DOES PHONETIC DETAIL GUIDE SITUATION-SPECIFIC SPEECH RECOGNITION?

Sarah Hawkins

Centre for Music and Science, University of Cambridge, UK
sh110@cam.ac.uk

ABSTRACT

The evidence is overwhelming that listeners use both statistical patterns about abstract properties of their language, and its phonetic detail, to make informed predictions and decisions. Listeners seem to learn about new phonetic detail when it does not contradict other important cues to communicating meaning. Though we do not always use available phonetic detail, we probably monitor it constantly. An eye-tracking experiment shows that we use weak, subphonemic information to predict the morphemic status of words. This and other data suggest that we use phonetic detail when it is relevant to the task at hand, even when there is little apparent advantage in doing so. Since perceptual learning about speech seems central to interpreting or systematising a speaker's or group's use of phonetic detail, these various observations encourage speculation about how to model speech perception as part of a biologically-grounded, situated theory of human interactive behaviour.

Keywords: phonetic detail, active perception, learning

1. INTRODUCTION

ICPhS-2007 included two special sessions which addressed abstractionist vs. episodic perception, and roles of phonetic detail in communication. In this paper, I try to draw these overlapping threads together by discussing perceptual learning of phonetic detail, and some of its implications for speech perception as part of a general theory of situated human communication and interaction.

The episodic-abstract debate seems essentially a polarisation of the need to balance two facts. The first is that speakers provide, and listeners exploit, all sorts of phonetic detail that does not normally figure in phonology, nor serve to distinguish the citation forms of lexical items. This we called Fine Phonetic Detail (FPD), but in Saarbrücken, where we tried to define FPD, we argued that a better term is Phonetic Detail (PD), because much of it is easily audible [8, 18]. We argued that, when we have moved beyond the type of debate exemplified by episodic-abstractionist polarisation, PD might be renamed Phonetic Information because it is linguistically-structured information that can make or break successful communication, but that might not affect communication in another context, and may not be central to distinguishing phonemes in citation-form words [8]. The second fact is that listeners do generalise from exemplars or episodes to what we can describe as more abstract or superordinate forms. But all animal behaviour involves generalisation: it is adaptive, necessary to survival. The issue for me is not that we generalise, but when and what we generalise, and when and why we do not. In other words, the crucial issue is to know what to do with phonetic detail in models of human communication (see [15] esp. p.481-484).

I start from the claim that it is now abundantly clear that the details of the speech signal exert strong influences on most if not all aspects of speech communication, see e.g. *Phonetica* Vol. 6, and [15, 20, 21, 30, 31, 33, 37, 51]. I discuss this claim with respect to studies that directly or indirectly address perceptual learning about PD. Section 3 outlines a very few non-speech factors that affect spoken communication. Section 4 describes an experiment that shows listeners use FPD to predict a word's morphemic structure. Section 5 outlines some of the general theoretical principles.

2. PERCEPTUAL LEARNING

2.1. Learning about imperfect speech

The rapidity of perceptual learning for speech has long been known about, but is enjoying renewed interest. Categorical perception studies began with practice trials until responses stabilized. Usually 6-10 were enough. So they explored classification of sounds when listeners were familiar with the range of variation. Nonetheless, especially in the 1970s and 1980s, a good deal of attention was devoted to how much and by what means phoneme category boundaries could be induced to shift [43].

Critical appraisal of the role of the phoneme in speech perception has encouraged researchers to revisit this question, this time under the name of perceptual learning, a term which makes explicit connections with non-speech perceptual processes. The earliest of these new studies used some form of distorted speech, thereby exploring what we all have personal experience of: that, with exposure to an unusual pronunciation, we tend to stop noticing it and have less trouble understanding the speaker.

Norris et al. [38] showed that after hearing a word list in a voice that has an ambiguous fricative in place of /s/ or /f/, listeners shift their category boundary for /s/-/f/ in a direction appropriate for that voice. They proposed that perceptual learning operates over phonemic representations, the driving force being knowledge of lexical meaning.

Later work using vocoded speech indicated that lexical access is not always necessary [23], and that generalisation is poor across types of vocoder source excitation, but good across differing frequency bands [24], indicating that learning remains close to the physical signal yet abstracts patterns that can be generalised to other stimuli.

Crucially, by tightly controlling the phonemic structure of isolated words and nonwords, which [23] had not, Dahan and Mead [10] convincingly conclude not only that lexical access is not needed for phonological perceptual learning, but that the primary linguistic level of learning is allophonic—in fact, coarticulatory: like speech errors, codas generalise better to codas than to onsets, and vice versa, and it helps to have the same vowel context. They proposed “that the process by which adult listeners learn to interpret distorted speech is akin to building phonological categories in one’s native language…categories and structure emerge from the words in the ambient language without completely abstracting from them.” [10]: Abstract. This is consistent with my position e.g. [14, 15, 20].

2.2. Perceptual learning about more natural allophonic detail

Distorted speech stimuli allow good experimental control and can isolate factors that are difficult to examine in natural speech. But with a signal as complex as speech, it is important to discover what listeners normally attend to. This section reviews a few learning studies that exploit natural allophonic variation in isolated words and connected speech.

Perceptual learning in classification tasks can use sub-phonemic information e.g. [2]. But studies that used long-term repetition priming to investigate whether specific allophonic details are retained in memory provide mixed results. In long-term repetition priming, pairs of tokens that either share, or do not share, some aspect of phonetic form, are presented, separated by many intervening tokens before a response is required. If priming is not found for those tokens that share the same aspect of form, it is assumed to have been forgotten. McLennan, et al. [35] studied flapping. They used tokens of words like *atom* and *Adam*, which contained either an intervocalic flap, or non-flapped [t] or [d]. Flapped *atom/Adam* primed careful *atom/Adam* in most of the repetition priming tasks used, suggesting that allophonic information was not being used. In contrast, when the task was an easy lexical decision (the nonwords were intentionally very un-wordlike e.g. /jʌʃðʌtʃ/), priming was sensitive to the specific allophone in the prime. [35] suggest that allophonic detail is used only when the judgment task is easy or captures performance in early stages of processing.

McLennan et al.’s account of why allophonic detail should affect early and shallow processing more than late or deep processing is compelling for the particular stimuli and allophonic variants used. However, it seems to lack generality. First, results of intelligibility in noise experiments suggests that PD is indeed used when the task is difficult cf. [18, 22, 27], and though we know little about whether PD benefits early or late decisions in these paradigms, they do allow late processing. Second, allophonic effects could arise in tasks implicating deeper processing, if different manipulations of allophonic detail were used. In natural speech, there can sometimes be subtle differences in flapped tokens according to the underlying voicing status of the consonant [28]. In McLennan et al.’s experiments, the same flapped tokens were used to instantiate both /t/ and /d/ and indeed were chosen as the most ambiguous from a set of flapped tokens. In consequence, a flap was uninformative about voiced/voiceless status, and the reason it did not play a role in tasks requiring a deeper level of processing may have been because it did not actually help listeners to do the task.

Similar considerations apply to Sumner and Samuel [52]’s study of whether priming is affected by whether words are pronounced with variants of word-final /t/: [t], glottalised and unreleased [ʔt̚], and glottal stop [ʔ]. Immediate semantic priming was not affected by allophonic variant: *flute* with

[ʔt̚] or [t] primed *music* equally well. But in long-term repetition priming with a lexical decision task, strong priming was only found with [t]: if the first presentation of *flute* had [t], lexical decision to its second presentation was faster if the second presentation also had [t]. If the first presentation had [ʔt̚] or [ʔ], response to the second presentation of *flute* was not faster, even if both presentations had the same variant. The authors concluded that information about non-canonical allophones (those other than [t]) is not held in long-term memory, at least for tasks about processing isolated words.

Again, [52]'s negative results may be due to the role of allophonic detail in these tasks. Allophonic detail can signal aspects of linguistic structure, and therefore assist in understanding meaning, but it does not always do so. In [52], the word-final variants of /t/ did not change the word's meaning. English plosive releases inform about position in word, speech style, position in conversational turn, and speaker indices, but these factors are not relevant in [52]'s design, which presented isolated words. So there was little reason for listeners to remember details of plosive release. Instead, the key task-relevant information was the phonemic identity of the final consonant, since filler items were nonsense words like *floop*. This may explain why only [t], which would have the strongest place of articulation cues, showed long-term priming. Linguistic relevance may be a crucial factor in the extent to which systematic variation is retained and used e.g. [39, 49].

None of the above studies have tested whether speaker-specific allophonic variation is learned about when it informs about contrasting linguistic structure that changes meaning. Smith [48] found listeners learned speaker-specific pronunciations of allophonic cues to word boundaries (and hence also syllable position), and exploit this learning in difficult listening conditions.

2.3. Limits to perceptual learning in speech perception

Animals need to structure their sensory world in order to interact effectively with the physical world. This requires a fine balance between assuming, and thus predicting, that many things tend to be stable, even if they seem different (perhaps in different lighting conditions) yet adapting appropriately to new conditions even if they are unexpected. To avoid chaos, adaptation must be limited. Such adaptation may be short-term or long-term, and the organism may not know which is the case when first confronted with the new sensory input. So part of skilful adaptation requires assessing whether the conditions should be treated as new, or a temporary aberration. This is an acute issue for people listening to speech, because there are so many possible parameters of lawful and indeed helpful variation, including speaker-indexical differences, yet speakers do also make mistakes, sometimes more than once.

One challenge facing our understanding of the role of adaptation and learning in speech is thus to establish what discourages perceptual learning. There are some obvious factors, including random variation in a stimulus parameter. More interesting is what is not learned after systematic exposure to an easily audible, natural cue. Since this sort of approach tests the null hypothesis, there is little published literature on this topic.

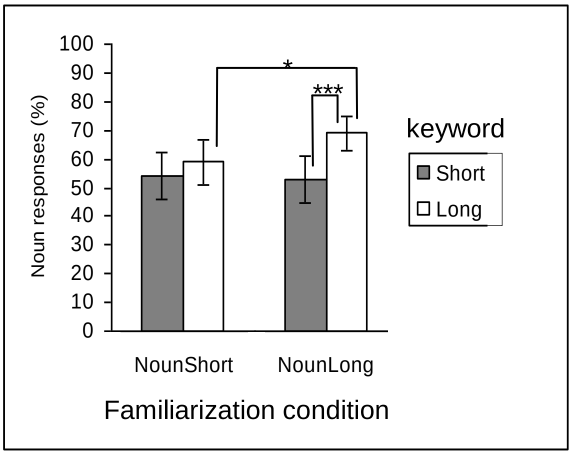
However, Barden [6] explored this issue, trying to approximate real-life conditions. One method she uses is to familiarize listeners with unusual pronunciations in a long story reporting adventures of a fictional eccentric naturalist/philanthropist. The chosen parameters might be due to a regional or idiosyncratic accent, and can be acoustically manipulated to produce other novel and familiar versions. This is harder than it sounds: many variables not of interest to the experiment must be controlled, and attention must be held (Barden tests comprehension). To assess learning, sections of the story that participants hear (familiarization) are interleaved with tests requiring an active response.

In one study, Barden asked whether listeners would adapt to durational differences between nouns and verbs. Stressed vowels were lengthened by 33% in the 326 nouns and shortened by 25% in the 326 verbs in the story, or vice versa.

To assess learning, 20 sentence beginnings like *The French cook....*, in which *cook* can be a noun or a verb, were used. These had been recorded as full, unambiguous sentences, matched for nuclear stress position, syllable number etc e.g. *The French cook attested it was pepper, The French cook a tasty sort of pepper.* A trained phonetician read the story and sentences, the latter with an intonation that is heard with both grammatical forms (H*L). Each sentence was truncated after the third word *(cook*), called the keyword, and made into two sets of test stimuli by changing the keyword's stressed vowel duration: 33% longer, and 25% shorter.

In a fully-crossed design, 56 listeners heard the story, 28 with nouns long (NounLong), 28 with verbs long (NounShort). Training and test were interleaved 10 times: 5 minutes for the first part of the story, and about 52 s for each of the other nine parts. At test, listeners heard a truncated sentence and typed an ending as quickly as possible, without regard to its plausibility.

Figure 1: Percent noun responses in the sentence completion task, by keyword length and familiarization condition. Error bars: 95% confidence re mean. Pairwise comparisons: * p < 0.05; *** p < 0.001. From [6].



Responses (Fig. 1) were scored for the keyword being interpreted as a noun or verb. An interaction was predicted: NounLong listeners should interpret more long keywords as nouns and more short keywords as verbs; the opposite for the NounShort group. The interaction was significant in a mixed-model analysis ($\chi^2(1)$ = 6.2, p = 0.02), but only NounLong listeners had the predicted pattern. NounShort differences were nonsignificant.

Barden reasoned that the interaction was not as predicted because the test confounds grammatical class and position in the syntactic phrase. English, nouns are always at the end of a syntactic noun phrase, while verbs may or may not end the verb phrase (*The boy draws*; *The boy draws the dog*). So listeners could have learned an association between duration and word class, or duration and phrase position, i.e. perhaps they learned that the speaker marks phrase ends by exaggerated lengthening (NounLong) or by no lengthening (NounShort).

A follow-up study confirmed that this interpret-ation is probably right. It was identical to the first study except that keywords were syntactic-phrase-final & followed by another word: *I made [Peter/ Pete a] plan although...*The keyword is *plan*; the bracketed words are the two forms of the sentence pair. Responses were at chance for all 4 conditions.

This suggests that people adapt to new speech patterns within pre-existing frameworks cf. [7]. A pattern that violates a powerful principle is less likely to be learned fast. English phrase-final lengthening is not just strong but helps resolve syntactic ambiguities e.g. [29, 42]. The noun-verb distinction, in contrast, is accompanied by many phonetic differences which can affect perception (e.g. stress pattern, syllable number, vowel quality, distribution of phonological features, see [46] and references therein), but none by itself strongly cues grammatical class. Taken together, they do [11, 12], but Barden did not manipulate them.

In Barden's studies, then, listeners adapted to durational differences reflecting prosodic structure, and this knowledge influenced their interpretation of syntactic structure. But it did not override strong prosodic rules. What Barden has shown is a limit to plasticity that makes perfect behavioural sense.

3. TALKING IN EVERYDAY SITUATIONS

Section 2 shows that perceptual learning in speech is an adaptive response to task demands, guided and limited by how the listener weights existing knowledge. This section briefly reviews how listeners react to information in a broader context.

Participants in a conversation engage in a wide range of speech and non-speech behaviours, apparently unconsciously, that serve to align each other more closely (and presumably sometimes less closely). These range from phonetic properties expressing pragmatic functions [32, 40] to imitative behaviour that indicates attitudes and group cohesion, only some of which involve speech.

Hay and Drager [21] show that even apparently inconsequential exposure to national symbols can affect judgments of vowel quality, while Babel [3] showed that personal attitudes and liking affects the degree of vowel accent mimicry.

We have long known that seeing lips move can influence phoneme perception [34] and facilitate intelligibility, though appreciation of the many factors influencing this process are only recently becoming understood [25]. An increasing body of research is now showing that head, hand and other body gestures are coordinated with speech to enhance and indeed at times change perceived meaning. When head nods and eyebrow raising align with prosodically prominent syllables, speech intelligibility increases [1]. Recent work [47] shows that a speaker's hand gestures can reflect whether the listener has understood, and may be a more reliable index of understanding than words.

These types of work serve the dual purpose of (1) giving phonetics a determining role in a general theory of human communication, rather than being simply a carrier of information whose details have no intrinsic importance, and (2) making more plausible the hypothesis that meaning is represented dynamically and multimodally in the brain. These points are returned to in Section xx.

4. EYE-TRACKING FINE DETAIL OF A MORPHOLOGICAL DISTINCTION

I proposed a way of conceptualizing how speech is processed, called Polysp [14, 15, 20]. Like many other models, Polysp emphasises prediction and exploitation of multiple sources of information, both sensory and knowledge-based. Perhaps its most important claim is that phonetic detail is mapped directly to the relevant linguistic unit(s) rather than obligatorily going through intermediate stages: if the sensory signal fits expectations for the auditory pattern of a higher-order unit, then that unit can be identified before, or at the same time as, lower-order units. This section reports an eye-movement experiment that tests this claim.

I discussed the claim with respect to the distinction between true and pseudo-prefixes that have the same phoneme sequence, as in *mistimes* vs *mistake*s and *discolour* vs *discover*. The distinction is standard in linguistics, and reflects a difference in morphological productivity: *mis*- and *dis*- are productive (true) prefixes in *mistimes* and *discolour* (*times/colour* mean roughly the opposite when *mis-/dis*- is added) whereas they are not prefixes in *mistakes* and *discover* (*takes/cover* do not mean the opposite). When they are pseudo-prefixes, *mis*- and *dis*- are part of the stem, though they share the same initial phoneme string with the productive prefixes. The phonemic identity may be because these syllables were once true prefixes, but they are no longer for today's speaker-listeners.

Acoustic differences have been discussed many times [14, 15, 20] so are only summarised here. Though the first four phonemes are the same in each pair, the first syllables differ in rhythm due to small differences in the acoustic properties of their component segments. Both are metrically weak; pseudo prefixes are more reduced. One reliable difference is the duration of aperiodicity for [s] relative to the duration of periodicity of [ɪ]: about 1:3 in true prefixes, but 1:6 in pseudo prefixes [5]. Another is that when a voiceless stop begins the second syllable of the word, its VOT is long after the true prefix, and short after the pseudo prefix.

Combined, these properties create a heavier beat on the true morphemic syllables, but most people are unaware of the difference until it is pointed out to them, and many non-native speakers tell me they cannot hear it. Further, the difference is somewhat malleable: while clear in spontaneous speech, it may be much reduced in read speech because, except for VOT, the pseudo prefix is sometimes read with a heavier beat. Nonetheless, the difference is reliable enough to be perceptually salient in intelligibility-in-noise tests [4], and to allow acceptable automatic classification [41].

The true vs pseudo-prefix distinction thus makes an especially apposite test of the claim that phonetic detail can be interpreted in terms of levels of linguistic structure 'higher' than the phoneme. The difference is present in clear speech, hard to notice and perhaps to hear (both types of syllable are unstressed) and as noted, the phonemes in the target syllables are identical. Near-minimal pairs exist in the later syllables, allowing good control of prosodic and segmental structure. Eye tracking is a good test, because eye movements reflect on-line prediction, and are sensitive to phonetic detail e.g. assimilation [13], prosodic phrase position [44], and gradience in probability of occurrence [9, 36]. But in those studies, subjects identified lexical items to which there are either no other available cues, or no especially dominant ones. In contrast, subjects in the present study do not need to listen closely, nor to use the fine detail of unstressed *mis*- or *dis*- to distinguish the words. They can merely wait until the second syllable, when all will become abundantly clear. So there is little payoff in this experiment for early prediction. Thus, if eyes do move in the right direction as the target syllables are heard, this would be strong evidence for a role for prediction that exploits knowledge of morphological structure affecting syllabic fine detail, rather than phonemes or words.

A further level of interest was built in. While *mis*- and *dis*- words share at least the same first 4 phonemes, the prefixes *re*- and *ex*- follow the same rules, but syllabic reduction in the pseudo prefix changes the vowel phoneme in the target syllable:

repeal /rɪˈpi:l/ or /rəˈpi:l/, but *re-peel* /ri:ˈpi:l/; *extravagance* /əksˈtravəgəns/ but *ex-trampoliner* /əksˈtrampəli:nə/. Thus *dis- mis-* true vs pseudo syllables contrast in morphemic but not phonemic status; *re- ex-* syllables contrast in both.

4.1. Method

Participants (Ps) were 30 native English speakers at the University of York (21 women), mean age 21 yrs, range 18-32, normal or corrected-to-normal vision, no history of speech or hearing problems,.

128 stimuli were made from 32 pairs of target words differing in true vs pseudo prefix status of their first syllable: 7 *dis-* (*discolour/discover*), 4 *mis-* (*mistypes/mistakes*), 16 *re-* (*re-store/restore*), and 5 *ex-* (*expatriate/expatiate*). Each word was placed in a prefix sentence which was identical to its pair before the target word, and sometimes after it, and could be illustrated by a picture: e.g. *It was difficult because Sam distrusted/distracted him*. Sentences that differed after the target word had the same prosodic structure and number of syllables: *A swan displaces water when it lands, A swan displays its plumage to its mate.*

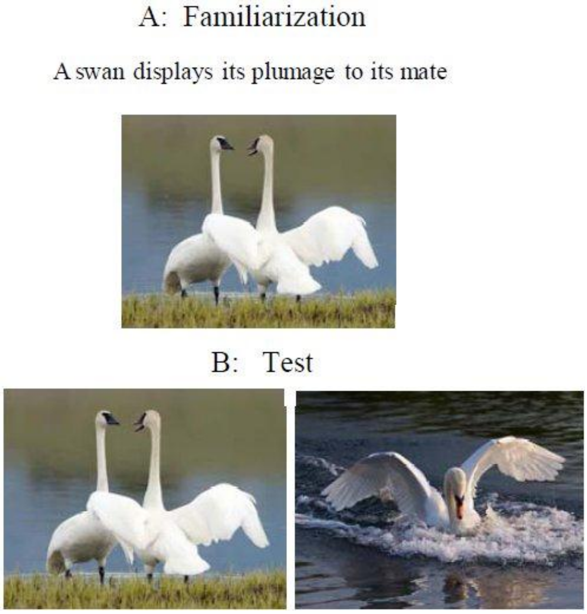
These 64 sentences were recorded in 6 random orders by a near-RP speaker while he looked at its picture, to minimise reading effects. Some of these were then spliced to make a ‘match’ and a ‘mismatch’ for each member of a pair, as follows.

An original sentence was cut at the end of the target syllable (or just before the burst of the next stop, if there was one) and spliced either to the rest of another token of the same sentence (match), or else to the rest of another token of the paired sentence (mismatch), making 4 sentences from an original pair. E.g. To make a match sentence for *A swan displaces water when it lands,* the beginning of one token, *A swan disp-* was spliced to the end of another token of the same sentence, *-laces water when it lands.* To make a mismatch sentence, that same sentence beginning was spliced to *-lays its plumage to its mate* from a token of *A swan displays its plumage to its mate.* Likewise, a match and a mismatch sentence were made from this second original sentence (*A swan displays its plumage to its mate*). Thus one sentence pair produced 4 experimental sentences, totalling 128 stimuli (32 pairs x 2 match x 2 mismatch). They were chosen from the 6 tokens for naturalness and the best match of f0, rhythm and loudness.

There were 67 fillers. 30 were ‘r-res’ sentence pairs used in another experiment (same speaker) [22]. Pairs were identical except for one word, which in turn differed only in whether it contained /r/ or /l/ e.g. *rams/lambs*. Matches and mismatches were made using the same principles as above. The other 37 fillers pairs mostly mimic the prefix ones in some way e.g. *We liked the description of the fantastical dragon, We liked the description of the balloons over mountains* have pseudo *dis-* four syllables before the apparent target word. These were recorded twice; the most natural of each was chosen, and not spliced.

The procedure was novel because the stimuli were complicated. Stimuli were blocked in 33 sets, each of 6 familiarization (F) trials and 3 test (T) trials—1 prefix, 1 r-res, and 1 filler sentence. Each F trial presented one picture with the descriptive sentence above it (Fig. 2A). Ps read the sentence silently; there was no sound. They clicked on the picture to move to the next one; minimum display time was 2.5 s. Each T trial began with calibration, then 2 of the previous 6 pictures appeared, with no writing (Fig. 2B). After 2 s one of the 4 possible sentences was played, and Ps clicked on the correct picture as quickly and accurately as possible, with ‘correct’ being the sentence that is heard after the splice point. Each picture was replaced by a blank screen 0.5 s after P clicked on it, before the next trial.

Figure 2: Screen displays in the eye movement study.



The design was as follows. All matched stimuli were presented on one day, and all mismatched on another; all fillers appeared on both days. Order of presentation was counterbalanced between two groups of 15 Ps. On Day2, half the Ps heard the same picture described as on Day1 (but different (mis)match); half heard the other sentence.

Ps were tested alone in a quiet room with an Eyelink 2000 remote desktop and headphones. Tests were 6-8 days apart for most Ps, longer for a few. Sessions took 45-50 minutes, including two practice runs that used extra stimuli of each type.

4.2. Results

Figure 3 shows the main results, for *re- ex-* in the top panel, and *mis- dis-* below. Proportion of looks to the true (tr, *displaces*) or pseudo (ps, *displays*) target picture is plotted against time, aligned at the splice point (0 ms); for clarity, only the time window of interest is shown. Looks above the y-axis zero are to ps targets; looks below it are to tr targets; so as the lines break away from this zero, listeners are looking towards a target. Vertical dotted lines show the splice point and 200 ms before and after it. The lag between stimulus onset and a look is about 200 ms. Since target syllables also begin about 200 ms before the splice point on average, looks between 0 and +200 ms mainly reflect the target syllable; looks after +200 ms begin to show post-splice influences. Responses to hearing matched stimuli are labelled tr-tr for true and ps-ps for pseudo target words. Responses to hearing mismatches are ps-tr and tr-ps: the first part is prefix type, the second is target word type.

As expected, ps-ps and tr-ps rise towards the right hand side of each panel, while tr-tr and ps-tr fall. This reflects when the sentence continuation has its effect, from the target's second syllable. For matched *re-ex-* targets, these changes start early and are gradual; they are later and more abrupt for the mismatched targets tr-ps and ps-tr. For *dis-mis-* the main change towards the target is abrupt for all conditions, with less variation in when it starts. But there are weak trends in expected directions in the 0 to +200 ms window, and the mismatched curves lag behind the matched curves after +200 ms (tr-ps later than ps-ps; ps-tr later than tr-tr).

These points are supported by linear mixed model analyses of the data aggregated over each trial, for trials in two time windows (0 to 800 ms, and 200 to 600 ms) and transformed into log odds. The larger window gives a general picture while the smaller, later window offers a closer look at interference due to mismatching. It is used because looks to the target are only reliable after 200 ms, when listeners start responding to information after the splice. So looks after 200 ms should mainly reflect the clear, post-splice target information. Hence, a delay for mismatch relative to match after 200 ms should be due to conflicting information heard before the splice. Significance levels are one-tailed as hypotheses are directional.

In the 0 to 800 ms window, *re-ex-* differed from *dis-mis-* (t=2.37, p<0.01); and match from mismatch in a model without the nonsignificant match/type interaction (t=3.08, p<0.001). As predicted, listeners more reliably use the prefix type for *re-ex-* (t(1128)=2.74, p=0.003) than *dis-mis-* (t(627)=1.59, p=0.057). But disruption due to mismatched *dis-mis-* lingers into the 200-600 ms window: after the wrong prefix, looks to target are delayed (t=1.87, p=0.03; t-tests over trials t(663) = 1.79, p<0.04; over subjects t(115)=1.86, p<0.03).

4.3. Discussion of eye movement experiment

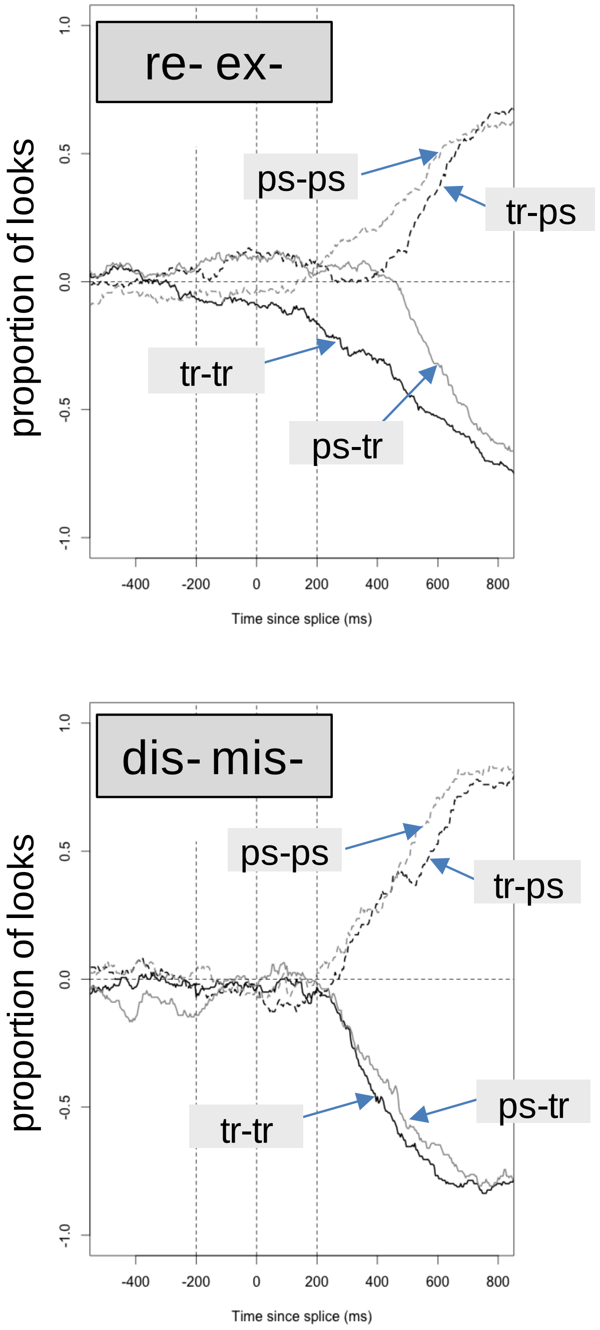
These preliminary analyses confirm that very fine detail that reflects only morphological status does guide on-line perceptual processes even in good listening conditions. Coarticulatory cues from the following syllable are too weak to allow the competitor word to be identified above chance when the splice is before the burst [4].

That the differences are more reliable for *re-ex-* can be interpreted as due to the presence of both morphological and phonemic information, whereas *dis-mis-* contrasts in only one class, morphological. Indeed, this was the reason for the expected difference between the two sets. However, there are also greater acoustic differences in the *re-ex-* set than the *mis-dis-* set. The range of durations in the critical syllables is greater, and there may be early cues in that there seems to be some prosodic reorganisation of unstressed syllables before *re-*. This may explain the gradual, early start for *re-ex-*. So although we can attribute the 'better' results for *re-ex-* to more information distinguishing true from pseudo sets, we cannot in these data distinguish influences of an abstract phonemic contrast from influences of acoustic differences that may stem from other causes, e.g. prosodic restructuring in a run of unstressed syllables before a very weak *re-*.

The reason for the weaker *dis-mis-* effect may be because it is indeed hardly attended to. After all, as noted above, we are looking for a small effect from a small contrast in acoustic pattern over a short stretch of signal that listeners need take no notice of, for the following syllable disambiguates.

Further, though the sentences were picturable, the scenes were often complex because the target words are not easily imageable. Looked at this way, it is surprising we see anything at all. However, the small effect immediately after the splice point in Fig. 3 obscures the fact that the two subject groups differed: on Day1, those hearing mismatches looked to the target faster than those hearing matches. So the weak mismatch effect many reflect subject differences.

Figure 3: Looks to True & Pseudo target words. See text.



5. CONCLUDING REMARKS

I suggested in the Introduction that a crucial issue is what to do with phonetic detail in models of how humans process speech. Part of the answer seems to be to pay attention to situational and functional context, or (only partly facetiously) the natural habitat of the detail in question. Section 2 shows that listeners who know a great deal about phonetic detail do not necessarily use it, in behaviourally measurable ways, unless it is relevant to the task. One message is to make tasks relevant to the detail. Using allophones that typically mark discourse functions or speaker identity to test identification of isolated words tells us what listeners do with those allophones in that type of situation, but not what listeners do with them in their natural habitat: detail may be situation-specific.

Another part of the answer seems to be to pay attention to phonetic context and knowledge about phonetic distributions. This is not the same as the natural functional habitat described above. There, the sensory signal is being interpreted in a time-dependent process that involves individuals and events in addition to the listener and signal. Here, we have probabilistic distributions based on the listener's own past experience: knowledge about transitional probabilities between linguistic units, coarticulation, syllable structure etc. Sect. 2 gives examples for acoustic correlates of grammatical class [46] and syllabic context [10]; there are many others. These distributions are comparatively static over the course of a few utterances in familiar situations, but, as we know, they can change very fast in some situations. Section 2 provides two messages about such linguistic knowledge-sources: co-occurrences between several factors, each one only weakly predictive of some linguistic contrast, can produce strong effects on perceptual decisions and learning; and strong rules, such as those for phrasal rhythm, may constrain and refocus restructuring that arises from new experiences.

These cautionary remarks must themselves be contextualised. Sections 3 and 4 try to do just that. Section 3 shows that the same types of principles apply in other kinds of communication: seemingly infinite varieties of influence, yet more system than chaos if one looks in the right framework. Section 4 shows that listeners do use subtle detail in real time to predict meaning even when not forced to, and, crucially, that detail may distinguish contrasts that are non-phonemic yet still central to linguistic (as opposed to interactional) structure. This, to the best of my knowledge, is the first demonstration of this point, and it validates one of Polysp's strongest claims. The demonstration probably required connected speech, and may require the task to be tied to situated meaning, rather than lexical identification.

The implication is that listeners exploit details of the signal pragmatically, depending on the type of speech, ambient listening conditions, and their construal of the task. Key concepts are memory, attention, prediction, and pattern matching [16, 17]. This very active process uses multiple sources of information—expectation, multi-modal sensation. It may involve embodiment (storing memories or concepts in terms of the modalities in which you experience them) and entrainment of one's own neural rhythms to rhythms from another source e.g. another person's speech. Such processes involve complex, distributed, dynamically recombining circuits in the brain. We are very far from being able to model this for speech. Starts have been made in several disciplines, but there is typically an inverse relationship between the complexity and sophistication of the system, and the amount of speech/language accounted for. Some thought-provoking recent studies are [26, 45, 50]; other references can be found in my papers cited here.

Clearly, processes of speech perception are fundamentally situation-specific. But this paper's title asks whether phonetic detail guides situation-specific speech recognition. The answer is yes, very often e.g. in social interaction. Phonetic detail is less necessary in isolated word recognition, yet listeners still monitor it and learn from it. Further, social interaction is a type of situation. So it may be more accurate to say that how listeners see the situation guides their use of phonetic detail.

6. ACKNOWLEDGEMENTS

I am grateful to many people for allowing me to report their work. Section 2 is adapted from work by R. Smith [48] and K. Barden [6]. The eye-tracking work is being done with M. Clayards, G. Gaskell, M. Piccolino-Boniforti. I thank E. Tham for running subjects, and G. Altmann for advice and the use of his eye tracker. I thank G. Docherty, R. Ogden, and members of S2S for education and inspiration. All errors are my responsibility.

7. REFERENCES

[1] Al Moubayed, S., Beskow, J., House, D., Granström, B. 2010. Audio-visual prosody: Perception, detection, and synthesis of prominence. In Esposito, A.E.A. (ed.), *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*. Berlin: Springer, 55-71.

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

[3] Babel, M. 2010. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *http://faculty.arts.ubc.ca/mbabel/Babel_dissarticle_manuscript.pdf*

[4] Baker, R. 2008 *The Production and Perception of Morphologically and Grammatically Conditioned Phonetic Detail*. Ph.D. thesis, University of Cambridge.

[5] Baker, R., Smith, R., Hawkins, S. Submitted. Phonetic detail that distinguishes prefixed from pseudo-prefixed words. *J. Phonetics*.

[6] Barden, K.J. 2011. *Perceptual Learning of Context-Sensitive Phonetic Detail.* Ph.D. thesis, U. Cambridge.

[7] Best, C.T., McRoberts, G.W., Goodell, E. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *J. Acoust. Soc. Am.* 109, 775-794.

[8] Carlson, R., Hawkins, S. 2007. When is fine phonetic detail a detail? *16 ICPhS* Saarbrücken, Paper ID 1721.

[9] Clayards, M., Tanenhaus, M.K., Aslin, R.N., Jacobs, R.A. 2008. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108, 804-809.

[10] Dahan, D., Mead, R.L. 2010. Context-conditioned generalization in adaptation to distorted speech. *J. Exp. Psych: HLM* 36, 704-728.

[11] Durieux, G., Gillis, S. 2001. Predicting grammatical classes from phonological cues: An empirical test. In: Weissenborn, J., Höhle, B. (eds.), *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition*. Amsterdam: John Benjamins, 189-229.

[12] Farmer, T.A., Christiansen, M.H., Monaghan, P. 2006. Phonological typicality influences lexical processing. *Proc. Nat. Acad. Sci.* 103, 12203-12208.

[13] Gow, D.W., McMurray, B. 2007. Word recognition and phonology: The case of English coronal place assimilation. In Cole, J., Hualde, J.I. (eds.), *Papers in Laboratory Phonology 9: Phonology and Phonetics*. Berlin: Mouton de Gruyter, 173-200.

[14] Hawkins, S. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. *J. Phonetics* 31, 373-405.

[15] Hawkins, S. 2010. Phonetic variation as communicative system: Perception of the particular and the abstract. In Fougeron, C., Kühnert, B., d'Imperio, M., Vallée, N. (eds.), *Laboratory Phonology 10: Variability, Phonetic Detail and Phonological Representation*. Berlin: Mouton de Gruyter, 479-510.

[16] Hawkins, S. 2010. Phonological features, auditory objects, and illusions. *J. Phonetics* 38, 60-89.

[17] Hawkins, S. 2012 To appear. The lexicon: Not just elusive, but illusory? In Cohn, A.C., Fougeron, C., Huffman, M.K. (eds.), *The Oxford Handbook of Laboratory Phonology*. Oxford: Oxford University Press.

[18] Hawkins, S., Local, J.K. 2007. Sound to sense: Introduction to the special session. *16th ICPhS* Saarbrücken.

[19] Hawkins, S., Slater, A. 1994. Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *ICSLP-1994*, 1, 57-60.

[20] Hawkins, S., Smith, R.H. 2001. Polysp: A polysystemic, phonetically-rich approach to speech understanding. *Italian J. Linguistics* 13, 99-188.

[21] Hay, J., Drager, K. 2010. Stuffed toys and speech perception. *Linguistics* 48, 865-892.

[22] 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 Communication* 52, 1038-1055.

[23] Hervais-Adelman, A.G., Davis, M.H., Johnsrude, I.S., Carlyon, R.P. 2008. Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *J. Exp. Psych.: HPP* 34, 460-474.

[24] Hervais-Adelman, A.G., Davis, M.H., Johnsrude, I.S., Taylor, K.J. 2011. Generalization of perceptual learning of vocoded speech. *J. Exp. Psych.: HPP* 37, 283-295.

[25] Jiang, J., Bernstein, L.E. 2011. Psychophysics of the McGurk and other audiovisual speech integration effects. *J. Exp. Psych.: HPP* doi: 10.1037/a0023100.

[26] 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, e8622.

[27] Kalikow, D.N., Stevens, K.N., Elliott, L.L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337-1361.

[28] Kwong, K., Stevens, K.N. 1999. On the voiced-voiceless distinction for writer/rider. *Speech Communication Group Working Papers-RLE* MIT, 1-20.

[29] Lehiste, I., Olive, J.P., Streeter, L.A. 1976. Role of duration in disambiguating syntactically ambiguous sentences. *J. Acoust. Soc. Am.* 60, 1199-1202.

[30] Local, J.K. 2003. Variable domains and variable relevance: Interpreting phonetic exponents. *J. Phonetics* 31, 321-339.

[31] Local, J.K. 2007. Phonetic detail and the organisation of talk-in-interaction. *16th ICPhS* Saarbrücken, Paper ID 1785.

[32] Local, J.K., Walker, G. 2005. Methodological imperatives for investigating the phonetic organization and phonological structures of spontaneous speech. *Phonetica* 62, 1-11.

[33] McGettigan, C., Warren, J.E., Eisner, F., Marshall, C.R., Shanmugalingam, P., Scott, S.K. 2010. Neural correlates of sublexical processing in phonological working memory. *J. Cognitive Neuroscience* 23, 961-977.

[34] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264, 746-748.

[35] McLennan, C.T., Luce, P.A., Charles-Luce, J. 2003. Representation of lexical form. *J. Exp. Psych: LMC* 29, 539-553.

[36] McMurray, B., Tanenhaus, M.K., Aslin, R.N., Spivey, M.J. 2003. Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access. *J. Psycholinguistic Research* 32, 77-97.

[37] Niebuhr, O., Kohler, K.J. 2011. Perception of phonetic detail in the identification of highly reduced words. *J. Phonetics* doi:10.1016/j.wocn.2010.12.003

[38] Norris, D.G., McQueen, J.M., Cutler, A. 2003. Perceptual learning in speech. *Cog. Psych.* 47, 204-238.

[39] Nygaard, L.C., Burt, A.S., Queen, J.S. 2000. Surface form typicality and asymmetric transfer in episodic memory for spoken words. *J. Exp. Psych: LMC* 26, 1228-1244.

[40] Ogden, R. 2006. Phonetics and social action in agreements and disagreements. *J. Pragm.* 38, 1752-1775.

[41] Piccolino Boniforti, M.A., Ludusan, B., Hawkins, S., Norris, D. 2010. Same phonemic sequence, different acoustic pattern and grammatical status: A model. *AISV* Naples, Italy, 279-291.

[42] Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C. 1991. The use of prosody in syntactic disambiguation. *J. Acoust. Soc. Am.* 90, 2956-2970.

[43] Repp, B.H., Liberman, A.M. 1987. Phonetic category boundaries are flexible. In Harnad, S. (ed.), *Categorical Perception: The Groundwork of Cognition*. Cambridge: Cambridge University Press. 89-112.

[44] Salverda, A.P., Dahan, D., Tanenhaus, M.K., Crosswhite, K.M., Masharov, M., McDonough, J. 2007. Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition* 105, 466-476.

[45] Schroeder, C.E., Wilson, D.A., Radman, T., Scharfman, H., Lakatos, P. 2010. Dynamics of active sensing and perceptual selection. *Curr. Op. Neurobiology* 20, 1-5.

[46] Sereno, J., Jongman, A. 1995. Acoustic correlates of grammatical class. *Language and Speech* 38, 57-76.

[47] Sikveland, R.O., Ogden, R. 2010. Intersubjectivity, phonetics and gesture in the design of turns at talk. *http://sites.google.com/site/rao1york/downloads*.

[48] Smith, R., Hawkins, S. Under revision. Production and perception of speaker-specific phonetic detail at word boundaries. *J. Phonetics*.

[49] Sommers, M.S., Barcroft, J. 2006. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *J. Acoust. Soc. Am.* 119, 2406-2416.

[50] Stephens, G.J., Silbert, L.J., Hasson, U. 2010. Speaker-listener neural coupling underlies successful communication. *Proc. Nat. Acad. Sci* 107, 14425-14430.

[51] Strange, W. 2010. Automatic selective perception (ASP) of first and second language speech: A working model. *J. Phonetics* doi:10.1016/j.wocn.2010.09.001.

[52] Sumner, M., Samuel, A.G. 2005. Perception and representation of regular variation: The case of final /t/. *J. Memory and Language* 52, 322-338.