



Roles and representations of systematic phonetic fine-detail in understanding speech

Sarah Hawkins

Dept. of Linguistics, University of Cambridge, UK
sh110@cam.ac.uk

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 directly mappable onto meaning itself. 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, acoustic-phonetic fine detail that is frequently overlooked or seen as random variation in fact systematically signals many linguistic distinctions. 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 (e.g. *other, mother, bother, rather, father, gather, wither, breathe, writhe, soothe...*). Connected speech processes also differ between content and function words, in systematic ways that could facilitate perception.

Moreover, properties of some segments, such as English /r/, can spread over long stretches of speech, up to several syllables; these 'resonance effects' are relatively subtle, but they are perceptually salient and, when included in synthetic speech, they significantly increase its intelligibility in noise.

I suggest that such systematic variation in the speech signal performs two functions. First, it provides perceptual coherence, so that the signal sounds as if it comes from a single talker, 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 and grammar

as well as meaning and 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 disregard 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 the 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 episodic and may be only partly formed.

I outline the beginnings of 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 episodic 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. 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 main route between sensation and meaning.

To assess Polysp’s worth, two issues need early attention. One is to identify a plausible neural basis for the proposed processes. In this paper, I explore the compatibility of Polysp’s linguistic structures with Hebbian cell assemblies and synfire chains (cf. Pulvermüller, 1999). The other issue is how to constrain and test Polysp. Experiments are in progress to test the contribution of episodic memory to specifically linguistic behaviour. For example, Rachel Smith (this volume) has recently shown that inconsistent allophone usage disrupts word spotting more when the speaker is familiar rather than unfamiliar to the listener. Ideally, Polysp should be implemented computationally. Especially relevant are the issues of how to integrate long- with short-domain information about segmental identity; and how to use the phonetic fine detail in the signal, since most computational models effectively use abstract categories earlier than Polysp assumes they should. I hope that one outcome of TIPS is that we will make progress in addressing these issues.

Hawkins, S., and Smith, R. (2001) Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Journal of Italian Linguistics—Rivista di Linguistica* 13, 99-188.
Pulvermüller, Friedemann (1999) Words in the brain’s language. *Behavioral and Brain Sciences* 22, 253-336.