Proceedings of the 15th International Congress of Phonetic Sciences (2003) 293-296.

Contribution of fine phonetic detail to speech understanding

Sarah Hawkins
Department of Linguistics, University of Cambridge, U.K.
E-mail: sh110@cam.ac.uk

ABSTRACT

This paper first summarizes some phonetic observations that challenge theories of speech perception and speech understanding which assume the speech signal is converted to an abstract form, such as features or phonemes, before it can be understood, and that 'knowledge-driven' processes are largely 'top down' rather than signal-driven. Systematic variation in fine phonetic detail is shown to indicate linguistic structure of all types. Second, it outlines a model of speech understanding, Polysp, in which the speech signal is multi-purpose, with the same properties feeding different meaning systems, and in which processes of speech understanding are polysystemic, with the speech signal mapped directly to meaning whenever possible. The units of formal linguistic analysis are seen as self-organizing, emergent properties of such a system, rather than essential steps on the way to understanding, and they need not be completely systematic in any one person's mind.

1. SYSTEMATIC VARIATION IN PHONETIC DETAIL SIGNALS LINGUISTIC STRUCTURE

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 form before it can be understood, but instead, it may be directly mapped onto meaning itself.

Consider the following example. There are many ways to say that you do not know something; the form a speaker chooses depends on what 'extra' information he or she wishes to convey. Spoken with a neutral intonation and 'ordinary' voice quality, tempo, and rhythm, the sentence *I don't know* typically conveys little more 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 phonetically by the particular segmental realisation (choice of 'word forms') and a wide range of other properties. The expanded form, *I do not know,* is often accompanied by some unusual voice quality, tempo and rhythm, and typically implies emphasis with some negative attitude such as impatience. The person who says *Dunno* tells us that he or she is, or has been, content not to know, or is indifferent to the listener's wish for information; *dunno* can only be used in informal situations, or to convey insolence. 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 fine 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 it to have the intended meaning.

Likewise, within the linguistic system itself, much acoustic-phonetic fine detail systematically signals linguistic distinctions, but has been seen as random variation because the emphasis in phonetics and speech perception is on lexical distinctiveness—and hence on phonemic contrast—rather than on the susbsytems that make up a whole linguistic system. For example, whereas many English function words begin with /ð/ (*this, that, then, these…*), no content words do; /ð/ can only occur in the middle or at the end of a small set of content words (*other, bother, rather, gather, wither, feather, heathen, breathe, writhe, soothe (soothing soothes soothed…)…*).

Connected speech processes also differ between content and function words, in systematic ways that could facilitate perception [1,2]. So, for example, when /n/ or /l/ occurs before /ð/ in English, as in *ban these, in that, still the, all that,* the sequence is often produced with long dental nasals or long dental laterals, and there is no audible friction. Such pronunciations are never used for similar sequences in which the fricative is voiceless, and thus begins a content word rather than a function word e.g. *ban thought, in thatch, still thorough, all thanks.* The acoustic realisation of these sequences with function words is complex, and listeners are attuned to them; presumably, they provide consistent acoustic-phonetic (i.e. low-level) evidence for the existence of a word boundary followed by a function word. Local [3] provides other examples with the phoneme /m/, which differs in realisation depending on whether it occurs as part of the pronoun+*be* system, viz. *I'm*, or in content words such as *lime, time, lamb, sham.* Within the same speech register (degree of formality), /m/ varies much more in *I'm* than at the end of content words. Indeed, as

long as the syllable contains a low vowel and nasalisation, or [m] alone, almost anything seems acceptable for *I'm*. This is not at all the case for content words with final /m/. The suggestion is that, whereas phonetic/phonological contrasts with content words are potentially unlimited, *I'm* need only contrast phonetically with other grammatical contenders whose phonetic forms are very different: *you're, he's, we're* and so on. So, at any given tempo and degree of formality, *I'm* allows commensurately more variation; and the variation is itself informative, because it is lawful within the larger linguistic system.

The above examples distinguish *grammatical functions* and as far as is known seem to affect, for the most part, local parts of the signal. However, some cues to *phonemic distinctions* operate over much longer time scales, and these pose a different type of challenge to theories of speech perception. I offer two examples. First, spectral and durational cues to coda voicing extend not only into the preceding vowel [4,5] but also into the onset of the syllable [6,7], and they are perceptually salient [8,9]. Second, properties of some segments, such as English /r/, can spread over several syllables [10,11,12]. These 'resonance effects' [13] are subtle, but they are perceptually salient and, when included in synthetic speech, they significantly increase its intelligibility in noise [10,11,14].

2. A MODEL OF SPEECH UNDERSTANDING THAT USES PHONETIC DETAIL

I suggest that systematic variation in the speech signal of the types described above performs two functions. First, it provides perceptual coherence, so that the signal sounds as if it comes from a single talker, using a consistent speech style, or register, and forming a single perceptual 'stream', to use that term loosely. Some aspects of perceptual coherence result from the way the vocal tract works; others are language-specific: the distinction is immaterial to an individual speaker/listener. Second, the systematically varying speech signal offers information about all levels of formal linguistic analysis—pragmatics, grammar, meaning, phonology—simultaneously and in richly complex ways, rather than just providing cues to word identity, with prosody a partly independent 'tack-on'.

Traditionally, models of how we understand speech neglect this richness in the physical signal, assuming that the input to the lexical decision and identification process is already abstract (for example phonological features or phonemes), and distinguishing levels of mental process that parallel levels of formal linguistic analysis (for example, phonemes before syllables before words before grammar). In contrast with these approaches, I propose that listeners need to retain the detailed acoustic information until an utterance has been understood: this acoustic fine detail, combined with the listener's understanding of general environmental conditions, allows simultaneous multiple access to many linguistic levels, thus letting listeners rapidly build up a picture of the full meaning of the utterance to be understood, so that they can interact successfully with the speaker. It is this type of understanding that listeners aim for, rather than a complete linguistic description of a given utterance. The conclusion from these arguments is that speech patterns are stored in the brain as multimodal sensory memories, and form part of a complex network that represents an individual's knowledge, linguistic and non-linguistic: initial linguistic representation is exemplar-based and may be only partly formed.

I outline a model, Polysp (for POLYsystemic SPeech understanding), that uses rich, polysystemic linguistic structures like those of Firthian phonology to represent the type of structures proposed. Polysp involves exemplar memory of speech events, organised within a dynamic system in which phonetic categories behave like other cognitive and linguistic categories: they are emergent, context-sensitive, dynamic, and plastic throughout life. Recognition and understanding may arise via development of 'resonance' as in Adaptive Resonance Theory [15,16], though including more fine-grained detail than most of the ART models in the literature.

Here in outline is an example of how it might work. Figure 1 shows spectrograms and associated syllabic structures of the words *mistimes* and *mistakes,* spoken by the author (British English woman) in *I'd 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 the 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, which is heard as a rhythmic difference: the first syllable of *mistimes* has a heavier beat than that of *mistakes*.

These acoustic-phonetic differences in the words' first four phonemes arise because their morphological structure differs. The word *mistimes* contains the morphemes *mis+time*, which each have a separate meaning; and the meaning of *mistimes* is straightforwardly related to the meaning of each of the two morphemes. But the meaning of *mistakes* is not obviously related to the meaning of its constituent morphemes, and the word is best regarded as monomorphemic. This difference in the productivity of *mis-* is reflected phonologically in the syllable structure, shown on the right of Fig. 1. When *mis-* is non-productive, the /st/ cluster is ambisyllabic (*mistakes*), but the morpheme boundary in *mistimes* prevents /st/ from being ambisyllabic. Hence, in *mistimes*, /s/ is the coda of syllable 1, and /t/ is the onset of syllable 2. In contrast, the /s/ and /t/ in *mistakes* form both the coda of syllable 1 and the onset of syllable 2. In an onset /st/, the /t/ is always relatively unaspirated in English (cf. *step, stop, start*). The durational differences in /m/ and /ɪ/ arise because the morpho-logically-conditioned differences in syllable structure result in *mist* being a phonologically heavy syllable whereas *mis* is phonologically light, while both syllables are metrically

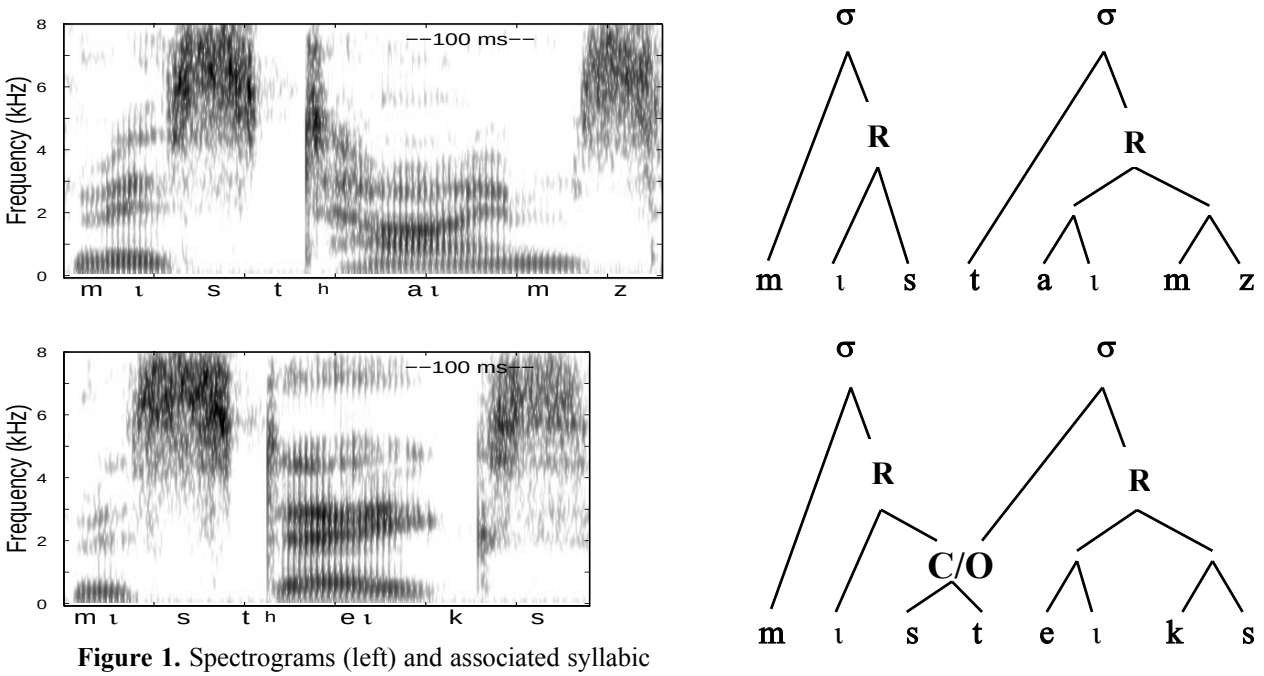

Figure 1. Spectrograms (left) and associated syllabic structures (right) of the words *mistimes* and *mistakes*.

weak. Thus the morphological differences between the words are reflected in structural phonological differences; and these in turn have implications for the phonetic detail of the utterances, despite the segmental similarities between the words. Complex though these influences are, listeners seem to keep track of them and use them, in that, when the various properties have the wrong relationships to one another, they are heard as unnatural, and would presumably be harder to understand.

Consider now how this and other information might be obtained from the signal. The listener first hears [m], which should activate [nasal] and possibly [labial] features, as well as 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, though [high front] will have been expected from coarticulatory evidence. When the aperiodicity of the [s] is heard, both syllable coda and syllable onset could be activated. Silence for the [t] closure will likewise activate features for stop and most likely alveolar. Simultaneously, the relative durations of the first three segments will begin to favour 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] duration is relatively long, then *mistakes* will rise in probability. 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 other words, before the [t] burst is reached, there will already be good evidence for whether or not the [st] cluster is ambisyllabic, and hence for the morphemic structure of the word. In consequence, the number of competing words will be reduced. 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.

Most of the above description has been offered as if the word was heard in isolation, but 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. For example, there would be little doubt about the rate of speech, so decisions about the relative durations of [mɪ] and [s] could be taken earlier.

This partial description of on-line identification of morpho-phonological and segmental structure focuses on segments partly for ease of description, and partly because there is no question that certain sounds function as anchor points, or islands of reliability, against which the quality of others can be evaluated. Other aspects of the signal provide reliable evidence for different style of structure. In these examples, [s], and the manner of articulation of the stops and nasals, taken with the vowels as syllabic nuclei, seem good candidates to provide a reliable skeletal structure. Their combined attributes contribute prosodic information. This skeleton, combined with predictions of grammatical class (such as the probability that a verb will follow the name *Tess*), could make spoken word identification extremely fast without confirming the exact identity of each sound segment.

In other words, identification takes place in a probabilistic way, using all possible sensory information to flesh out linguistic structure at all possible levels in parallel. The spectro-temporal relationships between different parts of

the signal are all-important in the identification of structure: each element is identified relative to others in the environment and relative to the listener's expectations from past experience. These expectations may be derived from exemplar memory for speech and non-speech aspects of communication, with linguistic structures developing throughout childhood, and probably beyond, as emergent categories from these sensory events. The listener's aim is to arrive at meaning, not at a complete linguistic description, so he or she will jump to the most probable meaning as soon as possible, as long as the evidence overall is a good enough match to the expected sound pattern. 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.

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 speech 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 main route between sensation and meaning. These ideas are developed in [17,18].

REFERENCES

[1] J. Kelly and J.K. Local, *Doing Phonology,* Manchester: University Press, 1989.

[2] S. Manuel, "Speakers nasalise /ð/ after /n/, but still hear /ð/," *Journal of Phonetics,* vol. 23, pp. 453-476, 1995.

[3] J. Local, "Variable domains and variable relevance: Interpreting phonetic exponents," *Journal of Phonetics,* vol. 31 (3/4), 2003, (in press).

[4] W.V. Summers, "F1 structure provides information for final-consonant voicing," *Journal of the Acoustical Society of America,* vol. 84, pp. 485-492, 1988.

[5] C.G. Wolf, "Voicing cues in English final stops," *Journal of Phonetics,* vol. 6, pp. 299-309, 1978.

[6] J.P.H. van Santen, J.S. Coleman and M.A. Randolph, "Effects of postvocalic voicing on the time course of vowels and diphthongs," *Journal of the Acoustical Society of America,* vol. 92(4/2), p. 2444, 1992.

[7] S. Hawkins and N. Nguyen, "Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English," *Journal of Phonetics,* vol. 32, 2004 (in press).

[8] S. Hawkins and N. Nguyen, "Perception of coda voicing from properties of the onset and nucleus of *led* and *let*," in *Proceedings of the 7th International Conference on Speech Communication and Technology,* P. Dalsgaard, B. Lindberg, H. Benner Eds., vol. 1, pp. 407-410, 2001.

[9] S. Hawkins and N. Nguyen, "Effects on word recognition of syllable-onset cues to syllable-coda voicing," in *Papers in Laboratory Phonology VI,* J.K. Local, R.A. Ogden, R.A.M. Temple, Eds., pp. 38-57. Cambridge University Press, 2003.

[10] P. West, "Perception of distributed coarticulatory properties of English /l/ and /ɹ/," *Journal of Phonetics,* vol. 27, pp. 405-426, 2000.

[11] A. Tunley, *Coarticulatory influences of liquids on vowels in English,* PhD diss., U. Cambridge, 1999.

[12] S. Heid and S. Hawkins, "An acoustical study of long-domain /r/ and /l/ coarticulation," in *Proceedings of the 5th Seminar on Speech Production: Models and Data,* pp.77-80. Kloster Seeon, Bavaria, Germany, 2000.

[13] J. Kelly and J. Local, "Long-domain resonance patterns in English," in *International Conference on Speech Input/Output: Techniques and Applications*. Conference publication No. 258, pp.304-309. London, Institutae of Electrical Engineers, 1986.

[14] S. Hawkins and A. Slater, "Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech," in *Proceedings of the 3rd International Conference on Spoken Language Processing,* vol. 1, pp. 57-60, 1994.

[15] S. Grossberg and C.W. Myers, "The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects," *Psychological Review,* vol. 107, pp.735-767,2000.

[16] S. Grossberg, "Resonant neural dynamics of speech perception," *Journal of Phonetics,* vol. 31 (3/4), 2003, (in press).

[17] S. Hawkins and R. Smith, "Polysp: A polysystemic, phonetically-rich approach to speech understanding," *Italian Journal of Linguistics—Rivista di Linguistica*, vol. 13, pp. 99-188, 2001.

[18] S. Hawkins, "Roles and representations of systematic phonetic fine detail in speech understanding," *Journal of Phonetics*, vol. 31 (3/4), 2003 (in press).