Sādhanā Vol. 36, Part 5, October 2011, pp. 555–586. © Indian Academy of Sciences

Phonetic perspectives on modelling information in the speech signal

S HAWKINS*

Centre for Music and Science, Faculty of Music, University of Cambridge, Cambridge, CB3 9DP, UK
e-mail: sh110@cam.ac.uk

Abstract. This paper reassesses conventional assumptions about the informativeness of the acoustic speech signal, and shows how recent research on systematic variability in the acoustic signal is consistent with an alternative linguistic model that is more biologically plausible and compatible with recent advances in modelling embodied visual perception and action. Standard assumptions about the information available from the speech signal, especially strengths and limitations of phonological features and phonemes, are reviewed, and compared with an alternative approach based on Firthian prosodic analysis (FPA). FPA places more emphasis than standard models on the linguistic and interactional function of an utterance, de-emphasizes the need to identify phonemes, and uses formalisms that force us to recognize that every perceptual decision is context- and task-dependent. Examples of perceptually-significant phonetic detail that is neglected by standard models are discussed. Similarities between the theoretical approach recommended and current work on perception–action robots are explored.

Keywords. Phonetic detail; Firthian prosodic analysis; Polysp; phoneme; linguistic function; interaction.

1. Introduction

This paper reviews some common assumptions about the information available from a spoken signal, especially strengths and limitations of analyses that focus mainly on phonological features and phonemes. These are widely regarded as essential units of human speech perception and production, and, by extension, as central to speech recognition and synthesis systems. However, traditional models of speech perception, which attempt to map cues derived from the speech signal onto discrete, abstract phonological units, cannot fully account for listeners' abilities. This is not surprising if one considers that features and phonemes were developed, for quite different purposes, as abstract theoretical constructs within particular schools of theoretical linguistics. An alternative theoretical approach, the Firthian prosodic analysis (FPS), places more emphasis than standard models on the function of an utterance, de-emphasizes the need to identify phonemes, and uses formalisms that force us to recognize that every perceptual decision

*For correspondence

is context- and task-dependent. This leads to more complex linguistic structures than the ones in standard models, but they may better match current thinking in speech technology, as well as being more biologically plausible. Adopting this approach encourages re-evaluation of the information available from the speech signal.

Examples of variation in phonetic detail which systematically signals non-phonemic linguistic information such as the grammatical or morphological status of a stretch of sound are given. Other examples indicate the discourse function of the utterance. Some of these systematic differences in phonetic detail are relatively localized in the speech signal, while others stretch over several syllables. Both types can make speech easier to understand, presumably by increasing the signal's perceptual coherence and/or by directly indicating linguistic structure and pragmatic functions as well as phonological form. Although such phonetic detail influences phonological structure, it is not predictable from the phonological abstractions typical of current computational models and standard phonological theory.

We need models that simultaneously accommodate detail and abstraction, and use parallel streams of knowledge to drive interpretation of information in the incoming signal so that processing is maximally efficient. Such models may be achievable by combining linked hierarchical prosodic and grammatical linguistic structures with recent advances in robotics which use function-oriented, body-centred, knowledge-driven systems to perform visual tasks. This approach rejects the assumption that speech processing proceeds in strictly serial order, e.g., from the lowest phonological units to higher ones. Instead, sound can be mapped onto any level of linguistic structure, in parallel or in sequence, using context-dependent probabilistic processes. The important point is that there is no predetermined or rigid sequence: the process of understanding speech is governed by the properties of the particular signal in conjunction with the listener's construal of the particular situation, that is, with task demands. These could be modelled for specific communicative functions as long as pragmatic and phonetic knowledge are combined.

1.1 *Common Western assumptions about the acoustic speech signal*

Most speech research is characterized by the assumption that, to make sense of the acoustic speech signal, the first task of both human listeners and machines is to identify words; when words are identified, then grammar and meaning can be worked out in a relatively straightforward manner. This assumption entails that the first focus of a recognizer is on the phonological units that define lexical form, which, in common with most textbooks, normally gives centre stage to phonemes, with words defined as unique (or almost unique) strings of phonemes representing minimal units of independent meaning. The assumption that phonemes are the basic building blocks from which words are formed in turn entails another assumption which forms the focus of much comment in speech science textbooks, namely that the speech signal comprises 'essence' and 'variation', and that the task of the researcher, or of a machine recognizer, is to preserve the essence (phonemes or phonological features) and to discard the variation. This approach, as is well-known, results in many ambiguities in phoneme strings from which words are to be identified, e.g., /greɪtreɪn/ could represent the words *grey train, great rain,* or *great train.* Much ingenuity has been spent in researching how best to use 'higher-order knowledge' to disambiguate the resultant ambiguities in the phoneme strings.

The position argued for in this paper is that much of the information that would disambiguate the resultant phoneme strings is actually present in the physical signal, but, because it contributes little or nothing to phoneme identity, it is neglected by models that use the physical signal solely

or largely as a source of information about phonemes. The trick is to exploit this non-phonemic information in the physical signal to support and even guide the knowledge-based inferences which are sometimes called a 'language model' by speech technologists. It is argued that this is not just a sensible use of available information, but more closely reflects processes of human speech perception and is thus more likely to result in models whose performance approaches human performance across a range of tasks and listening conditions. The rest of this paper explores and develops this claim.

Figure 1a illustrates the basic assumptions in traditional psycholinguistic conceptualizations of the process of speech recognition. The sound signal is used to extract a string of phonemes (or feature bundles that closely correspond to phonemes) and 'higher-order' language knowledge is then applied to identify larger units, especially words (lexical items) and grammatical categories. These analyses are taken to be fundamental to assigning meaning to a spoken signal. In modelling human behaviour, a strong psycholinguistic focus has been on whether signal and knowledge operate autonomously or interactively. In autonomous models (e.g., Norris *et al* 2000), information flow is solely 'bottom-up' from the signal to higher-order levels, as illustrated by the arrows at the left of the labelled levels in figure 1a. In interactive models, illustrated by the arrows at the right of figure 1a (e.g., McClelland *et al* 2006), information flows in both directions, so probability weights of units at lower levels are directly influenced by probabilities of categories at higher levels.

The background theoretical assumptions made in speech technology are fairly similar to those illustrated in figure 1a, and most technology applications of course maximally exploit the top-down contribution. This is also the position adopted in this paper, in one sense in quite an extreme form: it is argued that the physical signal is never interpretable in the absence of an overarching structure or system of priors. This argument comes partly from the nature of the physical speech signal, partly from evidence for the so-called corticofugal system of rich innervation from the cerebral cortex to lower levels of the brain (see summaries in section 4.3 and in Hawkins (2010b)), and partly from the properties of successful perception–action robotics models as outlined in section 5. In this sense, the top-down vs bottom-up debate of cognitive psychology is irrelevant to the present approach. However, to develop systems that can handle unrestricted topics, multiple talkers, and spoken dialogue, it is worth examining the history of speech research and some linguistic-phonetic theory, to highlight the strengths and weaknesses of standard approaches, and to assess the value of recent linguistic-phonetic advances in research on speech. These recent advances identify acoustic-phonetic cues to grammatical and pragmatic function. They help to shift theoretical focus from identification of phonological and lexical form alone to the identification of form that indicates communicative function. Thus they promise to expand the use of signal-driven top-down knowledge to understand interaction between individuals.

No single psycholinguistic model conforms to that shown in figure 1a, partly because no single model of speech recognition deals with all represented levels, but also because those that address the same issues differ in detail. Typically, models deal either with phoneme or word recognition, or else with grammar. This restriction allows investigation of only a manageable number of parameters. However, such restricted focus has also encouraged the type of model illustrated, because it allows researchers to neglect signal properties and recognition processes that are difficult to account for or that simply fall outside the scope of theoretical interest. For example, most models which address spoken word recognition focus on sequential relationships between successive units of one or two types, typically phonemes and words, and on a few influences on those levels, such as word frequency, morphological structure, and transitional probabilities between phonemes. These few parameters produce simple models which allow

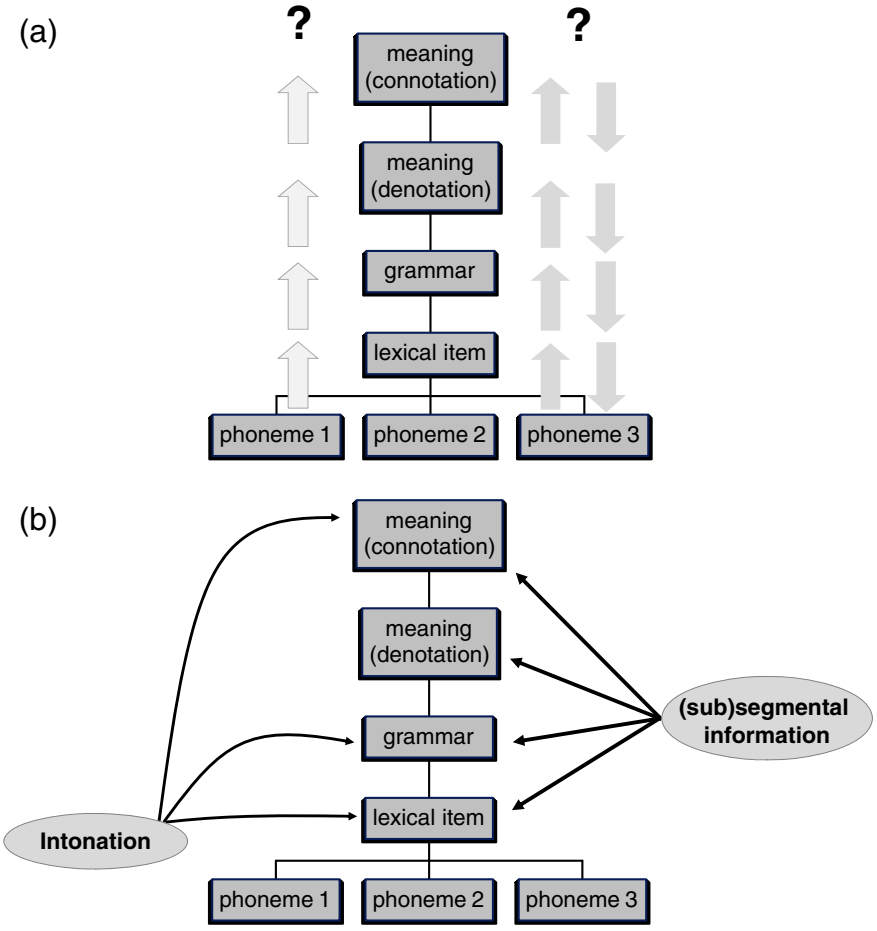


Figure 1. Illustration of the type of standard theoretical assumptions under discussion. Phonemes are identified from the acoustic signal (not shown) and what is known about how they combine, e.g., phonotactics, and transitional probabilities. From these, syllables (not shown), lexical items, grammatical relations, and aspects of meaning are successively identified. (**a**) gives the basics, with the emphasis being on information flow between 'levels', whether one-way or bidirectional, as indicated by the thick arrows at left and right sides respectively. (**b**) Intonation has always been known to affect many different levels, of which three are shown here. Recent research confirms that segmental and subsegmental phonetic information likewise informs about most or all of these levels, much of it in relation to attributes of the preceding signal(s), whether spoken by the same person or another talker.

clear hypotheses to be tested, giving broadly generalizable results, but at the cost of limited relevance to any given speaking situation.

Although all speech researchers are aware of this, there is less awareness that a significant cause of the lack of generality is the theoretical framework itself, with its strong emphasis on the abstract phonological form of an utterance rather than its function.

The most well-known neglected aspect is that intonation affects all levels in the hierarchy (as indicated in figure 1b). Intonation, and indeed many other aspects of prosody, are treated as

separate and quasi-independent from segmental identity, even though it is obvious that they are in fact strongly mutually dependent: segments need the right durations (Klatt 1976) and (to a lesser extent) pitch properties, and intonation not only has to be made with segments normally thought of as phones or phonemes, but there are tight, language-specific temporal dependencies between syllable structure and maxima and minima in the pitch contour (Post *et al* 2007).

Much less well known is that much phonetic information indicates, not phonemic identity, but higher-level linguistic units and hence sentence structure. Well known, but for the most part not yet comprehensively described, is that pragmatic function influences people's responses to another person's speech. Thus, the particular pronunciation of a phone (i.e., a sound that can be written as a single phoneme) or sequence of phones can provide valuable information about a wide variety of linguistic units, including morphological structure, informational focus, and interactional function. Although phonemically distinct aspects of these units are represented in standard word-recognition models—for example, those that reflect morphological structure, as in the English past tense /t d ɪd/ (*looked, sobbed, carted)* and plural /s z ɪz/ (*cats, dogs, horses*)—recent research confirms that many of these linguistic distinctions give rise to systematic phonetic yet non-phonemic differences that are perceptually salient. Some are salient in good listening conditions, while others are probably most influential in adverse conditions (e.g., Hawkins & Smith 2001; Hawkins 2003; Local 2003).

It is important to recognize that the word *systematic* is crucial. We are not talking about random variation, but about acoustic-phonetic contrasts that reliably distinguish crucial aspects of the spoken message. Such systematically-varying phonetic detail has in the past been called random, and hence ignored. However, that is because the model that guided research was too impoverished to be able to relate the detail to the relevant system. It is only recently that this systematic variation has been paid much attention by speech researchers. This paper summarizes some of the evidence. Section 1.2 gives more detail about phonemes and their limitations; the section is self-contained and can be skipped by readers for whom it holds no interest. However, it may be worth considering the message in figure 2.

1.2 *Phonemes as abstract and only partially adequate theoretical constructs*

Phonemic systems were designed to provide a simple and parsimonious system to describe the system of sound contrasts that differentiate words in a language or dialect. Thus a phoneme is a unit that distinguishes a particular word form from other similar ones, and hence that potentially distinguishes words that have different meanings. However, phonemes are abstractions away from the sound itself: you have to know the language or dialect in order to know how to pronounce the sounds themselves. This can be difficult to understand partly because there is obviously a relationship between actual speech sounds (the phonetic system) and the abstract system of sound contrasts (the phonological system), but it is not a simple relationship, and the details are a matter of theoretical definition and hence of debate within linguistics. One big advantage of phonemes is that they allow efficient communication between researchers about language and sound patterns—but only if the theoretical framework is well understood.

In practice, any given phoneme is normally described in phonetic terms, usually in the way the phoneme is produced in a stressed consonant-vowel (CV) or CVC syllable spoken carefully in isolation, such as a voiced bilabial stop, /b/, or a voiceless alveolar fricative, /s/. Further, the symbol used to represent the phoneme is as close as possible to the simplest form of the appropriate phonetic symbol in one or more Western European writing systems. However, as every speech researcher knows, pronunciation in other contexts can be very different from these so-called

canonical contexts, and as noted above, pronunciation even in canonical contexts can differ radically between dialects of the same language, between languages, and of course between different styles of speech.

These practices are reasonable for the purposes for which linguists have used phonemes, but they have given rise to some misconceptions. One such misconception is that the initial consonantal sounds of stressed monosyllables are the 'best' or most representative of a dialect's consonantal phonemes; they are not best in all contexts. Another is the dogma that phonemes, which are phonological concepts, and 'phones', which are phonetic and physical, and hence measurable, are largely isomorphic. This dogma of isomorphism between phonetics and phonology has increased in strength since the second world war, with the rise in acoustic-phonetic analysis and applications outside of linguistics. Finally, this dogma gives rise to the implicit or explicit assumption that sounds have an 'essence' and irrelevant 'variation', as described above.

Phonemic analysis thus provides a system of contrastive units of abstract phonology, devised to allow maximal information transmission using a minimal symbolic system. Although this appears to be ideal for engineering applications, it is not straightforward to apply in practice. First, phonemes are abstractions that do not actually exist, and whose relationship to the physical signal is a matter of agreed convention. Second, there is no agreed or perfect definition of the term phoneme. As a consequence, no language can be perfectly described in terms of phonemes; there are always unsatisfactory 'loose ends'. Third, phonemic analysis is not suitable for some languages, including Mandarin Chinese. Each of these points is illustrated briefly in the following sections; more discussion can be found in phonetics textbooks, e.g., (Jones 1967; Cruttenden 2001:42 (section 5.3); Lodge 2009:11).

1.2a *Phonemes are not clearly identifiable in movement or in the acoustic speech signal*: As noted, phonemes do not actually exist, but are abstractions—away from the signal—and are often not straightforwardly imposed on the signal. Most elementary speech science textbooks discuss the basic fact that a sequence of acoustic segments cannot be mapped to a sequence of phonemic segments in a one-to-one way. However, especially since the rise of acoustic-phonetic analysis, the viewpoint that there is a simple isomorphism between phonological and phonetic units of analysis has persisted. ('Phonological' means the sound system for expressing differences in word form that can differentiate meaning; 'phonetic' means the actual sounds that we use to express concepts, including word meaning, and to perform interactive functions.)

1.2b *Multiple and imperfect definitions of the term phoneme*: Less well known is that no definition of the term phoneme is perfect, and there is no perfect way to establish the phonemic system of a language or dialect. There are many definitions, operationalized as diagnostic criteria, but the ultimate criterion is the intuition of native speakers, with a tendency to emphasize similarity in place and manner of oral articulation over similarity in glottal/laryngeal behaviour.

For example, one good objective criterion of whether two sounds are *allophones* of the same phoneme is whether they are in complementary distribution, which means that each appears in distinct, mutually exclusive contexts, such as the beginning vs the end of a syllable. Thus, in most varieties of English, the phoneme /l/ is spoken as two rather different sounds, clear [l] in syllable onsets, and dark [ɬ] in syllable codas. (A coda is the consonants at the end of a syllable. A transcription of phonemes is indicated by slashes e.g. /l/, /pat/; a transcription of the sound, or phone, itself is indicated by square brackets, e.g., [l] or [t^hap], [t^haʔp] (see below for more explanation). As much or as little detail as is required to make the point at issue is used within

square brackets; phonemes never include all the detail.) These two 'allophones' of /l/ are fairly similar; but many other allophones of a single phoneme are much more dissimilar. For example, as discussed in section 1.2d, English voiceless stops /p t k/ are always aspirated in syllable onsets, thus [p^h], [t^h], [k^h], unless they occur after syllable-initial /s/; whereas in syllable codas, if voiceless stops are aspirated then the aspiration is much weaker than in onsets, and there are many other ways that coda stops can be pronounced too, none of which are possible in syllable onsets.

The principle is not how similar or dissimilar the sounds are, but whether they can distinguish words that are otherwise identical. For example, we have seen that [p^h t^h k^h p t k] are all allophones of English /p t k/ respectively; but in Hindi, Urdu, and many other languages of the Indian subcontinent, aspirated and unaspirated voiceless stops are separate phonemes, /p^h t^h k^h p t k/ because each distinguishes words. As another example, dental or alveolar (tongue-tip) nasal [n] and velar (tongue back against roof of mouth) nasal [ŋ] are two phonemes in English, /n/ as in *ban* and /ŋ/ as in *bang*; but in Hindi they are allophones of the same phoneme /n/, because [ŋ] only occurs before velar stops, and [n] never occurs before velar stops (e.g., Jones 1967). For the same reason, the retroflex nasal [ɳ] is a phoneme in its own right in Punjabi and Marathi, but is an allophone of /n/ in several other Indian subcontinental languages, including Bengali, Hindi, colloquial Sinhala, and Urdu: in these latter languages, it only occurs before retroflex consonants (and in a few loanwords, which are discounted for this purpose).

According to Jones (1967), these issues were not discussed by ancient Indian grammarians, but must have been understood, because, for example, in Tamil the phoneme /k/ has many distinct allophones in complementary distribution (i.e. they occur in different contexts), yet only one symbol is used for all of them in the orthography. It should be noted that these examples underline the context-free nature of phonemes: the context determines the particular (allo)phone, and the phoneme neglects context. To continue our previous examples with stops, in almost all English accents of the British Isles, USA, South Africa and the Antipodes, aspirated [p^h] in *peak* and unaspirated [p] in *speak* are allophones of the phoneme /p/; but if the acoustic [s] is removed from the waveform of a token of *speak*, native English listeners hear *beak* i.e., a different word, and hence a different phoneme, /b/.

All these distinctions rest on the sound *system*: there is no need for the sound space to be divided as it is in English or in any other language. Thus, for speakers of most languages of the Indian subcontinent, the phones [b], [p] and [p^h] are three distinct phonemes /b/, /p/ and /p^h/, i.e., all three can distinguish words in the same syllabic positions. However, for English speakers, these sounds comprise only two phonemes, written /b/ and /p/; the phone [p] does occur, but it is only heard as a /p/ after syllable-onset /s/; it is heard as a /b/ in most contexts. (Several languages of the Indian subcontinent, including Hindi, Gujarati, Sindhi and Urdu, have a fourth 'murmured' or 'breathy' stop at all places of articulation—the breathiness is symbolized by the diacritic [^{ɦ}], so the murmured labial stop is written [$b^{ɦ}$], /$b^{ɦ}$/. Similar to /b/, it is voiced during the closure, but differs from /b/ in having a noticeable period of breathy phonation (breathy voice) immediately after the release transient that is due to the release of the oral articulators. For English listeners, this difference in breathy vs normal phonation does not distinguish words.)

However, all languages have some sounds that pose classification problems. English has two sounds, the glottal fricative [h] and the velar nasal [ŋ], which should be classed as allophones of the same phoneme because they are in perfect complementary distribution: [h] only in syllable onsets (beginnings), and [ŋ] only in syllable codas (cf. *hang, song*). Yet these two sounds are never classified as allophones of one phoneme, mainly because native speakers do not 'feel' them to be related. The best explanation for them to be unrelated is that their articulations are too different, and possibly their sounds. Yet other sounds with very different articulations are readily

classified as allophones of other phonemes. In some northern English dialects, allophones of the voiceless alveolar stop /t/ are the voiceless glottal stop [ʔ] and the voiced alveolar approximant [ɹ], which is the standard English /r/ sound. Both these sounds are regularly used when the /t/ phoneme is in the middle of words or some phrases. An example is *I gorra leʔa* for *I got a letter.* Neither [ɹ] nor glottal stop [ʔ] can occur syllable-initially as the phoneme /t/; when [ɹ] does occur syllable-initially, it is classified as the phoneme /r/, not /t/. This criterion of native speakers' 'feelings' is not a comfortable one, but it is regularly applied. Jones (1967) discusses the case of colloquial Sinhala pre-nasalized voiced stops (he calls them compounds): the nasal is clearly present, but is very short, and rhythmically distinct from a sequence of nasal plus stop. The two types of sequence could be distinguished in terms of length, but Sinhala speakers consider the prenasalized (i.e., short nasal) variety as a single sound, and so the sequence is accorded phonemic status, without a length distinction. Similar considerations result in English /tʃ/ as in *church* having phonemic status.

1.2c *Phonemic analysis is unsuitable for some languages*: By definition, phonemes are context-free: each is independent of others. This does not mean that phonemes can be sequenced in any order—all languages constrain the possible combinations—but it does mean that there should be no inevitable dependencies between one phoneme and the next. In some languages, however, there are such strong sequential dependencies. Beijing (Mandarin) Chinese is one example. There are at least seven distinct mid-vowels in Beijing Chinese, but the particular quality used is determined by the structure of the syllable.

For example, one of them only appears in an open (CV) syllable. All other qualities occur either before or after glides, e.g., [tʲɛn] and [tʷɔ]. Although phonetically very distinct, the distribution of the mid-vowels is clearly complementary. The most convincing phonemic proposal is that there is only one mid-height vowel phoneme in Beijing Chinese (Wiese 1997). This means that the single phoneme has at least seven distinct allophones determined by the syllable structure—that is, by preceding consonants. Those seven allophones include a huge range of very different sounds, five of which can be roughly represented by the vowels in the English words *hay, head, hut, ought,* and *hope.* (The other two are sounds that do not occur in standard English.)

An alternative is to reject phonemic analysis in favour of one that reflects syllable structure, and is thus context-sensitive rather than context-free. This approach is entirely consistent with Chinese being a tone language, for tone also applies to syllables rather than segments, but differentiates word meanings. Tone languages in the Indian subcontinent include Punjabi and some of the north-eastern Tibeto–Burman languages e.g., Manipuri and Bodo. However, the problem changes, but is not solved, when syllable structure is acknowledged, as shown in section 1.2d.

1.2d *Syllable structure*: Figure 2A shows the syllable structure and phoneme symbols for the English words *pat* and *tap*. A syllable comprises a nucleus, preceded by an onset and followed by a coda, both of which are optional. (Some languages do not allow codas, but most if not all languages of the Indian subcontinent do allow them.) Thus a syllable need only contain a nucleus. The nucleus usually has relatively high amplitude and well-defined spectral prominences, or resonances. This is because it is usually (but not always) a vowel sound, and usually but not always periodic and sonorant (i.e., with well-defined resonances, or formants). By definition, sounds that precede or follow the nucleus are called consonants. Consonants at the beginning of the syllable are in the syllable onset, and those at the end of the syllable are in the syllable coda.

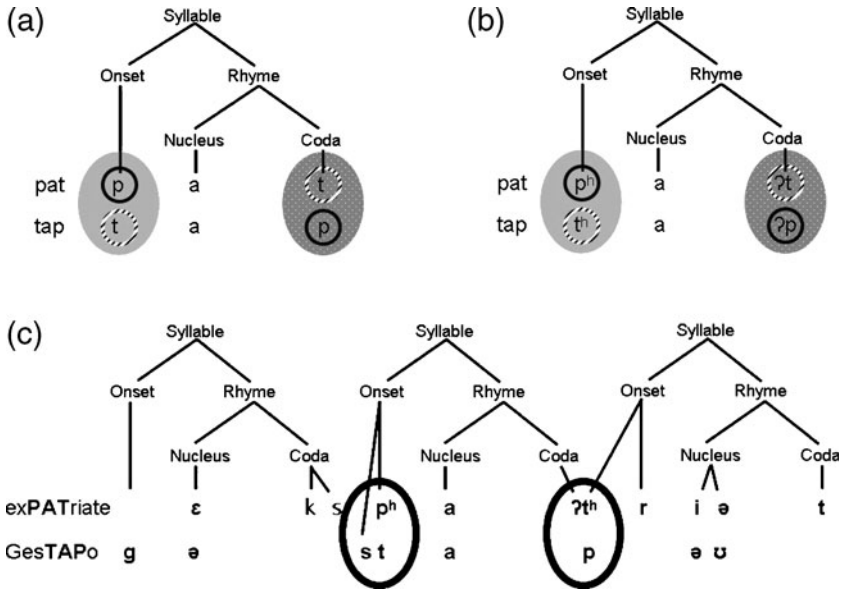


Figure 2. Syllable structures for *pat* and *tap*. A: phonemic symbols for the bilabial stop /p/ and the alveolar stop /t/. B: allophonic symbols for a clear pronunciation in standard English of these syllables when they are words spoken in isolation. C: allophonic symbols for the same syllables in longer words, *expatriate* (above) and *Gestapo* (below); to show the contrast with panel B, allophonic symbols are used for sounds in *pat* and *tap*, but the simpler phonemic symbols are used for sounds in the surrounding syllables. See text for further explanation.

Thus, as figure 2A shows, *pat* and *tap* have the same phonemes, /p/, /t/ and /a/, but in different orders. *Pat* has /p/ in the onset, and /t/ in the coda; *tap* has the opposite sequence. Figure 2B shows how their position as onset or coda consonants affects how these voiceless stops are produced. Onset /p/ and /t/ are both aspirated: superscript [^{h}] after p or t means that there is audible aspiration, or aperiodic flow of air through the partially open glottis, when the oral constriction is released. This happens in standard English whenever the stop sound is in the syllable onset (as long as it is not preceded by [s] in the same syllable). When these stops are in the syllable coda, they are not aspirated enough to warrant the use of [p^{h}] and [t^{h}]. However, they do not simply lose the aspiration. In contrast, very often, voiceless English stops have an additional closure made at the glottis, with the vocal folds held firmly together so that airflow ceases even before the oral closure is complete. The symbol [ʔ] means that the glottis is fully closed before or at the same time as the oral closure begins. This glottal closure can affect sound in a number of ways, the two main ones being disrupted periodicity, called creak (see section 3.2a), at the end of the vowel (more accurately, at the end of the preceding sonorant portion), and loss of formant transitions as the oral vocal tract closes, because when the glottis is closed there is no airflow to excite the vocal tract. An example can be seen in figure 4, top right hand panel (*let*).

Figure 2B shows that onset /p/ has as much in common with onset /t/ as it does with coda /p/. The onset stops are both voiceless and aspirated; they differ only in places of articulation (labial (lips) for /p/, alveolar (tongue tip against alveolar ridge) for /t/. Similarly, coda /p/ and /t/ differ in place of articulation, but are similar in that both are 'glottalized' before the closure and relatively unaspirated after it. This highlights the point made earlier that similarity in place

of articulation is a major determiner of whether two sounds are considered to be the same or different phonemically.

Thus, because of the way they co-occur in English, we cannot classify stop sounds by both place of articulation and the aspects of laryngeal behaviour which govern both aspiration, [^{h}], and glottal stop, [ʔ], and theorists favour the place of articulation. This point would be relatively trivial if that were the end of the story; but it is not. When these syllables occur in other contexts, their component sounds can differ systematically in yet other ways. Figure 2C shows *pat* and *tap* as syllables within the longer words *expatriate* and *Gestapo.* For simplicity, phonemic symbols are used for the segments of the other syllables in both these words, but the allophones for the consonants of *pat* and *tap* are shown; the pattern is not the same when these syllables are part of the longer words (2C) as when they form words in their own right (2B). Specifically, although /p/ in *expatriate* is aspirated, similar to /p/ in *pat*, the /t/ in *Gestapo* is not aspirated. The /t/ in the coda of *pat* in *expatriate* is first glottalized and then aspirated on release of the alveolar closure, [ʔt^{h}], while the /p/ in *Gestapo* is neither glottalized nor heavily aspirated: [p].

The unaspirated [t] arises because it follows an /s/ in the same onset. In English, voiceless onset stops are always unaspirated after /s/ in the same onset. However, the /p/ in *expatriate* also follows an /s/: why then is the /p/ aspirated? The answer lies in the morphological structures of the two words, which in turn affect their syllabic structure. In *expatriate,* /ɛks/ is a productive morpheme; its /s/ forms part of the first syllable, but when that syllable is removed from the rest, we are left with a meaningful word in its own right: *patriate* (dialects differ in pronunciation of the last two vowels). The same is not true for *Gestapo*. Although this is a German word, and *ge-* is a morpheme in German, *ge-* does not function as a morpheme in this particular word, certainly for English speakers, who are rarely aware when they first learn it that it is in fact an acronym. Thus, for *Gestapo*, the /s/ follows the normal English syllabification rules and is regarded as the first segment of the onset of the second syllable, *stap*.

In brief, differences in morphological (that is, grammatical) status affect syllable affiliation, which in turn affects the way in which component segments are pronounced. This point is developed further in section 2.2. In *pat* and *tap*, we have noted that aspects of laryngeal behaviour which govern both aspiration, [^{h}], and glottal stop, [ʔ]), are considered to be of secondary importance. Indeed, in phoneme theory as applied to English, the only laryngeal feature deemed important is whether or not the vocal folds vibrate. When they vibrate, they produce periodic excitation, and the sound is said to be ‘voiced’. However, even for this relatively simple concept, the real picture is actually very complicated. The term ‘voiced’ typically means different things in phonology and in phonetics, and the phonetic correlates of phonological voicing can affect all the segments of a syllable, not just the segment that is called voiced (Hawkins 2010b).

The points considered so far describe relationships between the actual sounds and the abstract phonemes in just two syllables, considering them first as words in isolation, and then as the second syllable of words with or without an initial productive morpheme. The conclusion is that *t* in *tap* has more in common with *p* in *pat* than it does with *t* in *pat* or *terrine:* the syllable structure indicates that *t(ap)* shares with *p(at)* all the other properties of syllable position, stress, rhyme features, etc., and only differs from it by place of articulation. It is well known that syllable structure is an inherent part of rhythmic (prosodic) structure, which is fundamental to word identity and intelligibility of spoken language. Thus, prosodic structure has systematic and perceptually-relevant effects on the articulatory and acoustic properties of segments (e.g., Turk & Shattuck-Hufnagel 2000; Fougeron 2001). For example, the higher the level of a prosodic boundary, the greater are the ‘articulatory strength’ and duration of the segments immediately after the boundary, as reviewed by Fougeron (2001).

1.3 *Summary: what phonemes can and cannot do for speech technologists*

The concepts discussed in section 1.2 are so enshrined in conventional thinking and techniques that phonemes are often not thought of as linguistic, but as 'real'. However, they are purely theoretical concepts with only partial adequacy even for the domains for which they were intended—which was parsimonious linguistic description, rather than elucidation, of the information contained within real speech.

By definition, phonemes are free of all context. Their connection with everyday speech is therefore relatively tenuous. Phonemes do not reflect much that is important about a spoken message. First, they lack information about the speaker's identity, mood, and purpose in speaking, and second, the only aspect of meaning that they reliably reflect are contrasts that serve to distinguish the formal phonological structure of words spoken in isolation. In formalizing the phonemes of a language, certain attributes of sounds are emphasized and others are deemphasized, for reasons that sometimes have little to do with what a human listener or a machine recognizer needs to do even to recognize isolated words. The value of phonemes is even more limited in recognizing connected speech, especially in dialogue and other types of interaction. Further, it has long been known that words are easier to identify in connected speech than in isolation (Miller *et al* 1951) and that sometimes a significant portion of the signal must be heard for accurate word identification (Pickett & Pollack 1963).

Machine speech recognition, on the other hand, avoids many of these problems, and in many ways is compatible with the theoretical position argued for in this paper. If the style of speech and intended applications are sufficiently narrow (e.g., if responses are single words or from a known set of acoustically dissimilar items), then a restricted focus such as that offered by phonemes can work well. For more complex applications, a different approach is necessary. Most-likely pathways through a phoneme lattice, together with probabilistic pattern-matching techniques such as hidden Markov models (HMMs) or Mel frequency cepstral coefficients (MFCCs), can exploit phonetic detail that may be neglected in phoneme-based models. These techniques may also more closely reflect human behaviour, since human listeners are sensitive to phonetic detail and to its statistical distribution. However, statistical techniques in the absence of any linguistic model can have limited success, because there must be an appropriate linguistic model to refer statistical outcomes to. Hence, for more complex applications, Jelinek is widely reported to have said 'Every time I fire a linguist, my recognizer's performance goes up.' He was right: the linguistic model being used for his more complex applications was inappropriate. The rest of this paper addresses properties that such a model needs.

2. Less common assumptions about the acoustic speech signal

2.1 *Overview*

The point from the preceding section is that many limitations result from the common assumption that the input to a language-understanding system should be a sequence of discrete, 'clean' units, be they phonemes, subphonemic features, or larger units such as syllables, all of which have equal status and are more or less equally informative about phonological form. Relatedly, it is assumed that anything else present in the speech signal is irrelevant to speech recognition.

The alternative view offered here is that much variation is systematic and perceptually salient, but does not necessarily help identify either citation-form words (i.e. carefully-pronounced individual words spoken in isolation) or phonemes, other than indirectly by narrowing the range of

probabilities of particular words or longer chunks of speech. This alternative view seems difficult to accept, so ingrained in Western scientific thinking is the traditional view, but the insights provided by the alternative make perseverance worthwhile.

It has long been known that phonetic detail informs listeners about the position of a segment in a syllable, and about syllable structure in general. Such so-called allophonic variation is briefly outlined in section 1.2d. More complete descriptions are available in any good textbook on acoustic phonetics, as well as textbooks on impressionistic phonetics of particular languages. So-called phonetic detail also informs listeners about a wide range of levels of linguistic analysis other than phonemes. This has been known and somewhat systematized since at least the 1930s but there has been a resurgence of interest in it since the late 1980s, especially in the interactional functions and perceptual salience of non-phonemic phonetic detail. Phonetic detail can convey speaker identity via phonetic contrasts (Nygaard & Pisoni 1998; Bradlow *et al* 1999; Allen & Miller 2004); phonetic detail that enhances prosodic distinctions can be localized (Keating *et al* 2004), while phonetic detail enhancing segmental distinctions can affect long stretches of sound e.g., traces of English /r/ (r resonances) can occur several syllables before the main /r/ segment and influence perception (West 1999; Heid & Hawkins 2000; Coleman 2003; Heinrich *et al* 2010). Even some well-researched distinctions which involve multiple acoustic properties, such as coda voicing, include phonetic detail that is less local than was believed until recently (Hawkins & Nguyen 2004), while traditionally 'paralinguistic' attributes of speech, such as voice quality in English and Finnish, can play linguistic rather than paralinguistic functions (Ogden 2004; Local 2007). Such variation produces systematic sound contrasts that have linguistic functions but do not all distinguish lexical forms (see *Phonetica* Vol. 61, especially Local & Walker 2005; Ogden & Routarinne 2005; Plug 2005).

Some aspects of phonetic detail do not need formal perceptual testing to establish their salience—it is perfectly clear that people can hear them—although it can be worth investigating their contribution to speech understanding and the effect on speech processing of using the 'wrong' version. Examples for English include grammatically-distinct variants of strong and weak auxiliary verbs such as *had* (cf. Ogden 1999).

Conversely, many instances of phonetic detail seem hardly noticeable to the casual listener. This is intriguing, because phonetic detail that is difficult to detect in silence can increase intelligibility in noise e.g., /r/-resonances (see above). Ignorance of phonetic detail, or of how it works in a particular language or dialect, may explain the disproportionate difficulty of understanding a foreign language in noise (Garcia Lecumberri & Cooke 2006; Heinrich *et al* 2010). Furthermore, not all observed phonetic detail is perceptually salient in all conditions. As yet, factors influencing salience are little studied and poorly understood. Presumably, similar to physical aspects of a stimulus, particular aspects of phonetic detail may be ignored if other stimulus attributes override them or if the task makes them irrelevant. These complex issues are not discussed here.

As noted in section 1.1, these observations are important because standard phonology and most perceptual models assume that lexical form is the most important thing to be distinguished, and that the citation form of the word is the most important lexical form. Yet citation forms are those that are least often heard; and, consistent with their restricted range of linguistic functions, the phonetic detail that they include is relatively restricted. As demonstrated by Local (2003), in natural, and perhaps especially conversational speech, different forms of phonetic detail are not only widespread, but enormously informative about the function of the utterance in discourse. Moreover, experiments demonstrate that speech which lacks natural detail is remembered less effectively, and that learning new sound categories is more effective when the stimuli include the variation typical of natural speech (Duffy & Pisoni 1992; Pisoni *et al* 1994).

2.2 *Detailed example: Phonetic detail accompanying True and False prefixes in English*

An example of non-phonemic phonetic detail is that the morphemic status of a syllable affects the pronounciation of its component segments. This point was introduced in section 1.2d, in describing figure 2C, and is developed further here. Figure 3 shows spectrograms of the words *mistimes* (above) and *mistakes* (below) spoken in the phrase *I'd be surprised if Tess* ____ *it*, with the main sentence stress (technically, the nuclear stress) on *Tess.*

The first four phonemes in these words are identical, /mɪst/, but there are clear differences between them. For example, in *mistimes,* the periodicity associated with /mɪ/ is longer in absolute duration and relative to the duration of the aperiodic /s/ that follows, compared with the periodicity associated with /mɪ/ in *mistakes.* Spectral differences in this first periodic section include, for *mistimes,* a more abrupt shift in formant frequencies at the segment boundary, slightly different formant frequencies consistent with a more extreme vocalic articulation, and higher amplitudes. Similarly, there are differences in the durations of the silence associated with the closure of the oral cavity for the /t/, and with the aspiration (or VOT) following the transient at the release of oral closure. The details need not concern us, except to note that they are different both in absolute terms and relative to the durations of the periodic and aperiodic segments in the earlier parts of the syllable.

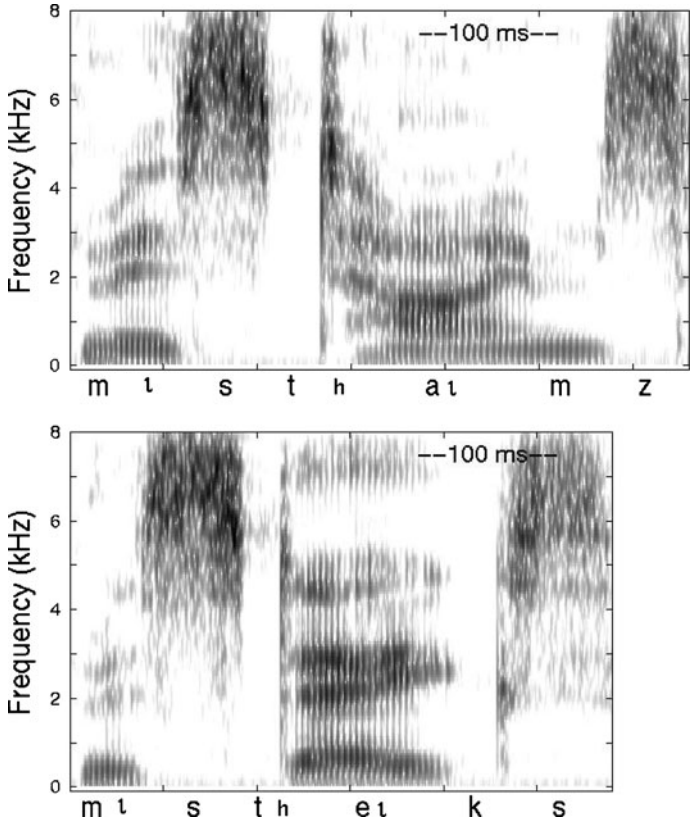
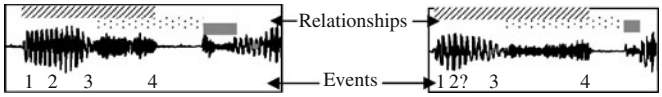


Figure 3. Spectrograms of *mistimes* and *mistakes* from *I'd be surprised if Tess*___ *it*. The first four phonemes (/mɪst/) are the same. Their acoustic differences, summarized in Table 1, produce a different rhythm that may signal that *mis* in *mistimes* is a productive morpheme, whereas *mis* in *mistakes* is not.

Table 1. Perceptual information thought to be available in the syllables *mist* as shown in figure 3, based on the literature. **Bold** font indicates potential perceptual units, including nodes in prosodic structure. The waveforms at the top roughly indicate the portion of the utterance the table refers to, and are those from which the first parts of the spectrograms in figure 3 were made. Numbers below the waveforms indicate short-time events and correspond to those in the leftmost column of the table. Patterned bars indicate the durations of adjacent acoustic segments in the waveform. The ratios of these durations, shown next to the labels in the corresponding bars in the table, provide non-phonemic linguistic information. Question marks (?) indicate low certainty/likelihood. Left: *mistimes*. Right: *mistakes*.



	˅tɛsmɪstaɪmzɪt		˅tɛsmɪsteɪksɪt	
Acoustic property	Cue	Perceptual correlate	Cue	Perceptual correlate
Events: see waveforms				
1. periodic, nasal	damped	new **syllable** (simple **onset**); new **morpheme**; **word**; poor segment identity	damped	same as *mistimes*
2. nasal-oral boundary + formant definition	Abrupt Clear	**features** for [m]; **phoneme** /m/? high front **vowel**?	Unclear Unclear	**features** for nasal? labial?? high **vowel**? front vowel??
3. frication start	rel. Late	syllable **coda** starts; **rhyme** has voiceless coda??; **features** for [s]; **phoneme** /s/? **syllable** is weak, light??	rel. Early	same as *mistimes* except **syllable** is weak
4. fricative-silence boundary	rel. Early	**phoneme** /s/; voiceless **coda**; **coda** ends? new **syllable**? **features** for [t]? **morpheme** ends?? productive **morpheme/** same **word??**	rel. Late	**phoneme** /s/; **features** for *mis*, maybe **dis**; **features** for [t]?; syllable **coda** continues??; **morpheme** continues ?? (is nonproductive)
Relationships				
Relative durations: sonorant sibilant	1:1	weak, light **syllable?** (light as used by Ogden *et al* 2000)	1:2	weak heavy **syllable** (heavy as used by Ogden *et al* 2000)
Relative durations: sonorant sibilant plus sibilant silence	1:1 2:1	weak, light **syllable**; productive **morpheme** *mis*? *(dis??)*; silence + intonation heralds new **syllable onset**, new **foot?**	1:2 3:1	weak, heavy **syllable** 1, strong **syllable onset** 2 of same **word** (monomorphemic monosyllable)?; defocussed **verb** *missed*??
transient + aspiration	Long	confirms: productive **morpheme** *mis (dis??)* new strong **syllable onset** [t^h]; **features** for [t^h]; **phoneme** /t/; new **foot**; new **morpheme**, same polymorphemic **word**	Short	new strong **syllable onset** [st]; new **foot**; Confirms monomorphemic **word** beginning *mis(t), vis, bis (dis?)*; **features** for [t]; **phoneme** /t/.

These differences are systematic. They depend on the productivity of the morpheme *mis-*. When a morpheme is productive, as *mis-* is in *mistimes*, it means that it changes the meaning of the word: *mistimes* means something similar to the opposite of *times*, or, alternatively, *times* is a word that means something about timing in its own right, and is more positive or neutral than *mistimes*, unless followed by another word with negative meaning such as *badly*. Likewise, for the productive morpheme *dis-*. When we say that something is *discoloured* we mean that its colours have been spoiled. In contrast, when these same phonemic sequences, *mis* and *dis*, are removed from the beginnings of some other words, the meaning is not a more negative version of the meaning of the rest of the word. Thus *mistakes* is not a more negative meaning of *takes,* and *discover* is not a more negative version of *cover*. (Some of these connected meanings may have been true when the words were first used, but they have lost that connection in present-day usage.)

The take-home point is that the pattern of segments within the syllable indicates its status either as a morpheme in its own right, or as just the first syllable of a polysyllabic monomorphemic word. Put another way, spectrotemporal differences in this first syllable, clearly visible in the spectrograms, result in different rhythms: a heavier beat when *mis* is a productive morpheme, as in *mistimes*. The reasons for this are described in (Ogden *et al* 2000). Measurements which show the reliability of these acoustic patterns for many such pairs are reported by Baker *et al* (2007) and Baker *et al* (under revision).

This example is valuable for four reasons. It illustrates phonetic detail operating when traditionally crucial variables (phonemes, word boundaries, foot structure, intonation, grammatical class) are constant i.e. it is controlled according to traditional criteria. Second, unlike communicatively important phonetic detail in fast or casual speech, it lies within the domain of current modelling: it arises from normal-to-slow clear speech with no 'missing phonemes'. Third, it cannot be dismissed as unnatural. The examples illustrated in figure 3 were read while the speaker role-played the part of a mother at a child's athletics meeting where the accuracy of the person timing the races was in dispute; Baker's data (Baker *et al* 2007, submitted) are obtained from scripted dialogues read fast and fluently by practised speakers; they sound quite natural. Fourth, the data illustrate the complex interactions in linguistic information that listeners can exploit: as explained in Ogden *et al* (2000), the phonetic detail signals morphological differences, via phonological differences realized segmentally and prosodically.

Recent work tentatively confirm that phonetic detail in such syllables facilitates their intelligibility in background noise (Baker 2008); further perceptual investigations are in progress. This systematic variation has implications for the nature of the input to models of word recognition which incorporate lexical competition (e.g., McClelland & Elman 1986; Norris 1994; Gaskell & Marslen-Wilson 1997; Norris *et al* 2000): with the exception of the aspiration associated with the /t/, the featural and phonemic information are identical in each pair; the phonetic detail is not. Other experiments exploring the perceptual salience of these syllables, and modelling, are in progress.

2.3 *Relevance of systematic phonetic detail to large-scale technical applications*

Table 1 shows information available from the utterances of /mɪst/ in figure 3. The two main points are: (i) every acoustic sound chunk signals more than one thing about abstract linguistic structure, with different levels of probability for different types of abstract units; (ii) phonetic detail signals things about the linguistic system that phonemes (similar to 'clean input') cannot signal,

and hence it follows that phonetic detail must be represented in memory and in abstract linguistic structure. To illustrate these points, consider, first, the second row in table 1. The acoustic cue is the presence of an abrupt shift in formant frequencies and amplitudes at the segment boundary between [m] and [ɪ] in *mistimes*, vs its absence in *mistakes*. The difference in information about linguistic structure is that, for *mistimes*, features for a bilabial nasal are signalled with high probability, and the phoneme /m/ and a following high front vowel with reasonably high probability. Whereas the features signalled for *mistakes* are altogether less certain, and the phoneme /m/ itself is unlikely to be identified.

As a second example, consider the penultimate row in table 1: the relative durations of the sonorant to the sibilant (that is, of the periodic segment to the aperiodic segment), and of the sibilant to the silence. As discussed for figure 3 and noted in table 1, these relative durations are very different for the productive morpheme in *mistimes* compared with the non-productive *mis* syllable in *mistakes*. Amongst other things, they indicate different types of 'phonological syllable' (light vs heavy). This distinction essentially governs the overall rhythmic difference and the location of the syllable and the foot boundaries. Crucially, it is the morphological difference that causes all these other differences. Thus, all of these signal properties affect the probability of *mis* being a productive morpheme, or not. At the same time, of course, these same acoustic segments help define the actual sound segments: if, for example, the aperiodicity we are calling /s/ is too short for the context, it may not be heard as /s/ but as /z/ or /tʃ/ (as in the first and last sounds of *church)*. If the periodicity of [ɪ] is too long, the syllable might be heard as *me*, or (depending on the rest of the context), there could be a rise in the probability of a word boundary between the nasal and the vowel.

For speech technologists, the above explanation may provide more linguistic detail than is wanted. A brief summary is that the fine detail in the signal can provide information about things that at present tend to be estimated from 'higher-level' language models, in this case morphological structure and word affiliation. The point is that a model should allow the same information to be derived with higher certainty from the signal itself. How to represent this context-sensitive information is discussed in section 4, after other types of systematically-varying phonetic detail have been illustrated.

3. Examples of phonetic indications of linguistic subsystems

This section briefly overviews types of systematic variation that occur in ordinary connected speech, and are largely undocumented in textbooks that deal with acoustic phonetics. Unless they are known about, their presence in a speech corpus can present technical challenges to recognition systems. More generally, their theoretical implications encourage the development of theoretical frameworks that automatically take care of them.

3.1 *Grammar and phonetic detail*

The morphological example discussed in section 2.2 represents an instance of fine phonetic detail that indicates grammar. Another grammatical subsystem concerns the English system of pronouns with auxiliary verbs (such as *had, are,* etc.), and in particular the phonetic processes, or patterns, that are associated with them and them alone when they are present in connected speech. One example (Local 2003) is for pronunciations of the syllable /aɪm/ as the function word(s) *I'm*, in comparison with pronunciations of the same phonemic sequence in the rhyme of

content words such as *lime, mime, time, crime,* and *grime.* In ordinary lexical items which are content words, syllable-final /n/ normally assimilates to the place of articulation of the following consonant in English. Thus sequences such as *the line broke* can be spoken more like *the* /laɪm/ *broke,* and *the line goes there* like *the* /laɪŋ/ *goes there.* In each case, the consonant at the end of the word *line* takes on the same place of articulation of the consonant in the onset of the next word, but retains its nasality. Speakers are usually unaware that they do this, but the process is well-attested in most dialects of English, and listeners 'compensate' for it (Gaskell & Marslen-Wilson 2001).

However, when the syllable-final consonant is /m/, such assimilation does not take place. Thus, speakers will say *the* /laɪŋ/ *goes there* for *line*, but not for *lime.* Content words such as *lime*, with a bilabial coda, retain their bilabial place of articulation. However, the same does not hold good for the pronoun+auxiliary sequence *I'm*. Local (2003) reports data indicating assimilation of place of articulation of the /m/ in *I'm* before the words *blowing, throwing,* and *going:* in each case, what we write as an /m/ is produced as a nasal with the same place of articulation as the onset of the next word—bilabial still for *blowing,* but dental for *throwing*, and velar for *going*. The claim is that it is only important that *I'm* contrasts with other sequences with which it is in competition, in this case the pronoun+auxiliary sequence. In English, these competing contrasts are acoustically different from one another: *you're, she's, he's, we're, they're, it's.* To retain its distinctiveness, all that is necessary is that *I'm* retains, most crucialllly, nasality, and, less importantly, a relatively open vowel (if it has a vowel—it need not). Although the vowel of *I'm* may not be pronounced in very fast speech, the place assimilation of its /m/ may occur even in clear speech. Importantly, in most utterances, no matter how clearly or how fast they are spoken, *I'm* will always be 'more reduced' (less clear) than the associated content words. So, if you are trying these examples out for yourself, you may find that you assimilate a content word's coda place of articulation in very fast speech; but notice that, when you do that, the *I'm* may be so reduced that it is little more than a nasal grunt.

Likewise, each of the other sequences in this set has it own rules of what must be retained for its identity still to be clear. So, for example, *she's,* spoken as one syllable, has spread lips throughout when it means *she is*, but rounded lips when it means *she was.* The rounded lips cause a subtle difference in vowel quality. The lip position for the otherwise similar 'phoneme' segments is entirely congruent with what would happen in a slow and very clear production of these sequences as two syllables/words; when *she's* is contracted into one syllable, the vowel quality due to lip-rounding may help listeners decide whether the verb is in the present or past tense.

These are just some examples within one subsystem of English, but the same type of difference is common. Ogden (1999:65) notes 'It is well known that in many languages the constraints on well-formed minimal words are different for (open) lexical systems than for (closed) functional systems, and different processes apply to function words, including auxiliaries (McCarthy & Prince 1995: 324, Iivonen 1996: 110, Casali 1997: 495f, 503).'

3.2 *Perceptually salient information about a single phoneme or feature can be distributed*

The preceding sections show that the acoustic properties of particular sound segments always indicate non-phonemic attributes such as position in syllable structure, and typically indicate longer parts of the linguistic structure too, for example the syllable's position in the prosodic foot and intonational phrase. Their acoustic properties can also indicate non-phonemic category membership such as grammatical status and morphological structure. This was shown by examining acoustic patterns of sound segments within syllables. In contrast,

this section shows that perceptually-salient information that is traditionally associated with a single segment or phoneme can be distributed over entire syllables or even groups of syllables.

3.2a *The phonological voicing distinction: a property of the whole syllable*: The first example is consonant voicing in the coda (end) of a syllable. Figure 4 shows four spectrograms of English words, each with a single stop in a syllable coda. In the top row are words with a voiceless coda stop: *buckeye* at the left, and *let* at the right. Words with the corresponding voiced stops are below: *bugeye* (left) and *led* (right). The letters a–f on the spectrogram of *bugeye* show well-established perceptually-salient acoustic properties that are dependent on coda voicing and can be read about in any acoustic phonetics textbook, and e.g., Klatt (1989). Compared with the voiceless /k/ in *buckeye*, the voiced /g/ in bugeye has (a) longer preceding syllabic nucleus (the vowel); (b) shorter closure (interval of low-frequency energy or no energy, the former being shown here), followed by (c) shorter voice onset time (VOT); (d) greater proportion of low-frequency periodicity in the closure; (e) lower F1 at the boundaries of the closure; (f) lower f0 at the boundaries of the closure. In addition, in English, voiced coda stops (and fricatives) typically have shorter and lower-amplitude aperiodic energy in the coda obstruent release burst at high frequencies, shown as (j) in the lower right hand panel of figure 4.

More recently, Hawkins & Nguyen (2004) confirmed that acoustic properties of the syllable onset covary with coda voicing. These are illustrated by the letters g–i in the spectrogram for *led*. Thus, in addition to the properties pointed out for *bugeye*, the syllable with the voiced coda,

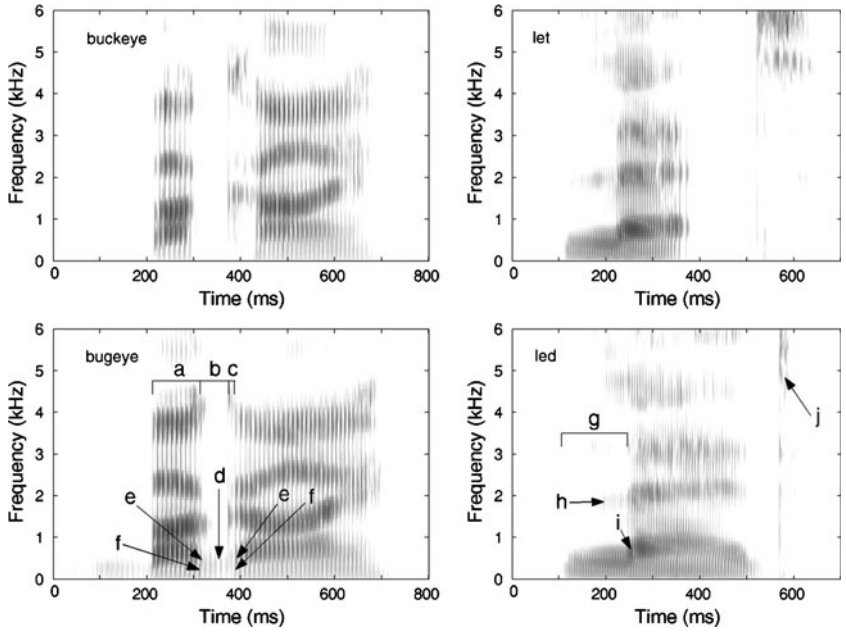


Figure 4. Left: Spectrograms of *buckeye* (above) and *bugeye* (below), man's voice. Right: Spectrograms of *let* (above) and *led* (below), woman's voice. The letters on the spectrograms show (left) well-established perceptually-salient acoustic properties that are dependent on coda voicing, and (right) less well-known differences in the syllable onset as well as the rhyme. See text for details.

led, has (g) longer [l]; (h) phonetically darker [l] due to lower F2 frequency; and (i) longer formant transitions at [l] release. Additionally, but not marked, the first two formants in the vowel follow different trajectories to produce more contrast in 'brightness' between the edges and the centre of the vowel when the coda is voiced. Just before the onset of /t/ closure in *let*, there is a region of more irregular voicing (from about 350–380 ms), caused by irregular but abrupt closing (and then opening) of the vocal folds. This is called glottalization (also laryngealization, vocal fry or creak), and is characteristic of some accents of English before voiceless stops. Although in acoustical terms, glottalization lowers f0 (noted in point (f) as associated with [+voice]), the perceptual effect of glottalization is very different from lowered f0 in more modal voicing. At a sonorant-obstruent boundary, glottalization is almost certainly associated with abrupt offset of all periodic excitation and absence of falling F1 transition, rather than a change to a continuing but low f0. Glottalization at a sonorant-obstruent boundary thus covaries with (the opposite properties of) points (a), (e) and (f) above to reinforce the voiceless properties of the coda. In connected speech, it probably functions as a cue to which syllable the obstruent is in. These differences together result in voiced and voiceless codas producing distinctive dynamic patterns throughout the entire syllable.

This information tends to surprise many people for at least two reasons. One is that the difference tends to be described as phonemic, and thus is thought of as due to the difference between two particular segments, whose strongest acoustic correlates are thought of as arising in the vicinity of the stop closures (or, in the case of fricatives, the aperiodic portions) together with any following aperiodicity. Although this works adequately (but not perfectly) for many cases of syllable-onset stops and fricatives, it is wholly inadequate for coda stops and fricatives. The strongest perceptual cue for English coda voicing is really the duration of the preceding vowel (or sonorant (and hence, normally, periodic) part of the syllable rhyme). The other reason is that the same term, voicing, is often used to describe a phonetic property and a phonological contrast. Terms do exist to distinguish them, e.g., the phonetic term phonation for periodicity associated with glottal vibration, and [voice] for the abstract contrast of phonological theory, but to go further into the issue here would go beyond the scope of this paper. For more information, see Hawkins (2010b).

Almost ironically, the end result is that the phonological contrast between voiced and voiceless stops (and fricatives) in English syllable codas is indeed associated with the presence of copious periodicity in the case of voiced codas, and its absence in the case of voiceless codas; but the source of the difference is spread across the entire syllable rather than resting only in the final consonantal segment, and it is most evident in the syllable nucleus—the vowel. Syllables whose codas are phonologically voiced can be called 'sombre': with long durations of periodic noise, short closures in the coda, a low spectral balance at syllable edges, and, where there is aperiodic noise, it is short and low amplitude. In contrast, syllables whose codas are phonologically voiceless can be called 'bright': with short durations of periodic noise, long closures in the coda (i.e., long periods of silence or low amplitudes relative to the periodic parts), a high spectral balance at syllable edges, and long, high-amplitude aperiodic noise.

3.2b *Spread of partial information about a segment over several syllables*: It has long been known that certain properties or features of individual phones can spread over several segments. This is normally called coarticulation of the relevant feature, one of the earliest and best-known examples being coarticulation of lip rounding, which Benguerel & Cowan (1974) showed may appear several phones before the phone segment that causes it, as long as the intervening phones do not demand active lip-spreading. Bell-Berti & Harris (1981) presented data supporting the

idea that the relevant unit could be durational—about 200 ms. This type of question remains unresolved, probably at least partly because the answer will depend on what type of prosodic and interactional structures are examined, as well as what particular sounds (Coleman 2003). However, as mentioned in section 2.1a, some properties, notably of British English /r/, can spread over significantly longer durations than 200 ms, and are perceptually salient in that they affect intelligibility in adverse listening conditions. Features that last for such long durations could be reproduced in synthesis systems, but cannot be modelled using current machine speech recognition technology. Nevertheless, their existence may interest researchers in fields such as auditory scene analysis, as they seem to influence the perceptual coherence of auditory and presumably of audio-visual signals.

3.3 *Summary*

This section has given examples showing that phonetic detail can systematically distinguish the grammatical function of a stretch of sound which ranges from one to several phonemes. Phonetic detail can also provide weak but consistence evidence of individual phonemes or features over the duration of a syllable or more. Other studies, referenced but not illustrated in section 2.1a, show that phonetic detail provides valuable information about the discourse function of specific words and phrases.

Although much of the information we have is about patterns in speech production, perceptual experiments suggest that human listeners are sensitive to most if not all of the measured distinctions when they are systematically distributed in the database, and, perhaps, relevant to the specific task. To be systematically distributed, the occurrence of an acoustic-phonetic contrast must closely correlate with a contrast that has communicative relevance. This might be present at any level of analysis from discourse structure and pragmatics, through many types of grammar, to syllable structure. In other words, humans use sensory cues that serve a useful purpose. For the speech technologist, these points could be relevant to efficient processing in applications that involve large-vocabulary connected speech, and/or multiple styles or rates of speech, or multiple talkers, especially in models that can adapt to individual talkers.

4. Representing systematic phonetic detail

The information reviewed above encourages revision of the standard viewpoint that the acoustic signal is disorderly—that is, with lots of variation that is random, or else irrelevant to the meaning of the message. Instead, these data encourage the view that much of the spoken signal which had been thought to be disorderly in fact reflects a rich range of linguistic and interactional information that is intrinsic to the meaning of the message. It had been thought of as random or irrelevant only because this type of information cannot be systematized by a linguistic model whose focus is to identify a linear string of phonemes, perhaps together with some information about word and syllable junctures and stress, and to do so by using as direct a relationship as possible between robust signal properties and phoneme string.

To move away from this standard school of thought, a radically different model is needed. There is no consensus on what this model should be, but most people now believe that it should be able to cope with probabilities of category membership, with a range of temporal domains, and with conversational and interactive speech, whose properties are typically very different from the read sentences typical of early speech databases. There is also general agreement that context-sensitivity is vital, and that one cannot necessarily generalize principles that work for

one type of speech, such as read sentences, to another type such as dialogues. In other words, that it was a mistake to assume that working with 'clear speech' genres such as isolated syllables or words, and read sentences, would establish the main principles of speech recognition which could then be readily applied to other speech situations. Speech technologists were amongst the first to recognize this: typical recognition applications use probabilistic pattern recognition, and any given application may be restricted to a small range of speech situations; synthesis systems may have different and essentially independent software to produce different types of speech.

Section 4.1 describes a linguistic theoretical approach, Firthian prosodic analysis (FPA) which provides information about much of the argument in this paper. As with section 1.2, it can be skipped by readers who only want to get the gist of this paper—a summary is offered in section 4.2; see also Clarke *et al* (2006). However, it is recommended that those who want to understand the full argument should read section 4.1. Section 4.3 outlines the principles of Polysp, a conceptual framework intended to guide research into a biologically-plausible model of speech processing.

4.1 *Comments on linguistic theory: Firthian prosodic analysis*

The theoretical linguistic viewpoint advocated here shares many principles with the speech technology approach, and also, it is hoped, with attributes of human memory and cognition. In place of the single all-purpose system that is the simplest form of the linear phoneme model, speech and language are seen as multiple, distinct subsystems each of which may combine 'the same phonemic' sound elements in ways that may be different in different sub-systems—as exemplified in section 2.2 for distinct syllabic patterns amongst morphological prefixes, and section 3.1 for connected speech processes that distinguish the English pronoun+*be* system from similar phoneme sequences in content words such as nouns and main verbs. In the system described here, this multiplicity of subsystems is called *polysystemic*, from Greek *many systems*—literally, many organized wholes. Equally important in this linguistic model, therefore, is emphasis on specifying the hierarchical structure that governs patterns of sound segments within any given subsystem. This provides their context sensitivity. Finally, each given sound chunk (utterance or part of an utterance) is represented in multiple systems, including metrical/prosodic, grammatical, and interactional, each linking to and affecting the others. For example, the status of *I'm* as a grammatical function word determines both its rhythmic properties in the hierarchical metrical-prosodic structure, and what connected speech processes that particular sound pattern can exhibit.

This approach is heavily influenced by the branch of linguistic phonetics called Firthian prosodic analysis. FPA was developed (with the name prosodic phonology) by John R. Firth and colleagues in London in the early-to-midpart of the twentieth century, largely through analyses of languages of India, South East Asia, China and Africa. The work was not completely explained to people outside the school of thought (so its technical terms are often misunderstood), much was not systematized, and little of it was published. All these factors made it relatively inaccessible, but in recent years it has enjoyed a resurgence of interest and more accessible scholarly input, especially with respect to researching phonetics of speech during interaction with a single or multiple interlocutors.

The reason for the resurgence of interest is that FPA offers insights that many other systems cannot. Autosegmental phonology (Goldsmith 1990) bears similarities with FPA, and there has been a debate about how distinct the two are (Ogden 1993; Goldsmith 1994; Ogden & Local 1994). The ideas described in the present paper need not be affected by this debate. I use FPA

because it is valuable in developing and testing hypotheses about processes of human speech understanding, from perception of simple syllables to understanding the meaning of utterances heard in special circumstances. It has shaped my ideas because I collaborate with the phoneticians John Local and Richard Ogden at the University of York, England, who have used their phonetic expertise and FPA's insights to produce robust, natural-sounding text-to-speech systems and are now researching pragmatic functions of systematic phonetic variation of conversational speech. Thus, I find FPA valuable as a conceptual and practical tool, without the need, yet, to engage in theoretical debate about its details.

FPA principles have been computationally implemented in XML for speech synthesis, and hence speech production (Ogden *et al* 2000), and a framework, Polysp, has been outlined for perception (Hawkins & Smith 2001; Hawkins 2003, 2010a, b), but not computationally implemented though (see Piccolino Boniforti *et al* 2010). A distinguishing property of FPA is its focus on relationships between structures and contrastive subsystems within those structures. FPA is irrevocably context-bound: just as was argued for the syllable structures shown in figure 2, no sound, and no other linguistic unit, is fully describable in isolation from its prosodic and grammatical structural context. Interactional functions and goals can also be included. As implied above, such rich structure contrasts with most other models. Although several models emphasize multiple, hierarchical levels and account well for particular domains (e.g., Mattys *et al* (2005) psychological model of word segmentation, Pierrehumbert's (2003) probabilistic phonology, and Hertz's speech synthesis (Hertz 1991; Hertz & Huffman 1992; Hertz 2006), none takes as comprehensive an approach as FPA, and they all accord relatively privileged status to the phoneme (or similar unit) and to phonological form, while neglecting grammatical and interactional function.

Figure 5 gives a flavour of what is meant by the above FPA descriptions. It shows two 'tree structures' that represent the prosodic attributes of *Tess mistimes it* (with True prefix, left) and *Tess mistakes it* (with Pseudo prefix, right). The highlighted parts of each structure essentially provide the same information for the True and Pseudo prefix as table 1 does. The light grey lines and labels in each tree represent the prosodic structural context that, with one exception irrelevant to the example, is the same for the two utterances, *Tess, -imes/-akes it*. (The exception is that the coda of *-imes* is voiced while that of *-akes* is voiceless.) Links to some other aspects of knowledge are shown by arrows. Note especially the link to a parallel syntactic tree, indicated by the large triangles in figure 5. Syntactic trees also link to words, which are not represented in the prosodic tree for English. Information about a word includes indication of its morphemic structure, since pronunciation is affected. Other links, e.g., to the limbic, motoric and wider memory systems, are important but not shown. For perception, dashed arrows can be thought of as indicating direction of information flow as the speech signal is processed; the tree itself represents information that is hypothesized, built up, and confirmed, from incoming phonetic detail meshing with predictions from existing knowledge (priors).

English words cannot be represented as a level in the prosodic tree because English word boundaries do not uniquely match prosodic units, particularly feet. (A foot comprises a stressed syllable and all following unstressed syllables. There are two feet in each prosodic tree in figure 5: the first beginning with *Tess* and the second with *times/stakes.*) Word boundaries can thus correspond to prosodic units in languages that have fixed syllable stress (e.g., Czech, Polish), but it is worth noting that the concept of word is weak in some languages, perhaps especially if their grammatical units significantly change either the beginning of the word, or its stem. For speakers of either of these types of language, the absence of an explicit word level in the prosodic tree seems likely to be less surprising than it tends to be for native speakers of languages such as English. Similarly, many languages, including those of the Indian

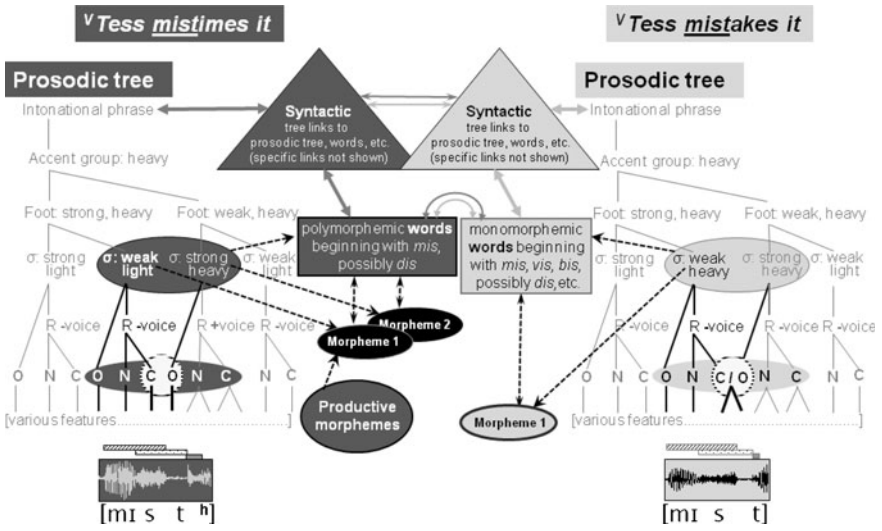


Figure 5. Partial structures (following Ogden *et al* 2000), describing differences between *mistimes* (left) and *mistakes* (right) in the utterances *Tess mistakes/mistimes it*, with nuclear (main) stress on *Tess*. Bars above the waveforms are as in table 1. Node labels and prosodic tree branches that are the same for both utterances are shown in light grey. Differences are highlighted and in darker shades: they are the prosodic tree structures corresponding to the two *mist* portions, which determine whether one or two morphemes are identified. Dashed arrows show the parts of the prosodic structure crucial to informing about the particular morphemic structure: two morphemes for *mistimes*, one for *mistakes*. Solid arrows show links to some other sources of linguistic knowledge, which, in a full system, will be specified in as much detail as the prosodic trees. See text for further discussion; also Hawkins (2010a).

subcontinent, which have weak or no stress distinctions, may have no foot level in their prosodic structures.

Interrelated structures such as those in figure 5 are complicated, but so is speech if all its functions are represented—a complete system would include links to associative semantic networks and to more general biological systems governing memory, attitude, and emotion, all of which affect understanding. For Hawkins' model Polysp, six points are central.

(i) Each prosodic tree is linked to a corresponding grammatical tree and to other information e.g., about words and word boundaries.

(ii) There will also be links to other structures, such as those indicating the place and function of an utterance within a discourse, and to non-linguistic knowledge systems. For example, certain prosodic properties, voice quality, and words are associated with specific emotional content. Representation of meaning is probably embodied, i.e., meaning is partially represented in multi-modal sensory neural systems and in responses of motor neural circuits that govern action.

(iii) Prosodic trees are the linguistic core of the representational framework, because they represent rhythm, which, in Polysp, binds other knowledge together (e.g., Large & Jones 1999; Grossberg 2003, especially §3).

(iv) The principle of linking nodes in prosodic structure to nodes in other structures allows great power: potentially, any currently relevant contextual influence can introduce phonetic variation. For example, there is no place in these linguistic structures for phonemes, because

phonemes are by definition context free. However, links can be made to phonemes, which can be useful in some circumstances.

(v) For the same reason, there is significant redundancy in this system. Although redundancy is normally avoided in linguistic theories, it is almost certainly an advantage for powerful, multi-faceted adaptive biological/cognitive systems such as language, as well as for successfully performing even simple language tasks in adverse listening conditions. There is no question that redundancy in the acoustic speech signal contributes to speech robustness. Moreover, redundancy can be argued to be necessary because language performs many different functions; in this approach, each type of function (deictic, emotional, interactive) can be seen as having its own quasi-independent system.

(vi) Phonetic information is distributed at all levels of the structure. For example, in English, when a consonant in the syllable coda is voiced, the feature [+voice] applies to the syllabic rhyme, not just its coda, with consequences for the entire syllable, since properties of a syllable's rhyme normally dominate those of its onset (Ogden *et al* 2000). This point is illustrated in figure 5 by the [+voice] associated with the rhyme *-imes* in *mistimes,* compared with the [−voice] for the rhyme *-akes* in *mistakes*, which are pronounced /aɪmz/ (voiced consonants) and /eɪks/ (voiceless), respectively.

In figure 5, elements linked by a vertical line connecting two or more levels dominate those that are at the same level but are linked to the same node at the next higher level by an oblique line. Thus the fact that a syllable rhyme dominates the onset of the same syllable is represented by the vertical line from each Syllable (σ) node to its Rhyme (R) node, while there is an oblique line from each Syllable to its Onset (O). This type of dominance relationship means that properties of the dominant element override or otherwise influence properties of the less dominant element. Thus in this case, and as noted in point (vi) above, properties of the Rhyme influence the output properties of the Onset rather than the other way around. Similarly, the stressed syllable that occurs at the beginning of a Foot dominates the unstressed syllables that follow it. The vertical line also indicates the beginning of other major prosodic boundaries, including the entire Intonational Phrase, and what it dominates. See Ogden *et al* (2000) for more, including an implementation in XML.

Point (vi) above merits elaboration. A listener is seen as placing feature values on the nodes of a general tree or linked set of trees. Trees are computationally implementable metaphors for linguistic knowledge. The main difference between this and a standard model of human speech perception is that the phonetic detail maps onto every type of node, at any place in the tree, not just at its lowest level. Thus, one premise of Polysp is that there is no single or rigid sequence to identify the properties of a tree: the process of perception is governed by the properties of the particular signal in conjunction with the listener's construal of the particular situation. Mapping the sounds of an utterance to linguistic structure is therefore massively parallel, and the order of mapping differs with listener, speaker, task and ambient conditions. Decisions depend heavily on previous decisions as long as the incoming signal is congruent with them, and a listener may sometimes understand the meaning of an utterance without having first identified all its linguistic structure (cf. Hawkins & Smith 2001; Hawkins 2003). This view deemphasizes the distinction between knowledge and sensation during speech perception.

4.2 *Polysystemic representation: a summary*

Figure 5 presents its information in a way that many non-linguists find unappealing. Its details do not all need to be understood. It simply illustrates the point that it is possible to

comprehensively and systematically describe the abstract properties of connected speech and its physical correlates, and that structural descriptions can differ even when the phoneme strings in a sequence of words are the same. That is, when the acoustic patterns in the speech differ, the structural description differs, regardless of whether the phonemes and the word boundaries are the same or different. Identical structures undergo identical processes, but structures that are different may be subjected to different processes, even, once again, when the phoneme strings and word boundaries are the same. As noted in the previous section, in FPA terminology, this multiplicity of subsystems is called *polysystemic*, from Greek *many systems*—literally, many organized wholes. Thus, much so-called variability in the speech stream can be systematically accounted for.

The polysystemic FPA model is thus fundamentally Bayesian. Speech is structured as multiple, hierarchically organized units, such as feet, syllable onset, terminal node, and so on. Units cannot be represented independently of their functional and linguistic context. Low-level phonological properties are no more important in determining the degree of similarity between units than any other level of structure: recall the point made in section 1.2d that *t* in *tap* has more in common with *p* in *pat* than it does with *t* in *pat* or *terrine.* The only thing that onset *p* and *t* do not share is the same place of articulation, but the determinedly segmental orientation of phoneme theory gives more importance to the place of articulation than to structural and prosodic properties, and ignores non-phonological influences such as grammar completely. FPA takes the opposite view. In other words: context is crucial. Thus, for any given unit U, if something in its context changes, then there will probably be consequences for the acoustic-phonetic patterns associated with U.

The extreme context sensitivity of FPA forces explicit relationships between elements that represent an utterance, which allows extremely accurate descriptions of the physical signal. This arises from the polysystemicity: different constraints apply to different linguistic structures. Thus, in contrast with the standard textbook view outlined in the first paragraph of section 4—that speech is characterized by much unexplained variability—FPA assumes that the signal is largely orderly: when variation is discovered, the underlying structure that describes its cause is sought. Thus, this approach can in principle account for acoustic-phonetic variation that reflects many aspects of communicative function. A consequence of this alternative viewpoint is that the focus of an acoustic-phonetic model is on identifying all and any units that help to access the meaning and function of each utterance, rather than identifying only those units that identify lexical form, although they too, of course, are important. In connected speech, the extra units always include (minimally) aspects of grammar; in conversational speech, extra units must include pragmatic and connotative meanings and interactional functions, such as requests, (dis)agreements, signalling the continuation vs the end of a conversational turn, and various syntactic devices (e.g., Clift 2001; Tily *et al* 2009).

For listeners in a laboratory setting or for a simple speech technology application, recognition of isolated words may not require identification of much higher-order structure. However, human speakers do not neglect this structure when they produce isolated words, and it seems likely that listeners attend to it as well. An isolated monosyllabic English word is not just a syllable and word but also a foot, an accent group, and an intonational phrase. Recognition that an intonational phrase is complete depends partly on recognizing lawful syllable structure and contributes to knowing that you have heard the end of the word, which, for a listener or a recognition system, would encourage a search for words of just this type. Furthermore, as the same phonetic detail can contribute to word recognition and to higher-order units, it makes sense to model them together.

FPA's so-called 'nonsegmental' approach lends itself to recognition systems that exploit pattern perception. This is partly because it attends to differences of detail in the physical signal that can be systematically mapped onto standard linguistic units (syllables, feet, intonational phrases, etc.). Thus it allows for physical variation to be systematized onto abstract structures. The detailed relationships between units in these structures define domains of influence of different durations. Thus, the traditional separation of prosody and segments is not possible within the FPA system, yet patterns that signify distinct meanings or functions can be differentiated. The focus is shifted from formal phonological structure, to contextualized meaning and function of the utterance.

In short, FPA shares properties with several aspects of speech technology. In particular, it is fundamentally context-sensitive, which could be computationally modelled as Bayesian functional networks; and there are strong parallels between the success of task-specific applications in speech technology, and FPA's emphasis on language structure as sets of independent subsystems, in which superficially similar elements may be governed by different processes if they are in different subsystems. Furthermore, FPA shares with powerful perception-action robotics models the use of highly abstract control structures of numerous types to govern the fine detail of physical behaviour (see section 5).

4.3 *Polysp: a summary*

Polysp is a conceptual framework developed as a biologically-plausible guide for research into human speech perception/understanding. It is motivated primarily from phonetics and neuropsychology, but is compatible with many principles that are standard in speech technology. As it is described by Hawkins & Smith (2001) and Hawkins (2003), and some of its basic principles are enumerated earlier, only a brief outline is given here.

The FPA principles described above are taken wholesale into Polysp. It follows that Polysp is function-oriented, 'situated' (i.e., uses context-dependent classification) and therefore 'knowledge-driven'. Additionally, it is assumed that classification into communicatively-relevant units is governed by probabilities derived from past experience, while classification itself is done in terms of probabilities of category membership rather than binary yes/no decisions.

Hawkins (2010b, especially section 9) argues that much of the process of understanding a physical speech signal may involve active construction of physically ephemeral but psychologically real 'auditory objects' from the combination of neural feedforward and feedback information flow. This idea is neither new nor unique to Hawkins, although its application to the processing of phonetic detail in natural speech may be unique. Although these and related speculations need not concern speech technologists, some of the evidence and assumptions upon which they are based may be useful. For example, as mentioned in (ii) in section 4.1, it is assumed that much knowledge, including meanings of words and longer utterances, is embodied. Neuropsychological support for this assumption comes for e.g., from Pulvermüller (1999), who presented evidence that verbs may activate different parts of the human motor cortex depending on what part of the body they normally involve. However, these relationships are not always simple. For example, contextual meaning can modulate localized processing differences in the brain due to grammatical status (Tyler *et al* 2008; Raposo *et al* 2009). In complex systems that can learn, shared parameters allow the development of a rich structure (Moore 2007). So these assumptions of embodied meaning and temporary emergence of a sense of 'tangible objectness' for units such as words partially justify the complexity of the FPA-type models of multiple, interlinked structures (or neural circuits), including those that experience and convey attitude and emotion.

Whether they are valuable for speech technologists depends on the application. However, their potential relevance to speech technologists is illustrated by a promising line of research which involves model prediction of distinctive, distributed brain activation patterns that arise when an individual thinks about a specific word (Mitchell *et al* 2008; Just *et al* 2010). It is noteworthy that these complex patterns, which go far beyond simple concepts of embodiment, are just for single words. Patterns for connected speech and complex thought are likely to be more complex.

The linguistic core of Polysp is the hierarchical prosodic structures. These represent rhythm, which binds knowledge together and guides prediction. Each prosodic tree links to its own grammatical tree, its function and sequential place in discourse (alternatively, the task) and to other knowledge systems, both linguistic (e.g., words) and non-linguistic (e.g., properties of things and of people). Each linguistic unit is completely context-dependent, and phonetic information from the signal informs all structural levels and linguistic systems (grammatical, phonological, semantic and pragmatic, including interactional) in parallel, knowledge-driven, distributed systems. The latter involve feedforward and feedback (or calibration), and dynamic, Bayesian stochastic signal processing.

The Polysp approach offers a solution to the (arguably fruitless) polarization of the exemplar/abstractionist debate. The proposed polysystemic 'structures' are highly abstract. Each utterance type has a unique (possibly partially specified) structure which determines the possible types and range of phonetic variation; an experienced listener will build a huge number of them. These structures are essential to accommodate the detail, because phonetic variation is uninformative when it cannot be related to a structure.

Attention to detail in the sensory signal is crucial for the structures to be used effectively, but this detail may provide information about any level in any of the structures. Thus, the phonetic detail maps onto many different units, potentially at any level in the structure. For example, in (i) *use the TAP* and (ii) *use the COLD tap*, (main stress on *tap* and *cold*, respectively) specific phonetic properties of the *t*, in its context, indicate new syllable, new onset, same accent group and intonational phrase, and also contribute to indicating that in (i) the syllable is stressed (thus new foot), while in (ii) it is unstressed (thus same foot).

Acknowledgement of implicit learning in speech (as in every other behaviour) justifies modelling listeners as monitoring and hence in some sense processing details all the time, even when not necessary for the immediate task at hand—so that the individual listener stays optimally adaptive for the future.

Although, as explained above, these principles have limited applicability in standard speech analyses due to the theoretical assumption that the first and most important stage of processing is to identify phonological form (i.e., due to the quest for context-free abstract phonological units), they are in fact used in models of processing in other modalities. They are used in models of non-speech language processing, of action, of visual perception, and of visual perception-action, or robotics, as discussed in the next section.

5. Connections with perception-action robot models

Although the principles of Polyp are radically different from those of standard models of speech perception, they are very similar to those of recent models that control the actions of robots that process an environment visually and act appropriately in that environment. These models emphasize that behaviour proceeds through constant cyclical interaction between action and perception, and that all computations are task- (or function- or context-) dependent. Two especially interesting models are those of Dana Ballard's group (e.g., Sprague *et al* 2007)

and Deb Roy's group (Roy 2005a, b). Sprague *et al* model control of a virtual agent, Walter, who navigates a virtual environment undertaking simple tasks such as following a path, avoiding obstacles and picking up litter, directing his gaze so as to manage his concurrent and sometimes competing goals. Roy and colleagues model control of a mechanical robot, Ripley, whose sensory inputs and motor outputs afford adaptive interaction with objects in its environment.

These perception-action models work with both highly abstract parameters and the detailed sensory input together. The most abstract levels provide reference coordinates that represent non-linguistic context and emphasize the function of the action—they are task- and goal-oriented, and work with 'situated meaning', i.e., with the task-relevant general context. In addition to the obvious benefits of incorporating task goals, these high-level control parameters reduce the degrees of freedom problem inherent in controlling large numbers of low-level variables or microbehaviours. What drives the processing of the low-level sensory input is knowledge that has been learned via interaction with the physical world. Hence, the knowledge is embodied, or 'grounded' in experience, and therefore mediated by memory; expectation of what is normal strongly influences perceptual decisions. In consequence, top-down and bottom-up information are not distinguishable or discrete 'stages' of processing. Thus low-level sensory processes, as well as being context-sensitive and task-oriented, depend heavily upon prediction, use attentional resources efficiently, and respond quickly to unambiguous information. Furthermore, there is no primary unit of perception, the same sensory information contributes to more than one larger unit, and larger patterns usually matter more than smaller ones, though the larger patterns are crucially dependent on properties of the smaller ones.

All these properties of visual perception-action robots are also proposed, for independent reasons, as properties of Polysp, but, interestingly, most of them play no part in the speech systems used by either robot group. Roy's Ripley robot can respond to simple speech acts, while Yu *et al* (2003) model unsupervised learning of both form and meaning of simple words from observation of a human's body movements in natural tasks (see also Yu & Ballard 2004). Yet both groups use standard speech-processing tools, HMMs (Roy) or neural nets (Ballard), to extract a phoneme string, and include undiscussed acceptance of the primacy of phonemes as the important perceptual units of speech. In consequence, the sophistication of their vision-perception-action models is not carried through into their speech-understanding models. For example, phoneme-based word segmentation in Yu & Ballard's (2004) model performs at about 69% accuracy, whereas all other components perform with 80%–90% accuracy. As noted at the outset of this paper, such speech models can work well for small vocabularies using a limited number of tasks and grammatical structures with a single speech style, but they are unlikely to generalize well to other domains, especially to real speech situations.

As the robotics models and Polysp are fundamentally compatible, combining their individual strengths could result in significant progress in both. Phonetic detail maps flexibly to knowledge and concepts, but can be informative only when it is systematically related to a communicative function—in other words, to a structure whose properties are relatable to the general structure. The robotics models lack the types of rich linguistic structure described in previous sections that allow phonetic detail to be treated as information. As such detail can potentially provide information about any level in any structure or control subsystem, to include those structures could significantly strengthen the adaptive and pragmatic use of the speech signal. Conversely, phonetic detail will be attended and responded to, to the extent that the task and ambient conditions make it functionally relevant and usable—so the same acoustic-phonetic information can produce different responses in different task situations, for example. Hay & Drager (2010)

demonstrate that phonetic decisions can even be influenced by non-linguistic experiences such as seeing national symbols. Adding to the FPA/Polysp structures the capability to deal adaptively with changed goals and changed contexts should allow systematic exploration of which phonetic detail is used in which circumstances. This should benefit both speech technology applications, and further research into phonetic and gestural influences on communicative function in interaction.

6. Conclusion

When a system is restricted to identifying phonemes from the acoustic signal, it effectively discards all information that is not directly associated with identifying phonological or lexical form. Knowingly or unknowingly, the person who uses such a system adopts a theoretical model in which each phoneme is associated with an 'essence' or set of core features, around which there is variation which, on the whole, is judged to be less important to discovering the meaning of the message. This model derives from the efforts of certain schools of linguistics to reach a parsimonious description of the formal properties of language, or of a particular language. This is a valuable goal in itself, but its uncritical transfer to models of how spoken language is produced and understood has proved to be not wholly appropriate for any except the simplest applications with very restricted functions.

Models intended to address a more normal range of functions and listening conditions may not benefit from throwing away phonemic labels entirely, which would be necessary within a pure FPA approach, but they may benefit by noting the insights offered by FPA's polysystemicity and its extensions such as Polysp and work in conversational interaction, and by adopting those that are relevant to their particular applications. A first, modest, step would be to always identify the function and goal of an utterance within its context. Subsequent steps could include mapping sound to parallel, interacting hierarchical structures. Principles used successfully in perception–action robotics models seem particularly appropriate for this process. To achieve this, collaboration between artificial intelligence engineers and researchers into the phonetics of conversation could be invaluable.

References

Allen J S, Miller J L 2004 Listener sensitivity to individual talker differences in voice-onset-time. *J. Acoust. Soc. Am.* 116: 3171–3183

Baker R 2008 *The production and perception of morphologically and grammatically conditioned phonetic detail* (Cambridge, University of Cambridge)

Baker R, Smith R, Hawkins S 2007 Phonetic differences between *mis-* and *dis-* in English prefixed and pseudo-prefixed words. *16th Int. Congr. Phonetic Sciences* W J Barry, J Trouvain (eds.) (Saarbrücken: http://www.icphs2007.de/). 553–556, Paper ID 1507

Baker R, Smith R, Hawkins S Phonetic detail that distinguishes prefixed from pseudo-prefixed words. *J. Phonetics* (under revision)

Bell-Berti F, Harris K S 1981 A temporal model of speech production. *Phonetica* 38: 9–20

Benguerel A-P, Cowan H A 1974 Coarticulation of upper lip protrusion in French. *Phonetica* 30: 41–55

Bradlow A R, Nygaard L C, Pisoni D B 1999 Effects of talker, rate and amplitude variation on recognition memory for spoken words. *Percept. Psychophys.* 61(2): 206–219

Clark J, Yallop C, Fletcher J 2006 *An Introduction to phonetics and phonology* (3rd ed.): (Oxford: John Wiley and Sons Ltd.)

Clift R 2001 Meaning in interaction: The case of actually. *Language* 77(2): 245–290

Coleman J S 2003 Discovering the acoustic correlates of phonological contrasts. *J. Phonetics* 31: 351–372
Cruttenden A 2001 *Gimson's Pronunciation of English* (6th ed.) (Latest edition of *An introduction to the pronunciation of English* by A.C. Gimson. London: Arnold)
Duffy S A, Pisoni D B 1992 Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Lang. Speech* 35: 351–389
Fougeron C 2001 Articulatory properties of initial segments in several prosodic constituents in French. *J. Phonetics* 29: 109–135
Garcia Lecumberri M L, Cooke M P 2006 Effect of masker type on native and non-native consonant perception in noise. *J. Acoust. Soc. Am.* 119(4): 2445–2454
Gaskell M G, Marslen-Wilson W 1997 Integrating form and meaning: A distributed model of speech perception. *Lang. Cognit. Process.* 12: 613–656
Gaskell M G, Marslen-Wilson W D 2001 Lexical ambiguity and spoken word recognition: Bridging the gap. *J. Mem. Lang.* 44: 325–349
Goldsmith J A 1990 *Autosegmental and metrical phonology* (Oxford: Basil Blackwell)
Goldsmith J A 1994 Disentangling autosegments: a response. *J. Linguistics* 30: 499–507
Grossberg S 2003 Resonant neural dynamics of speech perception. *J. Phonetics* 31: 423–445
Hawkins S 2003 Roles and representations of systematic fine phonetic detail in speech understanding. *J. Phonetics* 31: 373–405
Hawkins S 2010a Phonetic variation as communicative system: Perception of the particular and the abstract, in C Fougeron, B Kühnert, M d'Imperio, N Vallée (eds.) *Laboratory Phonology 10: Variability, Phonetic Detail and Phonological Representation* Berlin: Mouton de Gruyter, 479–510
Hawkins S 2010b Phonological features, auditory objects, and illusions. *J. Phonetics* 38(1): 60–89
Hawkins S, Nguyen N 2004 Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *J. Phonetics* 32(2): 199–231
Hawkins S, Smith R H 2001 Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian J. Linguistics-Rivista di Linguistica* 13: 99–188. http://kiri.ling.cam.ac.uk/sarah/TIPS/hawkins-smith-101.pdf
Hay J, Drager K 2010 Stuffed toys and speech perception. *Linguistics* 48(4): 865–892
Heid S, Hawkins S 2000 An acoustical study of long domain /r/ and /l/ coarticulation, *Speech Production: Models and Data, and CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Munich: Institut für Phonetik und Sprachliche Kommunikation, Ludwig-Maximilians-Universität, 77–80
Heinrich A, Flory Y, Hawkins S 2010 Influence of English r-resonances on intelligibility of speech in noise for native English and German listeners. *Speech Commun.* 52: 1038–1055
Hertz S R 1991 Streams, phones and transitions: toward a new phonological and phonetic model of formant timing. *J. Phonetics* 19: 91–109
Hertz S R 2006 A model of the regularities underlying speaker variation. *Proc. Interspeech* (revised version) available from http://linguistics.cornell.edu/people/Hertz.cfm
Hertz S R, Huffman M K 1992 A nucleus-based timing model applied to multi-dialect speech synthesis by rule *2nd International Conference on Spoken Language Processing: ICSLP-1992*, Banff, Alberta, Canada, 1171–1174
Jones D 1967 *The phoneme* (Cambridge: Cambridge University Press, reissued 1976, 2009)
Just M A, Cherkassky V L, Aryal S, Mitchell T M 2010 A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE* 5(1): e8622
Keating P A, Cho T, Fougeron C, Hsu C-S 2004 Domain-specific articulatory strengthening in four languages. *Phonetic Interpretation: Papers in Laboratory Phonology VI*, J K Local, R A Ogden, R A M Temple (eds.), Cambridge: Cambridge University Press, 143–161
Klatt D H 1976 Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J. Acoust. Soc. Am.* 59(5): 1208–1221
Klatt D H 1989 Review of selected models of speech perception in W D Marslen-Wilson (ed.), *Lexical representation and process* Cambridge, MA: MIT Press, 169–226

Large E W, Jones M R 1999 The dynamics of attending: How people track time-varying events. *Psychol. Rev.* 106(1): 119–159

Local J K 2003 Variable domains and variable relevance: Interpreting phonetic exponents. *J. Phonetics* 31: 321–339

Local J K 2007 Phonetic detail and the organisation of talk-in-interaction. *16th Int. Congr. Phonetic Sciences* W J Barry, J Trouvain (eds.) (Saarbrücken: http://www.icphs2007.de/). 1–10, Paper ID 1785

Local J K, Walker G 2005 Methodological imperatives for investigating the phonetic organization adn phonological structures of spontaneous speech. *Phonetica* 62: 1–11

Lodge K 2003 A declarative treatment of the phonetics and phonology of German rhymal /r/. *Lingua* 113: 931–951

Lodge K 2009 *A critical introduction to phonetics* (London: Conitinuum International Publishing Group)

Mattys S, White L, Melhorn J F 2005 Integration of multiple speech segmentation cues: A hierarchical framework. *J. Exp. Psychol.: General* 134(4): 477–500

McClelland J L, Elman J L 1986 The TRACE model of speech perception. *Cognitive Psychol.* 18(1): 1–86

McClelland J L, Mirman D, Holt L L 2006 Are there interactive processes in speech perception? *TRENDS Cognit. Sci.* 10: 363–369

Miller G A, Heise G A, Lichten W 1951 The intelligibility of speech as a function of the context of the test materials. *J. Exp. Psychol.* 41: 329–335

Mitchell T M, Shinkareva S V, Carlson A, Chang K-M, Malave V L, Mason R A, Just M A 2008 Predicting human brain activity associated with the meanings of nouns. *Science* 320: 1191–1195

Moore R K 2007 Spoken language processing: Piecing together the puzzle. *Speech Commun.* 49(5): 418–435

Norris D G 1994 Shortlist: a connectionist model of continuous speech recognition. *Cognition* 52: 189–234

Norris D G, McQueen J M, Cutler A 2000 Merging information in speech recognition: Feedback is never necessary. *Behav. Brain Sci.* 23: 299–370

Nygaard L C, Pisoni D B 1998 Talker-specific learning in speech perception. *Percept. Psychophys.* 60: 355–376

Ogden R 1993 What Firthian prosodic analysis has to say to us. *Computational Phonology: Edinburgh Working Papers in Cognitive Science* 8: 107–127

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

Ogden R A 2004 Non-modal voice quality and turn-taking in Finnish. *Sound patterns in interaction*, E Couper-Kuhlen, C Ford (eds.), Amsterdam: Benjamins, 29–62

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

Ogden R A, Routarinne S 2005 The communicative functions of final rises in Finnish intonation. *Phonetica* 62(2–4): 160–175

Ogden R A, Hawkins S, House J, Huckvale M, Local J K, Carter P, Dankovicová J, Heid S 2000 ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Comput. Speech Lang.* 14: 177–210

Piccolino Boniforti M A, Ludusan B, Hawkins S, Norris D 2010 Same phonemic sequence, different acoustic pattern and grammatical status. A model. F Cutugno, P Maturi, R Savy, G Abete, I Alfano (eds.), *Parlare con le persone, parlare alle macchine: la dimensione interazionale della comunicazione verbale. VI Convegno Nazionale AISV - Associazione Italiana di Scienze della Voce.*, Naples, Italy, 279–291

Pickett J M, Pollack I 1963 Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Lang. Speech* 6: 151–164

Pierrehumbert J 2003 Probabilistic phonology: Discrimination and robustness, in R Bod, J Hay, S Jannedy (eds.), *Probability theory in linguistics* Cambridge, MA: MIT Press., 177–228

Pisoni D B, Lively S E, Logan J S 1994 Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception, in D Goodman, H C Nusbaum (eds.), *The development of speech perception: The transition from speech sounds to spoken words*, Cambridge, MA/ London: MIT Press, 121–166

Plug L 2005 From words to actions: The phonetics of *Eigenlijk* in two communicative contexts. *Phonetica* 62(2–4): 131–145

Post B, D'Imperio M, Gussenhoven C 2007 Fine phonetic detail and intonational meaning. *16th Int. Cong. Phonetic Sciences* W J Barry, J Trouvain (eds.), (Saarbrücken: http://www.icphs2007.de/). 191–196, Paper ID 1723

Pulvermüller F 1999 Words in the brain's language. *Behav. Brain Sci.* 22: 253–336

Raposo A, Moss H E, Stamatakis E A, Tyler L K 2009 Modulation of motor and premotor cortices by actions, action words, and action sentences. *Neuropsychologia* 47: 388–396

Roy D 2005a Grounding words in perception and action: Computational insights. *Trends Cognit. Sci.* 9(8): 389–396

Roy D 2005b Semiotic schemas: A framework for grounding language in action and perception. *Artif. Intell.* 167(1–2): 170–205

Sprague N, Ballard D, Robinson A 2007 Modeling embodied visual behaviors. *ACM Trans. Appl. Percep.* 4(2): Article 11

Tily H, Gahl S, Inbal A, Snider N, Kothari A, Bresnan J 2009 Syntactic probabilities affect pronunciation variation in spontaneous speech. *Lang. Cognit.* 1–2: 147–165

Turk A, Shattuck-Hufnagel S 2000 Word-boundary-related duration patterns in English. *J. Phonetics* 28: 397–440

Tyler L K, Randall B, Stamatakis E A 2008 Cortical differentiation for nouns and verbs depends on grammatical markers. *J. Cognit. Neurosci.* 20(8): 1381–1389

West P 1999 Perception of distributed coarticulatory properties of English /l/ and /ɹ/. *J. Phonetics* 27(4): 405–426

Wiese R 1997 Underspecification and the description of Chinese vowels, in J L Wang (ed.), *Studies in Chinese phonology* Berlin: Mouton de Gruyter, 219–249

Yu C, Ballard D H 2004 A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Trans. on Appl. Percept.* 1(1): 57–80

Yu C, Ballard D, Aslin R N 2003 The role of embodied intention in early lexical acquisition. *Meeting of the Cognitive Science Soc.* Boston, MA