Rachel Smith  
*University of Glasgow*

**Perception of Speaker-Specific Phonetic Detail**

**Abstract:** The individual speaker is one source among many of systematic variation in the speech signal. As such, speaker idiosyncrasies have attracted growing interest among researchers of speech perception, especially since the 1990s, when theories began to treat variation as information rather than noise. It is now a common assumption that people remember and respond to speaker-specific phonetic behaviour. But what aspects of speaker-specific behaviour are learned about and used to guide perception? Do listeners make full use of the richness of speaker-specific information available in the signal, and how can listeners’ use of such information be modelled? In this chapter I review evidence that processing of the linguistic message is affected by inter-speaker variation in a number of aspects of phonetic detail. Phonetic detail is defined here as patterns of phonetic information that are systematically distributed in the signal and perform particular linguistic or conversational functions, but whose perceptual contribution extends beyond signalling basic phonological contrasts (such as differences between phonemes or between categories of pitch accent). Following Polysp, the Polysystemic Speech Perception model of Hawkins and colleagues (Hawkins and Smith, 2001; Hawkins, 2003, 2010), I argue that people can learn about speaker-specific realisations of any type of linguistic structure, from sub-phonemic features up to larger prosodic structures and, potentially, conversational units such as speaking turns. Speaker-specific attributes may even, on a more associative basis, enable direct access to aspects of meaning. I discuss circumstances liable to promote or disfavour the storage of speaker-specific phonetic detail, considering issues such as the frequency and salience of particular speaker-specific patterns in the input, and listener biases in attribution of variation to possible causes.

1. **The changing role of the speaker in speech perception theories**

Individual speakers are a source of considerable variability in the realisation of linguistic categories. This much has been clear since the early days of acoustic phonetics: for example, Peterson and Barney (1952) measured formant frequencies of American English vowels spoken by adult male, female and child speakers, and demonstrated not only extensive within-category...
variation, but also between-category overlap, when vowel tokens were plotted in F1-F2 space. Very many speech production studies show that, while speakers behave consistently with one another in many ways, there is also a significant degree of variability among them. For example, Johnson et al. (1993) found variation in the degree to which speakers of American English recruited the jaw to produce low vowels; Borden and Gay (1979) observed some speakers to produce /s/ with the tongue-tip up and others with it down (for a few more examples among many, see Dilley et al., 1996; Fougeron and Keating, 1997; van den Heuvel et al., 1996).

The implications of this inter-speaker variability for perception have been interpreted in shifting ways over the years. In the 1970s and 1980s, the dominant assumption was that speaker variability had to be stripped away, or normalised, before sounds and words could be recognised. Halle (1985: 101) writes: “when we learn a new word we practically never remember most of the salient acoustic properties that must have been present in the signal that struck our ears; for example, we do not remember the voice quality of the person who taught us the word or the rate at which the word was pronounced.” Views such as Halle’s are often referred to as abstractionism: i.e. the assumption that the brain must store abstract linguistic units, in order to account for the compositionality of language (e.g. McClelland and Elman, 1986; Norris et al., 2000; Pisoni and Luce, 1987). According to abstractionist views, the perceptual details of individual utterances do not ordinarily form part of linguistic representation. (Nonetheless the perceptual details of spoken utterances can be remembered and accessed for some purposes, such as autobiographical memory.) With isolated exceptions (Klatt, 1979 and to a lesser extent Wickelgren, 1969), the idea that words are stored in the form of discrete symbolic units dominated psycholinguistics and speech perception research until the 1990s. Accordingly, researchers sought to develop the best algorithms to normalise the speech signal across speakers, and/or to identify properties of sounds that remained invariant across speakers (e.g. Stevens, 1989).

From the 1990s, this view encountered a radical challenge from exemplar (also known as non-analytic or episodic) approaches to speech perception. According to these approaches (e.g. Goldinger, 1996, 1998), individual exemplars or instances of speech are retained in memory. When a new speech signal is encountered, it is matched simultaneously against all stored
exemplar traces in memory, and each stored exemplar is activated in proportion to the goodness of match. The aggregate of these activations produces a response. There is no need for storage of abstract forms; linguistic categories are simply the distributions of items that a listener encounters, encoded in terms of values of parameters in a multidimensional phonetic space. Accordingly, information about the speaker need not be stripped away: it is assumed to be retained in memory, and to play a role in perception. Early work within the exemplar framework (e.g. Goldinger et al., 1991; Palmeri et al., 1993; Nygaard et al., 1994) showed that perception can be facilitated when conditions allow information about the speaking voice to be encoded and accessed (and, conversely, can be disrupted under less optimal conditions). This work emphasised global speaker characteristics like f0, vocal effort and rate (e.g. Bradlow et al., 1999; Schacter and Church, 1992; Church and Schacter, 1994; Nygaard et al., 1995).

Subsequently the pendulum swung back to a somewhat more categorical view that mixes elements of the abstractionist and exemplar approaches. This hybrid approach was motivated particularly by the need to explain how learning about one word may transfer to other words containing the same sound. For example, if listeners learn that a particular spectral profile is appropriate for a given speaker’s /s/ in the word mice, they will, assuming other conditions stay sufficiently constant, expect a similar spectral profile for that speaker’s /s/ in house, dice, miss, etc. (McQueen et al., 2006). Such patterns of generalisation across words may be difficult to explain in a purely exemplar framework, unless a degree of abstraction is assumed. Thus, Cutler et al. (2010) propose that speech is represented prelexically in terms of abstract phoneme categories, which are updated where relevant with specific information about how each phoneme is pronounced by individual speakers. Evidence supporting this position has come primarily from experiments focusing on idiosyncratic pronunciations of individual segments. A case in point is the line of research pioneered by Norris, McQueen and Cutler (2003) in which realisation of a fricative was manipulated to be ambiguous between [f] and [s]: after being exposed to the ambiguous fricative in words containing either [f] or [s] listeners shifted their perceptual category boundary between [f] and [s] to accommodate the new variant. Further research along similar lines has shown similar patterns of learning for idiosyncratic pronunciations of stops (Kraljic and Samuel, 2006) and
vowels (Maye et al., 2008; Dahan et al., 2008). Based on experimental results, some researchers have proposed that the prelexical representations that undergo retuning may be allophonic rather than phonemic (e.g. Mitterer et al., 2013; Reinisch et al., 2014). However, these proposals contain little detail on questions such as how many and how subtle allophonic variants would be separately represented. Thus the idea that adaptation focuses on phonemic categories remains the most fully-developed hybrid approach.

Recently, a new class of speech perception models has emerged that deal with probabilistic processing in terms of a set of statistical concepts known as Bayesian inference (Scharenborg et al., 2005; Norris and McQueen, 2008; Clayards et al., 2008; Feldman et al., 2009). Bayes’ theorem gives formal expression to the idea that under conditions of uncertainty, probabilistic inferences are made based on knowledge or expectation (‘prior probability distributions’) in combination with current evidence. While most Bayesian models of speech perception do not deal explicitly with speaker-related variation, Kleinschmidt and Jaeger (2015) propose a speaker-specific belief updating model, which involves inferences at multiple levels: inferences about which linguistic categories are being produced, inferences about who is speaking, and inferences about the mappings between acoustic cues and linguistic categories that the speaker is using. In Kleinschmidt and Jaeger’s words (2015: 151-2), “good speech perception depends on using an appropriate generative model for the current talker, register, dialect, and so forth. The listener never has access to the true generative model, but rather only their uncertain beliefs about that generative model. Thus, adaptation can be thought of as an update in the listener’s talker- or situation-specific beliefs about the linguistic generative model.” The notion of a linguistic generative model is very broad and carries no commitment to any specific linguistic unit or units as the object of belief updating. However, the modelling carried out so far within this framework focuses on distributions of individual acoustic cues to phonemic contrasts, e.g. VOT as a cue to voicing or spectral centre of gravity as a cue to fricative place of articulation.

In summary, any theory of speech perception must account in some way for inter-speaker variability. Current views favour some degree of retention of speaker-specific information in memory, rather than assuming all such information is stripped away during perception. In terms of the phonetic nature of speaker-specific information that is retained, most work has
focused on global prosodic attributes of a speaker, on idiosyncratic realisation of phonemes, or on speaker-specific distributions of individual cues to phonemic contrasts (see e.g. Samuel and Kraljic, 2009, for an overview). These choices may reflect either a theoretical commitment (e.g. Cutler et al., 2010), or simply be convenient for model-building. Either way, they present a rather restrictive picture of what speaker-specific behaviour can entail. The main purpose of this chapter is to argue, from phonetic and perceptual evidence, that a broader view of speaker-specific phonetics should be taken. To adopt the terms of Kleinschmidt and Jaeger (2015), this amounts to arguing that what is needed is a richer specification of the linguistic generative model about which listeners have speaker-specific beliefs.

2. Speaker-specific phonetic detail (SSPD)

Many dimensions of speaker-specific behaviour relate to linguistic structure and linguistic categories, but in ways that cannot be captured if speech is considered solely in terms of an inventory of phonemes and major intonational categories. Rather, there are dimensions of speaker-specific behaviour that involve phonetic detail. As defined by (among others) Local (2003) and Hawkins (Hawkins and Smith, 2001; Hawkins, 2003, 2010), phonetic detail refers to phonetic information that affects people’s responses but “is not considered a major, usually local, perceptual cue for phonemic contrasts in the citation forms of lexical items” (Hawkins and Local, 2007: 181). This type of information is “systematically distributed [according to linguistic/communicative function] but not systematically treated in conventional approaches” (ibid.). Thus, phonetic detail refers not to information that mainly distinguishes phonemes (such as /pa/ vs. /ba/), but to cues that distinguish other aspects of linguistic structure, such as prosodic structure (compare the unstressed /p/ in potato with the stressed /p/ in important); syllabic and morphological structure (/p/ is more heavily aspirated in the morphologically-complex word displease than in the mono-morphemic word displays; Smith et al., 2012); or pragmatic function (for Standard Southern British English, both [pʰ] and [p’] are possible allophones of /p/ in it’s a tap, but the ejective sounds more emphatic, definite, and final than the aspirated stop.
The range of aspects of linguistic structure that condition systematic variation in phonetic detail is extensive. Crucially for the present purposes, there is evidence of speaker-specific variation in many of them, henceforth termed *speaker-specific phonetic detail* (SSPD).

For example, speakers vary in the extent to which they coarticulate, and in the precise coarticulatory strategies that they use. Reviewing research in this area, Kühnert and Nolan (1999) comment that it is relatively scarce, and that “the high variability found in the data makes it difficult to distinguish between effects which should be considered as being idiosyncratic and effects which simply reflect the allowed range of variation for the phenomenon”. Nonetheless, they identify several experiments showing individual coarticulatory differences: among British English speakers in coarticulation of /r/ and /l/ with a following vowel (Nolan, 1983, 1985), and among both Swedish (Lubker and Gay, 1982) and English speakers (Perkell and Matthies, 1992) in the timing of movements for anticipatory lip rounding. Some of this variation may be due to an individual’s genetic (anatomical and physiological) inheritance, as suggested by Weirich et al.’s (2013) finding that tongue looping trajectories are more similar in monozygotic twins than in dizygotic twins or unrelated speakers (though see Nolan and Oh, 1996 for a demonstration of articulatory variability within identical twin pairs).

Speakers also vary in their “prosodic signatures”, i.e. the detailed phonetic means they use to index prosodic prominence and prosodic boundaries. With respect to prominence, individual speakers mark prominent as opposed to non-prominent words using different subsets of prosodic properties, such as lengthening, pausing, increased intensity, increased f0, location of an f0 peak, and formant frequencies (Dahan and Bernard, 1996; Mo, 2010). With respect to prosodic boundaries, speakers vary subtly in the way they mark boundaries between syllables and words (Lehiste, 1960; Quené, 1992; Smith and Hawkins, 2012). For example, Smith and Hawkins (2012) recorded speakers of Standard Southern British English producing phonemically-identical sentence pairs, such as “So he diced them” vs. “So he’d iced them”, “They also offer Mick stability” vs. “They also offer mixed ability”, and found variation in patterns of allophonic detail at word boundaries: different speakers used duration to differing extents to mark the contrast between word-initial and word-final allophones, and some speakers lenited word-final sounds more than others. Similar variation occurs in
the way speakers distinguish other types of prosodic domain, as shown by Redi and Shattuck-Hufnagel (2001) with respect to glottalisation, and by Fougeron and Keating (1997) for articulatory lengthening and strengthening. Across these studies, not only do different speakers preferentially use different properties to signal a distinction, but some speakers clearly distinguish all levels of the prosodic hierarchy, while others tend to “flatten” it, i.e. they fail to exploit the possible range of prosodic levels (Fougeron and Keating, 1997; Mahrt et al., 2012).

Furthermore, as outlined by Abercrombie (1967), Laver (1980) and Mackenzie Beck (2005) among others, speakers differ in their long-term settings of the larynx and the supralaryngeal articulators. These articulatory settings impart characteristic qualities that systematically colour vocal output, such as breathiness, creakiness, dentalisation, labialisation, denasalisation, and so on. In Abercrombie’s description (1967: 91), such settings result in “a quasi-permanent quality running through all the sound that issues from [a person’s] mouth”. Importantly, however, the auditory consequences of such long-term settings depend in complex ways on the segments of the message, and also on the prosody (Mackenzie Beck, 2005). Thus if a speaker has a labialised voice quality, this will be audible on many of his/her segments, but not equally on all: segments normally produced with spread lips (e.g. /sl/, /θl/, /il/) will be particularly susceptible, while segments that are ordinarily labialised may sound more extremely so (e.g. /ʃl/, /rl/, /ʤl/, /ʧl/). Likewise a creaky voice quality may be especially audible at points in an utterance where creak is not normally found (e.g. phrase-medially in sonorant stretches of speech), as well as being heard as more extreme creakiness in places where creak is usual (e.g. phrase-finally, before word-final voiceless stops, between abutting vowels). By considering articulatory settings, we see that the way a speaker pronounces one of their phonemes is rarely completely independent of the way they pronounce others, yet a setting does not alter all phonemes in the same way, and prosody plays a role too.

Speakers also vary in longer-domain characteristics such as their speech rate, articulation clarity, and patterns of speech reduction (Hanique et al., 2015). Some of these longer-domain characteristics interact with the realisation of particular segments or features: Theodore et al. (2009) found that speakers vary in the extent to which changes in speech rate alter their
characteristic VOT patterns. Looking beyond the prosodic hierarchy as usually defined, there are systematic patterns of phonetic detail that occur over speaking turns and other interactionally-relevant chunks of talk (see e.g. Ogden, 2012). It seems plausible that individual speakers might implement these in idiosyncratic ways, although research has not addressed this issue to date.

In summary, a speaker’s phonetic individuality amounts to much more than a collection of phoneme realisations and some average prosodic properties. Speakers demonstrably vary in a number of aspects of phonetic detail, including their long-term articulatory settings, their coarticulatory behaviour, and the way they implement linguistic distinctions relating to prosodic structure. If we are accustomed to thinking about speech primarily in terms of the phonemic contrasts that distinguish individual words (e.g. *bin* vs *pin*), these types of SSPD may appear trivial, unsystematic, and of limited relevance to segment and word identification. However, when we think about recognition of words in their broader context—that is, in meaningful utterances heard in the flow of ordinary interaction—these aspects of sound structure take on a much greater importance, because they contribute some of the “glue” that holds chunks of speech together and makes them sound coherent. They help to encode phonological structure as well as phonological system; they represent “prosodies” as defined in Firthian prosodic analysis (see e.g. Ogden, 2012), or what in other phonological frameworks might be called prosody-segment interactions. If we broaden the definition of the listener’s task to include grasping the semantic, grammatical, information-structural and interpersonal relations within an utterance and a conversation, we can see that the above types of phonetic detail could well play an important role in understanding the message. Therefore, there is a clear potential advantage for listeners in learning to interpret patterns of SSPD produced by individual familiar speakers. The next section discusses whether listeners do in fact learn about and use these types of SSPD.

3. **Evidence for use of SSPD in speech perception**

If listeners know about speaker-specific phonetic detail as defined above — as opposed to simply about how a speaker realises their phonemes, or about
their average vocal pitch, etc. — and if they use this knowledge in perception, two consequences can be expected. First, exposure to a person’s speech should lead to changes in task performance that are general to properties that are common across groups of sounds a speaker produces. For example, if listeners are exposed to Fred’s voiceless alveolar plosives, which are dentalised and have extremely long VOT, they may form expectations that Fred will produce other voiceless plosives with long VOT, and/or that he will dentalise other alveolar sounds. Thus responses to Fred’s long-VOT /k/ and /p/, and to his dentalised /d/ and /n/, should be primed (facilitated) by the prior exposure to his tokens of /t/. Second, exposure to a person’s speech should lead to changes in task performance that are specific to sounds that occur in particular structures. If Jill lenites word-final /d/ to an unusual extent, realising it as an approximant in unstressed function words like he’d and she’d, listeners might expect similarly lenited variants in her unstressed we’d and I’d, and possibly also in stressed tokens of these words and in stressed content words; but they would not necessarily expect to hear them in Jill’s pronunciation of word-initial /d/. If, on the other hand, listeners only adjust phoneme categories when accommodating perceptually to a speaker, simpler patterns of responding would be expected. Exposure to Fred’s /t/s should only affect subsequent responses to /t/, and not to /p/, /k/, etc.; while exposure to Jill’s word-final /d/ in he’d should affect responses to /d/ in all other contexts.

These questions about generalisation and specificity in perceptual learning about individual speakers have been addressed experimentally by a small number of studies. Some of these explicitly manipulate both linguistic structure and speaker identity: that is, they use multiple speakers, and test whether listeners learn to associate a structure-specific pattern with an individual speaker, and not with other speakers who do not produce the pattern. While such experiments represent the “gold standard”, they are quite hard to conduct, as the requirement to test listeners with multiple voices and multiple linguistic structures leads to very long experiments. Therefore, a shortcut is sometimes taken: Experiments test whether an unusual pattern can be learned from a single speaker’s voice. If it can, it is inferred that in adapting to this unusual type of speech, listeners have learned a potentially speaker-specific property (e.g. Barden and Hawkins, 2013; Poellmann et al., 2014). Clearly, only the former type of experiment
directly demonstrates perceptual use of SSPD. Nonetheless, the inference drawn from the latter type is probably valid. Dahan et al. (2008) used one voice to assess perceptual learning of a contextually-conditioned allophone, and inferred from their results that the learning might be speaker-specific; Trude and Brown-Schmidt (2012) conducted a similar experiment with multiple voices, and confirmed direct evidence of speaker-specific learning.

There is evidence for the first claim, i.e. for perceptual use of SSPD relating to general properties that are common across groups of sounds a speaker produces. This evidence relates to the feature [±voice]. Individual speakers differ in their characteristic VOT in voiceless stops (Allen, Miller, and DeSteno, 2003). Listeners can learn to associate a speaker with a characteristic pattern of VOT (Allen and Miller, 2004; Theodore and Miller, 2010; though under some circumstances learning about the realisation of [±voice] may generalise to other speakers, Kraljic and Samuel, 2007). Several studies using a range of learning paradigms have shown that speaker-specific learning about VOT generalises not only among words beginning with the same phoneme (e.g. /t/), but also, partially or fully, across place of articulation, i.e. to other voiceless stop phonemes (Kraljic and Samuel, 2006; Theodore and Miller, 2010; Nielsen, 2011).

There is also evidence for the second claim, i.e. that exposure to a person’s speech can lead to changes in task performance that are specific to sounds that occur in particular contexts or structures. Several studies address whether learning about how a speaker pronounces a sound in one position in the syllable or the word generalises to other positions. Smith and Hawkins (2012) tested the perceptual relevance of the individual differences in phonetic detail at word boundaries discussed in section 2 above. Tests of intelligibility in noise before and after exposure to a voice showed that familiarity with an individual speaker’s patterns helped listeners to segment and identify words in noise. The learning was speaker-specific, and the perceptual benefit was small, but robust. Some other work on transfer of speaker-specific learning across positions in syllable or word supports Smith and Hawkins’ findings: Dahan and Mead (2010) found, for a range of phonemes, that learning to understand noise-vocoded speech was specific to position in syllable. However, Jesse and McQueen (2011) found that learning of an unusual pronunciation of /s/ was not specific to position in syllable. The divergent results may be due to the different phonemes under
investigation, and/or to other aspects of the experiments. For example, if the critical acoustic information for the perception of /s/ is contained mainly within the fricative itself, rather than distributed across more than one segment, this may encourage generalisation across positions (Reinisch et al., 2014). Relatedly, Jesse and McQueen (2011) spliced the identical fricative across positions in syllable, whereas the syllable-initial and -final fricatives in Smith and Hawkins’ study exhibited natural variation in duration and spectral composition.

Other research shows perceptual learning of speaker-specific phonetic detail that relates to specific allophones rather than specific phonemes. Dahan et al. (2008) exposed listeners to a dialect in which /æ/ is raised to [eɪ] or [ɛ] before voiced velar stops (e.g. in bag) but not voiceless ones (e.g. back). They hypothesized that if listeners learned this pattern, they would obtain an advantage in recognising the words: they would be able to use the information in the vowel to resolve the lexical competition between bag and back earlier in the time course of the word. Listeners’ eye-tracking performance supported this hypothesis: listeners who had been exposed to the raised vowel identified bag, as opposed to its competitor back, earlier and more accurately than listeners who had been exposed to the standard variant of the vowel. Trude and Brown-Schmidt (2012) replicated Dahan et al’s finding, varying the voice heard in the test phase and thereby demonstrating that the learning was genuinely speaker-specific. A different type of allophonic variation was shown to be perceptually important by Mitterer, Scharenborg and McQueen (2013). They generated an ambiguous segment by averaging approximant /r/ and dark /l/. Learning about this ambiguous segment altered performance on an approximant-/r/-to-dark-/l/ continuum, but not on continua where the endpoints were trill /r/ and light /l/.

In another test of the structure-specificity of perceptual learning, Barden and Hawkins (2013) investigated perceptual learning of phonetic patterns related to morphological structure. Grammatical morphemes can be pronounced differently from the identical phoneme strings when these do not function as morphemes. For example, the phoneme sequence /mɪst/, when spoken in a prefixed word like mistimes, has a longer and more peripheral /ɪ/, shorter /s/, and a /t/ with longer VOT, than when spoken in a word that does not have a true prefix, such as mistakes or mystique (Smith, Baker and Hawkins, 2012). The re- of repaint (which decomposes morphologically
into *re + paint*) likewise has a more peripheral vowel than the *re-* of *report* (which does not decompose into *re + port*). Barden and Hawkins asked whether, if exposed to an idiosyncratic pronunciation of a prefix, listeners would learn to expect this pronunciation in prefixed but *not* non-prefixed words. They trained two groups of listeners with stories containing prefix *re-*, either realised unusually as */rɪ/* (Accent group), or realised normally as */ril/* (Control). Listeners then performed an intelligibility-in-noise test containing keywords with prefix *re-* (e.g. *republication*) and non-prefix *re-* (e.g. *renal infection*) pronounced as */rɪ/*. Listeners in the Accent group, who had been exposed to the */rɪ/* prefix, identified the unusually-pronounced keywords significantly more accurately than listeners in the Control group. The benefit was present for both prefixed and non-prefixed test words, but was significantly greater for prefixed words, suggesting the listeners associated the unusual pronunciation more strongly with the specific linguistic structure in which it had been encountered, though the learning did partially generalise to other structures.

Along similar lines, Poellmann et al. (2014) demonstrated that listeners could adapt to particular realisations of a prefix that are characteristic of fast casual speech. Listeners who were exposed to words beginning with the Dutch prefix *ver-*, realised as *[fɪ]*, showed improved identification of new *ver-* words realised with *[fɪ]*, compared to unexposed listeners. Listeners in this study may have been learning about prefix pronunciation, or speech style, or both. The data do not allow these possibilities to be distinguished, but regardless, they underscore that perceptual learning cannot solely concern phonemic categories.

At the other end of the spectrum, there is also evidence from a different line of research, that what listeners learn about a voice is restricted neither to its gross prosodic properties nor its segmental fine structure. Several experiments have investigated perceptual learning by applying different types of degradation to the speech signal. Remez et al. (1997), Remez et al. (2002) and Sheffert et al. (2003) used sine-wave speech, which lacks natural vocal quality and segmental-phonetic fine structure, but preserves enough of the time-varying spectro-temporal structure of speech to support word recognition. Adult listeners are surprisingly good at identifying personally familiar talkers from sine-wave replicas of their utterances (Remez et al., 1997). Moreover, adults can generalise knowledge of speaker-specific attributes
that has been learned from sine-wave replicas to both novel sine-wave samples and natural speech (Sheffert et al., 2003). Interestingly, pre-school children can also discriminate familiar cartoon voices from spectrally-degraded (in this case noise-vocoded) speech (Van Heugten et al., 2014). The acoustic basis for speaker identification from degraded speech samples is not yet clear, but it presumably must rely partly on global qualitative speaker characteristics such as formant spacing, which are preserved in sine-wave and (given sufficient spectral resolution) noise-vocoded speech. Local phonetic properties (such as segmental durations) probably also play a role, but their specific importance has not been tested.

In summary, listeners can learn many aspects of SSPD. Learning sometimes transfers across phoneme categories, as in the case of VOT in voiceless stops. Learning does not necessarily transfer to all members of a phoneme category: it may be specific to certain positions in word, or to certain morpho-lexical structures, such as prefixes. From the evidence so far, it is reasonable to assume that the patterns of transfer are not arbitrary, but principled, reflecting how general vs. how specific to particular linguistic structures the phonetic properties in question actually are.

4. Modelling the perceptual relevance of SSPD

What kind of a model can account for the data on perception of speaker-specific phonetic detail — i.e. for listeners’ ability to learn patterns that are specific both to an individual speaker and to a particular (type of) linguistic context? The preceding sections have shown that models that assume abstract phoneme categories, updated with speaker-specific information (e.g. Cutler et al., 2010), cannot fully do so, because some aspects of SSPD generalise across phoneme categories, while others are restricted to only some instances of a phoneme category. At the same time, some abstraction is needed, to account for the patterns of generalisation that have been shown in perceptual learning studies, as well as to explain why exemplar effects are not consistently found across all experiments (see e.g. Hanique et al., 2013).

Smith and Hawkins (2012) discuss a range of modelling approaches that have the potential to accommodate their data on speaker-specific word segmentation. Here, I focus on Polysp (Hawkins and Smith, 2001; Hawkins,
2003, 2010), which is not a computationally implemented testable model, but indicates the lines along which such a model could develop. Polysp stands for Polysystemic Speech Perception; the term ‘polysystemic’ reflects the idea that the phonology of a language involves a range of structures, within each of which different systems of contrast may operate, as opposed to a single monolithic phoneme system (Hawkins and Smith, 2001). The model takes a hybrid episodic-abstract approach, and posits that phonetically detailed episodes are stored in memory alongside abstraction in terms of rich linguistic structures. Exhaustive parsing of the signal into abstract linguistic categories is argued not to be needed if meaning can be accessed without it. This may be the case when listeners hear familiar chunks of highly reduced speech: for example, [ə̞̃ə̃ə̝̃] is (in some circumstances) an acceptable, if highly casual, realisation of the phrase I don’t know, and probably does not need to be mapped on to the three individual words in order for the listener to understand that the speaker lacks some knowledge or information. Access to meaning without parsing into abstract categories may also occur in situations where identifying a particular voice is sufficient to constrain the interpretation of a linguistic structure or meaning, as demonstrated in an eye-tracking study by Creel and Tumlin (2011). Nonetheless, in general, Polysp proposes that phonological knowledge is represented in terms of rich, hierarchical structures (incorporating prosodic and grammatical information) and this representation improves the process of pattern-matching between signal and memories. These structures are abstract (like phonemes) but are richer than a phoneme string, and as such allow for more complex phonetic detail to be represented. Speaker-specific information can potentially be associated with any unit(s) at any level(s) of the representation.
Figure 1. The utterance "then you go down to the bottom right" spoken by a young male Panjabi-English bilingual speaker from Bradford (taken from the IViE corpus, www.phon.ox.ac.uk/IViE/). Top panel: Wideband spectrogram and phonetic transcription of the utterance. Bottom panel: Representation of the utterance as a prosodic tree. IP = Intonational Phrase; AG = Accent Group; S = strong, W = weak; O = onset, R = rime, N = nucleus, C = coda. Each terminal node in the tree could further be associated with a bundle of distinctive features, not represented here.
Figure 1 represents an utterance spoken by a male teenage bilingual Panjabi-English speaker from Bradford (UK), performing a map task (taken from the IViE corpus, www.phon.ox.ac.uk/IViE/). The top panel shows a spectrogram and associated phonetic transcription, while the bottom panel shows a prosodic tree corresponding to the utterance. Different theoretical approaches would differ as to the details of the prosodic tree (e.g. Selkirk, 1986; Nespor and Vogel, 1986), but this does not matter for our purposes: the main point of the tree is to show that syllabic and prosodic structures are core to this representation of the utterance, while the phoneme string is not.

The prosodic tree gives a window on the opportunities afforded by the example utterance to learn about speaker-specific phonetic detail, that is rather different from the picture presented by a phoneme string. For example, the phonemic transcription, /ðɛŋ jə go daʊn tə ðə bɔtəm ræt/\(\text{t}i\)/, indicates that the utterance contains three instances of the phoneme /t/\(\text{t}i\)/. However, the narrow transcription and the spectrogram indicate that the speaker realises these in quite different ways, with an aspirated /t/ in to and glottal stops in bottom and right. This much may seem fairly banal—/t/ is well known to have considerable allophonic variation in English, with glottal stop prominent among the variants. What the tree also shows but the phoneme string does not, however, is the structural constraints on this speaker’s use of glottal stop. He uses it word-medially (at the juncture between a stressed and a following unstressed syllable in bottom), and word-finally (in right), but not word-initially (to). A different speaker might also use glottal stop for /t/ word-initially but foot-medially (in down to the). A different person again might only use it word-finally.

The spectrogram also indicates that the speaker produces word-initial voiced stops quite consistently, regardless of their place of articulation: the initial stops in go, down, bottom are all voiced throughout their closures. Moreover, as the narrow transcription indicates, he produces slightly retracted alveolar consonants in down and to (which is a typical feature for British Panjabi speech; cf Alam and Stuart-Smith, 2011 and Kirkham, 2011). He reduces weak syllables quite substantially: they are considerably lower in intensity than adjacent strong syllables, and are segmentally...
reduced, with a syllabic nasal in *bottom* and a very minimal trace of *the* in *down to the* (the word is realised merely as some extra duration on the /b/ of *bottom*).

In summary, even from the single utterance represented in Figure 1, it can be seen that the prosodic tree makes it possible to capture a number of systematic patterns which are not evident from a segmental transcription alone. Although the structures look complex at first sight, their value from the perspective of modelling SSPD is that they allow a great deal of information about the speaker to be represented, which has the potential to predict the speaker’s future behaviour in some detail.

The foregoing discussion suggests that listeners can construct speaker-specific representations that are highly detailed and complex, comprising knowledge of speaker variation at many linguistic levels. It seems reasonable to assume that full representations of this kind could be built up only with considerable exposure to a speaker’s voice. That is, for a highly familiar speaker, such as a partner, parent, or close friend, the listener’s stored representations could well be elaborated with probabilistic knowledge of the speaker’s typical patterns at many or all of these levels. For example, a listener highly personally familiar with the speaker in Figure 1 might have detailed knowledge of how the speaker produces syllable-initial /t/ in foot-medial weak syllables, such that the listener would be surprised if this speaker were to use a glottal stop for /t/ in *down to the shops*. But when a listener is merely casually familiar with a speaker, or is beginning to get to know them, the listener would not be familiar with all the systematics of the speaker’s idiolect. The speaker-specific representations would be much less detailed, and would support less confident predictions and inferences during speech understanding. Moreover, different listeners might construct quite different speaker-specific representations, because the complexity of the structures allows flexibility in mapping of phonetic patterns to representations. From limited data such as the utterance in Figure 1, a listener might abstract a generalization about how the speaker produces the phoneme /t/, or voiceless stops in general, or voiceless stops in weak syllables, and so on. There are numerous possibilities for the attribution of phonetic
variation to causes\(^1\), and only with appropriate exposure would the listener be able to disentangle these from one another and fine-tune their model of the speaker’s behaviour.

In this regard, a particularly interesting set of results was obtained by Eisner et al. (2013). They looked at word-final devoicing, i.e. the pronunciation of (for example) overload as overloat, which is a common pattern in Dutch, among other languages, but occurs to a much lesser extent in English. When native English listeners were exposed to a Dutch speaker devoicing word-final /d/, their perceptual responses showed overgeneralization across positions in syllable: that is, they became more willing to accept the speaker’s devoiced tokens as instances of /d/, not only in final position, but also in initial position, as in down pronounced as town. Interestingly, however, this overgeneralization did not occur if listeners were also exposed to the Dutch speaker’s actual (voiced) variants of initial /d/, nor when the stimuli were presented in a native English accent. These findings underscore the flexibility inherent in learning of speaker-specific pronunciation patterns. Learning generalised across positions in syllable when listeners had no reason not to expect a speaker to produce the same variant in all contexts (i.e. in the case of the unfamiliar Dutch accent). But learning failed to generalise when listeners were presented with direct evidence of the speaker’s allophonic variation (in the case where they heard the Dutch speaker producing voiced initial /d/ and devoiced final /d/). Learning also failed to generalise when listeners had a strong expectation about the patterns of normal allophonic variation in the variety they were hearing, as in the case where they heard the native English speaker: listeners know that native English speakers sometimes devoice word-final stops, but rarely word-initial ones, and did not show overgeneralization in this case.

Polysp does not make detailed predictions about how speaker-specific representations might be built up over the course of exposure, focusing

---

\(^1\) The issue here is reminiscent of the problem of the indeterminacy of translation, as discussed by Quine (1960). If we see a rabbit, and hear a speaker of an unknown language say “gavagai,” there are numerous possible meanings: e.g. Look, a rabbit. Look, food. Let’s go hunting. There will be a storm tonight. Look, a momentary rabbit-stage. Look, an undetached rabbit-part. See Kraljic et al. (2008) for a similar point about attribution of phonetic variation to causes, memorably illustrated using a Benny Hill joke.
rather on the form such representations might eventually take. However, a Bayesian model like that of Kleinschmidt and Jaeger (2015) could generate empirically testable predictions about how representations develop through exposure, if the model were expanded to express richer linguistic structure. In addition to mere exposure, there appear to exist cognitive biases which affect how a listener builds up a representation of a speaker’s behaviour. Kraljic and colleagues carried out an elegant series of experiments exploring the circumstances under which listeners are willing to interpret phonetic variation as speaker-specific. They found that an unusual pronunciation was more likely to be assumed to be speaker-specific if it could not plausibly be attributed to the phonetic context (Kraljic, Brennan and Samuel, 2008), or to an extraneous proximal cause (such as a pen in the speaker’s mouth; Kraljic, Samuel and Brennan, 2008). Moreover, a pronunciation was more likely to be attributed to speaker-specific behaviour if it was first encountered early on in exposure to the speaker: listeners seemed to assume that a speaker-specific characteristic should be stable, and thus if a pattern had not been encountered early in exposure, they preferred to attribute it to some more transient cause (Kraljic, Samuel and Brennan, 2008). Again, for a fuller consideration in a Bayesian framework of circumstances under which listeners may incline to rely on existing beliefs vs. develop new speaker-specific ones, see Kleinschmidt and Jaeger (2015).

In summary, hybrid models like Polysp allow perceptual learning of SSPD to be conceptualised in terms of speaker-specific modulations of rich linguistic (phonological, prosodic, grammatical) structural representations. These representations have the potential to account for some of the more complex perceptual responses to speaker-specific phonetic detail, which are harder to capture in phoneme-based models. However, the richness of the representations does create the potential for indeterminacy in attribution of phonetic patterns to causes. Various cognitive biases may be involved in resolving such indeterminacy, and more work is needed to understand these. Sufficient exposure must surely be needed — listeners cannot learn a pattern unless they hear it, obviously — but beyond this, it may be the case that some regularities are easily learnable, while others are more resistant to perceptual learning (similar arguments are made in research on the transmission of sound change: Milroy, 2007). I speculate that listeners will be better able to learn about SSPD in chunks of speech that are
rhythmically and prosodically salient, and predictable in terms of meaning, because meaning is known to guide perceptual learning (Davis et al., 2005). A more general prediction is that listeners may also vary in exactly what and how they abstract from a person’s speech: that is, we might expect listener-specific perception of speaker-specific phonetic detail. A listener’s ability and readiness to make and generalise speaker-specific perceptual adjustments in this way might even correlate with the degree of phonetic shift (accommodation) they produce in response to a conversation partner’s speech. These speculations remain to be tested empirically.

5. Conclusions and future directions

The present review has shown that speakers vary in the way they realise many complex aspects of linguistic structure, from coarticulation through context-conditioned allophony to marking of syllable and word boundaries, and casual speech reduction strategies. These patterns of individual variation can be learned about, and can facilitate performance in various laboratory tasks. A promising approach to modelling them is using hybrid models that assume some degree of exemplar or episodic storage, combined with flexible abstraction that allows speaker-specific attributes to be associated with any level of hierarchically-organised phonetic and prosodic structure.

Where might the study of the perceptual role of speaker-specific phonetic detail head next? First, more work is needed to develop models that make concrete predictions about how representations of speaker-specific phonetic detail are built up as a function of experience. Second, a critical approach to the concept of the speaker itself will help to move the field forward. The discussion so far has implicitly assumed i) that individual speakers behave stably in their production of any given linguistic structure, and ii) that the individual speaker is the main locus of interesting variation. Both these assumptions are almost certainly incorrect. Many factors contribute to variation within a speaker (such as his/her temporary physical and emotional state, the physical speaking/listening environment, the task he/she is engaged in, the structural constraints of conversation, and intersubjective aspects such as his/her affiliation with an interlocutor). Moreover, speakers are not islands, but cluster according to numerous variables (including
sex and gender, age, personality, regional accent, socio-economic status, occupation, participation in communities of practice, and so on). Thus an understanding of the perceptual “speaker space” must ultimately take into account both variation within a speaker, and commonality across groups of speakers who share similar personal or social characteristics.

Finally, an interesting avenue to explore is how speaker-specific phonetic detail simultaneously contributes both to listeners’ understanding of the linguistic message (in a lexical/linguistic ‘search space’), and also to recognition of a speaker’s individual identity and/or group affiliations (in ‘speaker space’). The interactions between these two domains have not been thoroughly explored (though see Mullennix and Pisoni, 1990, and Creel and Tumlin, 2011 for promising directions), and many outstanding questions remain about how the tasks of speaker identification and word identification are solved in parallel, in real time. For the future, the study of speaker-specific phonetic detail can be expected to play an important role in developing an integrated account of how listeners simultaneously perceive speakers’ personal and social characteristics, and their verbal messages.

References


