Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Phonetics 31 (2003) 373–405
Journal of Phonetics
www.elsevier.com/locate/phonetics

Roles and representations of systematic fine phonetic detail in speech understanding

Sarah Hawkins*

Department of Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, UK

Received 4 August 2003; received in revised form 29 September 2003; accepted 29 September 2003

Abstract

This paper aims to show how we can make progress in elucidating how people understand speech by changing our focus of inquiry from abstraction of formal units of linguistic analysis to a detailed analysis of global aspects of the communicative situation, of which speech is just one part. It uses evidence of (a) the communicative importance of fine phonetic detail and (b) exemplar memory for speech to explore the idea that, in certain normal, easy conversations at least, one may interpret the meaning of an utterance directly from the global sound pattern; reference to formal linguistic units of analysis, such as phonemes, words, and grammar, is incidental; circumstances dictate whether such reference takes place at all, and if it takes place, whether it does so after the meaning has been understood, before it has been understood, or simultaneously with the construction of meaning. The implications of this position are that speech perception does not demand early reference to abstract linguistic units, but instead, to flexible, dynamic organization of multi-modal (and modality-specific) memories; and that models of speech perception should reflect the multi-purpose function of phonetic information, and the polysystemic nature of speech within language. A preliminary model that reflects this theoretical position, Polsyp, is described. Polysp has intellectual antecedents in Hebbian principles, and current relevance to adaptive resonance theory (ART). Neuronal bases for initial processing of exemplars are briefly discussed. Hierarchical and more abstract processing arises in an ART-like, self-organizing dynamic system in which, once processing has begun, the sensory input is not effectively distinguishable from top-down knowledge. Understanding meaning is more important than identifying linguistic structure, and processing is strongly guided by rhythmic and attentional factors.

*Tel.: +44-1223-335052; fax: +44-1223-335053.
E-mail address: sh110@cam.ac.uk (S. Hawkins).

0095-4470/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2003.09.006

1. Introduction

Local (2003), in this volume, has explained how systematic variation in fine phonetic detail reflects a wealth of information about the structure of utterances that is normally neglected in modelling speech perception, and a range of other factors important in keeping conversations going. Naturally, some of these factors can be analyzed into the well-accepted units of formal linguistic analysis: intonational phrases, feet, syllables, phonemes, and allophones associated with particular positions in syllables of particular structure. But some of the important details are not readily accommodated by standard phonological-linguistic units; yet when they are systematically reflected in the speech signal, they, too can be crucially important to communication. Local shows how Firthian prosodic analysis (FPA) has the potential to systematize fine (and not-so-fine) phonetic detail into richly specified linguistic structures that represent the salient contrasts of speech used interactively.

Likewise, Goldinger and Azuma (2003), in this volume, have put the case for episodic (or exemplar-based) memories being at the root of speech perception, and for linguistic categories to be properties that emerge from processes modelled by adaptive resonance theory (ART), described in detail in this volume by Grossberg (2003).

Endorsing the general thrust of these points of view, my aim in this paper is to tie them together and to suggest some consequences for the way we conceptualize the processes of speech understanding. In particular, I argue that understanding the role of systematic variation in fine phonetic detail is the key to understanding how to model the transformation from specific memories to speech understanding. Accordingly, the first part of this paper addresses the central role of fine phonetic detail in speech understanding, and evidence for exemplar or episodic memory for speech. Later sections describe a framework, Polysp, that can help conceptualize the transformation from specific memory to structured linguistic information, and suggest neurological processes that might underlie some of these processes of speech understanding.

2. The central role of phonetic detail in the understanding of meaning

Many properties of the speech signal perform multiple roles, providing strictly linguistic information as well as traditionally non-linguistic or paralinguistic information about, for example, the speaker's identity, attitudes, and current state of mind, and contributing importantly to the broad connotative as well as the narrow denotative meaning of the utterance. An extension of this well-accepted premise is that the detailed phonetic signal is not a relatively arbitrary carrier of meaning that must be interpreted into some other (also arbitrary and meaning-free) formal system before it can be understood, but is rather more directly mappable onto meaning itself. Most native speakers of English have an impressively wide variety of ways of conveying the meaning of *I do not know*: Table 1 lists some of them (items 1–7). The form a speaker chooses depends on what 'extra' information he or she wishes to convey. The most common forms probably range between *I don't know* and *dunno* (items 3–5), each of which can be pronounced in a number of different ways. When spoken with a neutral intonation and 'ordinary' voice quality, tempo, and rhythm, the most neutral form, *I don't know* (item 3), typically conveys little more

Table 1
Ways to mean *I do not know* as an illustration of the centrality of phonetic fine detail to the understanding of full connotative meaning

1. *I…do…not…know*
2. *I do not know*
3. *I don't know*
4. *I dunno*
5. *dunno*
6. [a̰ə̆nːəʊ̆]
7. [ə̰ə̃ə̰]
8. but not [ə̰ə̃ə̰] or [m̩ m̩ m̩]

Versions 1–7 are all acceptable. Versions 3–5 are standard for normal relaxed interaction (if spoken without insolence). Versions 1 and 2 imply negative attitude (if spoken fluently and without absent-mindedness); version 6 is normally used between social equals, but can introduce a new opinion; version 7 is only possible when the participants are social equals and external contextual cues are very strong (intonation and stress are important, but have been omitted).

than that the speaker lacks knowledge. This neutrality is itself informative: the message is not loaded with significant broader meaning.

Most other ways of expressing lack of knowledge offer extra information, which the interlocutor must understand if the conversation is to be successful. The connotations are communicated by facial expression and body language, and also phonetically by the particular segmental realization (choice of 'word forms') and a wide range of other properties. In general, the more extreme forms can only be used in particular circumstances, and the more extreme the form, the more constrained the circumstances. The person who says *I dunno* or *Dunno* (items 4 and 5) tells us that he or she is, or has been, content not to know, or is indifferent to the listener's wish for information. *(I) dunno* can only be used in informal situations, or to convey insolence in more formal ones (or shame, if accompanied by low amplitude and lowered eyes), but nevertheless the range of acceptable situations and pronunciations is very wide indeed: *dunno* is only slightly more restricted than I don't know. In contrast, it is hard to say the fully expanded form *I do not know* (item 2) without conveying some degree of exasperation, often signalled by unusual voice quality, tempo and rhythm. The even more extreme form, *I…do…not…know* (item 1), with pauses between the words, is in most cases so rude that it can only be used when the listener does not seem willing to accept that the speaker really does not know. Items 1 and 2 can have the same types of unusual use of voice quality, intonation and tempo, but, when pauses separate the words as in item 1, 'normal' intonation, possibly with raised jaw ('gritted teeth') can be remarkably threatening too.

At the other extreme, it is possible—again, only in the right circumstances—to convey one's meaning perfectly by means of a rather stylized intonation and rhythm, with very weak segmental articulation, ranging from something like [a̰ə̆nːəʊ̆] to [ə̰ə̃ə̰] (items 6 and 7 in Table 1, intonation not transcribed). Utterances like item 6 are normally used between social equals, but they can function in other ways, for instance to signal the introduction of a new opinion, usually with some negative, grudging, or mildly contradictory connotation: "[a̰ə̆nːəʊ̆], seems a bit risky to me"; "[a̰ə̆nːəʊ̆], *I* thought it was pretty good". Maximally reduced forms like item 7, [ə̰ə̃ə̰], could allow successful communication between relaxed family members. For example, it could be said by B when A asks B where the newspaper is, and B does not know, but does not feel that she needs to

stop reading her book in order to help find it. Person A should understand from this that he should not expect help in looking for the newspaper, and should either stop talking to B, or introduce a more interesting topic. Item 8 demonstrates that the intonation pattern alone is not enough: at least the vowels must be there ([m̥ m̥ m̥] will not do), and the vowels must start more open (and probably more fronted) than they finish, just as in the more clearly spoken utterance. In consequence, at least in this situational context, [ə̰ə̰ə̰] (item 8) is nonsense whereas [ə̰ə̰ə̰] (item 7) is not.

In summary, the narrow meaning of these and other related forms is the same: the speaker lacks knowledge. But good communication demands that the wider meaning is recognized as well. Detailed phonetic structure, together with the whole range of the mutually-understood situational context, are crucial in providing this information; and one part requires the presence of the other parts that it normally occurs with for the utterance to have the intended meaning.

These examples underline the obvious point that we speak in order to be understood; and that formal linguistic analysis of speech into abstract phonological units like features, allophones, phonemes, and words only shows part of what it takes to be understood: not only is the rest of the context missing, but also, such units of formal linguistic analysis neglect information that is available in the speech signal alone that enables broad connotative meaning to be understood. That is, successful communication demands that speakers and listeners share a mutual understanding of and attend to the whole situation. This truism is often neglected in models of speech perception, and in the types of research questions asked (but cf. Lindblom, 1996; Lindblom, Brownlee, Davis, & Moon, 1992; Bradlow, 2002). Consider these statements, taken from an excellent dictionary of linguistics, and typical of the genre—indeed, broader than some definitions (Matthews, 1997).

- *Phonetics*. The study of the nature, production, and perception of the sounds of speech, in abstraction from the phonology of any specific language.
- *Phonology*. The study of the sound systems of individual languages and of the nature of such systems generally.

Meaning is nowhere in these definitions, nor is it mentioned in explanations that follow such definitions in this or other dictionaries, although it is implicit in phrases like ‘sound systems’. This absence of explicit attention to meaning has resulted in biases in how we design experiments and build models, notably by focussing attention on those aspects of clear speech that formally distinguish citation-form words from one another (cf. Local, 2003).

This paper takes the approach that meaning is ‘present’ in the details of the entire communicative situation, and that it is available to the participants in a conversation via detailed representations of different types of sensation, retained in working memory and subsequently encoded in longer-term memory. Part of this detail includes details of the spoken signal.

3. Neural mapping of sound to meaning?

How can sensations of speech sound be bound with other sensations to form composite memories linked to, or underlying, meaning? Evidence has accumulated over recent decades that cells from many different parts of the brain can contribute to a single memory, a single concept,

and the representation and/or processing of a single word. Equally, the same cells can contribute to more than one memory, concept, and word. Thus individual cells or groups of cells enter into a number of different functional groupings.

Groups of cortical cells that process spoken words are located in the cortical area which processes complex sound, and another area close to the main modality or modalities associated with the peripheral input or output involved with the memory of the particular word. For example, in naming pictured objects, words for hand tools strongly activate areas associated with hand movement, while words for objects that primarily involve vision activate the visual cortex more strongly (e.g., Martin, Wiggs, Ungerleider, & Haxby, 1996; see Coleman, 1998 for a review). The point is illustrated in Fig. 1. Panel A shows distributions of activation in the left hemisphere for visual vs action words. It illustrates that, as noted, word representations typically have two parts, a perisylvian part related to word form, and thus to phonetics or phonology, and a part distributed mainly in other parts of the brain that represents its semantic word properties, and thus reflects the individual's experience with that word/concept, particularly with its relevant sensory modality/ies. Thus, in Fig. 1A, vision words have additional representation in the visual cortex, while action words have additional representation in the motor cortex. Fig. 1B, which shows both hemispheres, illustrates that content words, e.g., concrete nouns, typically have rather rich networks of cortical connections in both hemispheres, whereas networks for grammatical function words are mainly restricted to perisylvian sites in the left hemisphere. Presumably function words have fewer associations with sensory memories. This somewhat tentative schematization is compatible with some other empirical and theoretical work, e.g., Warrington and Shallice (1984), Warrington and McCarthy (1987), Coleman (1998, 2002), Wise, Scott, Blank, Mummery, Murphy, & Warburton, (2001). It seems reasonable to conclude that language,

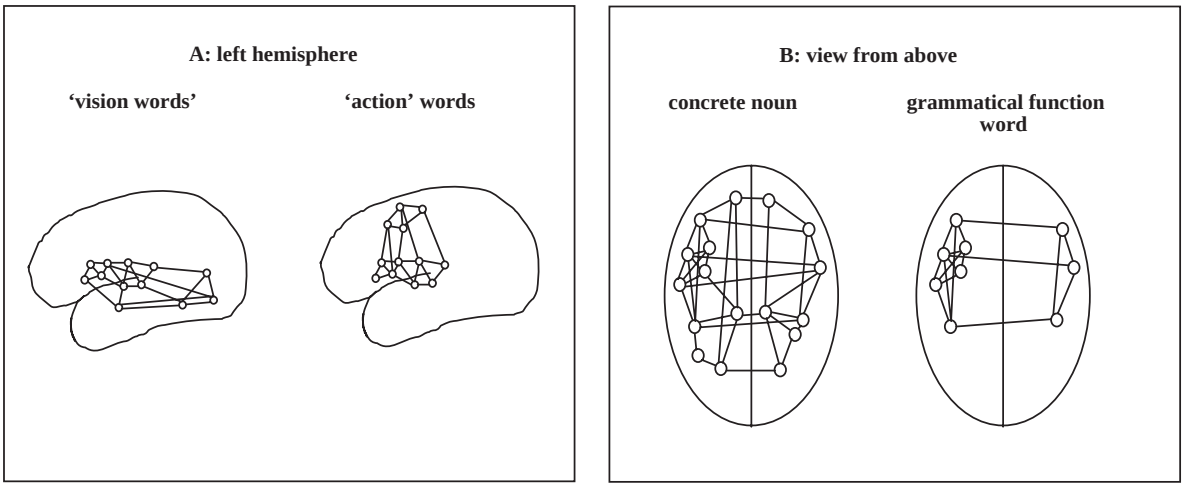

Fig. 1. Schematic diagrams of the brain showing representations of the type of functional cell groupings that may represent different types of spoken words. (A) View of the left cerebral hemisphere, showing distributions of activity during processing (such as naming pictures) of vision and action words. (B) View from above (left hemisphere on the left), showing different degrees of lateralization in the two hemispheres for content and function words. Adapted from Pulvermüller (1999).

including words, may be stored partly as modality-specific memories of actual sensations. (Pulvermüller (1999) and associated commentaries offer more discussion.)

Evidence of hierarchical processing in neural responses to speech and speech-like sounds is reviewed by Scott and Wise (2003; see also Davis & Johnsrude, 2003). The hierarchies identified correspond more to degrees of speech-likeness and degrees of meaningfulness than to the syntactically- or prosodically-structured trees of linguistic hierarchies. Evidence supporting hierarchical processing is to be welcomed, although speech processing may not always follow a hierarchical pathway. In suitably constrained contexts, a speech pattern may be understood before completing phonological and grammatical analyses. This might be the case for the more reduced forms of *I don't know* in Table 1, as discussed below. Moreover, neuropsychological experiments show that the time course over which the brain processes different aspects of linguistic structure is task-dependent and need not follow a rigid 'sound → lexicon → grammar → meaning' sequence: primary determinants of activation patterns are whether stimuli are meaningful and the task requires listeners to understand them (Démonet, Price, Wise, & Frackowiak, 1994; Coleman, 1998; Hickok & Poeppel, 2000; Scott, Blank, Rosen, & Wise, 2000; Pulvermüller, Assadollahi, & Elbert, 2001; Scott & Wise, 2003). For example, Pulvermüller et al. (2001) conclude that the earliest linguistically-related brain responses found so far, at about 100 ms after stimulus onset, reflect aspects of word meaning, whereas syntactic distinctions (mainly in word class) seem to appear later, about 120–150 ms after stimulus onset. Démonet et al. (1994) show that activation of perisylvian cortex happens later in phoneme monitoring than in lexical access, and other experiments show that syntactic and semantic processing can guide phonological decoding. The idea also motivates some computational models, e.g., Plaut and Kello (1999), although demands for computational tractability mean that the consequences of retaining phonetic detail may be lost sight of in the implementation (but see Kello & Plaut, 2003).

4. Exemplar memory: the basis of how fine phonetic detail is encoded?

It seems worthwhile hypothesizing that distributed neural representations of words may develop from (or even include) memories of actual experienced speech, and thus that the brain may store speech signals in some detail, at least for some appreciable amount of time. Mechanisms by which this could happen are called exemplar, or episodic, memory: actual instances are remembered in some detail. Some investigators, including Goldinger and Azuma (2003) refer to episodic memory, but others reserve a very specific meaning for this term (Tulving, 1972) and the more general term is probably exemplar memory (e.g., Nosofsky, 1988, 1991). Since the distinction between the two is not important for our purposes, this paper adopts the more general term, exemplar. Similarly, it is not clear whether exemplars are implicated in representation (e.g., storage of word forms) or in processing (e.g., understanding of words), but our understanding of the contribution of phonetic detail to exemplar-based memory is not yet at a stage where this distinction matters, and in any case there may be no distinction (cf. Coleman 1998, p. 300).

I suggest then that a person's speech is remembered in a form that gives access to exemplar memory for at least some aspects of it. We remember the general pitch range, the range of voice qualities and how they are usually used, the rhythms, the favored words and turns of phrase, and

all the idiosyncrasies of speech of a person we know, and these, along with their facial expressions, the way they move, and so on, form part of our percept of that person. They contribute to our knowledge of their likes and dislikes, of what makes them laugh or fires their enthusiasm, of what produces anger, and allow us to class them as like or unlike other people we come across. There seems no obvious reason why memories for words should be qualitatively different from other sorts of memories. They are developed from sensory percepts, and to the extent that they are abstract, the abstractions are developed from finding common factors amongst the many different pronunciations we have heard.

Moreover, the same attribute of the acoustic signal can contribute to many different abstractions. Within the linguistic system, systematic fine phonetic detail simultaneously contributes segmental (allophonic) and prosodic information (e.g., Fougeron & Keating, 1997; Smith & Hawkins, 2000; Keating, Cho, Fougeron, & Hsu, 2003). More broadly, raised pitch in a particular utterance can contribute to percepts of prosody, voice quality, lexical form, and information about the speaker's attitude to particular events or objects. We may attend more to some aspects of the utterance than others, depending on the task at hand: what we attend to is influenced by past experience and guides how the incoming information is organized (e.g., Burgess, Becker, King, & O'Keefe, 2001). Thus information derived from sensory input can and may be stored in more than one way, because, if memory is exemplar-based, then all sorts of information can be extracted from it, and used for particular tasks, as long as it was initially attended to and transferred into long-term memory.

Remez (2003), in this volume, doubts that speech is stored as exemplars, on the grounds that experiments show that auditory traces decay after about 400 ms (Pisoni & Tash, 1974; Howell & Darwin, 1977). There are at least two reasons to set this objection aside until the potential for exemplar representation has been better explored. First, a decay rate of around half a second seems a reasonable time during which an adult listener could take the first steps towards more long-term organization of the input, including abstraction. Average syllable rates in conversational speech, reaction times to words, and recent evidence that the brain can respond to the meaning of stimuli about 100 ms after presentation (Pulvermüller et al., 2001) seem compatible with a half-second decay time.

Second, it is possible that these early experiments did not get the best performance out of their participants. They involved simple, isolated, meaningless syllables using highly-controlled, over-simplified synthetic speech in a standard laboratory same-different task focused on phoneme identification. One of Howell and Darwin's (1977) experiments involved insertion of a tone between the syllables. Neither the sounds nor the tasks used are natural, and we might expect listeners to perform poorly. There would be little in existing memory to guide attention to salient properties of the stimuli for this particular task, or to allow these unfamiliar stimuli to be related to existing memories of speech. Real speech is more memorable than laboratory-standard synthetic speech (Duffy & Pisoni, 1992), possibly because it is more perceptually coherent, as discussed below (Section 5.2). Fig. 2 illustrates this point. When the stick figure in the top left of Fig. 2 is viewed alone, it is a person of unknown gender. Seen with the figure on the right, both, by convention, acquire gender: male on the left, female on the right. In this context, longish hair signifies the category female. But, as everyone knows and the photograph at the bottom of Fig. 2 illustrates, hair length is not always criterial of gender. The genders of the couple photographed are clear, but by virtue of properties like relative bone structure and musculature that each of us

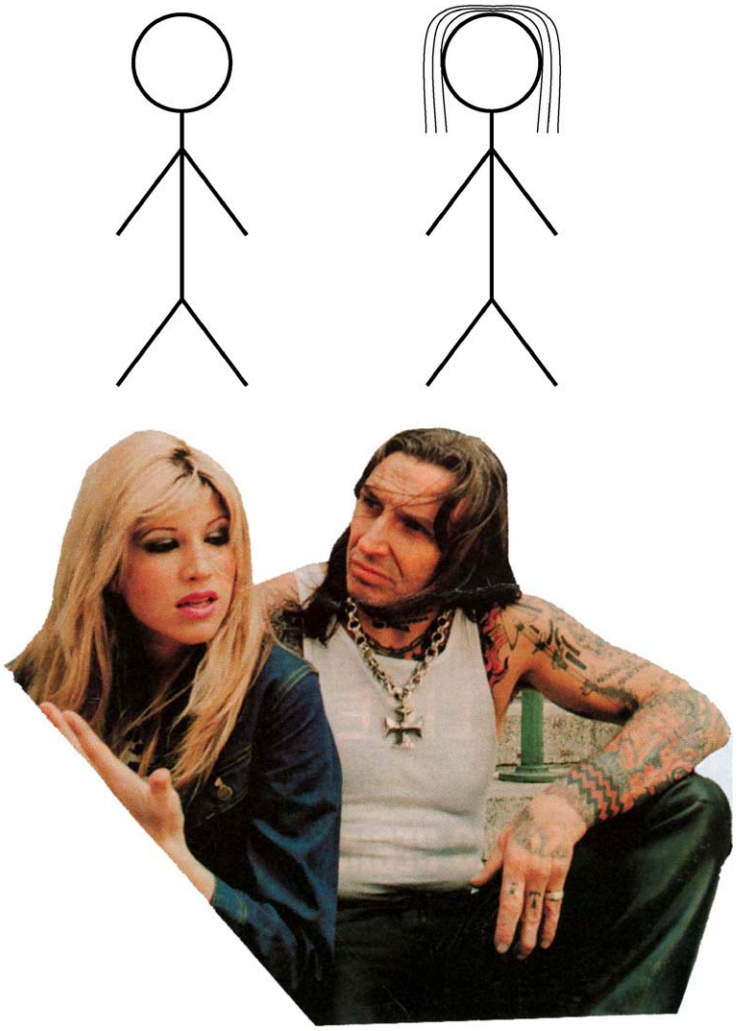

Fig. 2. Top left: A stick figure of a person. It has no distinguishing characteristics and might be interpreted as representing a man or a woman. Top right: when long hair is added to the same stick figure, the standard interpretation is that the one on the left is male and the one on the right is female. Hair length is criterial of gender for stick people. The photograph below shows that hair length is irrelevant to classifying the gender of real people, at least when they are hippies, whereas attributes such as bone and muscle structure are important. The stick figures are analogous to highly controlled synthetic stimuli, and the photograph to natural speech (from Hawkins & Smith, 2001).

has learned are more reliable than hair length. Extrapolating, the hair on the stick person is like a cue manipulated in the synthetic stimuli of standard, highly-controlled speech-perception experiments. While these experiments serve a useful purpose, we have gone too far in neglecting context: our controlled experiments in effect extol the primacy of 'no obvious context', despite the fact that most utterances are made within a context that is understood by all participants in the situation.

Although these arguments may represent more faith than science, they seem strong enough to make it reasonable to pursue, for the time being, the possibility of detailed acoustic–phonetic memory for meaningful speech, and to relate it to exemplar memory.

Goldinger (1998) has taken a strong stand that words are stored as detailed memory traces or exemplars. Lachs, McMichael, and Pisoni (2003) comprehensively review the evidence. Goldinger and Azuma (2003) have shown that word recognition, and phone- or syllable-monitoring, are affected to similar degrees by changing the speaker or the task, and that top-down 'knowledge' and bottom-up 'signal' interact in complex ways which suggest that, in terms of perceptual processes, it is an over-simplification to model the former as largely abstract and the latter as entirely sensory. However, these and related experiments (e.g., Nygaard, Sommers, & Pisoni, 1994; Nygaard & Pisoni, 1998; Sheffert, 1998) involve changes in signal attributes that are not directly part of the linguistic message. So, although the evidence for exemplar representations in speech processing looks good, these experiments do not show that exemplar memory traces are directly implicated in decoding the formal linguistic structure or meaning of the utterance; we only know that earlier experience with the speaker's voice or the task influences how well later tasks are done. This amounts to the claim that the more a given task resembles an earlier task, the better the listeners will perform in the later one. To establish that aspects of fine phonetic detail are encoded in exemplar memory, one has to manipulate particular types of fine phonetic detail: it must be detail that can differentiate between linguistic meanings rather than just between speakers, and it must be detail that is not retained in a linguistic description whose basic units are phonological feature-bundles or phonemes. Experiments by Rachel Smith (in preparation) indicate that exemplar memory traces do seem to be implicated in processing linguistic meaning.

In earlier work, Smith and Hawkins (2000) showed that segmental allophonic detail which distinguishes word, syllable, or morpheme boundaries goes a long way to explaining patterns of performance in speeded word-spotting experiments, in which listeners press a button when they hear a short word inside a longer word. Smith (2002), like Goldinger and Azuma (2003), manipulated listeners' familiarity with the speaker. Unlike them she also manipulated the allophonic appropriateness of a critical portion of a sentence, by cross-splicing with other tokens spoken by the same person. She predicted that listeners should be better at word-monitoring when the speech comes from a familiar compared with an unfamiliar speaker, as long as the allophones are those that the speaker normally uses. The crucial, connected, prediction was that when the speech was cross-spliced so that the speaker was the same, but the allophones were inappropriate for that particular word, then listeners would perform *worse* for familiar speakers than for unfamiliar speakers. In other words, Smith predicted an interaction between allophonic appropriateness and speaker familiarity: when the speaker was familiar, correct allophones would facilitate word recognition, but incorrect allophones would hinder it.

Smith used sentence pairs (underlined in the example below) with identical phonemic sequences but different (usually grammatical) structures in a critical part of them (in bold below), each preceded by suitable disambiguating contexts, to make pairs of two-sentence utterances as in:

He wanted the carrots to cook fast. So ***he diced*** them.

The top of the cakes had come out looking uneven. So ***he'd iced*** them.

Four versions of each sentence pair were made by cross-splicing just before the beginning of the target word. Cross-splices were always to another token by the same speaker, and the sound segments of the critical target word never included spliced material: the spliced material always immediately preceded the target word. The critical experimental manipulation was whether the splice was with another token of the same sentence (allophonic match), or with a token of the other member of the pair (allophonic mismatch). Table 2 shows an example, for the sentence pair *diced* and *iced*. When the target word was *diced*, an allophonic match was made by cross-splicing two tokens of the same sentence, *So he* from one token, and *diced them* from another. An allophonic mismatch was made by cross-splicing *So he* from *So he'd iced them* to *diced them* from *So he diced them*. Likewise, when the target word was *iced*, a match involved splicing *So he'd* from one token of *So he'd iced them* to *iced them* from another; a mismatch involved splicing *So he d* from *So he diced them* to *iced them* from *So he'd iced them*. (Obstruent–vowel splice points followed standard criteria; obstruent–obstruent splice points coincided with regions of greatest perceptual change between segments.)

Smith trained 40 listeners for $1\frac{1}{2}$ h with multiple tokens of 24 pairs of these types of utterances (i.e., each critical sentence with a disambiguating preceding sentence), spoken by two speakers, one male and one female. Utterances were presented in random order. To ensure that listeners attended to meaning (i.e., to ensure that they listened as normally as possible during the training), they answered questions about the general events described in each utterance after hearing it.

After this familiarization, the listeners undertook the word-monitoring task. In this, a trial comprised only the second sentence of each utterance (i.e., no disambiguating context), some spoken by these same two familiar speakers, and others by four unfamiliar speakers. All target words were cross-spliced, and the target word was always that intended by the speaker. In other words, the critical sentence always had the appropriate allophones for the written target word itself; it was the preceding context that was either correct or incorrect for the target word.

The target word appeared (in writing) on a computer screen, and then between one and seven candidate sentences were heard, the last of which was the critical (correct) one. Candidate sentences systematically resembled the critical sentence in some respect, so that listeners had to be quite attentive. The task was to press a button as fast as possible when the target word was heard (not embedded in another word). For example, if the target word was *iced*, a listener might hear: *So he dialled from a payphone. So he'd iron your clothes if you asked him to. You denied having enticed them. So he'd iced them*. The trial ended after the critical sentence was played (the number of candidate sentences varied from trial to trial. There were 'filler trials' in which the target word was never heard).

Table 2
Splicing principles to produce four stimuli from two original sentences, as used by Smith (2002)

Target word	Allophonic	
	Match	Mismatch
diced	[so he]$_{\text{diced-1}}$ [diced them]$_2$	[so he]$_{\text{iced}}$ [diced them]
iced	[so he'd]$_{\text{iced-1}}$ [iced them]$_2$	[so he d]$_{\text{diced}}$ [iced them]

Subscripts $_1$ and $_2$ indicate different tokens of the same sentence (same speaker). Subscript $_{\text{iced}}$ indicates speech from a sentence containing *iced*; subscript $_{\text{diced}}$ indicates speech from a sentence containing *diced*.

It was predicted that, when the talker is familiar, allophonic matches would be facilitative, whereas mismatches would be disruptive. When the talker is unfamiliar, allophonic match/mismatch was expected to have less or no effect. The pattern of reaction times to press the button broadly supported these predictions, except that disruption only occurred when the mismatching context included the onset of the alternative word of the pair. That is, disruption was observed for familiar speakers' *[so he d]$_{diced}$ [iced them]*, in which /d/ of *diced* is heard but the target is *iced*, but there was no disruption for *[so he]$_{iced}$ [diced them]*, in which *diced* is the target and is intact; all that is 'wrong' in the latter case is that the preceding word, *he*, is appropriate for *he'd* but not necessarily for *he diced*. Presumably this difference between the two contexts is because disruption of the stimulus is greater at the beginning of these words (cf. Cole & Jakimik, 1980). Note that part of the systematic fine-detailed differences between this and similar pairs includes the quality and dipthongization of /aɪ/, so *iced* does not sound exactly like *diced* without the /d/, even when there is a /d/ in the coda of the preceding word. Likewise, /i/ can sound quite different in *he (diced)* and *he'd (iced)*.

The conclusion is that disruption to a familiar situation can produce worse performance, as Goldinger and Azuma have shown. The novelty in Smith's results is that 'familiarity' can involve linguistically relevant speech material, and that the recognition of words can be affected in predictable ways when the speaker-specific allophones are coherent. Smith's observed effects are small and subtle, and it remains to be shown that they generalize to more speakers than just the two she used. Nevertheless, her data offer reasonably clear support for the claim that exemplar memory for linguistic properties—here, allophonic detail—can be implicated in linguistic processing. More generally, this and later work by Smith indicates that these data are also compatible with the view that categories 'emerge' through learning, and that this type of learning, or accommodation to new information, happens routinely and fast, for subsequent experiments showed its effects can be observed after about 8 min of training (Smith, in preparation).

5. Roles of systematic variation in fine phonetic detail

5.1. The range of roles

What sort of fine phonetic detail is linguistically relevant? There is much still to learn here, because rigorous empirical work in the area that is directly relevant to perception models is relatively recent, requires painstaking observation and measurement, and is thus slow. (Automatic, statistical techniques can help, but need to be informed by *appropriate* theory and close listening, cf. Coleman (2003) in this volume.) Local (2003) demonstrates in this volume that certain types of fine phonetic detail can systematically reflect not just the phonemic content but the wider phonological and grammatical structure of the message. Some systematic differences in phonetic fine detail are relatively localized in the speech signal. Others stretch over several syllables. Both types can make speech easier to understand.

For example, it is well-known that local (short-range) phonetic detail, which tends to be relatively complex acoustically, can indicate word boundaries in strings of identical phonemes (e.g., Nakatani & Dukes, 1977). Short-range effects can also indicate grammatical function. The best-known example is that /ð/ is word-initial only in function words, and is subject to different

connected-speech processes than initial /θ/ in content words (Kelly & Local, 1986; Manuel, 1995; Local, 2003). Local (2003) shows how word-final /m/ is realized differently depending on whether it occurs in the pronoun + *be* system, viz. *I'm*, or in content words (*lime*, *shame*, *loom* etc). Variation for *I'm* is considerably greater than for content words with final /aɪm/. The variation is informative, because it is lawful within the linguistic system. Ogden (1999) discusses processes unique to auxiliary verbs.

Variation also reflects aspects of discourse function and lexical properties, much of which is perceptually salient (e.g., Hawkins & Warren, 1994; Docherty, Milroy, Milroy, & Walshaw, 1997; Bard, Anderson, Sotillo, Aylett, Doherty-Sneddon, & Newlands, 2000; Hay, 2000; Brown, 2002; Pierrehumbert, 2002; Local, 2003).

Long-range detail includes (a) phonetic properties that keep a conversation going by signaling things like agreement, (quality of) disagreement, and opportunities for the listener to talk (Local, 2003), and (b) phonetic features of particular phonological contrasts. These latter include long-range resonance effects due to the presence of /r/ or /l/ (Kelly & Local, 1986; Hawkins & Slater, 1994; West, 1999, 2000; Tunley, 1999; Heid & Hawkins, 2000), voicing of coda obstruents (Coleman, 2003; Hawkins & Nguyen, 2004), and the feature [anterior] (Coleman, 2003). These and others are discussed by Hawkins and Smith (2001), Coleman (2003), Local (2003), and Hawkins and Nguyen (2004). There may be other influences that we do not yet know about, although Coleman's (2003) work suggests that the list may be relatively short.

To summarize, these types of systematic variation in fine phonetic detail indicate linguistic function and category identity at all levels of linguistic and interactional analysis, not just units of phonological analysis, and they provide perceptual coherence. See Hawkins and Smith (2001), Pierrehumbert (2002), and Local (2003) for more detailed discussions.

5.2. *Perceptual coherence*

As I define it, a speech signal is perceptually coherent when it appears to come from a single talker because its properties reflect the detailed vocal-tract dynamics particular to that talker. The term 'perceptual coherence' or similar is used by a number of individuals independently (e.g., Grunke & Pisoni, 1982; Remez, 1994, 2003; Remez, Rubin, Berns, Pardo, & Lang, 1994; Hawkins, 1995, 1996). There is no simple definition, possibly because, although we intuitively understand the concept, it is part hypothesis, part factually based: we do not know exactly what properties make speech perceptually coherent, but we do know from many different types of work that small perturbations can change its perceived coherence (e.g., Huggins, 1972a, b; Darwin & Gardner, 1985).

Perceptual coherence is the 'perceptual glue' of speech (cf. Remez & Rubin, 1992). It is rooted in the sensory signal but relies on knowledge; the two are not distinct in this respect, but feed each other. It underlies the robustness of natural speech and determines why it sounds natural. Conversely, it may be the key to understanding a number of the challenges in producing good synthetic speech, such as why synthetic speech can be very intelligible in good conditions but tends not to be as robust as natural speech in difficult listening conditions, and why speech synthesized by concatenating chunks of natural speech that preserve natural segment boundaries can give an impression of sounding natural even if it is not very intelligible.

Many phenomena indicate the importance of perceptual coherence to a signal. They include the McGurk effect (McGurk & MacDonald, 1976), general vowel-to-vowel coarticulation (e.g., Alfonso & Baer, 1982), and complex dynamic changes at abrupt segment boundaries that affect the amplitude envelope and nature of excitation at the boundary (Heid & Hawkins, 1999). Many of them operate in relatively short domains. Examples from longer domains are the long-domain resonance effects mentioned above and summarized in Coleman (2003). The effect of including them in synthesized sentences often goes unnoticed in good listening conditions, yet when these sentences are heard in noise, their intelligibility can improve by around 15% (Hawkins & Slater, 1994). The auditory impression is hard to describe; utterances just seem to 'fit' better when they include resonance effects—i.e., they are more coherent. More technically, one could say that when resonance effects are present, formant relationships match the particular accent and style of speech better, so the signal fits together and sounds as if it comes from one person, using a consistent accent and style—i.e., it sounds more natural. Another long-domain coherence effect is systematic variation in English onset /l/s due to coda voicing, discussed by Hawkins and Nguyen for production (2004) and perception (2001, 2003).

What is the mechanism underlying perceptual coherence? There are probably many ways to look at it, and there may be more than one type of process underlying the concept. One possibility is Gestalt-type properties, as advocated by Bregman (1990) and implemented in many models and applications, e.g., Cooke (2003) in this volume; see also Cooke and Ellis (2001). There is some question as to whether these processes are basic properties of the sensory system, or more abstract (cf. Remez, 2003). The nature of speech means there are bound to be speech-specific attributes of perceptual coherence, some of which are presumably central rather than peripheral. Whether one calls them abstract or not may be partly definitional. However, some attributes of perceptual coherence relevant to speech perception are unlikely to be speech-specific and might arise from low-level psychoacoustic processes, see Shamma (2003) and Moore (2003) in this volume. Moore (2003) suggests that sounds could be represented internally as a particular type of spectro-temporal excitation pattern (STEP) with at least three dimensions (frequency, amplitude, time). A STEP is essentially a vector of values in three-dimensional space, built up by a series of 'looks' (Viemeister & Wakefield, 1991). Stored STEPS could serve as templates or prototypes against which incoming STEPs are compared. Moore discusses experimental evidence suggesting that the looks contributing to a STEP are not simply integrated over time in the mathematical sense, but that intelligent (i.e., knowledge-driven) central processes affect how they are selectively combined for specific tasks. How this concept might apply to speech as opposed to the simpler stimuli of psychoacoustics is not yet understood, but the concept of intelligent use of information available from multiple looks, not necessarily adjacent to one another, promises to be fruitful, perhaps particularly for the concept of perceptual coherence in speech. ART models have obvious relevance in this connection.

In sum, perceptual coherence is fundamentally dynamic and signal-dependent, yet crucially interacts with knowledge. Whatever the mechanism(s) underlying perceptual coherence, it seems indisputable that fine phonetic detail contributes to it, and that the concept of perceptual coherence has implications for how we should represent fine phonetic detail, and what a linguistic category is.

6. Linguistic categories: dynamic, relational, and therefore plastic

A category can only 'exist' by virtue of its context. And, as indicated by Fig. 2, the defining features of a category in one context may not be those that define it in another. Although phonological categories are clearly seen as contrastive, there is a tendency to think of phonetic categories as absolutes, especially when dealing with understanding at the level of words or higher. But phonetic categories are just as relational as phonological ones, as many experiments demonstrate. Interestingly, the relational attributes of phonetic categories are more often controlled for than investigated in their own right. Text-book examples include range and rate effects in categorical perception experiments, coarticulatory effects, and phenomena like the Ganong effect. Many influential theoretical approaches to explaining perception of phonetic/phonological segments seem also to be predicated on the idea that phonetic categories, once identified, are immutable. For example, though both the motor theory of speech perception (Liberman & Mattingly, 1985) and relative invariance theory (Stevens, 1998) emphasize and demonstrate the role of local context in identifying features or phonemes, neither offers processes whereby long-domain influences have cumulative effects, or later decisions change earlier ones. See Hawkins and Smith (2001), Section 4 for further discussion.

In summary, 'units' are functionally inseparable from 'context'; the context and the signal together determine whether the whole percept is coherent or not, and hence what each 'unit' is. Together with the preceding evidence on exemplar memory and learning, this suggests that linguistic categories, including phonetic ones, are dynamic, relational, i.e., context-sensitive, and plastic, i.e., labile.

7. Polysystemic structural representation and Firthian prosodic phonology: the coherent representation of fine phonetic detail

If systematic fine phonetic detail is important to perception, even if only in adverse listening conditions, then we need to find a model that 'knows about' these details and that can use them to make sense of the signal when necessary. In other words, phonetic fine detail must be mapped onto a structure that reflects linguistically-relevant distinctions, so that it can be interpreted as meaningful. The above arguments about the fundamental nature of perceptual coherence and the influence of context on category identity suggest that these structures should let the 'same' piece of sensory signal take on a different significance *in the linguistic system* when placed in a different context.

Although we can identify a number of good candidate contributors of perceptual coherence, we do not yet know exactly what it is about the phonetics of a particular unit or linguistic structure that causes it to be perceived as a coherent unit. But it seems reasonable to assume that a lot of weak evidence that points to the same conclusion (e.g., the presence of an /r/ later in the signal) can build up over time to effectively become quite strong evidence (cf. Warren & Marslen-Wilson, 1987; Hawkins & Warren, 1994).

Because coherence seems to result from complex relationships between physical and language-systemic constraints experienced in a particular context, we need a linguistic model that allows clear predictions about the phonetic properties associated with the different phonological and

grammatical structures that exist in a particular language. It needs to systematize the complexity introduced by the diverse linguistic factors that influence speech production and acoustics.

It follows that we need a very different structure from the standard, serially-organized stages of most models. Furthermore, the approach adopted here is at odds with the approach of some models that can be characterized as '*x* must be stored this way, or else that way, but it cannot be stored both ways'. As noted early in Section 4, I suggest that information derived from sensory input *can* be stored in more than one way, because all sorts of information can be *extracted* from exemplar representations, reorganized/re-stored, and used for particular tasks, as long as it was initially attended to.

Firthian phonology, or Firthian prosodic analysis, FPA, (Firth, 1948, 1957; Palmer, 1968, 1970; Ogden & Local, 1994) provides a way of representing this detail in terms of formal, hierarchical linguistic structures (see also Local, 2003). FPA offers a way of conceptualizing the representation of speech and language in the brain as a *polysystemic structure*, by which is meant the following. First, there are different subsystems within each level of formal linguistic analysis, each of which can have its own phonetic properties, as when connected speech processes differ between grammatical and content words. Second, each piece of phonetic information in the signal can potentially supply information to more than one of these systems: to grammar and content, to linguistic message and speaker identity, and so on.

There is thus a separate representation, depictable for present purposes as a lattice or tree, for each type of unique linguistic structure. FPA can be thought of as a declarative computational model, comprising trees which contain all the necessary information to generate (for synthesis) or match (for recognition) appropriate pronunciations. However, that is where the resemblance stops. The structure that represents a given utterance is really a set of linked structures, each completely describing the utterance, but from a different point of view, e.g., a prosodic tree, a grammatical tree, sets of lexical descriptions, and so on. These different but linked systems can be thought of as one aspect of a polysystemic system. A prosodic tree in current use (ProSynth: Ogden et al., 2000) comprises Intonational phrase, Foot, Syllable, Onset/Rhyme, Nucleus/Coda, and Feature nodes. Phonemes are not represented: they cannot be, because phonemes are by definition *devoid of current hierarchical context*: their relational attributes are part of a different contrastive system.

In the terminology used by current Firthian phonologists, the tree's nodes contain features (phonological, grammatical, etc.) that relate to particular types of phonetic events. The exact acoustic nature of a particular phonetic event is immaterial at the phonological level, but if the phonetic event is to be realized, then it must be represented in the phonology, i.e., in the description of the linguistic structure. Thus some of the features are not conventional, and some features reside higher up in the tree than others, e.g., Ogden et al. (2000) specify coda voicing as a property of the rhyme.

The links between these different types of trees allow certain features from one type of tree to effectively migrate into another and influence the phonetic outcome. So grammatical features migrate into the prosodic tree, which is likely to be the main one for perception. Consider how grammatical status affects phonetic realization. As noted above, the rules for coarticulation before grammatical words like *the* and *that* are quite different from those before similar phonemic sequences in content words; auxiliary verbs have their own distinctive set of connected speech processes; and so on. So, in the polysystemic system, a node that specifies labiality and nasality in

the coda of a weak syllable in a function word is not at all the same thing as a node that specifies labiality and nasality in the coda of a strong syllable in a content word, even though they may both be regarded as the phoneme /m/ in more mainstream linguistics.

So a polysystemic system has (a) different (prosodic, grammatical) systems as described above, and (b) different ways of organizing the information conceptually. Syllable onsets, for example, form a natural class, of which onsets with bilabial obstruents in heavy syllables in bi-syllabic feet form a sub-class. Such classes are clearly identifiable from the linguistic structure and define the set of contrasts that much of the systematic variation in phonetic fine detail reflects. Thus, its position in linguistic structure means that the /t/ of *tap* has more in common with the /p/ of *pat* than with the /t/s of *pat* and *temerity*. It is not that phonemic contrast is not valid or useful, but rather that other contrasts, reflecting finer phonetic detail, are more informative for many linguistic purposes. These contrasts are more complete and more 'real' than phonemes, which offer only one sort of contrast, that of lexical form—and a rather limited set of lexical forms, at that.

The crucial point in this approach is that there are a great many unique structures, because of the number of fine distinctions that are made. If the structure differs, then the sounds may differ even if the phonemes are the same. And if the sounds differ systematically, then the structures must differ (even if the phonemes are the same) because the difference in phonetic realization must be represented in the linguistic structure.

8. Polysp (polysystemic speech perception, or understanding): Firthian prosodic analysis meets exemplar memory

8.1. Outline

The polysystemic Firthian-type tree-structure seems well-suited as the basis of a psychological model that uses all the systematic fine phonetic detail available from the signal, since there is a rich store of putative comparisons, which we can think of as templates associated with particular meanings. Its richness and diversity also make it ideal for a model based on exemplar memory: comparisons of input against template can be subtle and affected by what is currently being attended to. (I use template loosely, to mean any comparison or standard held in memory, against which sensory input can be evaluated for classification/interpretation. It seems reasonable to suppose that there are such standards, and for current purposes their nature is immaterial. A template might be an abstraction from the input, or the modal input in a class, and the metric some evaluation of best fit.) Hawkins and Smith (2001), developing Hawkins (1995, 1996) described such a system, calling it *Polysp*, for Polysystemic Speech Perception. Polysp is not yet specific enough to be a model. It is more a conceptual approach whose aim is to show the value of including fine phonetic detail into any model of how human brains understand speech.

In Polysp, the mapping of phonetic form onto linguistic structure depends on identifying constellations of properties, and connecting them with the appropriate structural description for the particular accent and speech style. The description is detailed and declarative, with a clear relationship between abstract formal structure and sound pattern.

Exemplars, rooted in multi-modal memory of previous experiences, are mapped onto a polysystemic structure, part of which represents Firthian-type linguistic structure, to give

simultaneous linguistic and paralinguistic information. Information about linguistic structure includes syllable structure and stress, phonological weight, word boundaries, and grammatical status, as well as segmental identity. Paralinguistic information includes the speaker's voice quality, attitudes, and emotions, and, secondarily, how they affect the listener's own attitudes and emotions. All this is organized into coherent clusters of relevant information depending on what seems important—to the listener—at the time. The multi-modal aspect of the exemplar memories is important: it allows a particular acoustic–phonetic property to contribute to a wide range of cognitive systems, and for associative connections between like objects to develop according to individual experience.

Evaluation is probabilistic. There is no necessary segmentation of signal into formal linguistic categories: segmentation and categorization emerge as a result of the way the brain naturally organizes and classifies like with like. Because categories are self-organizing and emergent, each individual develops somewhat different mental representations of language.

The perceptual coherence of the signal confirms the categories identified, or raises the likelihood of certain categories over others. The latter process could allow the listener to access meaning without needing the sort of confirmation that a complete linguistic analysis offers. Phonetic support for this claim is that a coherent signal contributes to the signal being understood when it has not all been clearly heard (e.g., Hawkins & Slater, 1994; West, 2000). Psychoacoustic support comes from the STEP concept, linked with selective interpretation of multiple 'looks' (Section 5.2).

In outline, the process of understanding takes place as follows. The listener first picks up information that maps onto different parts of linguistic structure and begins to organize it coherently. For example, from the speech signal of a given utterance, this information might include stressed syllables, unstressed syllables, sibilant fricative, open vowel, determiner, removable morpheme, weakly stressed auxiliary verb, and so on. Simultaneously, non-linguistic information of all types is organized. Putative linguistic structures are built up and matched with putative meanings from the very beginning. Each time a meaning is arrived at, there is a check that what linguistic structure has NOT yet been constructed *from the sensory signal* is compatible with the signal. Long-domain phonetic detail plays a role here, as well as coarser information like the amplitude envelope, which indicates stress pattern, jaw height, etc. This same sort of cyclic process could take place for recognition of linguistic units such as words, syllables or phonemes. Of these, words may normally be checked for; phonemes seem less likely to be checked for in normal conversations than in some laboratory tasks. When meaning is arrived at and accepted as correct, then either of two things might happen. The listener might 'fill in' the rest of the linguistic structure, checking that it fits the memory of the actual signal satisfactorily: this might be what an experienced listener does in good listening conditions, and in poor conditions where accuracy of understanding is important. Alternatively, the listener may simply stop mapping the signal onto formal linguistic structure—hence some parts of any given constructed ('perceived') structure may be more complete than others. This may be what a child does, or an adult in poor listening conditions when accuracy is less important.

8.2. Example process

Fig. 3 partially illustrates this process. It shows spectrograms and associated syllabic structures of the words *mistimes* and *mistakes*, spoken by the author (British English-speaking woman) in *I'd*

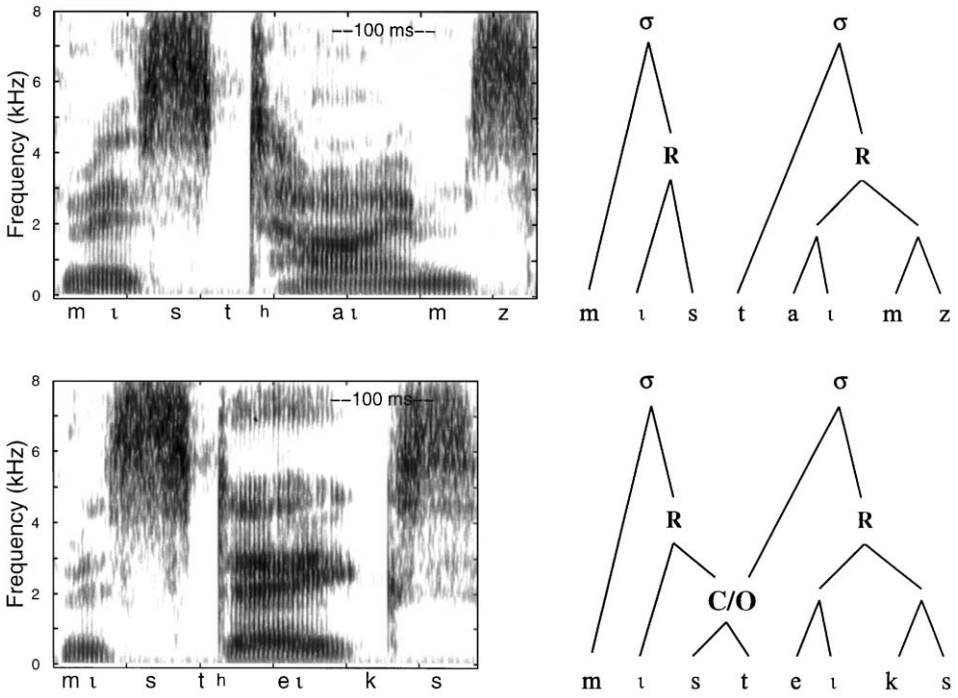

Fig. 3. Left: spectrograms of the words mistimes (top) and mistakes (bottom) spoken by a British English woman in the sentence *I'd be surprised if Tess…it* with main stress on *Tess*. Aligned at the onset of periodicity at word onset. Right: syllabic structures of each word (from Hawkins & Smith, 2001).

be surprised if Tess…it. Nuclear stress was on *Tess*. The beginnings of these two words are phonetically different in a number of ways, although the first four phonemes are usually thought to be same. The /t/ of *mistimes* is aspirated and has a longer closure, whereas the one in *mistakes* is not aspirated and has a shorter closure. The /s/ of *mistimes* is shorter, and its /m/ and /ɪ/ are longer. These differences in segmental duration are heard as a rhythmic difference: the first syllable of *mistimes* has a heavier beat than that of *mistakes*. As explained in Ogden et al. (2000), (or see Hawkins & Smith, 2001, or Hawkins, 2003), these acoustic–phonetic differences in the words' first four phonemes arise because *mis* is a productive (removable) morpheme in *mistimes* but not in *mistakes*. Following the Firthian, non-standard, analysis of, e.g., Ogden et al. (2000), this difference in the productivity of *mis-* is reflected phonologically in the syllable structure, shown on the right of Fig. 3. These structural phonological differences in turn affect the phonetic detail, despite the segmental similarities between the words. When the various properties have the wrong relationships to one another, they are heard as unnatural, and would presumably be harder to understand.

Mapping the sound pattern to the right linguistic structure proceeds as follows. The listener first hears [m], which should activate [nasal] and eventually [labial] features, as well as high probability of a new syllable and presumably a new word, since *Tess* is probably understood and expected in this context. The beginning of the [ɪ] segment confirms the presence of a new syllable, and eventually its vowel quality is also confirmed. At this point, *mistakes* and *mistimes* might be

equally probable (though see below for effects of context.) When the aperiodicity of the [s] is heard, both syllable coda and the next syllable onset could be activated (and hence weak probability of a second syllable), as well as moderately high probability of the morpheme *mis-*. Silence for the [t] closure will raise activation of features for stop and alveolar, and hence of the word *mist*, as well as of a following syllable. Simultaneously, the relative durations and spectral properties of the first two-to-three segments will begin to favor one morpheme over the other, and hence one word over the other: when the duration of [s] is short relative to what has gone before, then *mistimes* will be more likely than *mistakes*; when the [s] is relatively long, then the probability of *mistakes* will rise. Relatively abrupt transitions between /m/ and /ɪ/, with higher-amplitude formants, will favor *mistimes*. The time at which the [t] burst occurs will confirm the 'choice' made on the basis of the earlier evidence, so that by the time the definitive presence/absence of aspiration occurs, the right morphemic choice may already have been made. In consequence, the number of competing words will be reduced, compared with a system in which lexical items are stored as phoneme strings. Much of this information will in turn affect the accuracy with which the identity of the following diphthong is decided as it is heard in real time. Likewise, decisions based on speech rhythm will affect later decisions. In sum, fine phonetic detail in the first syllable will serve to increase overall certainty of decisions over a longer stretch of the signal.

The above description would apply if the word was heard in isolation. If it was heard in context, then the rhythm of the first part of the utterance would result in even earlier identification of the right syllable and morphemic structure. In the absence of hesitation pauses there would be little doubt about the rate of speech, so decisions about the spectral definition of [m] and [ɪ], and the relative durations of [mɪ] and [s], could be taken sooner because levels of certainty would rise earlier. Vowel quality might likewise be identified earlier due to coarticulation with *Tess*. In short, if we liken each spectrogram in Fig. 3 to a template or appropriate exemplar, aligned at the beginning of periodicity for /m/, and if the word *mistimes* was spoken fluently and with a preceding context, then, before the end of the periodicity for /ɪ/, the spectro-temporal properties of the spoken *mistimes* would match that of template *mistimes* better than that of *mistakes*. Thus the systematic fine phonetic detail could significantly reduce the probability of *mistakes* before the end of the second of the four identical phonemes in these words.

Reliable evidence for other aspects of structure provide an overall skeletal structure for each word. In Fig. 3, acoustic 'islands of reliability' include (the invariant properties of) [s], the manner of articulation of the stops and nasals, and certain spectrally-defined temporal regions within the vowels which confirm aspects of their manner of articulation as well as their function as syllabic nuclei (cf. Stevens, 1998, 2002). Their combined attributes contribute prosodic information: number of syllables, relative stress, etc. This skeleton, combined with predictions of grammatical class, e.g., a verb will probably follow *Tess*, could make tentative identification of spoken words extremely fast without confirming the exact identity of all sound segments.

In sum, identification takes place probabilistically, using all possible available information in parallel to flesh out linguistic structure at all levels. The spectro-temporal relationships between different parts of the signal are all-important: each structural element is identified relative to others in the environment and to the listener's expectations derived from exemplar memories. The listener aims to arrive at meaning, not at a complete linguistic description, so he or she will accept the most probable meaning as soon as the overall evidence matches the expected sound pattern well enough.

8.3. Some implications of Polysp's assumptions

There need be no difference in output between conventional and Firthian/Polysp models. But because Polysp is hierarchical and declarative, with clearer connections between different 'levels' of formal linguistic analysis, it is easier (a) to know whether the description of x is complete, (b) to classify like with like and distinguish them from unlike, to a more subtle (but perceptually important) degree and (c) to conceptualize how the perceiver may build up a picture of the meaning of the utterance from islands of reliability, or anchor points, that may be at several different levels of formal linguistic analysis. These and other properties mean that Polysp can in principle explain a number of well-known attributes of speech understanding that are typically under-emphasized by conventional models of speech perception.

It is fundamental to Polysp that identification of formal linguistic categories can take place at any point in the process—before meaning is understood, simultaneously with its understanding, or after it has been understood, as circumstances dictate. Linguistic and non-linguistic context guides experienced listeners in using global patterns appropriate to the accent and style, rather than searching for cues suitable for a 'canonical' pronunciation. Hence, both casual speech and exceptionally clear speech will be better understood by listeners who are experienced with it than by listeners for whom the language or accent is unfamiliar (Bradlow & Bent, 2002).

Given sufficient context and coherence, small parts of the signal can give strong cues to parts of structure at different linguistic levels. So the listener does not need to complete the entire formal 'jigsaw' in order to understand its message. Thus, at the moment an utterance is understood, its formal linguistic structure may have been only partially identified. This assumption allows the same basic processes to underlie infants'/children's and adults' speech understanding. However, whereas a young child may understand an utterance and move on without 'knowing' its linguistic analysis, adults will presumably identify its linguistic structure immediately its meaning is accepted. The common factors are the process (of understanding with incomplete structural analysis), and the fact that each listener, experienced or inexperienced, identifies the formal linguistic structure to the extent that his or her experience allows.

By including exemplar memories, Polysp inevitably implies that the mental structures corresponding to a linguistic system can differ between individuals. Brain-imaging studies (and much of the classical aphasia literature) support this view (Damasio, 1995; Pulvermüller, Assadollahi, & Elbert, 2001, p. 201). More abstract linguistic structures develop throughout childhood and beyond, as emergent categories from these sensory events. Thus, when a number of listeners hear the same utterance, their individual previous experiences may strongly influence the routes they take towards understanding its meaning. The greater their shared knowledge and common focus of attention, the greater the similarity in their processing, and the better they communicate. Though this common-sense point is speculative from the viewpoint of science, it seems worth emphasizing because it potentially links how isolated spoken syllables or words are understood with how any type of communication occurs: in Polysp, speech is not special.

Exemplar memory, learning/adaptation throughout life, the strongly relational basis of category identification, and probabilistic decision processes, mean that linguistic categories are plastic. Adaptation to new situations, accents, and individuals occurs because the distribution of

stored exemplars changes as the input changes. Even one's childhood accent may sound unfamiliar if it has not been heard for some years.

Although Polysp uses hierarchical structures, they arise through associative learning. In the right circumstances, Polysp allows linear analysis of speech. While hierarchical organization is probably a fundamental property of how the brain structures information, including language, many utterances can be understood without hierarchies—*representation* of something as a hierarchy need not necessarily entail that it always *functions* as a hierarchy. This is easily seen within semantic fields, where the precise topic and level of detail may determine the degree of hierarchical organization that need be attended to. Thus, eating habits of mammals and reptiles may be compared without systematically considering differences between subclasses of mammals (ruminants, rodents, etc.); or differences between mammalian subclasses may be discussed without considering individual members or their interrelationships.

Likewise for language development. A young child could successfully operate linguistically by substituting like with like in a sequence that has no, or very little, hierarchical structure (cf. Elman, 1993). The child needs only to be able to sequence items, and to connect them with common-sense understanding of relationships between events and objects in its physical world. She might learn parts of speech by associating actor, action, acted upon, as a sequence (in the correct word order for her language), and substituting equivalent 'things' in each place. She probably need hear rather few sentences like *the girl kicks the ball, the dog licks mummy*, *baby likes the milk*, in order to realize that *girl*, *dog*, *mummy*, *baby* comprise a class by virtue of being mutually substitutable in many contexts, and that *kicks*, *licks*, *likes* form another class. Rudimentary concepts of Noun and Verb can be considered to be established when each class has several members. When she notices that words like *the*, *a*, *this*, *that* are roughly interchangeable and co-occur with Nouns, she has the beginning of the concept Noun Phrase, and hence of hierarchical linguistic structure. Similarly for subcategorization of nouns. Viewed this way, there seems nothing mysterious about the construction of hierarchical relationships as the child's short-term memory span, cognitive abilities, and vocabulary increase: it is what brains do in order to organize complex relationships. There is thus much to value in models that allow 'flat' yet long structures, as well as hierarchies. For adults, some very casual speech seems a good candidate for non-hierarchical processing.

Finally, models that do not include fine phonetic detail are likely to invoke types of top-down knowledge and special processes in ways that do not reflect reality. If the hypothesis about the role of fine phonetic detail is correct, then the basis of many if not all processes that are typically seen as knowledge-driven and 'top-down' originate from the sensory input. The processes are indeed knowledge-driven, but the triggers to activate the relevant piece of knowledge are in the acoustics just as surely as the acoustic triggers that tell one that a 'discrete sound' is /ʃ/and not /tʃ/; (treating the [ʃ ~ tʃ] distinction as important is also knowledge-driven). In other words, compared with conventional models, Polysp makes less of the distinction between bottom-up and top-down processes, for it attributes more to the information-value in the sensory input, yet acknowledges that sensory information is useless in the absence of learned information about how it should be interpreted within the language (cf. Burgess, 2002; Grossberg, 2003; Pierrehumbert, 2003). There are commensurate gains in theoretical elegance.

9. How might Polysp be modelled?

The above outlines the general principles of Polysp, and some of their implications. This section suggests particular processes by which perception might take place, especially for conversational speech, with its wide variety of connected-speech processes. Continuous-activation models seem ideally suited to Polysp, provided that they allow close attention to the detailed sensory signal as it evolves through time. Two particularly promising approaches are Stevens' relative invariance and landmark approach, and Grossberg's and colleagues' ART models. The former provides the initial acoustic–phonetic triggers to begin classification. The latter provides the process by which larger linguistic structures are built up. However, there are alternatives (see below) and I use each in a way that its originator may not agree with.

ART models (e.g., Cohen & Grossberg, 1997; Boardman, Grossberg, Myers, & Cohen, 1999) can bind detailed current sensation with detailed exemplar memories to produce hierarchical linguistic structure. From a combination of current properties of the signal and existing knowledge, they use gradual establishment and decay of 'resonances' corresponding to self-organizing emergent, relational categories, and they have good potential for modelling long-domain effects as larger units established from consistent weak acoustic evidence. They are thus good at structuring information derived from widely different temporal windows, some of which overlap, while others could be non-adjacent. Identification of higher-order structure refocuses attention away from lower-order constituent units, and later-arriving sensory information allows earlier decisions to be revised. In this volume, Grossberg (2003) overviews these models, while Goldinger and Azuma (2003) develop their potential for exemplar-based speech perception. Hawkins and Smith (2001) discuss its relevance to Polysp.

Stevens' (1998, 2002) landmarks identify acoustic–phonetic properties that signify phonological distinctive features which reflect manner of articulation and thus distinguish between the syllable nucleus and its margins. The phonological features they include (vocalic, consonantal, continuant, lateral,[1] sonorant, and strident) mark parts of the signal that have status relatively high in Polysp's model of linguistic structure, including, vitally, those contributing to rhythm. Singly or in combination, they correspond closely to Zue's (1985) robust features. (Whereas most robust features are fairly steady-state—e.g., periodicity, strident/weak fricative, nasal, silence—Stevens' landmarks and relative invariants are typically dynamic; they define temporal regions in which critical events occur. This difference need not concern us here.) Stevens likewise distinguishes other phonological features, mainly reflecting place of articulation, by means of spectral changes that occur within a few milliseconds of one another and whose relationship is invariant. Polysp, by its nature, does not require invariant properties for all features at all times; but those that are present in a signal function as islands of reliability in the structure being built up en route to meaning.

In Polysp, speech rhythm is fundamental to the process of binding perceived features together into meaningful patterns. Rhythm can be partly predicted on the basis of what has gone before, and it partly 'presents itself' from the current signal, since the segmental properties themselves, including their amplitude envelope, 'carry' the rhythm. Thus the spectro-temporal structure of the

[1] Lateral is not included in the 2002 list, but was intended to be (K.N. Stevens, personal communication, 7 August, 2003).

speech signal itself plays a determining role on the course of processing (cf. also Jones, 1976; Hartley & Houghton, 1996; Large & Jones, 1999; Hartley, 2002; Grossberg, 2003). Rate and rhythmic changes signal potentially meaningful restructuring of an utterance's segmental details.

Information available from landmarks seems compatible with such restructuring: landmarks may allow chunks of speech to be warped into syllabic shapes, for matching against stored patterns that represent, or contain information relevant to, different speech styles. Rate-sensitive warping might be initiated or facilitated by detection of other changes that often accompany rate changes, e.g., lenitions and voice quality. Or landmarks' inherent auditory salience, sequenced in a new rhythm, may focus attention on alternative speech patterns. Tied with knowledge about how the language uses different styles and rates, landmarks offer promise as the link by which Moore's (2003) suggestions for the intelligent central use of STEP representations could operate.

ART processes of parallel working memories, sensitive respectively to transient and sustained properties of the acoustic signal, and interacting with one another, are designed to model a rather rate-invariant representation of speech by warping directly onto a single linguistic representation, partly to avoid the proliferation of forms due to different rates. They have been applied mainly to careful speech, but in principle seem compatible with alternative forms—indeed, ideally suited to building up representations from information that is weak, distributed over relatively long time domains, or dependent upon properties present in some other (earlier and/or later) part of the signal. This line of thought is worth pursuing, for decisions based on assumptions about memory capacity are open to question given our present state of knowledge, whereas adult native speakers of a language clearly do have a huge range of styles available to them, both as speakers and as listeners.

Thus rhythmic properties, together with just a few salient 'segmental' properties or features, may provide the coherence that links many related forms of words and phrases, including a range of reductions due to connected-speech processes in unstressed syllables, and very young children's production of unstressed syllables in approximately the right number and rhythm, but not with the right segments.

As an example, consider the various forms that an English speaker can choose from to indicate that he or she lacks knowledge (Table 1 exemplifies a few). All the reduced forms have nasalization and diphthongization. Nasalization is typically localized in the less reduced forms, and distributed in the more reduced ones. The diphthong(s) go from more to less open, and/or from relatively more to less front. Consistent with the default stress pattern in which *know* is more prominent than *I*, the essential diphthong resembles /əʊ/ of *know* more closely than /aɪ/ of *I*. The sound pattern of the relatively unreduced [ã̰ə̃nːəʊ̆], whose first syllable signifies *I* (and probably *don't* as well, when it includes a central vowel), somewhat resembles many pronunciations of *dunno*, which does not include *I*, such as [dənəʊ] and [d̰nəʊ]. But pronunciations that are writable as *dunno* tend to begin close and relatively abruptly, whereas when *I* is represented in the signal, the sound pattern normally begins less abruptly and is more open and front than the end of the utterance. An exception to this is [n̩nəʊ], in which a syllabic nasal consonant represents *I*. However, as far as I have observed, [n̩nəʊ] is unlikely to be said unless the shoulder is simultaneously shrugged vertically (often accompanied by a particular head movement and facial expression), in which case the shrug signifies *I* (rather than *you*, for example, which a vertical shrug can never signify). Evidently the nasal is an acceptable syllable-marker when the shoulder gesture provides the precise meaning.

These observations are culture-specific, but presumably all linguistic cultures have comparable conventions. The point of this example is that the reduced forms preserve salient properties of the unreduced forms, as Kohler (2003) discusses for German. Presumably these properties, which are all representable as phonological distinctive features, though not always as phonemes, form a 'family' of sound patterns that are all relatable to the broad range of meanings whose common attribute is that they concern the speaker's lack of knowledge, as developed in Section 2. The challenge seems not so much to find the common phonological features as to find a way of modelling their familial relationships, both with one another (for not all contrast in meaning), and with the range of meanings that they convey. It seems likely that this modelling must include selective attention to culturally- and contextually-determined salient properties of the signal.

To explain the step between structure and meaning is beyond the scope of this paper, especially as it is not clear what linguistic meaning is, in biological terms. Much simple meaning may be embodied, for example as affordances (e.g., MacWhinney, 1999, p. 218). For speech, see Fowler (1986), Best (1994, 1995). This Gibsonian approach is broadly compatible with exemplar-based models, as well as with recent models that assume that selective attention influences the structure of long-term memories (e.g., Craik & Grady, 2000; Burgess, 2002; Craik, 2003; Whittlesea, 2003). In Craik and Grady's (2000) hierarchical model of knowledge and memory, low nodes in the hierarchy represent contextual details, higher nodes represent commonalities among groups of related episodes, and yet higher levels represent 'abstract' or 'context-free' knowledge. Thus Polysp, like the models of Burgess (2002) and Craik and Grady (2000), among others, incorporates both exemplar and abstract memory. Its richly-patterned linguistic structures include abstract properties, yet are derived initially from exemplars and may be mapped onto meanings that are at least partially embodied, or non-abstract.

10. Possible neural mechanisms

This section considers principles of learning and neurobiological function that might underlie Polysp. The argument is necessarily speculative: I seek plausible general principles rather than particular mechanisms. If the general argument proves plausible, then work on its details will be worthwhile.

Associative learning models offer an appealing way of modelling how exemplar memory traces for words and phrases become associated with meaning via polysystemic representation of speech, language, and concepts. A number of promising approaches exist, often for non-speech behavior. The LAMINART model, summarized by Raizada and Grossberg (2003), which incorporates ART processes and is based mainly on neurophysiology of visual cortex, may be generalizable to speech perception. Burgess' sophisticated models of spatial navigation (e.g., Burgess, 2002; Burgess, Maguire, & O'Keefe, 2002; Hartley & Burgess, 2002) dealing with perception of complex patterns and efficient pattern completion, also have a general form (Burgess et al., 2001). Burgess proposes rapid (hippocampal) storing of episodic information, and slower abstraction of its meaning with respect to past experiences stored in the neocortex. For phonology, Burgess and Hitch (1999) propose a connectionist model based on Hebbian learning (Hebb, 1949) and decay over long and short time scales. However, this model is strongly phoneme-based and adapting it to the Polysp framework seems non-trivial. These and other models are valuable for being both

computationally implemented and rooted in neuroscientific research. Their multimodal relevance also appeals.

Pulvermüller's (1999, 2002) broad-based model of speech and language understanding offers one way of conceptualizing how a hierarchical formal linguistic structure could be built up from individual word or sound representations in memory. Pulvermüller proposes a version of Hebbian cell assemblies that fits well with Polysp. Hebbian cell assemblies are functional groupings of cells that become associated when they are repeatedly activated together, and thus develop into a functional unit. Flexibility of function is thus built-in, for the same neuron can take part in more than one functional group of neurons; and a structural hierarchy can be built up, in that low-level functional groups are organized into larger ones via associative memory.

Cell assemblies may involve no feedback, but one formulation involves feedback, or reverberation, which is currently thought to be necessary in some stages of perception. Pulvermüller proposes that such reverbatory assemblies or 'functional webs' represent speech and language in the brain. He suggests that cell assemblies could bind together by association at least acoustic/auditory and semantic aspects of words. Thus exemplar memory for speech and language could be stored in an associative network, and language acquisition would be initially associative. As experience increases, hierarchical linguistic structure would develop from increasingly complex associations between cell assemblies in a self-organizing system.

Full activation of an assembly is thought to be possible when only some of its neurons are activated by external stimulation. This satisfies three fundamental tenets of Polysp. First, all aspects of speech understanding involve both sensory and central ('top-down') information; to distinguish formally between the two in a model involving discrete 'stages' promises to be fruitless (Section 8.3). Second, perception can involve recognition of salient parts of the signal at many levels of linguistic representation. Third, at the instant when an utterance is first understood, the listener may have only constructed a partial representation of its linguistic structure: understanding need not always require a complete structural analysis of the signal.

A possible neural process that could underlie Hebb's cell assemblies, is one variant of the synfire chain (Abeles 1982, 1991), developed speculatively for phonological processing by Pulvermüller (e.g., 1999, 2002). A synfire chain is a functional grouping of neurons whose spatio-temporal firing characteristics seem suited to Polsyp's tenets. Pulvermüller's preferred variant, which includes feedback loops, can be thought of as equivalent to the phase-locked oscillator of dynamical systems approaches to perception, discussed in this volume by Port (2003) and Tuller (2003). As a neural substrate for Polysp, synfire chains might underlie aspects of perceptual coherence. But Pulvermüller's interpretation is different, and more work is needed to assess their potential role in speech processing; Hawkins and Smith (2001) offer a preliminary discussion.

Though the case for synfire chains as a substrate for speech is highly speculative, the general concept is not. Other evidence confirms that functional neuronal groupings can group into larger units, resulting in something like a hierarchical system. Griffiths, Buechel, Frackowiak, and Patterson (1998) provide evidence for hierarchical organization of pitch-related temporal events in the brain (see also Griffiths, Uppenkamp, Johnsrude, Josephs, & Patterson, 2001). They note encoding of fine temporal structure for pitch up to and including primary auditory cortex, with longer pitch sequences being responded to in cortical areas distinct from primary auditory cortex, and they propose that emergent temporal properties, such as pitch sequences in sound, derive from cortico-cortical connections from primary auditory cortex. Computer simulations by

Wrigley and Brown (1999) used firing patterns of synfire chains in a system comprising a number of auditory features, each of which is represented by a cluster of synchronized synfire chains. An auditory stream (Bregman, 1990) is represented by a population of synchronized synfire chain clusters (features). Different auditory streams arise when such synfire-chain clusters fire asynchronously.

Disregarding the specific neural mechanism and level of linguistic–phonetic analysis, Pulvermüller's proposals of a network of context-sensitive allophones are broadly compatible with Moore's (2003) STEP model and Johnson's X-MOD model (Johnson, 1997; Johnson, Strand, & D'Imperio, 1997), which themselves have antecedents in Klatts' LAFS (Klatt, 1979). Pulvermüller's view of Hebbian cell assemblies associatively binding acoustic/auditory and semantic aspects of words is not only compatible with Polysp, but parallels Coleman's (2002, 2003) proposal of associative links between semantic representations and paths in a phonetic space.

In sum, these models share with Polysp associative (and in most cases exemplar) memory, which allows connections to develop horizontally as well as hierarchically, functional neuronal groupings, and potential for multi-modal explanation. Pulvermüller's covers an unusually wide range of speech and language, taking into account psycholinguistic and brain research. However, none as yet accommodates information available from systematic fine phonetic detail.

11. Summary

I have suggested that systematic variation in fine phonetic detail plays a crucial role in how people understand ordinary conversational speech. Together with non-verbal information, it allows many aspects of spoken communication to take place swiftly and accurately because its various forms map onto all aspects of the communicative situation that are required for complete understanding, rather than mainly indicating only those aspects of meaning associated with lexical form. Additionally, systematic fine phonetic detail contributes to perceptual coherence, so that the signal sounds as if it comes from a single talker, forming a single perceptual 'stream'. To accept that fine phonetic detail may play such a central rather than a peripheral role in speech understanding entails work on how phonetic detail is represented in the brain. Polysp is proposed as a conceptual framework for this purpose.

Polysp assumes that the speech signal is first stored as multi-modal exemplar memories that are linked to non-linguistic as well as linguistic information. Thus the same sensory information feeds a number of different functional groupings, and a new utterance is processed for strict linguistic meaning, information about the speaker (her identity, personal characteristics, current mood), the general situation (pragmatics), and how all this information affects the listener.

Linguistic processing entails mapping the signal onto declarative, polysystemic structures, like those of Firthian Prosodic Analysis, which indicate meaningful, grammatical, and phonological relationships. However, the listener is concerned with successful communication, not construction of the correct formal linguistic analysis of an utterance, and will interpret speech meaningfully whenever the combined sensory and relevant knowledge about the situation allow. Mapping from sensation to structure takes place until meaning is understood; after that, either mapping ceases, or else it is completed from top-down knowledge and matched against the signal for probability of being correct. Phonetic categories behave like other cognitive and linguistic categories: they are

self-organizing, emergent, context-sensitive, dynamic, and plastic throughout life. Given these properties, the mental structures corresponding to a linguistic system can differ between individuals, depending on their experiences.

It follows that there is no one way to understand a speech signal: polysystemic linguistic structure can be identified by many routes, in different orders or in parallel. At times, the sensory speech signal can be mapped directly onto meaning, for it is just one—very important—type of sensory input concerned with communication. Other categories of formal linguistic analysis, such as phonemes and words, are by-products of the route between sensation and meaning; they are important and indeed necessary in certain situations, but they need not always be identified in order for the meaning to be understood.

Although there is no one route to understanding an utterance, Polysp allows the same basic processes to be used at all levels of linguistic maturity: babies, young children, adults, first and second language-users, and so on. The details will differ because the encoded experiences against which the input is mapped differ. In particular, not all processing of meaning depends on recognition of hierarchical linguistic relationships, especially when the listener is inexperienced (cf. Jusczyk (1993) for infants, but the comment also applies to adults processing foreign languages). Nevertheless, each linguistic category involved can only be interpreted meaningfully by virtue of its context: whether they are represented hierarchically or linearly, units are inseparable from context.

Polysp has not been computationally implemented, and it may not be necessary to do so in that a number of existing models are in principle capable of incorporating its tenets, and more may appear as evidence from a number of fields converges towards the same viewpoint (cf. Docherty & Foulkes (2000) for sociophonetics, and Bybee (2001) and Pierrehumbert (2003) for phonology, as well as the neurolinguistic evidence cited above). Avenues for Polysp's development were discussed however. Polysp's principles are broadly compatible with principles of associative learning and Hebbian cell assemblies. Stevens' principles of landmarks and relative invariance, coupled with processes like Grossberg's Adaptive Resonance Theory, could provide the route from exemplar memories to the organization of linguistic structure and understanding.

To assess Polysp's worth, two issues need early attention. One is to identify plausible neural bases for the proposed processes. The other is how to constrain and test Polysp, given that it is an inherently redundant system in which almost anything can trigger the 'fast track' towards understanding meaning. These topics are being addressed.

Acknowledgements

I thank John Coleman, Gerry Docherty, John Local, Noël Nguyen, Richard Ogden, Friedemann Pulvermüller, Sophie Scott, and Rachel Smith for helpful criticism and discussion. All errors and weaknesses, of course, remain my own.

References

Abeles, M. (1982). *Local cortical circuits: An electrophysiological study*. Berlin: Springer-Verlag.
Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.

Alfonso, P. J., & Baer, T. (1982). Dynamics of vowel articulation. *Language and Speech*, *25*, 151–173.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42*, 1–22.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman, & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA/London: MIT Press.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Baltimore: York Press.

Boardman, I., Grossberg, S., Myers, C., & Cohen, M. (1999). Neural dynamics of perceptual order and context effects for variable-rate speech syllables. *Perception and Psychophysics*, *61*, 1477–1500.

Bradlow, A. (2002). Confluent talker- and listener-oriented forces in clear speech production. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology VII (phonology and phonetics)* (pp. 241–273). Berlin: Mouton de Gruyter.

Bradlow, A., & Bent, T. (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, *112*, 272–284.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Brown, R. A. (2002). Effects of lexical confusability on the production of coarticulation. *UCLA Working Papers in Phonetics*, *101*, 1–34.

Burgess, N. (2002). The hippocampus, space and viewpoints in episodic memory. *Quarterly Journal of Experimental Psychology*, *55*, 1057–1080.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581.

Burgess, N., Becker, S., King, J. A., & O'Keefe, J. (2001). Memory for events and their spatial context: Models and experiments. *Philosophical Transactions of the Royal Society, London. B.*, *356*, 1493–1503.

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*, 625–641.

Bybee, J. L. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Cohen, M. A., & Grossberg, S. (1997). Parallel auditory filtering by sustained and transient channels separates coarticulated vowels and consonants. *IEEE Transactions on Speech and Audio Processing*, *5*, 301–318.

Cole, R. A., & Jakimik, J. (1980). How are syllables used to recognise words? *Journal of the Acoustical Society of America*, *67*, 965–970.

Coleman, J. S. (1998). Cognitive reality and the phonological lexicon: A review. *Journal of Neurolinguistics*, *11*, 295–320.

Coleman, J. (2002). Phonetic representations in the mental lexicon. In J. Durand, & B. Laks (Eds.), *Phonetics, phonology and cognition* (pp. 96–130). Oxford: Oxford University Press.

Coleman, J. (2003). Discovering the acoustic correlates of phonological contrasts. *Journal of Phonetics*, *31*, doi:10.1016/j.wocn.2003.10.001.

Cooke, M. P. (2003). Glimpsing speech. *Journal of Phonetics*, *31*, doi:10.1016/S0095-4470(03)00013-5.

Cooke, M. P., & Ellis, D. P. W. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, *35*, 141–177.

Craik, F. I. M. (2003). Commentary. In J. S. Bowers, & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 327–336). Oxford: Oxford University Press.

Craik, F. I. M., & Grady, C. L. (2000). Aging, memory, and frontal lobe functioning. In D. Stuss, & R. T. Knight (Eds.), *Principles of frontal lobe function*. Oxford: Oxford University Press.

Damasio, H. (1995). *Human brain anatomy in computerized images*. New York, Oxford: Oxford University Press.

Darwin, C. J., & Gardner, R. B. (1985). Which harmonics contribute to the estimation of first formant frequency? *Speech Communication*, *4*, 231–235.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, *23*, 3423–3431.

Démonet, J. F., Price, C., Wise, R., & Frackowiak, R. S. (1994). Differential activation of right and left posterior sylvian regions by semantic and phonological tasks: A positron-emission tomography study in normal human subjects. *Neuroscience Letters*, *182*, 25–28.

Docherty, G. J., & Foulkes, P. (2000). Speaker, speech & knowledge of sounds. In N. Burton-Roberts, P. Carr, & G. J. Docherty (Eds.), *Phonological knowledge: Conceptual & empirical issues* (pp. 105–129). Oxford: Oxford University Press.

Docherty, G. J., Foulkes, P., Milroy, J., Milroy, L., & Walshaw, D. (1997). Descriptive adequacy in phonology: A variationist perspective. *Journal of Linguistics*, *33*, 275–310.

Duffy, S. A., & Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, *35*, 351–389.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Firth, J. R. (1948). Sounds and prosodies. *Transactions of the Philological Society*, 127–152.

Firth, J. R. (1957). A synopsis of linguistic theory: 1930–1955. In: *Studies in linguistic analysis*. Philological Society (pp. 1–32). Oxford: Blackwell, Reprinted in Palmer (ed. 1968) *Selected Papers of J. R. Firth*. London/Harlow: Longman.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*, 3728–3740.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3–28.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Goldinger, S. D., Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*, doi:10.1016/S0095-4470(03)00030-5.

Griffiths, T. D., Buechel, C., Frackowiak, R. S. J., & Patterson, R. D. (1998). Analysis of temporal structure in sound by the human brain. *Nature Neuroscience*, *1*, 421–427.

Griffiths, T. D., Uppenkamp, S., Johnsrude, I., Josephs, O., & Patterson, R. D. (2001). Encoding of the temporal regularity of sound in the human brainstem. *Nature Neuroscience*, *4*, 633–637.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*, doi:10.1016/S0095-4470(03)00051-2.

Grunke, M. E., & Pisoni, D. B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception and Psychophysics*, *31*, 210–218.

Hartley, T. (2002). Syllabic phase: A bottom-up representation of the structure of speech. In J. A. Bullinaria, & W. Lowe (Eds.), *7th neural computation and psychology workshop*. Singapore: World Scientific.

Hartley, T., & Burgess, N. (2002). Models of spatial cognition. In *Encyclopaedia of cognitive science*. New York: MacMillan.

Hartley, T., & Houghton, G. (1996). A linguistically constrained model of short-term memory. *Journal of Memory and Language*, *35*, 1–31.

Hawkins, S. (1995). Arguments for a nonsegmental view of speech perception. In K. Elenius, & P. Branderud (Eds.), *Proceedings of the XIIIth international congress of phonetic sciences*, Vol. 3. (pp. 18–25). Stockholm: KTH and Stockholm University.

Hawkins, S. (1996). Perceptual modeling of connected speech. In A. P. Simpson & M. Pätzold (Eds.), *Sound patterns of connected speech: Description, models and explanation* (pp. 173–180). Universität Kiel: Institut für Phonetik und digitale Sprachverarbeitung, Arbeitsberichte nr 31.

Hawkins, S. (2003). Contribution of fine phonetic detail to speech understanding. *Proceedings of the 15th international congress of phonetic sciences* (pp. 293–296). CD-ROM.

Hawkins, S., & Nguyen, N. (2001). Perception of coda voicing from properties of the onset and nucleus of *led* and *let*. In P. Dalsgaard, B. Lindberg, & H. Benner (Eds.), *Proceedings of the 7th international conference on speech communication and technology (Eurospeech 2001 Scandinavia)*, Vol. 1 (pp. 407–410).

Hawkins, S., & Nguyen, N. (2003). Effects on word recognition of syllable-onset cues to syllable-coda voicing. In J. K. Local, R. A. Ogden, & R. A. M. Temple (Eds.), *Papers in laboratory phonology VI* (pp. 38–57). Cambridge: Cambridge University Press.

Hawkins, S., & Nguyen, N. (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics*, 32, in press, doi:10.1016/S0095-4470(03)00031-7.

Hawkins, S., & Slater, A. (1994). Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *ICSLP 94 (Proceedings of the 1994 International Conference on Spoken Language Processing)*, *1*, 57–60.

Hawkins, S., & Smith, R. (2001). Polysp: A polysystemic, phonetically rich approach to speech understanding. *Italian Journal of Linguistics—Rivista di Linguistica* 13, 99–188.

Hawkins, S., & Warren, P. (1994). Implications for lexical access of phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics*, *22*, 493–511.

Hay, J. B. (2000). *Causes and consequences of word structure*. Unpublished Ph.D. Dissertation, Northwestern University.

Hebb, D. O. (1949). *The organization of behavior: A neurophysiological theory*. New York: Wiley.

Heid, S., & Hawkins, S. (1999). Synthesizing systematic variation at boundaries between vowels and obstruents. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the XIVth international congress of phonetic sciences*, Vol. 1. (pp. 511–514). Berkeley, CA: University of California.

Heid, S., & Hawkins, S. (2000). An acoustical study of long domain /r/ and /l/ coarticulation. *Proceedings of the 5th seminar on speech production: Models and data, and CREST Workshop on models of speech production: Motor planning and articulatory modelling* (pp. 77–80). Munich: Institut für Phonetik und Sprachliche Kommunikation, Ludwig-Maximilians-Universität.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Neurosciences*, *4*, 131–138.

Howell, P., & Darwin, C. J. (1977). Some properties of auditory memory for rapid formant transitions. *Memory and Cognition*, *5*, 700–708.

Huggins, A. W. F. (1972a). Just noticeable differences for segment duration in natural speech. *Journal of the Acoustical Society of America*, *51*, 1270–1278.

Huggins, A. W. F. (1972b). On the perception of temporal phenomena in speech. *Journal of the Acoustical Society of America*, *51*, 1279–1290.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego/London: Academic Press.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*, 359–384.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention and memory. *Psychological Review*, *83*, 323–355.

Jusczyk, P. W. (1993). From general to language specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, *21*, 3–28.

Keating, P. A., Cho, T., Fougeron, C., & Hsu, C-S. (2003). Domain-initial articulatory strengthening in four languages. In J. Local, R. A. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology VI: Phonetic interpretation* (pp. 143–161). Cambridge: Cambridge University Press.

Kello, C. T., & Plaut, D. C. (2003). The interplay of perception and production in phonological development: Beginnings of a connectionist model trained on real speech. *Proceedings of the 15th international congress of phonetic sciences* (pp. 297–300). CD-ROM.

Kelly, J., & Local, J. K. (1986). Long domain resonance patterns in English. In *International conference on speech input/output: Techniques and applications* (pp. 304–9). Conference Publication No. 258. London: Institute of Electrical Engineers.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Kohler, K. J. (2003). Domains of temporal control in speech and language: From utterance to segment. *Proceedings of the 15th international congress of phonetic sciences* (pp. 7–10). CD-ROM.

Lachs, L., McMichael, K., & Pisoni, D. B. (2003). Speech perception and implicit memory: Evidence for detailed episodic encoding. In J. Bowers, & C. Marsolek (Eds.), *Rethinking implicit memory*. Oxford: Oxford University Press.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, *106*, 119–159.

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

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, *99*, 1683–1692.

Lindblom, B., Brownlee, S., Davis, B., & Moon, S-J. (1992). Speech transforms. *Speech Communication*, *11*, 357–368.

Local, J.K. (2003). Variable domains and variable relevance: Interpreting phonetic exponents. *Journal of Phonetics*, *31*, doi:10.1016/S0095-4470(03)00045-7.

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

MacWhinney, B. (1999). *The emergence of language*. Mahwah: Lawrence Erlbaum Associates.

Manuel, S. Y. (1995). Speakers nasalize /ð/ after /n/, but listeners still hear /ð/. *Journal of Phonetics*, *23*, 453–476.

Martin, C., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, *379*, 649–652.

Matthews, P. H (1997). *The concise English dictionary of linguistics*. Oxford: Oxford University Press.

Moore, B. C. J. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics*, *31*, doi:10.1016/S0095-4470(03)00011-1.

Nakatani, L. H., & Dukes, K. D. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, *62*, 715–719.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700–708.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 3–27.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics*, *60*, 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42–46.

Ogden, R. (1999). A declarative account of strong and weak auxiliaries in English. *Phonology*, *16*, 55–92.

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovičová, J., & Heid, S. (2000). ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language*, *14*, 177–210.

Ogden, R. A., & Local, J. K. (1994). Disentangling autosegments from prosodies: A note on the misrepresentation of a research tradition in phonology. *Journal of Linguistics*, *30*, 477–498.

Palmer, F.R. (1968) (Ed.). *Selected papers of J.R. Firth 1952–59*. London/Harlow: Longmans.

Palmer, F.R. (1970) (Ed.). *Prosodic analysis*. London: Oxford University Press.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology VII (phonology and phonetics)* (pp. 101–140). Berlin: Mouton de Gruyter.

Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probability theory in linguistics*. Cambridge, MA: MIT Press.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*, 285–290.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381–415). Mahwah: Lawrence Erlbaum Associates.

Port, R.F. (2003). Meter and speech. *Journal of Phonetics*, *31*, doi:10.1016/j.wocn.2003.08.001.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, *22*, 253–336.

Pulvermüller, F. (2002). *The neuroscience of language: On brain circuits of words and serial order*. Cambridge: Cambridge University Press.

Pulvermüller, F., Assadollahi, R., & Elbert, T. (2001). Neuromagnetic evidence for early semantic access in word recognition. *European Journal of Neuroscience*, *13*, 201–205.

Raizada, R., & Grossberg, S. (2003). Towards a theory of the laminar architecture of cerebral cortex: Computational clues from the visual system. *Cerebral Cortex*, *13*, 100–113.

Remez, R. E. (1994). A guide to research on the perception of speech. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 145–172). London: Academic Press.

Remez. R.E. (2003). Establishing and maintaining perceptual coherence: Unimodal and multimodal evidence. *Journal of Phonetics*, *31*, doi:10.1016/S0095-4470(03)00042-1.

Remez, R. E., & Rubin, P. E. (1992). Acoustic shards, perceptual glue. *Haskins Laboratories Status Report on Speech Research SR-111/112*, 1–10.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, *101*, 129–156.

Scott, S., & Wise, R. (2003). PET and fMRI studies of the neural basis of speech perception. *Speech Communication*, *41*, 23–34.

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

Shamma, S. (2003). Physiological foundations of temporal integration in the perception of speech. *Journal of Phonetics*, *31*, doi:10.1016/j.wocn.2003.09.001.

Sheffert, S. M. (1998). Contributions of surface and conceptual information to recognition memory. *Perception and Psychophysics*, *60*, 1141–1152.

Smith, R. (2002). Does memory for individual talkers help word segmentation? In S. Hawkins & N. Nguyen (Eds.), *Temporal integration in the perception of speech*. Proceedings of the ISCA TIPS workshop (p. 49). Available from: http://www.lpl.univaix.fr/~tips/.

Smith, R. (in preparation). The role of fine phonetic detail in word segmentation. Unpublished Ph.D. dissertation. Cambridge: University of Cambridge.

Smith, R., & Hawkins, S. (2000). Allophonic influences on word-spotting experiments. In A. Cutler, J. M. McQueen, & R. Zondervan (Eds.), *Proceedings of the workshop on spoken word access processes (SWAP)* (pp. 139–142). Nijmegen: Max-Planck Institute for Psycholinguistics.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA/London: MIT Press.

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

Tuller, B. (2003). Computational models in speech perception. *Journal of Phonetics*, *31*, doi:10.1016/S0095-4470(03)00018-4.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, & W. Donaldson (Eds.), *Organization of Memory* (pp. 381–403). New York: Academic Press.

Tunley, A. (1999). *Coarticulatory influences of liquids on vowels in English*. Unpublished Ph.D. dissertation, University of Cambridge.

Viemeister, N. F., & Wakefield, G. H. (1991). Temporal integration and multiple looks. *Journal of the Acoustical Society of America*, *90*, 858–865.

Warren, P., & Marslen-Wilson, W. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception and Psychophysics*, *41*, 262–275.

Warrington, E. K., & McCarthy, R. A. (1987). Categories of knowledge: Further fractionations and an attempted integration. *Brain*, *110*, 1273–1296.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*, 829–854.

West, P. (1999). The extent of coarticulation of English liquids: An acoustic and articulatory study. *Proceedings of the International Conference of Phonetic Sciences* (pp. 1901–4). San Francisco.

West, P. (2000). Perception of distributed coarticulatory properties of English /l/ and /ɹ/. *Journal of Phonetics*, *27*, 405–425.

Whittlesea, B. W. A. (2003). On the construction of behavior and subjective experience: The production and evaluation of performance. In J. S. Bowers, & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 239–260). Oxford: Oxford University Press.

Wise, R. J. S., Scott, S., Blank, C., Mummery, C. J., Murphy, K., & Warburton, E. A. (2001). Separate neural subsystems within 'Wernicke's area'. *Brain*, *124*, 83–95.

Wrigley, S. N., & Brown, G. J. (1999). Synfire chains as a neural mechanism for auditory grouping. http://www.dcs.shef.ac.uk/~guy/pdf/bsa99.pdf.

Zue, V. (1985). The use of speech knowledge in automatic speech recognition. *Proceedings of the Institute of Electrical and Electronic Engineers*, *73*, 1602–1615.