997b; Goldinger 1998; Pierrehumbert 2002, 2003; Lachs et al. 2003). Others, (e.g. Norris et al. 2000) have proposed counterarguments.

Each side’s argument focuses on a different level, and has limited data and/or a limited view of the goals of speech production/perception. Those arguing for exemplar models generally show that we attend to details and offer ways of classifying new exemplars with respect to existing distributions held in memory. But the focus has been mainly on computational feasibility: a linguistic exemplar model that is rich enough to account for how a language system might work has yet to be developed.

Those arguing against exemplar approaches typically espouse models that assume that the first stage of processing speech involves abstracting from the signal to feature bundles, phoneme strings, or similar. They include enough linguistic structure to account for certain types of linguistic knowledge, and show that processing which is interpretable as being at this abstract level has behavioural validity.

It is worth standing back from the details of both arguments to consider what is needed. Can we reconcile the perceptual importance of FPD with the apparently abstract nature of linguistic knowledge? It is clear we attend to details; it is clear we abstract to general categories: there seems no point polarising theories when the data support neither extreme.

For proponents of abstractionist approaches to speech understanding (e.g. Norris 1994; Norris et al. 2000), a major weakness of exemplar models is their inability to generalise—that is, to abstract.
However, all exemplar models have ways of assessing how similar one item is to another. The details of how this is achieved in computational models are less important to this discussion than the fact that it can be achieved. More important is the point that, to class two utterances as ‘the same’ or ‘different’ requires abstraction. The fact that existing models are inexplicit about what happens after this initial, minimal, abstraction need not mean that the general approach is wrong—or right. The problem is not inability to abstract, but lack of a suitable linguistic model to abstract to, and hence failure to explore how stored exemplars might inform human understanding. We have general methods and some specific examples (e.g. Hawkins and Smith 2001; Coleman 2002; Johnson 2006) but not a linguistically-sophisticated working model. The Polysp framework might help remedy this limitation.

Abstractionist models suffer from complementary weaknesses. They postulate clear processes but lack generalizability. Although views are changing, work in this genre is limited by unquestioned assumptions and unrealistic simplifications. Weaknesses include (1) the theoretical presuppositions or processes invoked to interpret the impoverished (phoneme/feature) input, despite the original signal containing much of this information, and (2) focus on a variety of speech perception that represents only a fraction of what human speech perception is actually about. Everyone has always known that these models have limited application, because the meaning of a message cannot be derived from the sum of its individual word meanings. Nevertheless, focussing on content word recognition has diverted attention from the main purpose of speech communication, which is normally some balance between understanding a whole message, and establishing or maintaining phatic communication between individuals: even innovative recent work asserts that the “real task” is word identification (cf. Scharenborg et al. 2005:878). Relatedly, the theoretical constructs of a mental lexicon, uniformly structured within and between individuals, and an identifiable stage of prelexical representation, seem rarely questioned (though see Gaskell (2002), Elman (2004)).

The assumption of “clean” (abstract) input suitable for describing lexical contrast in careful speech, but nothing more, encourages assumptions that do not generalise to most speech. For example, mechanisms are proposed to account for successful word identification when one phone(me) in a word is mispronounced while the others are appropriate for clear speech e.g. “[ʃɪɡərt]” for “how a drunk might say the word cigarette” (Scharenborg et al. 2005:911). Everyone would accept that there are usually more acoustic-phonetic cues to drunkenness than the realisation of local cues to a single phone(me). But the crucial point is this: the way phonemic realisations vary is informative in itself. Meaning, even word identity, can often only be gleaned in the context of the pronunciation of a significantly longer part of the utterance (e.g. Kohler 1998; Ernestus et al. 2002). The greater the reduction (in terms of phoneme count) from the canonical citation form, the more of the utterance listeners need to hear, up to the entire utterance Even simple intelligibility is not related to the number of syllables involved (Pickett and Pollack 1963). In other words, utterances spoken in more casual styles are coherent, but need their context to be understood.

Implications for theory are profound. For example, the Possible Word Constraint, PWC in the SHORTLIST model (Norris et al. 1997) is a sensible solution to the problem of finding “junk” sounds in putative word sequences during segmentation of continuous speech. In SHORTLIST, it operates, with other factors, when a sequence of consonants is found without a vowel and so cannot be readily mapped onto citation-form phonemic word representations. It is used, amongst other things, for identifying neologisms. But the PWC is unrealistic for normal speech: sound sequences that represent words but lack vowels abound (and, she’s) and are easy to understand in context. Indeed, they convey rich meaning. A more realistic formulation might be that the PWC operates when a sound chunk does not seem to be part of a syllable. But even this has problems, the most important being the assumption that when a sound chunk does not fit a known word pattern, it is nevertheless identifiable as a vowel or simple segment like [f]. This is often untrue. Broadly, neglect of FPD could result in mistaken application of the PWC to unstressed syllable reductions, consonant assimilations, and the realisation of consonant clusters especially (in Germanic languages) in codas and unstressed syllables. E.g. she should in a casual she should go now may be realised as [ʃʃʃ].

In sum, neglect of phonetic facts, of sophisticated linguistic models, and of the most normal forms and purposes of speech, have resulted in both exemplar and abstractionist models with limited applicability. The position explored here is that both exemplar representation and abstraction are necessary. The crucial issue is to know what to do with the detail. This has been discussed before
(Hawkins and Smith 2001; Hawkins 2003) as the theoretical framework Polysp (for POLYsystemic Speech Perception) but subsequent comments suggest that Polysp was seen by at least some readers as strongly pro-exemplar. In consequence, and also because it is part of a phonology conference, this paper gives more attention to the abstract attributes of Polysp’s hybrid approach.

3. Perception of the abstract: The representation of FPD

The above arguments demand that the representation of phonetic and phonological categories takes account of the functional and structural properties of the entire utterance. (An utterance in this sense has a particular discourse function or functions and can be anything from a word or phrase, to a string of phrases, sentences, or similar.) Firthian Prosodic Analysis (FPA: Ogden and Local 1994; Simpson 2005), takes this approach. Although complex utterances have not yet been completely described within the system, a fundamental principle is that communicative function determines speech form (pronunciation). Thus the context of a given form, or potential perceptual unit, must be as carefully represented as the form itself, and as influential in determining perceptual responses. Relatively, perception of lawful higher-order structure is fundamental to word identification, and interdependent with perception of lawful segmental structure. Realistically for normal listening tasks, syntagmatic relationships are vital.

Previous papers on Polysp (Hawkins, 2001; Hawkins, 2003) state that FPA is a type of declarative phonology. Both are polysystemic and assume congruence between types of linguistic structure. I use FPA because its attention to phonetic detail and its proof of concept in the Prosynth synthesis system, on which I worked, offered a valuable way to explore perception. To those who question this choice, I can only say that I am not a phonologist and cannot contribute to debates between competing phonological theories. But I can try to use general insights to further understanding about speech perception. At this stage, competing details seem relatively unimportant. Far more pressing is to explore the implications of a change in thinking about speech perception, away from early abstraction and strict separation of linguistic levels, and towards a flexible, process- and function-oriented model of human communication in which subsystems are congruent.

In FPA, a difference in FPD is reflected in different prosodic/grammatical structures: when the linguistic structures that describe two utterances differ, then their sounds differ. Polysp retains the polysystemicity but reverses the logic, so that in perception, a reasonable hypothesis is that if the sounds in two utterances differ, then one or more things in their structures differ. Thus in Polysp, small parts of the sensory signal (such as acoustically distinct segments) can only be processed in terms of their wider context. This contrasts with standard speech perception modelling, in which phonetic and phonological category members, once named, are treated as if independent of other elements. Polysystemic, function-oriented phonology does not allow this, and so may offer a fruitful way of thinking about how the physical signal may be decoded into a more abstract representation.

The basic principles are illustrated in Figure 1 with data discussed before in the non-standard framework of FPA (e.g. Ogden et al. 2000; Hawkins and Smith 2001; Hawkins 2003): the contrast in the first syllables of the words mistimes and mistakes. Other theoretical analyses differ in detail but agree that the difference depends on the productivity of the morpheme mis- and its phonological consequences. The point here is only that there is a morphologically-conditioned difference in the spectrotemporal pattern associated with the otherwise identical beginnings of these two words: both iambic bisyllabic verbs in identical intonational structures, and beginning with the same four phonemes /mis/. Spectrotemporal differences in this first syllable, clearly visible in the spectrograms, result in different rhythms: a heavier beat when mis is a productive morpheme, as in mistimes. These acoustic patterns occur for many such pairs (e.g. Baker et al. 2007).

This example is valuable for four reasons. It illustrates FPD operating when traditionally crucial variables (phonemes, word boundaries, foot structure, intonation, grammatical class) are constant i.e. unusually for FPD examples, it is controlled according to standard criteria. Second, unlike FPD in fast or casual speech, it falls well within the domain of current perceptual modelling: it comes from normal-to-slow clear speech with no “missing phonemes”. Third, it cannot be dismissed as unnatural. Although it was not spontaneous, neither was it read: the speech, though scripted, was natural in that the speaker imagined a detailed context concerning resolution of a dispute about scoring accuracy at an athletics meeting, culminating in I’d be surprised if \( \sqrt{Tess} \) ____ it. This elicitation method
Figure 1. Spectrograms of mistimes & mistakes from I’d be surprised if Tess ____ it. The first four phonemes (/mist/) are the same. Their acoustic differences, summarised in Table 1, produce a different rhythm that may signal that mis in mistimes is a productive morpheme, whereas mis in mistakes is not.

Table 1. Perceptual information thought to be available in the syllables mist shown in Figure 1, based on the literature. Bold font indicates potential perceptual units, including nodes in prosodic structure. The waveforms at the top roughly indicate the portion of the utterance the table refers to. Numbers below the waveforms indicate short-time events and correspond with those in the leftmost column of the table. Patterned bars indicate the durations of adjacent acoustic segments in the waveform. The ratios of these durations, shown next to the labels in the corresponding bars in the table, provide non-phonemic linguistic information. Question marks (?) indicate low certainty/likelihood. Left: mistimes. Right: mistakes.

<table>
<thead>
<tr>
<th>Acoustic property</th>
<th><em>tes mistaimzit</em></th>
<th><em>tes misteiksit</em></th>
<th>Cue</th>
<th>Perceptual correlate</th>
<th>Cue</th>
<th>Perceptual correlate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Events: see waveforms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. periodic, nasal</td>
<td>damped</td>
<td>new syllable (simple onset); new morpheme; word; poor segment identity</td>
<td>damped</td>
<td>same as mistimes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. nasal-oral boundary + formant definition</td>
<td>Abrupt Clear</td>
<td>features for [m]; phoneme /m/? high front vowel?</td>
<td>Unclear</td>
<td>Unclear</td>
<td>features for nasal? labial?? high vowel? front vowel??</td>
<td></td>
</tr>
<tr>
<td>3. frication start</td>
<td>rel. Late</td>
<td>syllable /s/; voiceless coda; coda ends; new syllable? features for [s]; phoneme /s/? syllable is weak, light??</td>
<td>rel. Early</td>
<td>same as mistimes except syllable is weak?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. fricative-silence boundary</td>
<td>rel. Early</td>
<td>phoneme /s/; voiceless coda; coda ends; new syllable? features for [l]? morpheme ends?? productive morpheme/same word??</td>
<td>rel. Late</td>
<td>phoneme /s/; features for mis, maybe dis; features for [l]?? syllable coda continues?? morpheme continues (is nonproductive)??</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relationships Relative durations:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sonorant : sibilant</td>
<td>1:1</td>
<td>weak light syllable? (light as used by Ogden, 2000)</td>
<td>1:2</td>
<td>weak heavy syllable (heavy as used by Ogden, 2000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative durations:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sonorant : sibilant plus sibilant : silence</td>
<td>1:1</td>
<td>weak, light syllable; productive morpheme mis? (dis??); silence + intonation heralds new syllable onset, new foot?</td>
<td>1:2</td>
<td>weak, heavy syllable 1, strong syllable onset 2 of same word (monomorphemic monosyllabic)?? defocussed verb missed??</td>
<td></td>
<td></td>
</tr>
<tr>
<td>transient + aspiration</td>
<td>Long</td>
<td>confirms: productive morpheme mix (dis??); new strong syllable onset [t’]; features for [t’]; phoneme /t/; new foot; new morpheme, same polymorphemic word</td>
<td>Short</td>
<td>new strong syllable onset [st]; new foot; Confirms monomorphemic word beginning mis(t), vis, bis (dis??); features for [t]; phoneme /t/.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
encouraged focus on meaning, avoiding pitfalls of read speech. Fourth, the data illustrate the complex interactions in linguistic information that listeners can exploit: the FPD signals morphological differences, via phonological differences realised segmentally and prosodically.

Recent work tentatively confirms that FPD in such syllables facilitates their intelligibility (Baker in prep). This systematic variation has implications for the nature of the input to models of word recognition which incorporate lexical competition (e.g. McClelland and Elman 1986; Norris 1994; Gaskell and Marslen-Wilson 1997; Norris et al. 2000): with the possible exception of the aspiration associated with the /t/, the featural and phonemic information are identical in each pair; the FPD is not.

Table 1 shows information available from the utterances of /mist/ in Figure 1. Restricting the focus to linguistics, some of the main points are: (1) acoustic cues to linguistic distinctions vary in duration and in how many acoustic segments they involve; (2) all cues signal more than one thing about linguistic structure; (3) for any given chunk of sound, degrees of certainty about category identity vary with the amount of the signal heard and its context, as well as the sound’s physical properties; (4) lack of clear evidence for a particular category, as with the reduced first syllable in mistakes, can be informative; (5) sound is not mapped to linguistic structure in strict serial order, from one level to another.

Partial prosodic trees distinguishing the bimorphemic and monomorphemic forms in the structure of the last two feet (three words) of the utterances from which these words were excised (nuclear stress on Tess) are at http://www.ling.cam.ac.uk/people/sarah/docs/CNS06trees.pdf. The principles and formalisms used are those of FPA, described in Ogden et al. (2000). The prosodic trees are just part of Polysp’s representation. Links to some other aspects of knowledge are shown. Note especially the link to a parallel syntactic tree. Other links, e.g. to the limbic, motoric and wider memory systems, are important but are not shown.

Such interrelated structures seem complicated, but so does speech, especially a full description that includes links to associative semantic networks and to more general systems of memory, attitude, and emotion, all of which affect understanding. For Polysp, five points are central. (1) Each prosodic tree is linked to a corresponding grammatical tree and to other information e.g. about words and word boundaries. (2) There will also be links to other structures, such as those indicating the place and function of an utterance within a discourse, and to non-linguistic knowledge systems. For example, certain prosodic properties, voice quality, and words are associated with specific affect. Meaning representation is probably embodied. (3) Prosodic trees are the linguistic core of the representational framework, because they represent rhythm, which, in Polysp, binds other knowledge together (e.g. Large and Jones 1999; Grossberg 2003 especially §3). (4) The principle of linking nodes in prosodic structure to nodes in other structures allows great power: potentially, any currently relevant contextual influence can introduce phonetic variation. For example, there is no place in these linguistic structures for phonemes, because phonemes are by definition context free. However, links can be made to phonemes (see Concluding Remarks). (5) Phonetic information is distributed at all levels of the structure. For example, the feature [voice] applies to the syllabic rhyme, not just its coda, with consequences for the entire syllable, since properties of a syllable’s rhyme normally dominate those of its onset (Ogden et al. 2000).

This final point merits elaboration. A listener is seen as placing feature values on the nodes of a general tree or linked set of trees. Trees are metaphors for linguistic knowledge. The main difference between this and a standard perception model is that the FPD maps onto every type of node, at any place in the tree, not just at its lowest level. There is no predetermined or rigid sequence: the process of perception is governed by the properties of the particular signal in conjunction with the listener’s construal of the particular situation. Mapping the sounds of an utterance to linguistic structure is therefore massively parallel, and the order of mapping differs with listener, speaker, task and ambient conditions. Decisions depend heavily on previous decisions as long as the incoming signal is congruent with them, and a listener may sometimes understand an utterance’s meaning without having first identified all its linguistic structure (cf. Hawkins 2003). This view deemphasizes the distinction between knowledge and sensation during speech perception.

Thus in the polysystemic approach, speech is structured as multiple, hierarchically organized units, which, by their nature, cannot be represented independently of their broad functional and
linguistic context. Low-level phonological properties are no more important in determining the degree of similarity between units than any other level of structure: \textit{t} in \textit{tap} has more in common with \textit{p} in \textit{pat} than it does with \textit{t} in \textit{pat} or \textit{terrine}, for \textit{t(ap)} shares with \textit{p(at)} all the other properties of syllable constituency, stress, rhyme features etc., and only differs from it by place of articulation.

In a laboratory setting, recognition of isolated words might not require identification of much higher-order structure. But neglecting it seems unlikely. An isolated monosyllable is not just a word but also a foot and an intonational phrase (IP). Recognition that an IP is complete depends partly on recognizing lawful syllable structure and contributes to knowing that you have heard the end of the word, which would encourage a search for words of just this type. Furthermore, since the same FPD can contribute to word recognition and to higher-order units, it makes sense to model them together.

In sum, the Polysp approach offers a solution to the polarization of the exemplar/abstractionist debate. The proposed polysystemic “structures” are highly abstract. Each utterance type has a unique (possibly partially specified) structure which determines the possible types and range of phonetic variation; an experienced listener will build a huge number of them. These structures are essential to accommodate the detail, because phonetic variation is uninformative when it cannot be related to a structure. Attention to detail in the sensory signal is crucial for the structures to be used effectively, but this detail may inform about any level in any of the structures. Thus, the FPD maps on to many different units, potentially at any level in the structure. For example, in (a) use the \textit{\textbackslash tap} and (b) use the \textit{\textbackslash cold tap}, specific phonetic properties of the \textit{t}, in its context, indicate new syllable, new onset, same accent group and intonational phrase, and also contribute to indicating that in (a) the syllable is stressed (thus new foot), while in (b) it is unstressed (thus same foot). Acknowledgement of implicit learning in speech (as in every other behaviour) justifies modelling listeners as monitoring and hence in some sense processing details all the time, even when not necessary for the immediate task at hand—so that the individual listener stays optimally adaptive for the future.

4. Context-dependent phonetic categories and the bottom-up vs top-down debate

Mental representations of phonetic categories in Polysp are necessarily dynamic, relational, and plastic. These properties of categories are accepted as fundamental in other areas of perception and cognition, but often seem to be lost sight of in speech perception, despite a large confirming literature. Classical categorical perception experiments show shifts in phoneme boundaries in response to diverse influences: removing stimuli from one end of the series, presenting stimuli from one end of the series more often than those from the other end, changing apparent speech rate, making stimuli such that there is a real word at one end of the series and a nonsense word at the other (e.g. \textit{dash-tash}), and embedding stimuli in biasing sentences. These influences are normally controlled rather than the focus of study. Making them the focus emphasizes context-dependency and plasticity of phonetic/phonemic category formation, and emphasizes the dynamics of perception (cf. Grossberg 2003; Tuller 2004). There is also good evidence that “better” phones exert stronger perceptual effects of many types.

Two implications of these data are: (1) a phonetic category cannot be defined independently of its context, with context defined in the widest sense; (2) the process of perceiving phonetic categories is likely to involve assigning probabilities to a range of possibilities. These conclusions point to the value of Bayesian models (probability of \textit{x} given the current context). Bayesian principles are compatible with current understanding of factors influencing neuronal firing (cf. Carpenter’s LATER model, \texttt{http://www.cudos.ac.uk/later.html}).

The fact that FPD can influence perception, and the re-emphasis on phonetic category identity as labile and context-dependent, casts a new light on the old debate about the relative importance of top-down vs bottom-up information. Instead of top-down information compensating for signal inadequacies, the view espoused here is that the signal is not interpretable in isolation from knowledge, and that the signal itself, or the context, can indicate what knowledge should be invoked, and when (cf. Roy 2005; McClelland et al. 2006; Sprague et al. in press). This view encourages the hypothesis that the neural representation of speech must include FPD, and hence that speech is partially represented as exemplars, interpreted with respect to abstract knowledge. Evidence in support of this view is discussed in the next two sections.
5. **Evidence from neuroanatomy and neurophysiology**

A model of speech perception gains credibility if its proposed functions are compatible with neuroanatomical facts. The anatomy and physiology of the mammalian auditory system is more compatible with the proposed approach than with a strictly serial feedforward cognitive model.

The cognitive processes required for speech comprehension probably rely on multiple cortical networks that operate in parallel, as instantiated by Polysp and several other models proposed over some decades, starting perhaps with aspects of Halle and Stevens’ (1962) analysis-by-synthesis model. In humans, this functional organization may map onto anatomically segregated, hierarchically-organized processing streams similar to those identified in macaque monkeys (Kaas et al. 1999; Davis and Johnsrude 2003) and illustrated in Figure 2. Consistent with these anatomical observations, neuroimaging studies suggest multiple, parallel, cascaded auditory streams of speech processing (Davis and Johnsrude 2003; Scott 2003; Scott and Johnsrude 2003; Scott and Wise 2003; Buchsbaum et al. 2005; Rodd et al. 2005). Pathways may be specialized to serve different processes or operate on different representations of speech e.g. articulatory, phonological and crossmodal (Scott and Johnsrude 2003), although the extent to which such streams are segregated or combined is still poorly understood.

![Figure 2](image.png)

**Figure 2.** Schematic views of the macaque brain seen from the left. Left: The area around the lateral fissure is partially dissected to expose the main auditory areas, which normally lie hidden inside it. Shaded areas indicate anatomically distinct regions. Core corresponds to human primary auditory cortex (PAC). Belt and parabelt areas probably correspond to regions of the human superior temporal gyrus (STG), perhaps including the planum temporale. (From Kaas et al. 1999). Right: Auditory–frontal projections. Red circles/arrows: caudal (more posterior) belt and parabelt regions and their reciprocal frontal connections. Pale blue squares/arrows: rostral (more anterior) belt and parabelt regions and their projections. Dark blues triangle/arrows: anterior temporal-lobe regions and connections. Larger squares and circles represent the contributions from the parabelt; these are greater than those of the belt. (From Scott and Wise 2003). Thus macaque cortical anatomy suggests four or five discrete, hierarchically-organized stages of auditory processing between the core and frontal cortex, with connections going in both directions (feedforward and feedback).

Moreover, the anatomical organization of the brain is not consistent with serial, feedforward models of speech perception. Information flow in the auditory system is not unidirectional. Cortical feedforward connections each have their feedback complement and there are, if anything, more feedback than feedforward connections (Pickles 1988; Pandya 1995; de la Motte et al. 2006). Feedback probably extends into the subcortical auditory pathway, possibly as far as the inner ear’s hair cells (Edeline 2003). Although almost nothing is known about this so-called corticofugal system in humans (though see Khalifa et al. 2001), and the evidence from non-human animals is mainly based on bats (e.g. Suga and Ma 2003) and controversial, it seems generally accepted that corticofugal projections down into the subcortical auditory pathway are modifiable by behaviourally significant events (Ian Winter, Jonathan Fritz, pers. comm.).

In the cerebral cortex, anatomy suggests converging influences from multiple higher stages of perception, which are removed from the stage in question by zero, one or more intervening stages (Kaas et al. 1999). This is congruent with Polysps’ claims that signal properties map to many levels of linguistic (knowledge/meaning) structure. Even core auditory cortex integrates information over multiple time domains (Nelken et al. 2003; Ulanovsky et al. 2004); parallel processing of multiple time domains is a basic tenet of Polysp.
Recent evidence converges to strongly suggest that feedback from higher-order areas affects earlier processing, with even primary auditory cortex (PAC) developing memory traces about the behavioural relevance of specific sounds (cf. Weinberger 2004). Experiments on single neurons in ferrets’ PAC show rapid plasticity in their response to auditory stimuli to which a response is required. Single neurons change their response to sounds—and, crucially, to similar but not identical sounds—that the ferrets have learned to associate with meaning (Fritz et al. 2005). Animals which learn the task less well show weaker changes in neuronal response. Although this work is a long way from human speech perception, it may promise a basis for context-sensitivity, for the association of meaning with the classification of particular sounds, and for learning new sound categories.

It is good that neuroanatomical connections are compatible with the proposed type of processing, but equally important is so-called functional connectivity: evidence for brain regions operating in concert whether or not they share direct anatomical connections. The literature that addresses this topic is too vast to discuss here, but two points are useful. First, processing complex stimuli such as language typically involves many regions distributed throughout cortical, subcortical, and cerebellar regions. Second, Hebbian learning, a type of neuronal associative learning, may exemplify the development and structuring of abstract speech units. Hebbian learning involves functional linking of sets of neurons that together fire in a distinctive pattern to particular stimuli. Neurons that frequently fire together eventually become more closely associated, by developing physical and/or functional connections, so that they are more likely to fire in a coordinated way, given an appropriate stimulus and context: “cells that fire together, wire together”. Individual neurons may be members of many different such sets, called cell assemblies. Hebbian synapses in a cell assembly embody the notion of conditional probability. Several Hebbian postulates (reverberation, feedback), need more investigation, but the concept is generally valuable.

In sum, this brief review of brain anatomy and function is consistent with models of speech processing that are massively interactive, mixing knowledge and sensation such that it may be impossible to distinguish between them, yet can also segregate streams of information; and in which the contextualised meaning of a signal can fundamentally alter how that signal is categorised. These observations are broadly compatible with the Hebbian learning proposed as a basis for speech perception (Hawkins and Smith 2001; Hawkins 2003); see also Coleman (2002), Johnson (1997b; 1997a; 2006), and Pulvermüller (2002).

6. Neuroimaging of speech perception: evidence for interacting levels of knowledge?

6.1 Background

The burgeoning field of neuroscientific studies of speech and language processing may have raised more controversy than it has answered questions, as evidenced by the tentative and general nature of conclusions drawn from meta-analyses (e.g. Vigneau et al. 2006). Much of the reason for the confused picture is because the brain’s response to speech stimuli varies enormously with the type of stimulus, the experimental task, and the control/comparison tasks. These matters are themselves the focus of much debate. Methodological decisions are inevitably subject to theoretical preconceptions, and as there are as yet no neuroscientific studies that test how FPD is processed, this brief review only identifies experiments whose results are compatible with Polysp. Furthermore, measurement techniques vary in how accurately they reflect attributes of the response: there is usually a trade-off between spatial and temporal accuracy, which further complicates comparisons between experiments. Given this paper’s target readership and space limitations, this section mainly discusses work on functional Magnetic Resonance Imaging (fMRI), which measures blood oxygenation levels; fMRI gives relatively precise information about the location of a response at the expense of information about its temporal course.

To isolate the behaviour of interest, fMRI studies typically involve statistical comparison of two brain responses, usually to two different tasks. In subtractive designs, which are still the most common, a baseline task is necessary to avoid measurement artifacts. Such artifacts include patterns that reflect listeners “thinking their own thoughts” when not given much to do, showing a generalised orientating response when they hear a sound after not hearing anything, and simply responding to changes in general rather than to the experiment’s independent variables. Such behaviour can swamp...
the subtle effects of interest. Likewise, any response required of the listener must be required in both experimental and baseline tasks, and this too imposes significant challenges for the design of linguistically-sensitive experiments. Thus the choice of baseline stimuli and task is critical, and much debated (Hickok and Poeppel 2004; Scott and Wise 2004). Interpreting data is also challenging: these two reviews cover much common ground and largely agree on the methodological issues, but differ in their interpretation of the data.

Thus, there is disagreement in the literature about what type of speech processing occurs, and where in the brain different types occur, but there is general consensus that areas in the temporal lobes close to the PAC are involved in basic auditory perceptual processing, whereas areas farther away are more involved in higher-order processes of speech perception and understanding. There is also agreement that certain experimental methods engender responses that may not arise during speech processing under normal circumstances. For example, explicit decisions about phonemic identity typically activate regions of the left inferior frontal gyrus (LIFG), but are thought to reflect “executive” decision difficulty rather than sensory processes per se (Zatorre et al. 1996; Burton et al. 2000; Binder et al. 2004; Hickok and Poeppel 2004; Blumstein et al. 2005).

Several of the above papers note that LIFG processes are not normally involved in auditory comprehension (e.g. Hickok and Poeppel 2004). While this claim may apply to laboratory listening experiments, it may not apply so well to normal communicative situations. When speech is unclear, whether because of the way the speaker is talking or because of the listening conditions, one might expect executive phonetic decisions to be made in tandem with analyzing for meaning, or even following hypotheses about meaning (cf. Davis and Johnsrude 2003).

There is also ample evidence that non-auditory information is recruited and probably combined with auditory information at relatively early stages of perception. For example, motor cortex can be implicated in perceptual tasks, and there is evidence (Petitto et al. 2000) that the planum temporale, just posterior to Heschl’s gyrus (which includes PAC) processes sign language in ways analogous to the complex auditory processing that is commonly ascribed to it (Griffiths and Warren, 2002; 2004). Indeed the entire neocortex and some subcortical structures may be predominantly multimodal (Ghazanafar and Schroeder 2006) and thus, presumably, significantly abstract in conventional modelling terms. Nevertheless, the details of the spoken signal need to be accounted for, and in a way that allows them to inform abstract and non-phonological linguistic processing.

6.2 Evidence for distributed, interactive processing and context-sensitive categories

Assuming it is meaningful to distinguish perceptual processes themselves from processes like metalinguistic tasks that reflect executive decisions about them, then a strong form of Polysp requires evidence for knowledge affecting perceptual processes themselves. There is ample support for this concept in the literature on vision, but, outside Direct Realism, it has had less influence in speech and language (though see McClelland et al. 2006), probably because of the widespread assumption that the input units are well-defined from the beginning and hence that the main processing activities involve sorting them into formal units like words. There are as yet no imaging data that address more-recently established aspects of FPD, but some support is provided by experiments that address influences on (presumed) phoneme identification. Two fMRI experiments are described. One concludes that lexical knowledge affects early responses to phonetically ambiguous stimuli; the other that language experience probably affects relatively early stages of processing.

6.2.1 A lexical influence on phonetic processing: an fMRI study of the Ganong effect

Myers and Blumstein (2007), henceforth MB, conducted a Ganong design with two seven-step series of stimuli varying in VOT between /g/ and /k/: gift–kift, giss–kiss, and a third, non-word comparison, gish–kish. Stimuli were made from natural speech by cross-splicing. Identification and fMRI data were collected. Listeners showed expected shifts in the 50% identification boundary, with more /g/s) heard in gift–kift than in giss–kiss. The lexical boundary shift meant that responses to an intermediate stimulus in the VOT series differed depending on whether it formed a word or a non-word. E.g. /k/ responses to stimulus 5 were 90% in giss–kiss but only 50% in gift–kift.

fMRI BOLD activation was compared between these two conditions (stimulus 5 when lexically-influenced (gift) and ambiguous (g/kiss)). When the stimulus was heard as a word in the lexically-shifted gift-kift series, areas of the brain were strongly activated that are associated with
early auditory perception: largely in left posterior Superior Temporal Gyrus (STG) including Heschl’s gyrus (which includes PAC) and superior temporal sulcus; and in a smaller region of right posterior STG. Early speech processing probably involves core, belt and parabelt regions of STG around Heschl’s gyrus (e.g. Binder et al. 2000; Scott and Wise 2003; Hickok and Poeppel 2004; Scott and Wise 2004). Two other crucial comparisons did not show this pattern of activation. Hence MB argue that early perceptual responses to Stimulus 5 were directly mediated by its lexical status.

MB’s data are consistent with Polysp’s predictions and the literature. They support interactive models in which higher-level knowledge influences perceptual processes as soon as it is available (e.g. TRACE, McClelland and Elman 1986), and challenge models in which lexical effects are due to “post-lexical” decision-stage processes (e.g. MERGE, Norris et al. 2000). They suggest a parallel in human phonetic perception with online sentence comprehension and with the animal evidence mentioned above of neuronal plasticity in PAC in response to behaviourally-relevant sounds.

6.2.2 Influence of phonological knowledge/experience on segment/syllable perception

Jacquemot et al. (2003) exploited cross-linguistic differences to show that the syllable structure of one’s native language seems to affect perception of stimuli early in cortical processing. The stimuli, nonsense in Japanese and French, were two- and three-syllable sequences like *ebza*, *ebuza*, *ebuua* used by Dupoux et al. (1999 Expt. 3): French listeners heard two vowels/syllables in *ebza* and three in *ebuza*, whereas Japanese listeners, unused to consonant clusters, heard three in both, and found *ebza* difficult to distinguish from *ebuza*. Conversely, while both groups heard three syllables in *ebuza* and *ebuua*, Japanese listeners, for whom vowel length is distinctive, heard a phonological difference; French listeners did not. For both groups, performance was better (fewer errors, faster reaction times) for contrasts that involved a phonological difference.

Table 2. Experimental design used by Jacquemot et al. (2003). Stimuli were presented in triples; listeners were told the first two were identical, and judged whether the third was different.

<table>
<thead>
<tr>
<th>Trials (examples)</th>
<th>French</th>
<th>Japanese</th>
</tr>
</thead>
<tbody>
<tr>
<td>hear <em>ebza</em></td>
<td>Phonological</td>
<td>Acoustic</td>
</tr>
<tr>
<td>hear <em>ebuza</em></td>
<td>Acoustic</td>
<td>Phonological</td>
</tr>
</tbody>
</table>

In an fMRI experiment with Japanese and French listeners, Jacquemot et al. (2003) used Dupoux’s stimuli and design (Table 2), but a change/no change task. Each listener/language group heard both contrasts (*ebza*/*ebuza* and *ebuza*/*ebuua*). Each contrast was phonological in one language, but acoustic (phonetic) in the other.

Change vs no-change conditions were compared, pooled across language groups. Acoustic conditions mainly produced activation differences in frontal regions, and none at all in the left temporal lobe. Phonological conditions produced differences throughout the perisylvian cortex, especially in the left hemisphere. Only two small regions significantly distinguished phonological and acoustic conditions: an area of the left STG on the boundary between Heschl’s gyrus and the planum temporale—at the edge of PAC—and part of the left supramarginal gyrus in the parietal lobe. These regions are probably part of the ill-defined Wernicke’s area (Scott and Wise 2004) and may subsume a large number of functions. These data not only confirm what has been shown in other ways—that basic auditory discrimination of speech involves assimilation into the allowable phonological patterns of the listeners’ native language—but they suggest that some of this influence probably occurs very early in processing, at a stage which has been claimed to significantly precede word identification and is more restricted to acoustic features, general patterns, and perhaps “auditory objects” (cf. Griffiths and Warren 2004).

6.2.3 Summary: top-down influences on early auditory-phonetic processing

MB and Jacquemot et al. (2003) confirm that acoustic-phonetic categorization depends on the context (be it stimulus, language, etc) in which a sound is heard, and support the claim for interaction between knowledge and sensation. Such interaction occurs in regions of the cerebral cortex traditionally associated with acoustic-phonetic rather than higher-level linguistic processing, and probably happens early in processing, although this cannot be determined from fMRI data. However,
ERP and MEG data, which is sensitive to time, supports this inference (Sussman et al. 2002). Neither experiment shows whether the basis for the responses was featural, segmental, phonemic, or due to some more holistic process of template matching, perhaps at a syllabic or phrasal level. But both are compatible with holistic pattern matching, and there appears to be little evidence in the literature that demands interpretation in terms of phonemes, independent of intelligibility (Scott and Wise 2003; 2004). In short, both studies are compatible with the claim that abstract knowledge influences the processing of details in the signal, but that the details influence what abstract knowledge is accessed.

7. Concluding remarks

It has been argued that, while systematic phonetic variation in the spoken signal—FPD—can strongly influence perception, it cannot do so unless the listener is able to relate the variation in sound patterns to different functions/meanings. In context, some utterances may be understood from the pattern of FPD with little or no reference to the formal linguistic structures that they themselves can be related to (Ernestus et al. 2002; Hawkins 2003). Sensitivity to contextualised FPD is essential for this type of understanding. However, presumably most utterances are at least partially structured/analyzed into linguistically-relevant components before they can be understood. The value of the type of FPD that indicates formal linguistic structure is that it should make this process faster and less error-prone.

It is therefore proposed that linguistic information in the signal does not just map to “low-level” abstract units like phonemes, features, or words, but simultaneously to many units, as outlined in Table 1. Some units will be phonological, but many will not: non-phonological units known to be indicated by FPD include bound morphemes, function words, (probably many subsets of) content words, auxiliary verbs, pronouns, and so on. Others map to discourse and interactional function. Each may have its own subsystem of sound patterns, and therefore a distinct abstract structure, presumably subject to other influences such as predictability. Thus, during speech processing, no formal unit of linguistic structure is prior or more fundamental than any other: meaning (including function) is prior, and all potential units that allow meaning to be quickly understood are valuable.

The order in which linguistic units are accessed/identified during perception will depend on some combination of the following: what the spoken utterance is like (i.e. signal properties), its linguistic and conceptual complexity, the listener’s understanding of the situation, facility with the language, familiarity with the speaker, and engagement with and attention to the speaker and the speech; also, probably, chance. Experienced listeners will direct attention to higher or lower regions of the hierarchy of detail and linguistic form depending on task demands (its purpose and difficulty). However, the detailed signal will presumably always be monitored to some extent, for to fail to attend to it would risk not learning important new information, which would be non-adaptive. Yet, like other physical aspects of a stimulus, particular aspects of FPD may presumably be ignored if other stimulus attributes override them or the task makes them irrelevant.

Given the need to identify higher-order structure in order to understand the signal, phonemes are unlikely to be fundamental to understanding everyday speech. Scott and Wise (2004) argue the same point from neurological evidence. Phonemic categories, by definition, are context free. Thus phoneme identification is secondary: if you know the whole structure, you automatically know the phonemes, but not vice versa. Phonemes are compatible with Polysp to the extent that categories can be learned as “self-organizing” and “emergent”, related by virtue of shared attributes or occurrence: the brain naturally works with contrasts, and classifies like things together. The relationship between sounds and phonemic categories may thus be learned by association rather than because phonemes play a fundamental role in speech communication. It follows that phonemic representations might not be comprehensive or fully systematic. Assuming that the shared attributes of phonemes can be learned via any relevant experience, then those that share consistent articulatory, auditory or orthographic properties (for people literate in an alphabetic writing system), seem particularly likely to form categories. In English, candidates for systematic phonemic categories might be /s/, which varies little acoustically, and bilabial (but not alveolar) stops, which vary little in critical aspects of articulation. Other sounds, for example vowel qualities conditioned by certain consonants, might never be grouped appropriately into phonemes. Orthographic consistency presumably influences the system’s development. If the language allows it, then, a rough phoneme inventory could develop, with some
phonemes being better categories (having more psychological reality) than others, and individuals, and speakers of different languages, differing in how systematic their phonemic inventories are. The de-emphasis on phonemes relative to other aspects of linguistic structure does not preclude responding in terms of phonemes in appropriate circumstances, such as lab experiments requiring simple identification responses to clearly spoken isolated words. However, statistical significance does not prove category status: there seems no strong proof that phonemes are used rather than say, allophones, syllabic shape, or associations of sound with letters of the alphabet, and there are limits on how interchangeable allophones of the same phoneme can be in perceptual experiments.

Although barely touched on here, rhythm is proposed as one—possibly the main—overall binding principle. Phonological rhythm may be the principal binding process, but there may also be more general influences on rhythm and other alignment processes (cf., for conversation, Garrod and Pickering 2004; Wilson and Wilson 2005). Functionally, analysis must take place in multiple time domains, ranging from a few milliseconds to several seconds or minutes. Likewise, although the examples of the proposed structures given in this paper are hierarchical, presumably associative processes also influence the structuring of linguistic knowledge and behaviour.

A processing system that allows this type of behaviour is likely to use current information and past experience to generate hypotheses about what is most likely given the evidence, to have ways of testing and revising these hypotheses, and ways to learn from new experiences while not forgetting old ones (though possibly revising their status), so that it is appropriately adaptive. These proposed perceptual processes are dynamic and broadly Bayesian. They involve massive interaction and feedback, and include the ability to separate sources of information (e.g. message content, speaker identity, speaker affect) when necessary. These factors seem compatible with current understanding of anatomical and functional connectivity in the cerebral cortex, and with the plasticity of memory processes. Research on the molecular correlates of memory for taste in rats, for example, shows that the process of retrieving a memory changes it, so that when it is stored again, its molecular signature is not quite the same as it was before (Dudai 2006). Here is evidence, albeit indirect, that new sensation is routinely related to existing stored memories, and that it subtly but irrevocably changes those memories. This degree of plasticity, coupled with processes of perceptual and associative learning, plausibly support the argument that any given instance of sound is processed with reference to a rich variety of more abstract and established concepts, both linguistic and non-linguistic.

Computational modeling of this nature has not yet been achieved for normal speech, but some prerequisites are being used in systems of vision and action (e.g. Sprague et al. in press) and semantic processing (e.g. O’Keefe 2003; Roy 2005), and a start has been made with infants’ lexical learning (e.g. Yu et al. 2003). Common factors in these models include embodiment of concepts as distributed systems, and such strong knowledge-driven perception that sensation, though powerful, essentially interrupts top-down assumptions. Consistent with Polysp, what constitutes a relevant bottom-up cue is thus strongly task-dependent, although unambiguous bottom-up information should always dominate the response (Sprague et al. in press). If these models influence speech perception models, the abstract will be very abstract indeed, relative to current models.

The FPA ‘trees’ proposed represent some of the knowledge. The issue of generalisability between trees is interesting. If many separate trees are needed to capture the information that is available, then we have the same problem for perception that we have for synthesis in an all-prosodic (nonsegmental) system: if you haven’t specified values for the structure, then the synthesizer cannot produce it. So, for perception, if you fully understand a stretch of speech, then you know its structure. This is an interesting issue concerning sophisticated knowledge use, exemplified by learning a foreign language and understanding foreign-accented speech. Unknown/unexpected or wrong stress, overall rhythm, and so on, can be inordinately disruptive to understanding.

If the role proposed for FPD in speech perception is accepted, then none of the Polysp proposals seem controversial. If FPD can influence perception, it must be accounted for, and the rest follows logically. Polysp, and FPA trees, simply offer a way to think about how we structure the detail that is perceived so that we efficiently understand meaning. However, the consequences for how we think about experiments, models of speech perception, and phonology, are far-reaching.

To understand most variability, we should study connected speech in as natural situations as possible. Ideally, we should refocus attention from the so-called easy problem of identification of isolated words or syllables, onto understanding fluent speech—minimally, well-formed sentences
with a clear, contextualised meaning that speakers and listeners must attend to. It might also be valuable to replace standard metalinguistic tasks (what phoneme? is it a real word?) with more natural ones. These changes will be hard, but 50 years of the simpler approach have not allowed us to generalise to real speech, and have deflected attention from important variables like task dependency.

The point here parallels Section 4’s discussion about research on categorical perception: if a variable needs controlling in perceptual experiments, it should feature in perceptual models. When the literature that discusses task variables and baseline comparisons in brain imaging is read in this light, it is clear that a focus on broad contextualised responses rather than “pure” phonology could be worthwhile. But it is not clear how to represent this in either perceptual or purely linguistic models. Although the computational approaches cited above may help, these problems are as much philosophical as practical. They raise old questions, such as how independent (of other linguistic subdisciplines, of meaning, of action) is phonology? Such questions may only be answerable in the pragmatic sense of defining the purpose and hence the focus of current interest. If that is the case, then the phonology that is required to model processes that intervene between receiving and understanding speech must deal with systematic, and therefore informative, variability, whether it informs about language and meaning, speaker characteristics, or anything else. But the phonology that is required to represent the most abstract knowledge about how the sounds of a language pattern—and that may guide perceptual decisions—might be very abstract indeed. There seems no reason to assume that people cannot develop both types of representation, and presumably intermediate forms too.

Acknowledgements.
Supported by the Leverhume Trust. I thank Charlotte Jacquemot, Ingrid Johnsrude, Klaus Kohler, Richard Ogden, Rachel Smith and Jan Volín for valuable discussion. None of them is responsible for the views expressed.

References


Norris, Dennis G., James M. McQueen, Anne Cutler and Sally Butterfield 1997  The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology* 34.


Rodd, Jennifer M., Matthew H. Davis and Ingrid S. Johnsrude 2005 The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*.


Scharenborg, Odette, Dennis G. Norris, Louis ten Bosch and James M. McQueen 2005 How should a speech recognizer work? *Cognitive Science* 29: 867-918.


