Production and perception of speaker-specific phonetic detail at word boundaries

Suggested running title: Speaker-specific detail at word boundaries

Rachel Smith[a] and Sarah Hawkins[b]

[a]Glasgow University Laboratory of Phonetics, University of Glasgow, 12 University Gardens, Glasgow G12 8QQ, United Kingdom

[b]Centre for Music and Science, Faculty of Music, University of Cambridge, 11 West Road, Cambridge, CB3 9DP, United Kingdom

Rachel.Smith@glasgow.ac.uk

sh110@cam.ac.uk

Corresponding author:

Dr. Rachel Smith
Glasgow University Laboratory of Phonetics
University of Glasgow
12 University Gardens
Glasgow G12 8QQ
United Kingdom

Rachel.Smith@glasgow.ac.uk
Phone: +44 141 330 5533
http://www.gla.ac.uk/schools/critical/staff/rachelsmith/

Abstract

Experiments show that learning about familiar voices affects speech processing in many tasks. However, most studies focus on isolated phonemes or words and do not explore which phonetic properties are learned about or retained in memory. This work investigated inter-speaker phonetic variation involving word boundaries, and its perceptual consequences. A production experiment found significant variation in the extent to which speakers used a number of acoustic properties to distinguish junctural minimal pairs e.g. *So he diced them—So he'd iced them.* A perception experiment then tested intelligibility in noise of the junctural minimal pairs before and after familiarisation with a particular voice. Subjects who heard the same voice during testing as during the familiarisation period showed significantly more improvement in identification of words and syllable constituents around word boundaries than those who heard different voices. These data support the view that perceptual learning about the particular pronunciations associated with individual speakers helps listeners to identify syllabic structure and the location of word boundaries.

Keywords

perceptual learning, segmentation, word boundaries, phonetic detail, individual speech style, allophone

1. Introduction

Experiments show that individual speaker characteristics can affect speech processing in a variety of tasks, including serial recall (Goldinger, Pisoni, & Logan, 1991), explicit recognition memory (Palmeri, Goldinger, & Pisoni, 1993), word recognition in noise (Nygaard, Sommers, & Pisoni, 1994; Nygaard & Pisoni, 1998) and shadowing (Goldinger, 1998; Shockley, Sabadini, & Fowler, 2004). The speaker characteristics that have been investigated mostly involve the realisation of particular phonemes on the one hand (e.g. Norris, McQueen & Cutler, 2003), and on the other, very global properties such as mean f0 and aspects of intonation (Church & Schacter, 1994), and amplitude, vocal effort and rate of

speech (Bradlow, Nygaard, & Pisoni, 1999). Little is known, however, about the perceptual effects of speaker characteristics related to higher-level linguistic structure, such as the way a particular speaker realises boundaries between different types of constituents in the prosodic hierarchy, or the way he/she gives prominence to prosodic constituents. The present study investigates inter-speaker variation in the production of phonetic patterns at word boundaries, and tests whether listeners learn about such variation and use it to recognise words in noise.

Theories differ widely about the specific nature of the learning or memory processes that underlie speaker-specific effects on perception, especially with regard to the issue of units of representation. In the 1990s, so-called "non-analytic" approaches were developed, which departed from earlier "abstractionist" accounts of perception by assuming that holistic episodes or exemplars of words or longer stretches of speech are stored in memory (e.g. Goldinger, 1996, 1998; Johnson, 1997, 2006; Pisoni, 1997; Lachs, McMichael, & Pisoni, 2003). According to these approaches, when a new speech signal is heard, it is matched simultaneously against all stored exemplar traces in memory, which are activated in proportion to the goodness of match, and the aggregate of these activations produces a response. There is no need for the exemplars to be broken down into smaller units.

Non-analytic approaches reflect scepticism about the relevance to memory or perception of traditional linguistic units and categories such as features, phonemes, syllables, and so on. Instead of assuming a single linguistic unit as the focus of perception, they are flexible about the units that might (or might not) be involved, emphasising mainly the role of the individual perceiver's experience in linguistic processing and change. Nonetheless, in order to make predictions and carry out simulations, a widespread assumption has been that words are stored as exemplars. Johnson (2006: 492) justifies this choice as follows, "I choose to treat 'words' as exemplars because words lie at the intersection of form and meaning and thus generate coordinated patterns of activity in both sensory and higher level areas of cognition." Though a useful working assumption, this choice ignores the sublexical (phonological and morphological) structure of words, and words' participation in supralexical relations, as expressed in prosodic and grammatical organisation.

Non-analytic approaches' de-emphasis of the role of traditional phonological units has been viewed by some researchers as problematic (e.g. Pierrehumbert 2006; Goldinger 2007). A particular issue is that when word-sized exemplar storage is assumed, there is little attention to lexical items' internal phonological structure. This makes it difficult for non-analytic models to explain how people generalise from pronunciation or perception of particular words to the 'same' (or similar) sounds in other words. Cutler and colleagues, for example, have emphasised the need not to throw out the linguistic baby with the bathwater of abstractionism. Norris, McQueen, & Cutler (2003) demonstrated that after hearing a list with words that contained an ambiguous fricative in place of either /s/ or /f/, listeners shift their category boundary for the /s/-/f/ distinction in an appropriate direction. In a related priming task, McQueen, Cutler and Norris (2006) demonstrated that such perceptual learning about phonetic segments generalises to untrained words. To account for these data, the authors proposed that perceptual learning operates over abstract phonemic representations, and is probably speaker-specific (Eisner & McQueen, 2005; although Kraljic & Samuel, 2006 found a degree of generalization across speakers). While the units over which such learning operates were argued to be sub-lexical, Norris et al. (2003) proposed that knowledge of the intended lexical meaning was the driving force behind such learning (cf. Davis et al., 2005, but see also Hervais-Adelman, Davis, Johnsrude, & Carlyon., 2008, for indications that lexical access is not always necessary).

The question investigated in this paper is whether or not phoneme categories are the only aspect of sound patterns that might trigger speaker-specific perceptual learning. In principle, perceptual learning could also occur for aspects of syllabic and prosodic structure, and the ways in which individual speakers phonetically implement these: for example, the way a particular speaker realises boundaries between constituents such as syllables, words, and phrases, or the way he/she gives prosodic prominence to constituents. Such aspects of speech beyond phonological category structure are important for understanding meaning, especially functional and interactional meaning, and might be learned in a speaker-specific way.

1.1. Evidence regarding perceptual learning about allophonic detail

There is much evidence that the allophonic correlates of syllabic and prosodic structure are used perceptually to help word segmentation and identification (e.g. Ogden, Hawkins, House, Huckvale, Local, Carter, Dankovičová, & Heid, 2000; Davis, Marslen-Wilson, & Gaskell, 2002; Salverda, Dahan, & McQueen, 2003; Cho, McQueen, & Cox, 2007; Baker, 2008). Rather less research has tested the role of such properties in perceptual learning.

Some studies indicate that perceptual learning can under appropriate circumstances make reference to sub-phonemic information, either position-specific allophones or phonological features. Allen & Miller (2004) and Shockley, Sabadini & Fowler (2004) have shown that listeners can select the VOT appropriate for /t/ in a voice they have been exposed to; this could be interpreted variously as learning about a phonemic category, /t/, or a feature, [+ spread glottis], or a regional accent (e.g. Yorkshire, Minnesota), or an individual speaker characteristic ('breathy here and there'). Nielsen (2011) investigated spontaneous imitation of VOT, and found that after exposure to word tokens with lengthened VOTs, listeners extended VOT in their own production of words. Crucially, this finding generalised across place of articulation, i.e. after hearing extended VOT in words with initial /p/, subjects lengthened their own VOT not only in the exposed words with /p/, and novel words with /p/, but also—though to a lesser extent—in new words with initial /k/. This finding is consistent with the idea that perceptual learning about sublexical units makes reference to both features and phonemes, or articulatory or syllabic patterns. Kraljic and Samuel (2006) also showed that perceptual learning about the voicing contrast can be featural in nature. Kraljic and Samuel used a modified version of Norris et al. (2003)'s paradigm, exposing listeners to words containing ambiguous stops in place of either /t/ or /d/. They observed a subsequent boundary shift not only for a /t/-/d/ continuum, but also for a /p/-/b/ continuum where the sounds tested shared a feature, but no phoneme, with the exposed words.

Crucially, in an experiment investigating generalisation of perceptual learning from noise-vocoded speech, Dahan and Mead (2010) convincingly conclude that the primary linguistic level of

adaptation (learning) is allophonic and indeed coarticulatory: codas generalise better to codas than to onsets, and onsets better to onsets than to codas, and it helps to have the same vowel context too. Dahan and Mead "propose that the process by which adult listeners learn to interpret distorted speech is akin to building phonological categories in one's native language, a process where categories and structure emerge from the words in the ambient language without completely abstracting from them." (Dahan & Mead 2010: Abstract).

In contrast, studies using long-term repetition priming have typically not found strong or reliable evidence for the preservation of allophonic variation in memory for spoken language (McLennan, Luce, & Charles-Luce, 2003; Sumner & Samuel, 2005). In this paradigm, pairs of tokens that either share, or do not share, some aspect of phonetic form, are presented, separated by large numbers of intervening tokens. If better priming is found for those tokens that do share the aspect of form relative to tokens that do not share the aspect of form, it is assumed that the source of variation was retained in memory over at least the time period investigated. We discuss the long-term repetition priming studies in some detail because they raise methodological and interpretative issues that are important to the motivation for the current experiment; readers who wish to skip the detail should proceed to the final paragraph of this section.

One instance of allophonic variation that has been investigated is flapping. McLennan et al. (2003) conducted priming studies using 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 information about allophonic realisation was not being used (cf. Luce, McLennan & Charles-Luce, 2003). In contrast, when the participants' task was an easy lexical decision task (with very un-wordlike nonwords), the degree of priming was found to be sensitive to the allophonic realisation of the word. McLennan et al. offer an account of this finding in terms of adaptive resonance theory (for a succinct description for phoneticians, see Grossberg, 2003) and suggest that processing dynamics are such that allophonic detail is used only when the judgment task is easy or captures performance at an early stage of processing.

McLennan et al.'s account of why allophonic detail should affect early more than late processing is compelling for the particular stimuli and allophonic variants used. However, we speculate that allophonic effects could arise in more difficult tasks or later in processing, if different manipulations of allophonic detail were used. First, intelligibility-in-noise tests show that phonetic detail benefits listeners in difficult listening conditions (Ogden et al., 2000; Heinrich, Flory & Hawkins, 2010), and as yet we know little about whether this benefit affects early or late decisions. Second, and returning to McLennan et al. (2003), in natural speech, there can sometimes be subtle differences in flapped tokens according to the underlying voicing status of the consonant (Charles-Luce, 1997; Kwong & Stevens 1999). 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, the presence of a flap was uninformative regarding 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 a related experiment by Sumner and Samuel (2005), which investigated whether the extent of priming was affected by whether words were pronounced with different variants of word-final /t/: plain (non-glottalised) [t], glottalised and unreleased [ʔtˀ], and glottal stop with no supralaryngeal articulation [ʔ]. In an immediate semantic priming experiment, Sumner and Samuel found that priming was not affected by allophonic variant: for example, a token of *flute* with a glottalised unreleased [ʔtˀ] primed the related word *music* just as well as a token with a plain [t] did. However, in a long-term repetition priming experiment with a lexical decision task, strong priming was only found for plain [t]: for example, if the first presentation of *flute* had a plain [t], the lexical decision response to the second presentation of *flute* was faster if the second presentation also had a plain [t]; but if the first presentation had [ʔtˀ] or [ʔ], the response to the second presentation of *flute* was no faster if the second presentation matched the first presentation than if it did not. Sumner and Samuel conclude that

information about non-canonical allophonic variants (i.e., those other than plain [t]) is not retained in long-term memory, at least for these tasks that are about processing words heard in isolation.

Again, the negative results of Sumner and Samuel (2005) may be due to the role that allophonic detail played in their tasks. Allophonic detail can inform listeners about aspects of linguistic structure, and therefore assist in understanding meaning, but it does not always do so. For example, in Sumner and Samuel's (2005) experiment, the choice of one or other word-final variant of /t/ does not change the meaning of the word. The way in which plosives are released can be informative about, e.g., position in word, speech style, position in conversational turn (Local, 2003), or socio-indexical characteristics of the speaker, but these factors are not relevant in Sumner and Samuel's design, which presented isolated words; therefore, there was little reason for listeners to remember the details of plosive release for an extended period. Instead, the key task-relevant information was the phonemic identity of the final consonant, since filler items were nonsense words like *floop*, and this could explain why only plain [t]s (which presumably had the strongest place of articulation cues) showed long-term priming.

What none of the above studies have tested is whether allophonic variation and, in particular, speaker-specific allophonic variation, is learned when it does actually play a role in giving information about linguistic structure, and contributing to decisions about meaning. As other studies have suggested, linguistic relevance may be an important factor in the extent to which systematic variation is retained and used (e.g., Sommers & Barcroft, 2006; Nygaard, Burt, & Queen, 2000). Allophonic variation at word boundaries can indicate differences in meaning, as in the case of "junctural minimal pairs", i.e. identical phoneme strings that differ in the placement of word boundaries, e.g. *grey train—great rain* (Wyld, 1913; Jones, 1931; Lehiste, 1960; Hoard, 1966; Gårding, 1967; Umeda & Coker, 1975; Rietveld, 1980; Quené, 1992, 1993; Cruttenden, 1994). If speakers differ in how they use allophonic detail at word boundaries, this speaker-specific detail could potentially be useful for segmenting and identifying words, and might therefore be learned as a listener becomes familiar with a speaker.

1.2. The phonetics of word boundaries

The phonetics of word boundaries are subject to many influences: the words' segmental composition and syllabic structure, speech rate and degree of casualness, frequency (Cooper & Paccia-Cooper 1977), and the prosodic and grammatical groupings in which the words participate. For example, the higher the level of a prosodic boundary, the greater the 'articulatory strength' and duration of the segments immediately after the boundary (see Fougeron, 2001 for a review). Moreover, the presence of a boundary has greater effects on segment durations for pitch-accented than non-pitch-accented words (Turk & White, 1999; Turk & Shattuck-Hufnagel, 2000). Similarly, patterns of assimilation—which can help to indicate word boundaries—differ between function and content words (Local, 2003). Finally, statistical distributions of word and phrase usage can affect pronunciation, including, of course, in combination with all the above factors (Bybee, 2006; Jurafsky, Bell and Girand, 2002; Hay and Bresnan, 2006). The general point is that multiple factors influence how segments in a particular word or syllable are pronounced. In the present case, we manipulate the pronunciation of two words by placing a given phoneme (or short phoneme string) either at the end of one word or at the beginning of the next word. The primary variable is thus word juncture—whether a word boundary falls before or after the critical phoneme(s)—and this is marked phonetically by standard allophonic differences which reflect syllable structure. Thus by varying the location of a word boundary in identical phonemic strings, we inevitably and simultaneously vary syllabic structure and hence allophonic quality. Other influences on allophonic quality such as sentence prosody are held constant.

Word-boundary contrasts involve a wide range of phonetic properties. Among them are certain well-known allophonic patterns (e.g., for English, aspiration, glottalization, flapping, and clear and dark variants of /l/), and durational regularities (e.g. longer word-initial than word-medial or -final allophones). Less well-understood aspects include differences in intensity contour, spectral differences between word-initial and non-initial vowels (Lehiste, 1960) and differences in voice quality and spectral balance of consonants (Umeda & Coker, 1975). Word-boundary-related phonetic variation has been shown to

contribute to lexical identification, both in simple forced-choice identification tests (Lehiste, 1960; O'Connor & Tooley, 1964; Hoard, 1966) and in on-line priming and gating tasks (Gow & Gordon, 1995; Davis, Marslen-Wilson, & Gaskell, 2002). Therefore, several models of word segmentation and lexical access incorporate it as a potentially important influence on segmentation, e.g. Shortlist (Norris, McQueen, & Cutler, 1997), Good Start (Gow & Gordon, 1995), and the hierarchical model of Mattys, White & Melhorn (2005).

Between-speaker variation in the phonetics of word boundaries has not been systematically investigated. However—as is often the case with individual variation in both segmental and prosodic parameters—it is mentioned in passing in many speech production papers. For example, Lehiste (1960) noted a range of between-speaker differences in her study of English junctural minimal pairs. Quené (1992) investigated the duration of Dutch vowels in CV_1C # V_2C *vs.* CV_1 # CV_2C phrases (where # denotes a word boundary), and found that for some speakers, V_2 was longer in word-initial than non-initial position, while for others the opposite pattern was found. Fougeron & Keating (1997) examined articulatory variation at the beginning of hierarchically-structured prosodic domains, and observed that while all speakers tended to increase lengthening and strengthening of segments in successively higher prosodic domains, speakers varied considerably in the particular pairs of domains that they systematically distinguished.

The goal of this paper is to test whether non-phonemic phonetic detail at word boundaries is learned about in a speaker-specific way. We first present a production study to test whether different speakers of a single accent vary in the patterns of phonetic detail they use to achieve junctural contrasts, and then a perception experiment which explores the consequences of the variation for intelligibility.

2. Production experiment

The main focus of the production experiment is on inter-speaker variation in durational relationships around word boundaries. Realisations of word-initial vs. -final consonants and word-final

vs. non-final vowels are also briefly addressed. These properties do not exhaust the possible areas of inter-speaker variation, but were chosen as representative of the range of production variables and because of their likely perceptual importance. For example, durational relationships are basic to linguistic coherence (e.g. Klatt, 1976) and to individuals' idiosyncratic speech rhythms, and have been reliably shown to underpin word-structure differences in a number of experiments in Dutch and English (Lehiste, 1960; Lehiste, 1972; Davis et al., 2002; Salverda et al., 2003; Kemps, Ernestus, Schreuder & Baayen, 2005; Kemps, Wurm, Ernestus, Schreuder & Baayen, 2005; Hay & Bresnan, 2006).

We expected to find the general patterns documented in the previous section (e.g., onset consonants realised with longer duration and less lenition than coda consonants; syllable- and word-final vowels realised more peripherally than non-final vowels). Crucially, we also expected that individual speakers would vary in the extent to which they produced junctural distinctions and in the combinations of cues used to achieve them.

2.1. Method

2.1.1. Materials

There were 24 pairs of experimental sentences, arranged in four groups of six sentences each. They were phonemically identical but differed (usually grammatically) in a critical portion, underlined, e.g. *So <u>he diced</u> them* vs. *So <u>he'd iced</u> them*. These critical portions are shown in Table 1, and the complete sentences are in Appendix A. In each pair, between one and three phonemes had ambiguous word affiliation. The members of each pair are denoted Early Boundary and Late Boundary (henceforth EB, LB respectively) according to where in the phonemic sequence the first word boundary falls (e.g., for /hiːdaɪst/, EB is *he # diced*, LB *he'd # iced*). Each experimental sentence was embedded in a meaningful context (Appendix A), e.g.

EB: He wanted the carrots to cook fast. *So <u>he diced</u> them*.
LB: The top of the cakes had come out looking uneven. *So <u>he'd iced</u> them*.

Table 1 shows the four groups of six sentence pairs, designed as follows. In Group D, e.g. *So <u>he diced</u> them—So <u>he'd iced</u> them*, /d/ was either in the syllable onset of word 2 (EB) or else was the reduced form of the auxiliary verbs *had* or *would*, suffixed to word 1 (LB, e.g. *he'd*). In Group S, e.g. *Those are <u>cat size</u>—Those are <u>cat's eyes</u>*, /s/ was in the syllable onset of word 2 (EB) or was a suffix to word 1 (LB, e.g. *cat's*, *likes*). Phonologically, these Group D and S suffixes constitute an 'appendix' to the syllable according to Ogden (1999). In Group A, e.g. *<u>That surprise</u>—<u>That's a prize</u>*, a weak syllable, /sə/ or /və/, occurred as the initial syllable of word 2 (EB); or the weak syllable's initial /s/ or /v/ formed the end of word 1, while word 2 was *a* or *are* (both /ə/ in the non-rhotic accent used), and word 3 was a monosyllable (LB). In Group T, e.g. *I <u>lay Steve's</u> costume out in the wings—I <u>laced Eve's</u> costume out in the wings*, /st/ was the onset of word 2 (EB), or was at the end of word 1, where /t/ was a past tense marker (LB).

The 8 tokens of each of the 48 experimental sentences (each with its context) were randomised, and this list of 384 experimental sentences + contexts was interleaved randomly with 216 filler sentences, with the restriction that tokens of the same sentence, or its pair, did not appear next to one another. The complete sentence list thus had (8(48) + 216 =) 600 sentences.

Table 1 Critical portions (Early Boundary—Late Boundary) of the phonemically-identical sentence materials, and measurements of segmental durations made (see text for details, this section and section 2.2.1). The four groups are defined by the phonological and/or grammatical nature of the difference between the members of the pair.

Group D **Measurements: D:Prec Syl | D:d | D:FolSyl**		**Group S** **Measurements: S:PrecSyl | S:s | S:FolSyl**	
he diced—he'd iced	hiː|d|aɪst	*cat size—cat's eyes*	kat|s|aɪz
she dyed—she'd eyed	ʃiː|d|aɪd	*collect skulls—collects gulls*	lɛkt|s|{k/g}ʌlz
we dared—we'd aired	wiː|d|ɛəd	*Pete stole—Pete's dole*	piːt|s|{t/d}əʊl
we dread—we'd read	wiː|d|rɛd	*eat sweet—eats wheat*	iːt|s|wiːt
we drank—we'd rank	wiː|d|rank	*like psalms—likes arms*	laɪk|s|ɑːmz
I drove—I'd rove	aɪ|d|rəʊv	*Pat sawed—Pat's awed*	pat|s|ɔːd
Group A **Measurements:** **A:PrecSyl | A:s | A:ə | A:C | A:FolSyl**		**Group T** **Measurements:** **T:PrecSyl | T:s | T (T:cl, T:VOT) | T:FolSyl**	
that surprise—that's a prize	ðat|s|ə|p|raɪz	*lay Steve's—laced Eve's*	lɛɪ|s|t|iːvz
Ralph surpasses— *Ralph's are passes*	ralf|s|ə|p|ɑːsɪz	*whack Stan's—waxed Ann's*	wak|s|t|anz
		Mick stability—mixed ability	mɪk|s|t|əbɪlətiː
Ruth sustained— *Ruth's are stained*	ruːθ|s|ə|s|tɛɪnd	*sly stroll—sliced roll*	slaɪ|s|t|rəʊl
		eye strain—iced rain	aɪ|s|t|rɛɪn
that salute—that's a lute	ðat|s|ə|l|uːt	*play strangers—placed*	plɛɪ||s|t|rɛɪndʒəz
that surround— *that's a round*	ðat|s|ə|r|aʊnd	*rangers*	
say veneer—save an ear	sɛɪ|v|ə|n|ɪə		

2.1.2. Speakers

The six speakers all had similar accents of Standard Southern British English. Table 2 shows their demographic characteristics. The speakers did not know the purpose of the experiment, except for RS (the first author) who was chosen because she has a fast, casual reading style suitable for these materials. The decision to use speakers of different ages and genders was for purposes of another experiment (Smith, 2003, 2004) but in this experiment allowed for examination of variation in the realization of the forms under investigation.

Table 2 Speakers in the production experiment (age in years) and degree of phonetic training.

	Young	Mid	Old
Male Phonetic training:	MJ (27) some training	JR (35) trained phonetician	PF (53) no training
Female Phonetic training:	RS (25) trained phonetician	SC (45) no training	AK (54) no training

2.1.3. Recordings

Recordings were made in a double-walled IAC booth using a Sennheiser MKH40 P48 condenser microphone and a Sony 55ES High Density Linear A/D D/A recorder, and were digitised at 16 kHz to a Silicon Graphics machine running the commercial version of ESPS and *xwaves+* (ESPS is now available with WaveSurfer from http://www.speech.kth.se/software/#esps). Speakers read as naturally and informally as possible. Because the critical parts involve frequently-occurring grammatical morphemes, they were easy to speak in a natural way. Each speaker was allowed to self-correct as often as necessary during the recordings, and no token was accepted unless it sounded (a) natural (b) fluent (c) prosodically appropriate and (d) acceptable for the intended meaning to the experimenter (the first author), who is a trained phonetician. Difficult cases were listened to by the second author, also a trained phonetician. The criterion for prosodic appropriateness was that there be no significant prosodic differences between the two members of any pair, or between speakers, as determined by the authors' auditory judgments, in that the realisations would be transcribed using identical symbols for stress and intonation. Where necessary, extra tokens were recorded to achieve a total of 8 tokens per sentence per speaker.

2.2. Analysis

2.2.1. Durational measurements

Durations were measured using *xwaves.* For all sentence groups, the durations of the syllable or syllable fragment preceding and following the critical part were measured (henceforth PrecSyl, FolSyl). For example, for *So he diced them—So he'd iced them,* the critical part was /d/, PrecSyl was /hiː/ and FolSyl was /aɪst/. Thus PrecSyl either corresponded to an entire syllable and word (/hiː/ in *he diced*) or to the onset and nucleus of a syllable (/hiː/ in *he'd iced*). FolSyl normally corresponded either to an entire syllable and word (/aɪst/ in *he'd iced*) or to the rhyme of a syllable (/aɪst/ in *he diced*). In some cases, the FolSyl measurement also included an onset consonant e.g. /rɛd/ in *we dread / we'd read.*

Measurements of the critical part differed between sentence groups, and are listed in Table 1. For Groups D (*he diced—he'd iced*) and S (*cat size—cat's eyes*), the critical part was a /d/ or /s/ respectively, and its duration was measured (/d/ from start of closure to start of periodic vocalic formant structure, henceforth D:d, /s/ from onset to offset of frication, henceforth S:s). For the other sentence groups, the critical part contained more than one segment. In Group A (*that surprise—that's a prize*), the critical part was a CVC sequence e.g. /səp/ for this example. Durations of the C, V and C were measured separately (henceforth A:s, A:ə and A:C). In Group T (*lay Steve's costume—laced Eve's costume*), the critical part was a /st/ cluster, and durations of the /s/, /t/ closure, and VOT were measured separately, henceforth T:s, T:cl, T:VOT. Predictions were that the components of the critical part would be longer when word-initial than word-final, to different extents for different speakers.

VOT was treated differently in Groups D and T, i.e. for /d/ and /st/ respectively. The measure for /d/ included VOT, i.e. closure duration and VOT were not measured separately. VOT and closure duration are likely to be subject to different influences, e.g. VOT is subject to aerodynamic influences

related to segmental composition, in addition to the syllabic/prosodic influences that are our main focus. Nevertheless, we did not measure VOT separately for three reasons: 1) there are no phonemic differences within each sentence pair; 2) we expected the syllabic/prosodic influences on VOT to be in the same direction as those for the closure; 3) we (correctly) expected VOT to vary relatively little in /d/ within most pairs, but to be perceptually important in /dr/ sequences, because it is long in /dr/ but short in the other cases. In contrast, for /st/ clusters we measured VOT and closure duration separately, in case orthography caused a bias towards shorter VOT in LB than EB items (/t/ corresponds to orthographic <ed> in LB items, e.g. *laced Eve's costume*, but to orthographic <t> in EB items, e.g. *lay Steve's costume).* The difference was on the cusp of significance ($p = 0.050$), but in the opposite direction from that predicted; it turned out to be driven by only one speaker, with the other five speakers showing no difference.

An index of speech rate was calculated: Critical Phrase Duration = PrecSyl + critical part + FolSyl. Mean Critical Phrase Duration differed considerably among the speakers (RS: 553 ms; JR: 563 ms; MJ: 606 ms; AK: 658 ms; PF: 679 ms; SC: 735 ms). The speakers with phonetic training tended to speak slightly faster than the non-trained speakers. Impressionistically, they were slightly more fluent readers, presumably due to practice with similar reading tasks. To control for these rate differences, statistical analyses on individual syllables and segments used relative durations, expressed as proportions of Critical Phrase Duration. Statistical results are, however, very similar when absolute durations are considered. That is, of 48 terms tested in the statistical models, 40 had the same significance status (and the same direction of difference, if a difference was present), regardless of whether absolute or relative durations were tested. Of the remaining eight terms, five were significant in the analyses of relative, but not absolute durations; and three were significant in the analyses of absolute, but not relative duration.

Thus, the choice to use relative durations did not artificially inflate the significance of the results. Moreover, only two of the eight divergences related to the critical part of the phrase itself; most related to the preceding or following syllable.

2.2.2. Non-durational measurements

Group D was chosen for investigation of non-durational allophonic differences. Initial observation indicated, amongst other things, that the critical /d/ had a range of realizations from canonical stops to voiced fricatives, while the preceding vowel varied in quality.

For /d/, voicing and the presence of formant structure in the F2-F3 region during the constriction portion were examined using wideband spectrograms, and categorised either as continuous through the constriction portion, or not. Frication during the constriction portion was examined using the waveform and spectrogram, and categorised as present or absent. No attempt was made to distinguish partial vs. continuous frication, or frication that occurred at the beginning vs. end of the stop closure period Predictions were that /d/ would be shorter and more weakly articulated in word-final than –initial position, and therefore that voicing, formant structure and frication during closure would occur more frequently in word-final than –initial tokens. This pattern was also expected to be speaker-dependent.

For the five sentences where the vowel preceding /d/ was /iː/, the frequencies of F1 and F2 were measured in the middle of the steadiest part of F2, using Praat's formant tracker (Boersma & Weenink, 2006), with 12-pole Burg lpc analysis below 5.5 kHz, and a 25-ms Gaussian window, step size 6.25 ms. Inaccurate values were hand-measured using 25-ms fft spectra in conjunction with the wideband spectrogram. The difference between F1 and F2 frequencies (henceforth F2-F1) was calculated in Bark

(Traunmüller, 1990). F2-F1 was expected to be larger, suggesting a more peripheral vowel, when /iː/ was word-final (e.g. *he*) than non-final (*he'd*).

2.2.3. Statistical analysis

The durational and formant measures were analysed in R using mixed-effects models (Baayen, 2008), which allow multiple random effects to be specified, e.g. for subjects and items. For each variable, the model-fitting procedure was as follows: first a full model was fitted, with Sentence as random effect, and Speaker (MJ, JR, PF, RS, SC, AK), Boundary Position (Early, Late) and their interaction as fixed effects. Speaker was treated as a fixed effect because of the constraints that had gone into speaker selection. Predictors that did not significantly contribute to the model were incrementally removed until the simplest model had been found. Log-likelihood tests were used to check model fit as predictors were removed. Sensitivity of the results to outliers was investigated by fitting models with and without the removal of outliers with a standardised residual at a distance greater than 2.5 standard deviations from zero (or 2 in the case of the two variables with the least normally distributed residuals, A:ə and S:PrecSyl); results were consistent and those reported are with outliers removed. Outliers constituted no more than 3% of the data (5% for A:ə).

The measures of voicing, formant structure and frication for intervocalic /d/ were analysed using generalized linear mixed-effects modelling with logistic regression, following the same model fitting procedure.

Pairwise comparisons, where reported, use a Bonferroni—Holm procedure (Holm, 1979). The Bonferroni—Holm procedure is a stepwise procedure, based on ordered *p*-values. First, the comparison with the most extreme (i.e. smallest) *p*-value is evaluated, using alpha level 0.05/*n*, where *n* = number of

comparisons. If the null hypothesis is rejected for this comparison, the next most extreme *p*-value is evaluated, with alpha level 0.05/*n*-1. This procedure continues until a comparison is reached for which the null hypothesis has to be accepted, after which no further comparisons are evaluated.

Tables detailing the statistical results are presented in Appendices; differences reported in the main text as statistically significant always had $p < 0.05$ or better, and in most cases $p < 0.001$.

2.3. Results and Discussion

2.3.1. Durational results

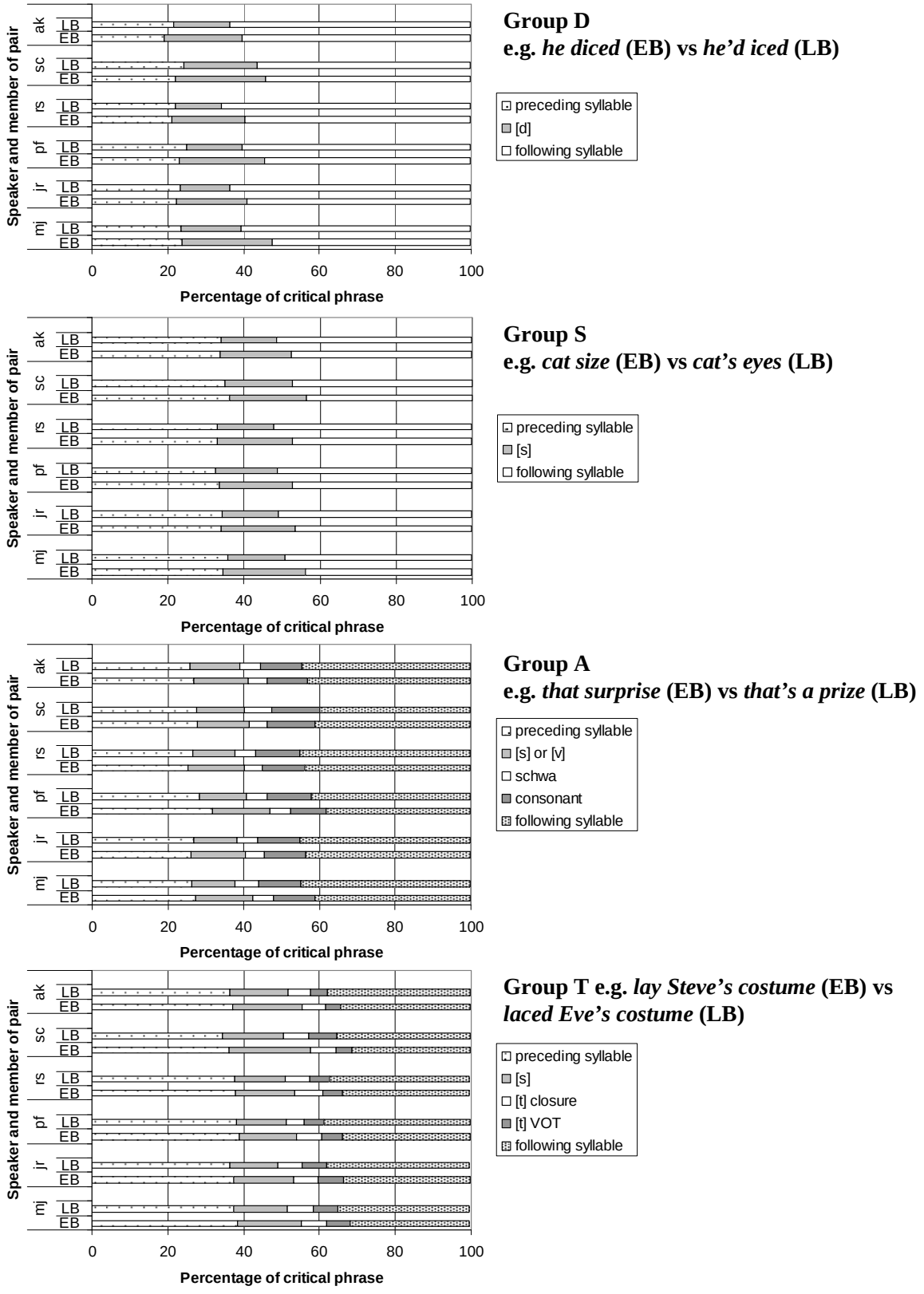
Though the main focus of the experiment was the extent of speaker-specific variation in the marking of word boundaries, we first describe results for general patterns due to word boundary placement. As Figure 1 and the tables of statistical results in Appendix B show, these were consistent with the literature. As expected, consonants at the crucial word boundary in the critical phrase were always proportionately longer when word-initial than word-final (i.e. EB > LB for variables D:d, S:s, A:s, T:s). The rhyme of the second word in the critical phrase was likewise almost always proportionately longer when word-initial than when non-initial. For example, /aɪst/ was longer in *he'd iced* than in *he diced* (FolSyl: LB > EB). The duration of the onset+nucleus of the first word in the critical phrase was more variable, with only two sentence groups showing any systematicity across speakers, and even then, the difference in these two groups was in opposite directions. In pairs like *lay Steve's costume* vs. *laced Eve's costume*, /leɪ/ was longer in *lay* than *laced* (T:PrecSyl: EB > LB); in contrast, in pairs like *he diced* vs. *he'd iced*, /hiː/ was longer in *he'd* than *he* (D:PrecSyl: LB > EB).The difference between the T and D groups may be due to a number of factors: e.g. the presence of a voiceless complex coda in *laced*, and the

location of the morpheme boundary in the case of *he'd*, in conjunction with *he* being a metrically weak (i.e. unstressed) function word.

Speaker-specific patterns, which are the main focus of our investigation, modulated the above general trends to a considerable extent. Figure 1 shows the speaker-specific durational results; statistical results are in Appendix B. For 13 out of the 16 durational variables investigated, significant interactions were found between Speaker and Boundary Position, indicating significant inter-speaker variability in the way that the word boundary contrasts were realised. (The exceptions were S:FolSyl; A:ɔ; T:PrecSyl.)

Specifically, for some variables, all speakers made word boundary distinctions in qualitatively the same way, but differed in the magnitude of the distinction. All speakers had longer onset than coda consonants, but to differing extents (Appendix B, Table B2: D:d, S:s, A:s and T:s). For example, /s/ was 53 ms (or 56%) longer when in onset than coda position for MJ, but only 23 ms (17%) longer for SC (S:s). Onset /d/ was 38 ms (60%) longer than coda /s/ for PF in sentence group D, but only 21 ms (42%) longer for JR (D:d). Similarly, in two sentence groups, the word rhyme of the second word was shorter for all speakers when the critical consonant belonged in the onset of the second word (e.g. *diced*), than when it did not (e.g. *iced*); but speakers differed in the magnitude of this difference. For example, /aɪst/ was 30 ms (13%) longer in *iced* than *diced* for MJ, but only 9 ms (3%) longer for SC (Appendix B, Table B2: D:FolSyl, T:FolSyl).

Figure 1 Mean durations in the four sentence groups, expressed as percentages of total measured sequence. "Preceding syllable" refers to the syllable (or syllable fragment) before the part of the phrase whose word affiliation is ambiguous; "following syllable" refers to the syllable (or syllable fragment) after the critical part. For example, for *So he diced them—So he'd iced them* (Group D), "preceding syllable" refers to /hiː/ and "following syllable" to /aɪst/.



Other durational variables showed more extensive speaker-specific variation, in that not all speakers made a significant durational distinction according to the word boundary placement. This was the case for the onset+nucleus in pairs like *he* vs. *he'd* (D:PrecSyl), which half the speakers distinguished durationally and half did not. Some of the critical consonant segments—especially those that were not in absolute word-initial position—also showed idiosyncratic variation. The closure of /t/ (T:cl) was 16 ms (57%) longer in *lay S<u>t</u>eve's costume* than *lace<u>d</u> Eve's costume* for speaker PF; and the VOT of /t/ (T:VOT) was 24 ms (80%) longer in *lace<u>d</u> Eve's costume* than *lay S<u>t</u>eve's costume* for speaker SC. PF's consonants were also significantly longer (17 ms or 21%) when they were word-initial (e.g. /p/ in *prize*) than when syllable-initial but word-medial (/p/ in *surprise*; A:C), but no other speaker showed this pattern.

Finally, for a third group of the most idiosyncratic variables, a significant word-boundary-related difference went in one direction for some speakers, and the opposite direction for others. Word fragments like /raɪz/ were longer in monosyllables (*that's a p<u>rize</u>*) than disyllables (*that surp<u>rise</u>*) for most speakers, but the other way round for speaker SC (A:FolSyl). For PF, fragments like /ðat/ and /kat/ were longer when word-final (in *<u>that</u> surprise* and *<u>cat</u> size*) than when non-final (in *<u>that</u>'s a prize* and *<u>cat</u>s' eyes*), but for speakers RS and MJ the opposite pattern was found, and the remaining speakers showed no difference (A:PrecSyl; S:PrecSyl; Appendix B, Table B2).

Some speakers were more systematic than others in their use of duration to distinguish word boundaries. PF was the speaker who had significant differences for the most variables (15 of the 16 tested, whereas the other speakers had between 10 and 12). PF also had significantly larger durational differences than other speakers for five variables, compared to four for MJ, two for SC, RS and JR, and only one for AK (Appendix B, Table B2). Interestingly, this systematicity does not relate

straightforwardly to age or speech rate. PF was not the slowest speaker, and indeed had a speech rate very similar to that of AK, who showed least systematicity and is the same age as PF. SC, the slowest speaker, was amongst the less systematic in marking her word boundaries: she had smaller EB-LB differences than other speakers for five variables. It is possible, of course, that the mean speech rate of each speaker is not the most informative predictor; instead the speech rate of the particular sentence token might predict the use of durations to distinguish minimal pairs. When analyses were re-run to take account of this, i.e. by including the total duration of each critical phrase as a predictor, the interactions between Boundary Position and Speaker remained significant, supporting our conclusion that rate is not the only or main contributor to the results.

In summary, the durational results show considerable speaker-specificity. The main patterns—lengthening of word-initial relative to word-final consonants, and longer rhymes for syllables without a consonantal onset than with one—are exhibited to some extent by all speakers, but implementation of them is variable across speakers in terms of the magnitude of the distinctions. Other durational patterns are used systematically by only some speakers. Certain speakers appear generally more systematic in their use of durations than others, in a way that is not predictable from speech rate alone.

2.3.2. Consonant realisation

Figure 2 shows spectrograms illustrating the extremes of /d/ realization observed, and some intermediate types. The descriptions refer to the period of maximum oral constriction in each case, i.e. the closure in a canonical case and its equivalent in other cases: a) a “canonical /d/”, partially voiced, with no formant structure or frication during its closure, b) /d/ with continuous voicing and partial formant

structure; c) /d/ with continuous voicing, partial formant structure and partial frication; d) a "fully lenited /d/", with continuous voicing, continuous formant structure, continuous frication and no burst.

As expected, coda /d/ (LB, e.g. *he'd*) was continuously voiced significantly more often than onset /d/ (EB, e.g. *diced;* Boundary Position, χ^2 (1) = 159.3, $p < 0.0001$). Table 3 shows that this pattern was found for all speakers, but the speakers also differed significantly in the overall extent to which they used continuous voicing (Speaker: χ^2 (5) = 269.1, $p < 0.0001$). Table 3 shows that PF and JR had the highest proportions of continuously voiced tokens (80% and 73%), RS, MJ and AK voiced an intermediate number of tokens (42%, 28%, 19%), and SC voiced the least (4%). AK and SC only ever used continuous voicing in coda /d/, and never in onset /d/, while JR and PF used continuous voicing not only in nearly all their coda /d/s, but also in many of their onset /d/s. The inter-speaker differences cannot be a simple function of the duration of the closure: PF had much longer closures than JR (83 ms vs 59 ms), but very similar closure durations to AK (75 ms) and MJ (83 ms), who voiced significantly less during /d/ closure.

Figure 2 Spectrograms illustrating four variants of /d/ in *He was pleased that we'd aired them* (shown: /wiːdɛə/ and part of following /d/). The progressive weakening of articulation is shown from top to bottom: a) partial voicing, speaker AK; b) continuous voicing and partial formant structure, speaker RS; c) continuous voicing, partial formant structure and partial frication, speaker AK; d) continuous voicing, continuous formant structure, continuous frication, speaker PF.

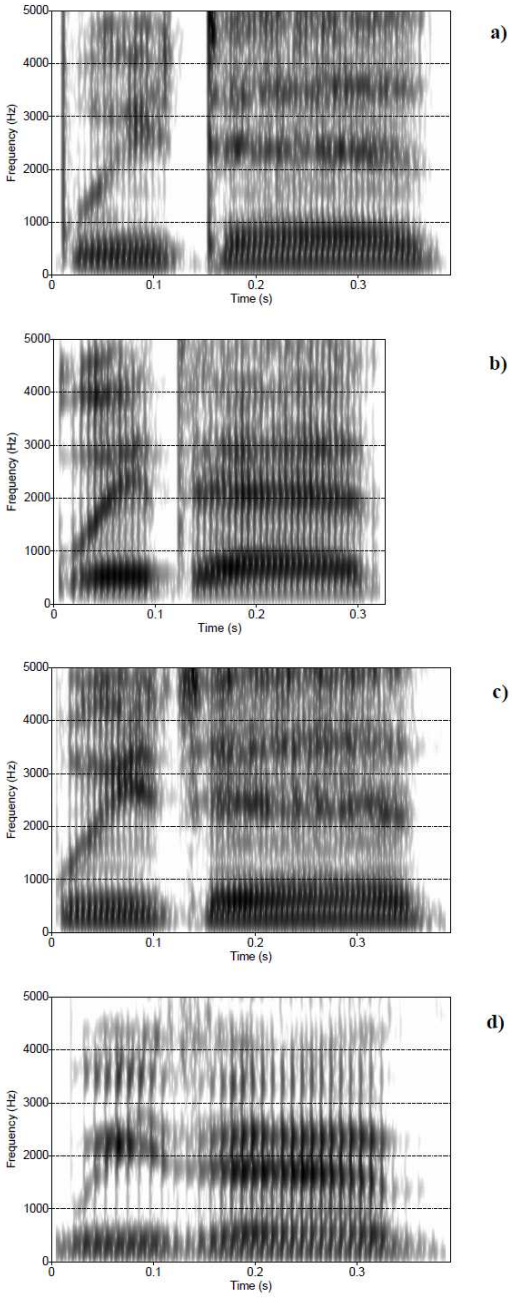


Table 3 Allophonic variation in tokens of /d/ in pairs like *he diced* vs. *he'd iced*. Percentage of tokens that exhibited continuous voicing during closure, continuous formant structure during closure, and presence of frication during closure; and significant pairwise comparisons among speakers.

	MJ	JR	PF	RS	SC	AK
	total % EB LB	total % EB LB	total % EB LB	total % EB LB	total % EB LB	total % EB LB
continuous voicing	28 8 48	73 48 98	80 66 94	42 6 77	4 0 8	19 0 38
	Inter-speaker differences: all speakers differ significantly from all other speakers ($p < 0.05$), except PF = JR and AK = MJ					
continuous formant structure	6 4 8	22 8 36	19 9 29	7 0 14	0 0 0	0 0 0
	Inter-speaker differences (all $p < 0.05$): (SC = AK) < (RS = MJ) < (PF = JR)					
presence of frication	19 4 33	17 4 30	15 2 27	10 2 19	2 2 2	6 0 12
	Inter-speaker differences: SC < MJ, SC < JR					

Continuous formant structure during the constriction was also significantly more prevalent in coda than onset /d/ (Boundary Position: χ^2 (1) = 22.6, $p < 0.0001$). Significant individual differences were found here too (Speaker: χ^2 (5) = 51.2, $p < 0.0001$), e.g. Table 3 shows that female speakers SC and AK exhibited continuous formant structure in none of their tokens; male speakers JR and PF did so in 22% and 20% of their tokens. As above, these differences cannot result straightforwardly from closure duration.

As expected, frication was very infrequent in onset /d/, yet occurred in about a fifth of coda /d/s (2% vs 21% of tokens, Boundary Position: χ^2 (1) = 47.5, $p < 0.0001$). Significant individual differences were again found (Speaker: χ^2 (5) = 21.6, $p < 0.001$). In general the male speakers fricated their /d/s more frequently than the females, but the only significant inter-speaker pairwise comparisons were between SC (2%) *vs.* JR (17%) and MJ (19%).

2.3.3. Spectral properties of /iː/

Figure 3 Bark frequencies of F1 vs. F2-F1 for /iː/, for each speaker separately. Early Boundary (EB): circles and solid-line ellipse. Late Boundary (LB): triangles and dashed-line ellipse. Ellipses contain the 70% of data points closest to the mean. Solid symbols indicate the mean of each distribution.

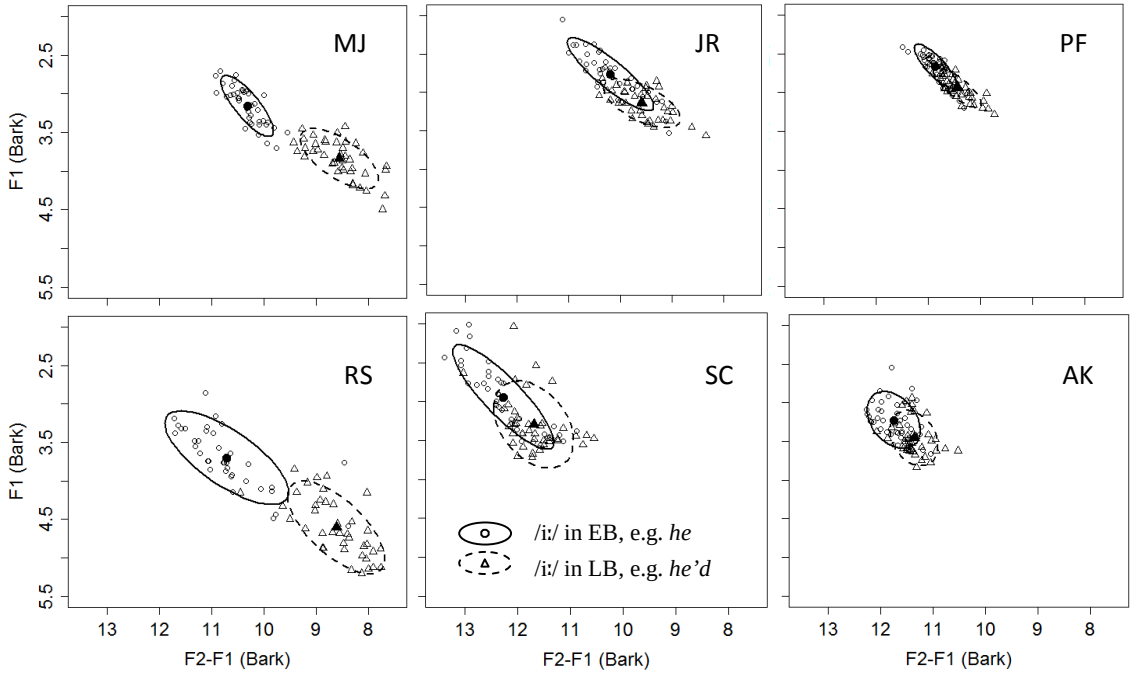


Figure 3 shows formant frequency measurements in Bark for /iː/ in *he('d), she('d), we('d).* As expected, the F2-F1 difference in Bark was significantly greater in EB (*he)* than LB (*he'd)* for all speakers (Appendix B, Tables B1 and B2), reflecting a more peripheral articulation in the former case i.e. in the monomorphemic, open syllable. As Appendix B: Table B2 shows, RS and MJ have a significantly greater EB-LB difference than any of the other speakers (mean 2.1 and 1.8 Bark respectively, as opposed to between 0.4 and 0.6 Bark for the other speakers). RS and MJ's LB vowels sound closer to /ɪ/ than /iː/. They are two of the three speakers who did not distinguish the duration of pairs like *he* vs. *he'd*, suggesting a possible trade-off between spectral and temporal marking of the word boundary. RS and MJ

are also the two youngest speakers; the two oldest speakers, AK and PF, have more or less the smallest spectral differences.

2.4. Discussion

The production experiment examined proportional durations, consonant realisation and vowel formant frequencies at word boundaries in natural, meaningful minimal pairs, with main focus on inter-speaker variation. Although the results showed the trends that are expected from the literature on mean data, what the experiment has also shown is that there are systematic patterns that distinguish speakers. Quantitative, and in several cases qualitative, inter-speaker variation was the norm for the variables investigated, though the extent and type of variation depended on the particular phonetic attribute: some variables were more consistently used than others.

A number of the durational variables were used consistently by all speakers. Most involved well-known syllable- and word-initial vs. -final consonantal contrasts (e.g. Lehiste, 1960) and support the view that initial position is phonetically strong. Others involved syllable rhymes, which were proportionately longer in onsetless syllables than after a consonantal onset, congruent with results from a corpus study by van Santen (1992). The consistency in the way speakers make these durational distinctions may be because they are important for word segmentation, and/or because they reflect basic aspects of the articulatory organisation of the syllable (Krakow, 1999). Nevertheless, significant inter-speaker variation in the *magnitude* of the contrasts was observed, and may be perceptually important. Allen, Miller and DeSteno (2003) report a similar pattern of quantitative inter-speaker variation for VOT in word-initial stops, which contributes to perception of speaker identity (Allen and Miller, 2004).

The spectral variable measured, the peripherality of /iː/ in open versus closed syllables, was also used consistently by all speakers: all speakers had a larger F2-F1 difference when the vowel was in an open syllable, e.g. *he,* than in a closed syllable, e.g. *he'd* (cf. Stevens & House, 1963). However, the behaviour of the two speakers with the most extreme differences goes beyond the expected degree of

centralisation of /iː/ in the contexts used: their *he'd* vowels sounded closer to /ɪ/ than /iː/. This quality difference probably relates to the morphemic status of the words, in that this open-closed syllabic distinction is also a morphemic distinction: *he* is monomorphemic while *he'd* is bimorphemic. Ogden (1999) analyses similar forms (e.g. (h)ɪz, ʃɪz, ðɪv for *he's, she's, they've*) as resulting from phonological fusion of non-syllabic weak clitic forms of auxiliaries with weak forms of the pronouns that host them. Because /iː/ and /ɪ/ do not contrast in these function words, unlike in comparable content word pairs such as *heed—hid*, considerable variation in the vowel quality is possible. In the present data, systematic variation that accompanies the presence or absence of the clitic also appears to be age-related. Only the youngest speakers in the cohort had extreme vowel quality differences. The oldest speakers had the smallest differences, and the mid-age speakers had intermediate differences.

The variables related to realisation of onset and coda /d/ were fairly consistent across speakers. In coda as compared to onset stops, all speakers had a higher probability of voicing and formant structure, and all but one had a higher probability of frication. The patterns for voicing and formant structure are consistent with acoustic and laryngoscopic observations by Umeda & Coker (1975). The pattern for frication is consistent with lenition in the form of an incomplete closure or a slowly made or released closure. (It could also arise due to higher oral pressure, but oral pressure seems unlikely to be higher in a short, word-final /d/ than a longer, word-initial /d/.) There were also significant speaker differences, and some indication of gender differences, with voicing, formant structure and frication generally more prevalent among male than female speakers. The greater prevalence of voicing and formant structure for male speakers may be a consequence of males' larger oral tracts, which can maintain a transglottal pressure drop for a longer duration. There is no obvious anatomical explanation, however, for why males had more frication than females during part or all of the constriction, and sociological causes may be more plausible than physiological ones. More work is needed to fully understand this pattern.

Finally, for some variables, qualitative as well as quantitative variation was found among speakers, i.e. not all speakers produced a significant contrast, or different speakers did so in opposite directions. These variables typically concerned durational adjustment processes for which the evidence in the literature is mixed, such as polysyllabic shortening in right-headed words (i.e. pairs like *prize* vs. *surprise*; Turk & Shattuck-Hufnagel, 2000). Inconsistently used variables such as these have interesting implications for perception: listeners may tune in to those which a speaker uses systematically, and learn to ignore those that are not informative.

As well as differences between speakers, the results might also reflect differences in the speech style (e.g. the degree of casualness) that the speakers either adopt habitually, or that they adopted for the recording. Although an interpretation of the data in terms of speech style cannot be ruled out, it does not correspond very well to our perceptual impression. While some pairs of speakers did differ in style—in particular, SC sounded more careful than JR—other speakers, such as AK, MJ and PF, all sounded similar in speech style, but differed in patterns of phonetic detail. Therefore, we consider the speaker interpretation more likely, and adopt it in the remainder of the paper. However, we acknowledge that whether the variation is about speaker or style or both is not critically important for the broad perceptual aim of this research (i.e. to understand whether or not listeners can carry out perceptual learning about the phonetic implementation of syllabic and prosodic structure, as well as about phonemic categories).

2.5. Summary

The present study demonstrated inter-speaker variation in durational and some non-durational properties. However, it far from exhausts the scope for inter-speaker variation in properties affected by word juncture. Other parameters include pitch, intensity, phonation quality, and additional spectral properties. Ultimately, it will be important to assess whether the numerous variables involved in word boundary distinctions can be captured by a 'speaker space' with lower dimensionality. For example,

Krakow (1999) observes two patterns for syllable-final consonants in casual speech, one in which the primary consonant constriction is weakened or lost, and another in which it is strengthened, i.e. shows articulatory organization more like that of an initial consonant. A speaker's preference for one or other of these types of organization might give rise to particular clusters of differences.

What matters most for the present purpose, however, is that systematic inter-speaker differences in phonetic detail at word boundaries exist. Speakers display subtly (and, in some cases, not-so-subtly) different phonetic patterns around word junctures. If the set of variables investigated here is representative, some speakers also mark boundaries more systematically than others. Similar fine-grained differences in VOT have demonstrable perceptual relevance (Allen and Miller 2004). It is plausible that exposure to a voice will result in learning those cues that assist with word segmentation in that voice. The aim of the perception experiment was to test this hypothesis by testing speech intelligibility in noise. Embedding speech in noise has been shown to produce sufficient errors of word identification for gross and subtle influences on processing to be assessed (Miller, Heise, & Lichten 1951, Heinrich et al., 2010). Speaker familiarity helps word identification in noise (Nygaard, Sommers, & Pisoni, 1994; Nygaard & Pisoni, 1998) while correct phonetic detail improves intelligibility of synthetic speech in noise (Ogden et al., 2000). Therefore, familiarity with a speaker's allophonic patterns at word boundaries is also expected to improve segmentation in noise.

3. Perception experiment

3.1. Overview and Design

The perception experiment consisted of a pre-test, familiarization phase and post-test. In the pre- and post-test, subjects heard the sentences from Experiment 1 in background noise without their preceding contexts, and typed (in standard orthography) what they heard; in the familiarization phase they

heard the same sentences in context and without noise, and answered questions about their meaning. Performance between pre-test and post-test was compared.

Each participant heard the same voice in the pre- and post-tests. Two crossed factors were manipulated: Voice heard in tests (speaker PF or MJ) and Familiarization Voice (Same or Different to the test voice). Thus the critical question was whether word segmentation in the post-test was facilitated by having had exposure to the same voice in the familiarization session.

3.2. Method

3.2.1. Materials

Materials were the 48 sentences (24 pairs) from the production experiment. Tokens spoken by two of the male speakers PF and MJ were used. These two speakers were selected because they were the most suitable of the six speakers, in that, of the four who spoke with a similar moderate degree of casualness (RS, MJ, AK, and PF) they were the two who shared the same gender, had the most similar speech rates, and were naïve as to the purpose of the experiment. Their realization of word boundaries showed a mixture of similarities and differences (details in Appendix C). First, both speakers showed some basic durational differences between onset and coda consonants, but PF also produced several additional subtle durational differences that MJ did not. Durations are known to be important for identifying morphological structure and word juncture (e.g. Lehiste, 1960; O'Connor & Tooley, 1964; Hoard, 1966; Gow & Gordon, 1995; Davis, Marslen-Wilson, & Gaskell, 2002; Kemps, Ernestus, Schreuder, & Baayen, 2005; Kemps, Wurm, Ernestus, Schreuder, & Baayen, 2005) and also to be learnable aspects of idiosyncratic behaviour (Allen & Miller, 2004; Nielsen, 2011). Second, the two speakers differed with respect to the spectral structure of vowels and details of stop consonant realisation, whose role in juncture perception and speaker adaptation is less well tested, but which are—in common with many other phonetic properties—likely to be learned about where relevant to the task (cf., for vowels, Sidaras, Alexander, & Nygaard, 2009).

Stimuli for the pre- and post-test were made from tokens of the 48 sentences (e.g. *So he diced them*). For each speaker, one token of each sentence was selected at random from the eight originally recorded in the production experiment. This chosen token was mixed with randomly-varying cafeteria noise at an average signal-to-noise ratio of +2 dB (average amplitude of sentence:average amplitude of noise). The noise was ramped up to its maximum amplitude over 5 s before the sentence began, continued at this average amplitude for 15 s after sentence offset and was then ramped down to zero over a further 5 s. The 15 s response period was established in pilot tests as necessary for the response to be typed.

Stimuli for the familiarization session were tokens of the same 48 sentences, each in its disambiguating context (e.g. *He wanted the carrots to cook fast. So he diced them.*). For each speaker, six different tokens of each context+sentence were selected at random from the seven remaining after random selection of the pre- and post-test token.

In each phase (pre-test, familiarization and post-test), sentences were presented in pseudo-random order with the constraint that members of a sentence pair never occurred adjacent to one another.

3.2.2. Participants

Participants were 80 speakers of Standard Southern British English (SSBE), 20 in each of four conditions: 1) Test Voice MJ, Same Familiarization Voice; 2) Test Voice MJ, Different Familiarization Voice (i.e. PF), 3) Test Voice PF, Same Familiarization Voice; and 4) Test Voice PF, Different Familiarization Voice (i.e. MJ). All were aged between 18 and 35 and were students or staff of the Universities of Cambridge or Glasgow. 72 participants were tested at the University of Cambridge in a sound-treated room, and 8 at the University of Glasgow in a quiet room.

3.2.3. Procedure

Participants were tested individually using high-quality Sennheiser headphones and a PC laptop running DMDX. Participants did the pre-test (48 items, 25 minutes), then the familiarization session (288

items, 40 minutes), and finally the post-test (48 items, 25 minutes). The pre- and post-tests were each preceded by one practice item and the familiarization session by two practice items.

In the pre- and post-tests, participants' task was to type what they heard into a computer. The words appeared on the screen as they typed, and they were instructed to type as many words as they had understood. A short break occurred after half the items.

In the familiarization session, participants heard each sentence in its disambiguating context. They were told that they would hear descriptions of events, and should judge whether it was LIKELY or UNLIKELY that an event involved the object, person, emotion or idea specified by a question displayed on the computer screen immediately after they had heard the sentence. Participants responded by pressing one of two labelled keys on a keyboard. Assignment of responses to hands was counterbalanced over participants. The comprehension questions appeared on the computer screen for 3 s, and each took the form *Does the event involve X?* A short break occurred after every 20 items, and a longer self-paced break half-way through the familiarization session. Participants were offered self-paced breaks between pre-test, familiarization and post-test.

3.3. Results

3.3.1. Analysis

All analyses were carried out using generalized linear mixed-effects modelling with logistic regression. Logistic regressions are performed on ratios of correct to incorrect responses, from which odds ratios can be calculated. However, for ease of understanding, the Figures and Tables, and the text below, are expressed in terms of percentages of correct responses.

Responses were measured in three different ways: Words, Boundary, and Boundary2, as illustrated in Table 4. Words (W) measures the percentage of words correct in the entire sentence. To be scored as correct, words had to be typed in the same order as in the actual spoken sentence. Obvious mis-

spellings and homophones were scored correct, morphological variants incorrect. The examples in column 3 of Table 4 show number of words correct for the two six-word sentences illustrated; for the presentation of the results, values were summed over all responses, and converted to a percentage of the total number of possible correct words.

Table 4 Examples of the three scoring systems used in the intelligibility experiment, applied to one pair of sentences: (I) EB *It may have been eye strain* and (II) LB *It may have been iced rain.* 'Words' scores indicate the number of words correct for that sentence. 'Boundary' and 'Boundary2' scores for a single sentence are binary: 1 for correct, 0 for incorrect. In both cases, a correct score (1) indicates that the correct phoneme sequence was provided with the word boundary correctly located within it. The correct phoneme sequence was the final syllabic constituent of the first word (Word1End), and the initial syllabic constituent of the second word (Word2Start). Under the Boundary criterion, all other responses were scored as incorrect (0). In contrast, all incorrect Boundary2 responses contained the correct phoneme sequence, but with the word boundary misplaced within that sequence; any other incorrect response was excluded from the Boundary2 analysis. The difference between Boundary and Boundary2 is thus that Boundary2 was restricted to correct phoneme strings in the vicinity of the word boundary and focussed entirely on the location of the boundary. See text for further explanation.

I. *Response to sentence "It may have been eye strain"*	Phoneme string	Words (# of 6)	Boundary Word1End	Boundary Word2Start	Boundary2 Word1End	Boundary2 Word2Start
It may have been eye strain	/aɪ str/	6	1	1	1	1
He may have been high strung	/aɪ str/	3	1	1	1	1
He may have seen mice trained	/aɪs tr/	2	0	0	0	0
He may have seen my drain	/aɪ dr/	2	1	0	1	excluded
He may have seen my Dane	/aɪ d/	2	1	0	1	excluded
He may have seen my rain	/aɪ r/	2	1	0	1	excluded
It may have been the test train	/ɛst tr/	4	0	0	excluded	0
It may have enticed strain	/aɪst str/	3	0	1	0	1
It may have been an A train	/ɛɪ tr/	4	0	0	excluded	excluded

II. *Response to sentence "It may have been iced rain"*	Phoneme string	Words (# of 6)	Boundary Word1End	Boundary Word2Start	Boundary2 Word1End	Boundary2 Word2Start
It may have been iced rain	/aɪst r/	6	1	1	1	1
He may have been high strung	/aɪ str/	3	0	0	0	0
He may have seen mice trained	/aɪs tr/	2	0	0	0	0
He may have seen my drain	/aɪ dr/	2	0	0	excluded	excluded
He may have said it might rain	/aɪt r/	3	0	1	excluded	1
He may have seen my rain	/aɪ r/	3	0	0	excluded	1
It may have been the test train	/ɛst tr/	4	1	0	1	0
It may have enticed strain	/aɪst str/	3	1	0	1	0

The Boundary and Boundary2 measures were introduced because the number of words correct in the sentence does not necessarily tell us much about listeners' use of phonetic detail to segment the critical words. Boundary and Boundary2 allow us to test not only whether familiarity with a voice allows subjects to identify more words, but also whether it helps them segment phoneme sequences which can be parsed in more than one way. Boundary (B) reflects the percentage of responses that contained all the correct phonemes, in the correct sequence, in the correct syllabic constituent abutting the word boundary: that is, the correct phoneme(s) in the nucleus or coda of the word before the boundary (/aɪ/ in the case of *eye strain*, /st/ in the case of *iced rain*), and the correct phoneme(s) in the onset or nucleus directly after the word boundary (/str/ in the case of *eye strain*, /r/ in the case of *iced rain*, and, not illustrated in Table 4, /aɪ/ in the case of *he'd iced*). However, phonemes before and after the critical word boundary were scored separately: there was no requirement for both to be correct in order to obtain a correct score for one of them, and therefore no requirement for the whole sequence to be correct across the word boundary. Thus B has two subparts: B Word1End was scored correct if a response contained all and only the correct phoneme(s) in the last syllabic constituent before the word boundary, and B Word2Start was scored correct if a response contained all and only the correct phoneme(s) in the first syllabic constituent after the word boundary. Columns 4 and 5 of Table 4 show the numerical score assigned to each response illustrated under the B criterion: either 1 (correct) or 0 (incorrect). Although, as with the W score, Table 4 shows the B scores that each particular response would be assigned, observed scores are expressed in the rest of this paper as a percentage of all possible responses.

The B measure is position-sensitive in that for a correct score to be achieved the listener must not only identify the correct phoneme(s) but assign them to the correct position in syllable. Nevertheless, improvement on this measure could conceivably be due to a general improvement in phoneme

identification, such as might be observed if exposure to a speaker's voice led to phoneme-category retuning. Therefore the third measure, Boundary2 (B2), was introduced as a more stringent test of whether or not position-sensitive allophonic quality is encoded in a way that could in fact be mediated by context-free phonemic categories. The B2 measure was identical to B in the way correct responses were defined, but differed in the way incorrect responses were defined. B2 excluded from the count of incorrect responses those cases where some or all of the correct phonemes were simply not reported at all, or, if reported, were in the wrong sequence. Thus, certain types of error led to a score of 0, while others that would score 0 under criterion B were excluded altogether from the B2 analysis, with the result that the ratios and percentages of correct scores were calculated relative to a smaller total number of observations for B2 than for B. Specifically, to be included as an incorrect B2 response, the response had to contain the correct phonemes in the correct order in the vicinity of the critical word boundary, but some or all of those phonemes would be in the wrong position relative to the word boundary. Further, a phoneme could be duplicated on each side of the boundary, as long as the right sequential order was maintained. Thus, the error was that not all the critical phonemes were in the correct syllabic constituent: one or more was in the onset/nucleus of the post-boundary word when they should have been in the nucleus/coda of the preceding word, or vice versa. Hence, for a B2 score of 0, all the phonemes were right, but one or more of the allophones were wrong.

Since the description of B and B2 criteria is complicated, we discuss them in some detail in this and the next paragraph. In this paragraph, we work through some of the examples in Table 4; the next paragraph compares the consequences of the two measures. The second rows of Table 4.I and 4.II give the correct answer to the heard stimulus (I: *it may have been eye strain*, II: *it may have been iced rain*) and so both answers score maximally correct under all three criteria, W, B and B2. The second answer,

He may have been high strung, has three of the six words correct for both sentences, so the W score is 3/6 in both cases. As a response to *eye strain*, the phonemes around the word boundary, /aɪ str/, are all present, correctly sequenced, and correctly aligned relative to the word boundary, so this response scores 1 for all four B and B2 criteria in Table 4.I. But as a response to *iced rain* (Table 4.II), /aɪ str/ scores 0 under all four B and B2 criteria because, though the right phonemes are in the right sequence, /st/ is on the wrong side of the word boundary—and hence would be spoken with the wrong allophones. Finally consider the answer in row 5 of each part of Table 4: *He may have seen my drain*. Two of the words are correct, so W = 2/6 in both cases. The first of the two words around the critical boundary, *my*, lacks a coda and has the correct nucleus /aɪ/ for *eye strain*, so Word1End scores 1 (correct) for both B and B2 in Table 4.I. But in Table 4.II, *my* for *iced* scores 0 (incorrect) for Word1End under criterion B, and it is excluded from the B2 Word1End analysis because it is the coda /st/ that must be right in this case, rather than the nucleus /aɪ/. Likewise, *drain* scores 0 for B Word2Start as a response to both *eye strain* and *iced rain* (Tables 4.I and 4.II respectively); although *drain* does contain /r/, it lacks /s/, and (perhaps as a consequence) the stop has been written as the wrong phoneme too: /d/ instead of /t/, a point we return to below. Under criterion B2, *drain* is of course excluded from both Word2Start analyses (Table 4.I and 4.II) because it does not contain all the right phonemes.

B and B2 test subtly different, but nevertheless closely related predictions. Because both are position-sensitive, any improvement observed on them after exposure to a voice implies that listeners are sensitive to the particular way that position-in-word, and hence allophonic quality, is encoded by the speaker that they are adapting to. The difference is that if the B measure shows improvement with exposure, then it could reflect phoneme retuning; if the B2 measure shows improvement with exposure, then it must reflect context-sensitive allophonic learning, without necessary recourse to context-free

phonemes. Specifically, improvement on the B measure could conceivably be due to a general improvement in phoneme identification. In contrast, the difference between correct and incorrect B2 measures reflects solely the placement of the word boundary within a correct phonemic sequence, because both phoneme identification and phoneme sequence are constant in B2. So any improvement observed on the B2 measure would be more appropriately modelled in terms of learning about categories that directly reflect syllable and word structure—in short, improvement on B2 must reflect learning about allophonic identity, but not phonemic identity.

In summary, the W measure is a standard word-based measure of speech intelligibility in noise; the B measure allows allophonic accuracy around the critical word boundary to be assessed; the B2 measures is a more stringent test of allophonic accuracy which is independent of phoneme identification in that it was conducted only on responses in which (context-free) phoneme identification was correct, but (context-sensitive) allophonic identification might or might not be.

As noted, the B2 analysis is reduced in statistical power compared with the B analysis, because a subset of incorrect responses is excluded from the analysis. In fact, the B2 dataset was only 75% the size of the B dataset. Comparisons between the two analyses must therefore be made cautiously (see below).

Generalized linear mixed-effects modelling was applied to the results of each measurement criterion (W, B, B2). First, a full model was fitted, with predictors Test (Pre-test vs. Post-test), Voice heard in tests (MJ vs. PF), Familiarization Voice (Same vs. Different as tests), Boundary Position (Early vs. Late) and all their interactions. Following the same procedure as in the production experiment, non-significant predictors were incrementally removed until the simplest model had been found; Appendix D shows the fixed effects included in this final model. Also in the full model were control variable Sentence Group, and random effects for Subject and Sentence. The inclusion of both sets of random effects ensures

that significant results, where obtained, are robust across participants and items. In all modelling, we found greater variances associated with items than participants. The crucial prediction was of an interaction between Test and Familiarization Voice: more improvement from pre-test to post-test was expected when the voice heard in the familiarization period was the Same as that heard in the tests.

Tables detailing the statistical results are again presented in Appendices; differences reported in the main text as statistically significant always had $p < 0.05$ or better, and in most cases $p < 0.001$.

3.3.2. Main results

Figure 4 and Table 5 show the percentages of correctly-reported Words and B and B2 syllable constituents at Word1End and Word2Start. Relevant statistical results are in Appendix D. Results were broadly similar for all measures. The most important result is that, as predicted, the improvement from pre-test to post-test was slightly but significantly greater when the Same Voice was heard in both familiarization and tests, than when a Different Voice was heard, as reflected in interactions between Test and Familiarization Voice.

The increase in improvement for the Same Voice relative to the Different Voice condition was very similar regardless of whether the B or B2 measure was considered (B Word1End: 5.1%, B2 Word1End: 4.6%; B Word2Start: 4.5%, B2 Word2Start: 4%). Because of the reduction in power for the B2 analyses, when comparing statistical results for B and B2, it is more appropriate to consider the estimate sizes for the crucial interaction between Test and Familiarization Voice, rather than their *p*-values. These estimates can be expressed as odds ratios (Rice 1995), where an odds ratio of 1 corresponds to no difference between conditions, and larger odds ratios reflect larger differences. The estimated odds ratios from the logistic regression models for the interaction of pre/post Test with Familiarization Voice

were: for Word1End, 1.26 and 1.21 for B and B2 respectively; and for Word2Start, 1.24, and 1.20 for B and B2 respectively. Appendix D1 shows that neither of the B2 values reaches statistical significance (p = 0.096 for B2 Word1End and p = 0.1 for B2 Word2Start) whereas the equivalent p values are 0.017 and 0.036 for the B measure, but this difference reflects the reduction in statistical power from the smaller B2 sample. What is important is that the odds ratios for the same-voice advantage using the B2 measure are slightly smaller, but essentially comparable in magnitude to those in the B measure.

Figure 4 Percentage of correctly reported Words in whole sentence (a); B syllable constituents (b, c); B2 syllable constituents (d, e), at pre- and post-test. Error bars represent one standard error.

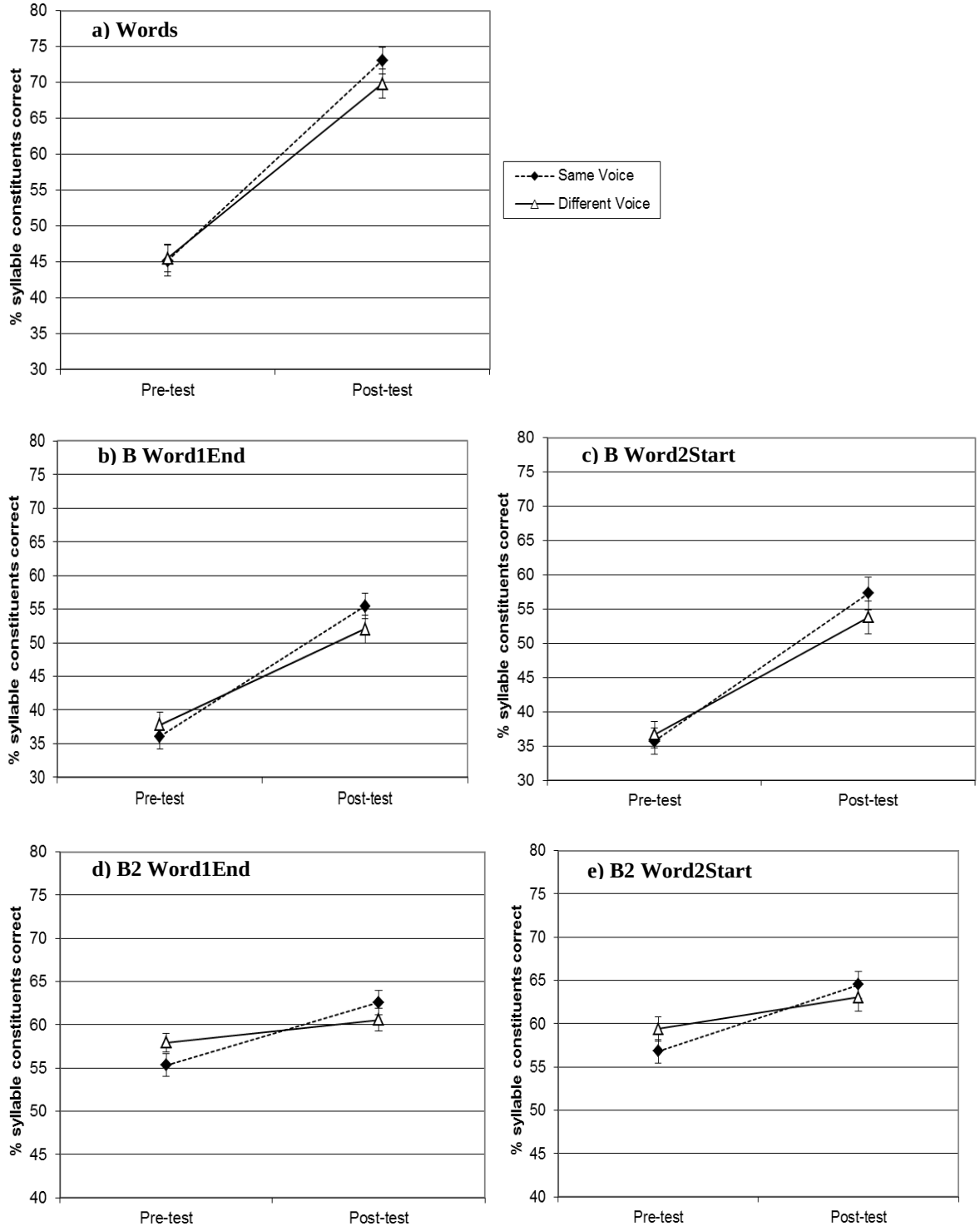


Table 5 Percentage of correctly reported Words and B and B2 syllable constituents, in pre-test and post-test according to Early or Late Boundary Position (top half of table, pooling across Voice) and Voice tested (bottom half of table, pooling across Boundary Position).

		Mean % correct responses			*Percentage change (improvement) from pre-test to post-test*
		Overall mean	Pre-test	Post-test	
Words	Early Boundary	59	46	72	*+26*
	Late Boundary	58	45	71	*+26*
B Word1End	Early Boundary	47	37	56	*+19*
	Late Boundary	45	37	52	*+15*
B2 Word1End	Early Boundary	55	49	60	*+11*
	Late Boundary	65	67	64	*-3*
B Word2Start	Early Boundary	54	45	63	*+18*
	Late Boundary	38	27	48	*+21*
B2 Word2Start	Early Boundary	68	66	70	*+4*
	Late Boundary	54	49	58	*+9*
Words	PF	69	56	81	*+25*
	MJ	48	34	62	*+28*
B Word1End	PF	54	45	62	*+17*
	MJ	37	29	45	*+16*
B2 Word1End	PF	56	59	65	*+6*
	MJ	62	53	58	*+5*
B Word2Start	PF	56	46	66	*+20*
	MJ	36	26	45	*+19*
B2 Word2Start	PF	65	61	68	*+7*
	MJ	56	53	58	*+5*

Taken together, therefore, the B and B2 measures suggest that the Same-Voice advantage for identifying syllable constituents at word boundaries is attributable only in small part to an overall improvement in phoneme identification. In consequence, as discussed further below, retuning of context-free phonemic categories is unlikely to be the best way to model the observed sensitivity to syllable structure and position in word.

The other significant effects do not weaken this main finding, as Table 5 and Appendix D show. As results for B and B2 measures were mostly similar, the following text does not distinguish them, except where specifically indicated. First, subjects in all conditions improved considerably from pre- to post-test, presumably because familiarization provided useful experience of the test sentences in meaningful contexts (Test, Appendix D; cf. Davis et al., 2005). Improvement was greater on B than B2 measures, suggesting that from pre- to post-test listeners got better partly because they became able to identify sounds that they had missed altogether, or whose phonemic identity they had misidentified, at the pre-test. Second, responses to PF were more accurate than to MJ overall (lower half of Table 5), although improvement was about the same for each speaker, within each measure. Third, responses to Early Boundary items differed from responses to Late Boundary items. Responses to Early Boundary items were very slightly but significantly more accurate than to Late Boundary items (by 1-2%) for Words and B Word1End, but the opposite pattern was found for B2 Word1End, suggesting that if listeners were able to identify the correct word-final phonemes for Late Boundary items at all, they were also quite accurate at identifying their position in syllable. In contrast, for Word2Start, responses to Early Boundary items were considerably more accurate than to Late Boundary items (B: by 16%; B2: by 14%) (upper half of

Table 5). That is, Late Boundary word beginnings, which generally corresponded to onsetless syllables (e.g. *iced*), were especially poorly identified.

Finally, there were some differences among the W and B measures, which took the form of significant interactions involving Boundary Position (Appendix D). These patterns all reflect that the prosodically weaker consonants in Late Boundary items had less perceptual salience than Early Boundary items, and were thus more difficult to identify, especially in the less intelligible voice (MJ's). Thus, for Word1End and Word2Start, the accuracy advantage for PF over MJ was greater for Late than Early Boundary sentences (Appendix D). Because they were more difficult initially, Late Boundary sentences generally had the scope to improve more as participants learned about how they were spoken. Thus, for Word2Start, Late Boundary sentences improved significantly more than Early Boundary sentences from pre- to post-test (Table 5; Test x Boundary Position interaction, Appendix D). For Words, greater improvement was again found for Late than Early Boundary sentences for PF's voice, but the opposite pattern was found for MJ's Late Boundary items (Test x Voice x Boundary Position interaction, Appendix D). The difficulty of these items in MJ's speech may have been so great that learning about them was slower.

Taken together, these relationships all support the idea that the prosodically weaker consonants in Late Boundary sequences make them more difficult to identify than Early Boundary sequences. Importantly, however, the Same-Voice advantage was constant across all measures and was not affected by these differences involving Early vs. Late Boundaries.

3.4. Discussion

The perception experiment showed that 40 minutes' exposure to sentences in a speaker's voice led to more improvement in understanding novel tokens of those sentences in background noise than did exposure to the identical sentences in a different voice. This Same-Voice advantage was significant not only for identifying words, but also for identifying allophones—that is, phonemic sequences in their correct syllabic and sequential context. Specifically and crucially, the Same-Voice advantage was significant for segmenting difficult phonemic sequences, as reflected in listeners' ability to assign segments to the correct structural position as a syllable coda or a syllable onset, and hence to the correct critical word (e.g. to correctly identify the /d/ in *we'd rank* as a word-final /d/; Figure 4). The results for word identification parallel those of Nygaard & Pisoni (1998), though that study used multiple speakers, and a longer familiarization phase; the results for difficult word segmentation are novel. The results provide evidence of learning of speaker-specific phonetic detail at word boundaries: evidently, an advantage of a familiar voice during word segmentation emerges when, as here, the task is difficult, yet the stimuli are sufficiently similar to a previous experience to make prior knowledge systematically useful. This advantage presumably contributes to the intelligibility benefit that familiarity with a voice provides to listeners in segregating a single talker from background noise (Newman & Evers, 2007).

Poorer performance was found for Late Boundary than Early Boundary sentences. In other words, coda consonants and onsetless syllables were identified less accurately than onset consonants (cf. Pickett, Bunnell & Revoile, 1995; Ohala, 1996). When a consonant has ambiguous syllable and word affiliation, English listeners are more likely to parse it as belonging to a syllable or word onset. The words in the Early and Late Boundary phrases had roughly equal frequencies and phonological probabilities (two-tailed sign tests: frequency using Kučera and Francis (1968), $p = 0.15$; phonological probability using Coleman's (2000) prosodic probabilistic grammar, $p = 0.54$). However, syllables with onsets have a

higher type-frequency overall in English than syllables without (O'Connor & Tooley, 1964), so the better performance for Early Boundary sentences may reflect listeners' implicit knowledge of this pattern. It seems reasonable to speculate, however, that much of the difference reflects the fact that, in Fougeron and Keating's (1997) sense, consonants were normally 'stronger' in onset than in coda position (Figure 1 and Table 3).

Poorer absolute performance was found for speaker MJ than PF. It is possible that PF was understood better than MJ because he marks his word boundaries particularly clearly, but other factors may also be at play: influences on intelligibility are complex and poorly understood (Markham & Hazan, 2004). Possible partial explanations are that PF's lower pitch made his voice easier to segregate from the cafeteria noise, and that PF is a professional teacher, while MJ is less used to public speaking. Regardless of the intelligibility difference between the speakers, the advantage due to familiarization with a specific voice was of equivalent magnitude for both.

What are the implications of these results for the kind of learning that took place? They must reflect some degree of abstraction over experience, because test and familiarization never used identical *tokens* of sentences. While previous studies have mostly assumed that listeners abstract to phonemic categories, the present results suggest a more complex picture. Although phonemic retuning may be part of the story, the data do not support the view that the listeners were learning *only* about categories as abstract as phonemes. Learning exclusively about phonemes would require assuming that listeners first generalise from specific contexts to context-free phonemic categories, and then throw away structure-specific information. Not only does this seem unlikely given this particular task, but it would not help in the assignment of sounds to the correct position in syllables and words. The results showed that exposure to a specific speaker's voice improved listeners' ability to identify phonemes in specific positions in

syllables and words, and the natural conclusion is that the learning that took place was sensitive to position in syllable.

4. General Discussion

This study investigated production and perception of speaker-specific variation in patterns of phonetic detail at word boundaries. The production experiment found quantitative and qualitative variation in six speakers' productions of junctural minimal pairs. Speakers varied in the extent and way in which they used acoustic durations, /d/ realisation and vowel formant frequencies to mark various types of word boundary. The perception experiment used two speakers who differed in a range of these dimensions, and investigated whether familiarization with a single voice led to learning about their speaker-specific properties. Familiarization with the specific voice improved participants' ability to identify words and syllable constituents at word boundaries in hard-to-segment, phonemically-identical sequences presented in background cafeteria noise.

In production, we found that speakers varied in multiple aspects of their phonetic realisation of word junctures. This information has of course been known since measurements of speech first began, but while traditionally group trends were sought, we report differences and exploit them in a perceptual learning task. However, the differences and their implications are of interest in their own right. While initial position in syllable and word was phonetically stronger than final position for all speakers, the variation beyond this basic regularity was complex, and further work is required to elucidate factors governing it. Interestingly, rate of speech, degree of lenition of /d/ consonants, and degree of durational differentiation between positions in syllable, patterned partly together, but partly separately. For example, speaker JR spoke fast, had relatively little durational differentiation between positions in syllable, and

generally lenited his /d/s. Speakers PF and MJ both spoke more slowly than JR, and at similar rates to each other, but PF showed the greatest durational differentiation between positions in syllable, yet also lenited his onset and coda /d/s much more than MJ and comparably to JR. Thus, we doubt that a single parameter such as rate or 'style' governs all aspects of the variation. Further, while we have framed our investigation in terms of idiosyncratic (indexical) differences, some aspects may be micro-dialectal or socially stratified. For example, the variation we observed in the vowel in *he'd, she'd,* etc, seems plausibly to be an age-graded change in the small set of contracted auxiliary verbs, related to their grammatical and/or metrical properties, while some aspects of the /d/ realisation seem to pattern with gender.

Although production models lacking an exemplar component can explain individual differences in production by invoking individual differences in phonetic implementation of linguistic units, we consider that our findings are most naturally accounted for in a model with an exemplar component, specifically a hybrid exemplar-abstract model such as those proposed by Pierrehumbert (2002, 2006) or Walsh, Möbius, Wade and Schütze (2010). For example, Pierrehumbert (2002) presents a model that involves storage of exemplar chunks of speech, combined with prosodic parsing/tagging of these chunks, and mapping of them to labelled phonemic and lexical categories. Production goals are selected by sampling regions from within the exemplar space corresponding to the selected label(s). Lexical knowledge acts as an attentional bias on the selection of particular exemplars for production; this approach allows for fine-grained allophonic differences to develop between phonemically similar structures (e.g. the words *r<u>ea</u>lign* and *r<u>ea</u>lise* exhibit different patterns of glottalisation at the hiatus, due to the morphological relationship between *realign* and *align*). Walsh et al. (2010) likewise present a

hybrid model in which production targets are generated from exemplar distributions labelled at multiple levels corresponding to units of different sizes (e.g. phonemes and words). In general, a larger unit (a word), will influence production more than its constituents (phonemes), as long as the word is frequent enough to have acquired a critical mass of associated exemplars; if not, production will be influenced mainly by the constituent phonemes. Thus, in hybrid exemplar models, production targets *can* be generated on the basis of phonemic categories alone, but will normally be influenced also by larger units, especially for experienced speakers. This idea accords fairly well with the principles of adaptive resonance theory (Grossberg, 2003), of task dynamics (Saltzman, 1995), where control of behaviour becomes more global as skill develops (for an embodied formulation, see Simko and Cummins, 2010), and indeed of perception-action robotics models (e.g., Sprague, Ballard and Robinson, 2007; Roy, 2005).

Our present findings support hybrid approaches in that they indicate that production is influenced by multiple levels of structure (e.g. syllable and word identity, grammatical and prosodic structure). Further, they suggest that the influences of these different levels vary according to individual speaker. Some speakers (e.g. PF) showed stronger influences from syllable structure and prosodic shape, but did not show, for example, the influence of grammar noted above for MJ and RS. We would expect that attentional and situational biases could alter the weightings of the various influences, as suggested by the intriguing data of Goldinger & Azuma (2003), who found variation in word production according to whether the experimenter's introduction generated an implicit bias to attend to syllables or segments.

For perception, our results provide direct support for an exemplar component in modelling. The perception experiment demonstrated a voice familiarization advantage that was small but consistent

across speakers and measures. Given that people do have experience throughout life with a variety of speakers, any processing benefit due to experience with a particular speaker, no matter how small that benefit, indicates that even the adult speech perceptual system is remarkably flexible. Furthermore, the effect we observed is potentially important because the stimuli were more natural than those used in many experiments on phonetic aspects of perceptual learning. For example, Allen & Miller (2004) tested inter-speaker differences in VOT that were exaggerated to roughly twice the range found in Allen, Miller and DeSteno's (2003) production study. Their study, and those of Norris, McQueen & Cutler (2003) and Eisner & McQueen (2005) also used isolated words, whereas the present stimuli comprised meaningful connected speech read very fluently in a casual style, and the emphasis in the perceptual tasks was on accessing sentence meaning. Moreover, unlike other studies which use fluent read speech (Nygaard, Sommers, & Pisoni, 1994; Nygaard & Pisoni 1998), the stimuli were also controlled enough to permit conclusions about the phonetic nature of perceptual learning: in this case, that learning about the particular pronunciations associated with individual speakers' voices helps correct identification of allophones at word boundaries, and hence facilitates access to meaning via lexical identification.

The B and B2 analyses allow us to explore to what extent the data can be reconciled with accounts based on the retuning of sub-lexical, specifically phonemic, category representations (Norris, Cutler, & McQueen 2003; Eisner & McQueen 2005). The data are compatible with this category retuning account, as long as it is modified so that perceptual learning is not based *solely* on adjustments to phonemic representations, but can also involve aspects of syllabic and prosodic structure (Hawkins & Smith 2001; Hawkins 2003; Hawkins 2010; Pierrehumbert 2002, 2006). By this account, listeners would not only store phonemic categories and learn about speaker-specific realisation of these (Norris et al. 2003), but would additionally store information about prosodic structures and processes, and learn about

speaker-specific phonetic implementation of these. For example, listeners undoubtedly know about strengthening of syllable-onset consonants and lenition of syllable-coda consonants; they might learn that such strengthening or lenition tends to be particularly pronounced for a particular speaker. Similarly, listeners will know that pitch-accenting a syllable produces various phonetic changes in that syllable, and might learn that certain of these changes tend to be particularly large or small for a given speaker.

Thus, while perceptual learning and category retuning may sometimes focus on phonemic categories, many other sound categories are also relevant. Like many others (e.g. Saffran, Newport and Aslin, 1996) we have argued that phonemic categories may be learned as "self-organizing" and "emergent", related by virtue of shared attributes or occurrence: the brain naturally works with contrasts, and classifies like things together (Hawkins, 1995; Hawkins & Smith, 2001). The relationship between sounds and phonemic categories may be learned very gradually as a speaker's language system matures, by association rather than because phonemes play a fundamental role in speech communication. Hawkins (2010:500-501) argues that it follows that phonemic representations might not be comprehensive or fully systematic. Assuming that the shared attributes of phonemes can be learned via any relevant experience, then those that share consistent articulatory, auditory or orthographic properties (for people literate in an alphabetic writing system), seem particularly likely to form categories. In English, candidates for systematic phonemic categories might be /s/, which varies little acoustically across contexts (although it is indexically variable, e.g. Stuart-Smith, 2007), and bilabial (but not alveolar) stops, because bilabial stops vary little in critical aspects of articulation. Other sounds, for example vowel qualities conditioned by certain consonants, e.g. those in *whole* and *hope* in several varieties of British English, might never be grouped appropriately into phonemes. Orthographic consistency presumably influences the system's development and functioning (cf. Ranbom & Connine, 2011). If the language allows it, then a rough

phoneme inventory could develop, with some phonemes being more clearly defined as categories and accessible to consciousness than others, and with individuals, and speakers of different languages, differing in how systematic their phonemic inventories are. The implication for perceptual learning is that learning may be about phonemic categories when those categories are particularly systematic, but may be about allophonic, featural, syllabic, prosodic or grammatical categories when these are systematic. These arguments are compatible with those made by Nielsen (2011), Dahan & Mead (2010) and Kraljic & Samuel (2006), who all found generalization of perceptual learning to non-phonemic categories, namely features or positions in syllable, and who emphasise the flexibility of perceptual learning as the key to understanding the range of units that may be involved.

What emerges clearly from the present work and other studies on indexical variability in production and perception is that when a listener encounters the mixture of personal and linguistic information in an individual's voice, understanding the message can be facilitated by learning about how the individual speaks, in order that phonetic detail can be effectively mapped to linguistic and personal perceptual dimensions. The results that we have observed demonstrate that speaker-specific variation in phonetic detail that indicates word boundary location exists, and is learned about. Our results suggest that by widening the range of phonetic properties investigated in perceptual learning studies, we will be able to discover more about the aspects of speaker variation that help listeners to decode messages, and hence more about the types of processes involved.

Acknowledgements

Parts of this work were done in partial fulfilment of the requirements for the PhD degree awarded to the first author by the University of Cambridge. Parts of it were presented in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, August 2007. We thank Harald Baayen, Mirjam Ernestus and Sam Miller for statistical help, Naomi Hilton for assistance testing subjects, and Pam Beddor, Gerry Docherty, Gareth Gaskell, Jane Stuart-Smith and several anonymous reviewers for helpful comments on previous drafts. Financial support was provided by the Arts and Humanities Research Board and Newton Trust at the University of Cambridge, and the John Robertson Bequest at the University of Glasgow.

Appendix A: experimental sentences and precursors

Critical portions of experimental sentences are underlined.

Group D

D1. EB: We lined up all the wines from the cheapest to the most expensive. Then we drank them in order.
LB: We decided that we'd mark all the tests first. Then we'd rank them in order.
D2. EB: I bought a new car stereo in time for the trip. I drove all over Europe playing music.
LB: I thought I would become a travelling bard. I'd rove all over Europe playing music.
D3. EB: He wanted the carrots to cook fast. So he diced them.
LB: The top of the cakes had come out looking uneven. So he'd iced them.
D4. EB: Her teenage son insisted all his T-shirts had to be black. She dyed them resentfully.
LB: She hadn't been able to afford the jeans in the shop window. She'd eyed them resentfully.
D5. EB: All writers are terrified when their book first comes out. We dread the reviews.
LB: We expected the play to be unusual. We'd read the reviews.
D6. EB: John thought they needed a challenge. He was pleased that we dared them.
LB: All Mark's clothes had been damp. He was pleased that we'd aired them.

Group S

S1. EB: There are some people out there with a morbid streak. Apparently the gentlemen collect skulls.
LB: We have a dedicated bird collector living next door. Apparently the gentleman collects gulls.
S2. EB: Most people would have borrowed money to pay for it. Pete stole money.
LB: She eventually realised what the cheque on the table was. Pete's dole money.
S3. EB: Chocolate and sugar are comforting when you're in danger. That's why the airmen eat sweet products.
LB: Carbohydrates are good for endurance. That's why the airman eats wheat products.
S4. EB: "The Lord is my Shepherd" is popular at St. Mary's. The congregation certainly like psalms.
LB: The pastor and his flock are gun-toting NRA members. The congregation certainly likes arms.
S5. EB: For a while the fallen trees blocked access. But Pat sawed them.
LB: To begin with they were unimpressed. But Pat's awed them.
S6. EB: Among the dog baskets, Sarah was surprised to see some much smaller ones. She said, "Those are cat size".
LB: I wanted to know what the little lights in the road were. She said, "Those are cat's eyes".

Group A

A1. EB: "The film was sentimental, but one scene was genuinely moving," she said. "That surprise for the child."
LB: "See the mountain bike over there?" he said. "That's a prize for the child."
A2. EB: I agree that Simon excels. But Ralph surpasses.
LB: Geoff's grades are all distinctions. But Ralph's are passes.
A3. EB: Don't say venereal. Say veneer for me.
LB: Don't eat the whole of the chocolate rabbit. Save an ear for me.
A4. EB: John gave up pretty early. Ruth sustained.
LB: I'm not sure whose trousers to borrow. Ruth's are stained.
A5. EB: He hasn't worked with the Gurkhas before. It's no wonder he didn't recognise that salute.
LB: He doesn't know anything about music. It's no wonder he didn't recognise that's a lute.
A6. EB: There was still one thing she didn't like about the fireplace. That surround.
LB: That song's not a fugue. That's a round.

Group T

T1. EB: They'd told him not to leave the house. But he went for a sly stroll.
LB: There were lots of fancy sandwiches on the table. But he went for a sliced roll.

T2. EB: After each show I arrange the props backstage, ready for the next day. And I lay Steve's costume out in the wings.
LB: Just before the curtain went up, I did the actors' make-up in the green room. And I laced Eve's costume out in the wings.

T3. EB: His parents take the idea of corporal punishment too far. They whack Stan's legs quite painfully.
LB: I don't recommend the King's Road salon. They waxed Ann's legs quite painfully.

T4. EB: I don't know what caused my migraines. It may have been eye strain.
LB: You must have misread the sentence: hail isn't acid rain. It may have been iced rain.

T5. EB: It's not just that his foster parents are well off. They also offer Mick stability.
LB: The yoga centre has a few beginners' classes this year. They also offer mixed ability.

T6. EB: They'll have a game of football with anyone that's willing. They even play strangers in the park.
LB: The local authority took all sorts of measures to combat crime. They even placed rangers in the park.

Appendix B: Statistical results for production experiment

Table B1 Mean durations, and results of mixed-effects modelling for relative durations in the four sentence groups, and for vowel formants in Group D. Statistically significant effects ($p \leq 0.05$) are shaded grey. ndf = numerator degrees of freedom; ddf = denominator degrees of freedom.

				Boundary position		*Speaker*		*Boundary position x Speaker*	
	means Early, Late Boundary		*ddf*	*F* (ndf 1)	*p*	*F* (ndf 5)	*p*	*F* (ndf 5)	*p*
	ms	% of critical phrase							
Group D (e.g. *he diced* vs. *he'd iced*)									
PrecSyl	96, 101	21.9, 23.3	554	24.3	<0.0001	20.2	<0.0001	2.3	0.04
/d/	92, 63	21.4, 15.0	547	751.2	<0.0001	34.2	<0.0001	3.2	0.007
FolSyl	246, 264	56.5, 61.5	548	371.1	<0.0001	66.6	<0.0001	6.4	<0.0001
Group S (e.g. *cat size* vs. *cats' eyes*)									
PrecSyl	248, 241	34.1, 34.1	542	0.1	0.82	18.1	<0.0001	2.4	0.03
/s/	136, 103	19.9, 15.5	548	521.0	<0.0001	9.1	<0.0001	7.2	<0.0001
FolSyl	340, 362	45.8, 50.3	569	198.9	<0.0001	21.2	<0.0001	n/a	n/a
Group A (e.g. *that surprise* vs. *that's a prize*)									
PrecSyl	214, 201	27.5, 26.9	552	1.9	0.17	30.2	<0.0001	4.3	0.0007
/s/ or /v/	113, 92	14.6, 12.1	553	238.8	<0.0001	3.4	0.01	6.4	<0.0001
/ə/	39, 45	5.1, 5.9	543	44.3	<0.0001	8.8	<0.0001	n/a	n/a
Consonant	86, 90	10.9, 11.5	556	21.6	<0.0001	18.3	<0.0001	5.1	0.0001
FolSyl	325, 329	41.7, 43.4	553	33.0	<0.0001	57.6	<0.0001	6.6	<0.0001
Group T (e.g. *lay Steve's costume* vs. *laced Eve's costume*)									
PrecSyl	226, 212	37.6, 36.8	569	18.3	<0.0001	20.6	<0.0001	n/a	n/a
/s/	104, 83	17.2, 14.1	552	410.6	<0.0001	96.1	<0.0001	8.9	<0.0001
/t/ closure	40, 35	6.7, 6.2	564	11.8	<0.0006	9.4	<0.0001	5.2	0.0001
/t/ VOT	31, 35	5.2, 5.9	551	4.0	0.05	23.2	<0.0001	6.8	<0.0001
FolSyl	216, 237	32.9, 36.7	551	500.1	<0:0001	37.3	<0.0001	2.7	0.02
Group D /i:/	Early, Late Boundary means (Bark)								
F2-F1	11.0, 10.0		461	665.8	<0.0001	504.7	<0.0001	63.8	<0.0001

Table B2 Pairwise comparisons for variables where a significant interaction between Boundary Position x Speaker was found. Only comparisons significant with the Bonferroni—Holm procedure are reported (see text for details).

	EB > LB	*LB > EB*	*Differences among speakers in size of EB-LB difference*
Group D (e.g. *he diced* vs. *he'd iced*)			
Preceding syllable	--	AK (p = 0.0001) PF (p = 0.0091) SC (p = 0.001)	AK > MJ (p = 0.0034)
/d/	all speakers (p < 0.0001)	--	PF > JR (p = 0.0014)
Following syllable	--	all speakers (p < 0.0001)	MJ > {AK, JR, SC} (all p < 0.0002)
/i:/ F2-F1	all speakers (p < 0.0001)	--	MJ > {AK, JR, PF, SC} (all p < 0.0001) RS > {AK, JR, PF, SC} (all p < 0.0001)
Group S (e.g. *cat size* vs. *cats' eyes*)			
Preceding syllable	PF (p = 0.0217)	MJ (p = 0.0126)	PF ≠ MJ (p = 0.0007)
/s/	all speakers (p < 0.0001)	--	MJ > {AK, PF, SC} (all p < 0.0025) JR > PF (p = 0.0021)
Group A (e.g. *that surprise* vs. *that's a prize*)			
Preceding syllable	PF (p = 0.0007)	RS (p = 0.0314)	PF ≠ JR (p = 0.001) PF ≠ RS (p = 0.0001)
/s/ or /v/	all speakers (p = 0.006)	--	{JR, MJ, PF, RS} > AK (all p < 0.005) {MJ, RS} > SC (both p < 0.002)
Consonant	--	PF (p < 0.0001)	PF > {AK, JR, MJ, SC} (all p < 0.002)
Following syllable	SC (p = 0.0173)	AK (p = 0.0075) JR (p = 0.0127) MJ (p < 0.0001) PF (p = 0.0012) RS (p = 0.0361)	SC ≠ {AK, JR, MJ, PF, RS} (all p < 0.002)
Group T (e.g. *lay Steve's costume* vs. *laced Eve's costume*)			
/s/	all speakers (p < 0.0001)	--	SC > {AK, JR, MJ, PF, RS} (all p < 0.0017)
/t/ closure	PF (p < 0.0001)	--	PF > {AK, JR, MJ, RS, SC} (all p < 0.0028)
/t/ VOT	--	SC (p < 0.0001)	SC > {AK, JR, MJ, PF, RS} (all p < 0.0036)
Following syllable	--	all speakers (p < 0.0001)	PF > MJ (p = 0.0007)

Appendix C: Similarities and differences between the speakers used in the perception experiment

Table C1 Summary of the differences in phonetic detail at word boundaries between the two speakers whose voices were heard in the perception experiment, PF and MJ.

Category	Grp	Specific variable	PF	MJ
PF makes a statistically significant distinction that is not significant for MJ				
Durations	D	/hiː/ longer in *he'd* than *he* by	6 ms (6%)	2 ms (2%), n.s.
	A	/ðat/ longer in *that surprise* than *that's a prize*) by	41 ms (17%)	8 ms (4%), n.s.
	A	/p/ longer in *prize* than *surprise* by	17 ms (21%)	3 ms (4%), n.s.
	T	/t/ closure longer in *lay Steve's* than *laced Eve's* by	16 ms (57%)	0 ms (0%), n.s.
Both speakers make a statistically significant distinction in the same direction, but magnitude is greater for PF				
Durations	T	/iːvz/ longer in *Eve's* than *Steve's* by	27 ms (11%)	18 ms (9%)
Consonant realisation	D	Continuous formant structure present during /d/	19% overall	6% overall
Consonant realisation	D	Continuous formant structure present more often in *he'd* than *he* by	20 percentage points	4 percentage points
Both speakers make a statistically significant distinction in the same direction, but magnitude is greater for MJ				
Durations	S	/s/ longer in *cat size* than *cat's eyes*	26 ms (23%)	53 ms (56%)
Spectral	D	/iː/ F2-F1 difference greater in *he* than *he'd* by	0.4 Bark	1.8 Bark
Consonant realisation	D	Continuous voicing present during /d/	80% overall	28% overall
Consonant realisation	D	Continuous voicing present more often in *he'd* than *he* by	28 percentage points	40 percentage points
Both speakers make a statistically significant distinction, but in opposite directions				
Durations	S	/kat/ in *cat* vs *cat's*	longer by 10 ms (4%)	shorter by 3 ms (1%)
Speakers do not differ significantly				
Rate	All	Mean duration of critical phrase:	679 ms	606 ms
Durations	D	/d/ longer in *he diced* than *he'd iced* by	38 ms (60%)	37 ms (57%)
	D	/aɪst/ longer in *he'd iced* than *he diced* by	20 ms (8%)	30 ms (13%)
	A	/s/ or /v/ longer in *that surprise* than *that's a prize*	23 ms (22%)	27 ms (33%)
	A	/raɪz/ longer in *that's a prize* than *that surprise*	22 ms (7%)	22 ms (7%)
	T	/s/ longer in *lay Steve's* than *laced Eve's*	14 ms (17%)	17 ms (22%)
Consonant realisation	D	Continuous frication present during /d/	15% overall	19% overall
	D	Continuous frication present more often in *he'd iced* than *he diced*	25 percentage points	29 percentage points

Appendix D: Statistical results for perception experiment

Table D1 Statistical results for the perception experiment. Only significant predictors and non-significant main effects that participated in significant interactions are reported. Terms not in the best fitting model are labelled "—". An asterisk (*) in the df column indicates that χ^2, df and significance levels are for the named factor plus any interaction terms in the table that included that factor. Marginally significant effects ($0.05 < p \leq 0.1$) are underlined. Nonsignificant effects are labelled n.s..

Effect	**Words**		
	χ^2	**df**	***p***
Test (pre/post)	3434.4	5*	<0.0001
Voice (MJ/PF)	110.7	4*	<0.0001
Familiarization Voice(Same/ Different)	20.6	2*	<0.0001
BoundaryPosition (EB/LB)	88.6	4*	<0.0001
SentenceGroup (D/S/A/T)	—	—	—
Test x Voice	12.1	2*	0.002
Test x Familiariz-ation Voice	19.5	1	<0.0001
Test x Boundary Position	13.9	2*	0.001
Voice x Boundary Position	10.7	2*	0.005
Test x Voice x Boundary Position	10.7	1	0.001

Effect	**B Word1End**			**B Word2Start**			**B2 Word1End**			**B2 Word2Start**		
	χ^2	**df**	***p***	χ^2	**df**	***p***	χ^2	**df**	***p***	χ^2	**df**	***p***
Test	245.3	2*	<0.0001	351.3	3*	<0.0001	51.7	3 *	<0.0001	38.0	3 *	<0.0001
Voice	75.0	2*	<0.0001	95.9	2*	<0.0001	28.5	3 *	<0.0001	54.9	2 *	<0.0001
Familiarization Voice	6.0	2*	0.05	5.1	2*	n.s.	2.8	2 *	n.s.	2.6	2 *	n.s.
Boundary Position	11.5	2*	0.003	251.1	3*	<0.0001	99.6	4 *	<0.0001	134.9	3 *	<0.0001
Sentence Group	9.1	3	0.028	8.3	3	0.039	8.6	3	0.036	5.5	3	0.1399
Test x Voice	—	—	—	—	—	—	—	—	—	—	—	—
Test x Familiariz-ation Voice	5.7	1	0.017	4.4	1	0.036	2.8	1	0.096	2.6	1	0.109
Test x Boundary Position	—	—	—	7.0	1	0.008	32.2	1	<0.0001	3.6	1	0.056
Voice x Boundary Position	7.9	1	0.005	8.5	1	0.003	3.3	1	0.068	8.2	1	0.004

References

Allen, J.S., Miller, J.L., & DeSteno, D. (2003). Individual talker differences in voice-onset time. *Journal of the Acoustical Society of America, 113,* 544-552.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 116,* 3171-3183.

Baayen, R. H. (2008). *Analysing linguistic data: A practical introduction to statistics.* Cambridge: Cambridge University Press.

Baker, R., (2008). Grammatical and probabilistic influences on the production and perception of fine phonetic detail. Unpublished Ph.D. dissertation, University of Cambridge.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception.. In Strange, W. (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research* (pp. 167-200). York Press.

Boersma, P., & Weenink, D. (2006). Praat: doing phonetics by computer (Version 4.5) [Computer program]. Retrieved October 26, 2006, from http://www.praat.org/

Bradlow, A. R., & Bent, T. (2008) Perceptual adaptation to non-native speech. *Cognition, 106,* 707-729.

Bradlow, A., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics, 61,* 206-219.

Bybee, J. (2006) From usage to grammar: The mind's response to repetition. *Language* 82, 711-733.

Charles-Luce, J. (1997). Cognitive factors involved in preserving a phonemic contrast. *Language and Speech, 40,* 229-248.

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics, 35,* 210-243.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20,* 521-533.

Coleman, J. (2000). Candidate selection. *The Linguistic Review, 17,* 167-179.

Cooper, W. E., & Paccia-Cooper, J. (1980). *Syntax and speech.* Cambridge, MA: Harvard University Press.

Cruttenden, A. (1994). *Gimson's introduction to the pronunciation of English* (5th ed.). London: Arnold.

Dahan, D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 704-728.

Davis, M. H., Marslen-Wilson, W., & Gaskell, M. G. (2002). Leading up the lexical garden-path: segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 218-244.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., McGettigan, C. (2005), Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134,* 222-241.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics, 67,* 224-238.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101,* 3728-3740.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics, 29,* 109-135.

Gårding, E. (1967). *Internal juncture in Swedish.* Lund: CWK Gleerup.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 152-162.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166-1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251-279.

Goldinger, S.D. (2007). A complementary-systems approach to abstract and episodic speech perception. In J. Trouvain & W.J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences (*pp. 49-54). Saarbrücken.

Goldinger, S.D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics, 31,* 305-320.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 344-359.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics, 31,* 423-445.

Hawkins, S. (1995). Arguments for a nonsegmental view of speech perception. In, K. Elenius & P. Branderud, (Eds.) *Proceedings of the 13th International Congress of Phonetic Sciences 3, (*pp. 18-25). Stockholm.

Hawkins, S., & Smith, R. H. (2001). Polysp: a polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics-Rivista di Linguistica, 13,* 99-188.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31,* 373-405.

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

Hay, J., & Bresnan, J. (2006). Spoken syntax: The phonetics of giving a hand in New Zealand English. *The Linguistic Review, 23,* 351-379.

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

Hervais-Adelman, A., Davis, M.H., Johnsrude, I.S., & Carlyon, R.P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 460-474.Hoard, J. E. (1966). Juncture and syllable structure in English. *Phonetica, 15,* 96-109.

Holm, S. A. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp.145-165). San Diego: Academic Press.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34,* 485-499.

Jones, D. (1931). The 'word' as a phonetic entity. *Le Maître Phonétique, 3rd series 36,* 60-65.

Jurafsky, Daniel, Alan Bell, and Cynthia Girand (2002) The role of the lemma in form variation. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology 7* (pp. 1-34). . Berlin: Mouton de Gruyter.

Kemps, R. J. J. K., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005). Prosodic cues for morphological complexity: The case of Dutch plural nouns. *Memory & Cognition, 33,* 430-446.

Kemps, R.,Wurm, L., Ernestus, M., Schreuder, R., & Baayen, R.. (2005). Prosodic cues for morphological complexity in Dutch and English. *Language and Cognitive Processes, 20,* 43-73.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59,* 1208-1221.

Krakow, R. A. (1999). Physiological organisation of syllables: A review. *Journal of Phonetics, 27,* 23-54.
Kraljic, T. & Samuel, A.G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13,* 262-268.
Kucera, H., & Francis, W.N. (1967). *Computational Analysis of Present-day American English.* Providence: Brown University Press.
Kwong, K. & Stevens, K. N. (1999). On the voiced-voiceless distinction for writer/rider. *Speech Communication Group Working Papers Research Laboratory of Electronics,* Massachusetts Institute of Technology, volume 9, pp. 1–20.
Lachs, L., McMichael, K., & Pisoni, D. B. (2003). Speech perception and implicit memory: Evidence for detailed episodic encoding. In J. S. Bowers, & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 215-235). Oxford: Oxford University Press.
Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica, 5, Supplement,* 5-54.
Lehiste, I. (1972). *Suprasegmentals*. Cambridge, MA: MIT Press.
Local, J. K. (2003). Variable domains and variable relevance: Interpreting phonetic exponents. *Journal of Phonetics, 31,* 321-339.
Luce, P. A., McLennan, C. T., & Charles-Luce, J. (2003). Abstractness and specificity in spoken word recognition: Indexical and allophonic variability in long-term repetition priming. In Bowers, J. & Marsolek, C. (Eds.), *Rethinking implicit memory* (pp. 197-214). Oxford: Oxford University Press.
Markham, D., & Hazan, V. (2004). The effect of talker- and listener-related factors on intelligibility for a real-word, open-set perception test. *Journal of Speech, Language, and Hearing Research, 47,* 725-737.
McLennan, C.T., Luce, P.A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 539-553.
McQueen, J.M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30,* 1113-1126.
Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41,* 329-335.
Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134,* 477-500.
Newman, R.S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics, 35,* 85-103.
Nielsen, K. Y. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39,* 132-142.
Norris, D., McQueen, J., Cutler, A., & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34,* 191-243.
Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204-238.
Nygaard, L.C., Burt, S.A., & Queen, J.S. (2000). Surface form typicality and asymmetric transfer in episodic memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1228-1244.
Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5,* 42-46.
Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60,* 355-376.
O'Connor, J. D., & Tooley, O. M. (1964). The perceptibility of certain word-boundaries. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (Eds.), *In honour of Daniel Jones: Papers contributed on the occasion of his eightieth birthday, 12 September 1961* (pp. 171-176). London: Longmans.
Ogden, R. A. (1999). A declarative account of strong and weak auxiliaries in English. *Phonology, 16,* 55-92.

Ogden, R. A., Hawkins, S., House, J., Huckvale, M., Local, J. K., Carter, P., Dankovicová, J., & Heid, S. (2000). Prosynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language, 14,* 177-210.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America, 99,* 1718-1725.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309-328.

Pickett, J. M., Bunnell, H. T., & Revoile, S. G. (1995). Phonetics of intervocalic consonant perception: Retrospect and prospect. *Phonetica, 52,* 1-40.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology 7* (pp. 101-139). Berlin: Mouton de Gruyter.

Pierrehumbert, J. (2006). The next toolkit. *Journal of Phonetics, 34,* 516-530.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego: Academic Press.

Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics, 20,* 331-350.

Quené, H. (1993). Segment durations and accent as cues to word segmentation in Dutch. *Journal of the Acoustical Society of America, 94,* 2027-2035.

Ranbom, L. J., & Connine, C. M. (2011). Silent letters are activated in spoken word recognition. *Language and Cognitive Processes, 26,* 236-261.

Rice, J.A. (1995). *Mathematical Statistics and Data Analysis* (2nd edition). Belmont, CA: Duxbury.

Rietveld, A. C. M. (1980). Word boundaries in the French language. *Language and Speech, 23,* 289-296.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence, 167,* 170–205.

Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35,* 606-621.

Saltzman, E. (1995). Dynamics and coordinate systems in skilled sensorimotor activity. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 149–173). Cambridge, MA: MIT Press.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90,* 51-89.

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception and Psychophysics, 66,* 422-429.

Sidaras, S.K., Alexander, J.E.D., & Nygaard, L.C. (2009). Perceptual learning of accented speech. *Journal of the Acoustical Society of America, 125,* 3306-3316.

Simko, J., & Cummins, F. (2010). Embodied task dynamics. *Psychological Review, 117,* 1229-1246.

Smith, R. (2003). Influence of talker-specific phonetic detail on word segmentation. In Solé, M.J., Recasens, D. & Romero, J. (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences.* 1465-1468.

Smith, R. H. (2004). *The role of fine phonetic detail in word segmentation.* Unpublished Ph.D. thesis, University of Cambridge.

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. *Journal of the Acoustical Society of America, 119,* 2406-2416.

Sprague, N., Ballard, D., & Robinson, A. (2007). Modeling embodied visual behaviors. *ACM Transactions on Applied Perception, 4,* Article 11.

Stevens, K. N., & House, A. S. (1963). Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research, 6,* 111-128.

Stuart-Smith, J. (2007).Empirical evidence for gendered speech production: /s/ in Glaswegian. In J. Cole and J. Hualde (Eds.), *Change in Phonology (Laboratory Phonology 9)* (pp. 65-86). Berlin: Mouton de Gruyter.

Sumner, M. & Samuel, A. G. (2005). Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language, 52,* 322-338.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America, 88,* 97-100.

Turk, A.E., & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics, 27,* 171-206.

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics, 28,* 397-440.

Umeda, N., & Coker, C. H. (1975). Subphonemic details in American English. In G. M. Fant and M. A. A. Tatham (Eds.), *Auditory analysis and perception* (pp. 539-564). London: Academic Press.

van Santen, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication, 11,* 513-546.

Walsh, M., Möbius, B., Wade, T., & Schütze, H. (2010). Multilevel exemplar theory. *Cognitive Science, 34,* 537-582.

White, L., & Mattys, S.L. (2007). Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró & M.-J.Solé (Eds.), *Segmental and Prosodic issues in Romance Phonology. Current Issues in Linguistic Theory series* (pp. 237-257). Amsterdam: John Benjamins.

Wyld, H. C. (1913). *Collected papers of Henry Sweet.* Oxford: Oxford University Press.