part number
part title

chapter number

Effects on word recognition of syllable-onset cues to syllable-coda voicing

SARAH HAWKINS & NOËL NGUYEN*

X.1 Introduction

It is well known that syllables in many languages have longer vowels when their codas are voiced rather than voiceless (for English, cf. Jones, 1972; House & Fairbanks, 1953; Peterson & Lehiste, 1960; for other languages, including exceptions, see Keating, 1985). In English, the durational difference in stressed syllables can be 100 ms or more, and it is well-established as one of the strongest perceptual cues to whether the coda is voiced or voiceless (e.g. Denes, 1955; Chen, 1970; Raphael, 1972). More recently, van Santen, Coleman & Randolph (1992) showed for one General American speaker that this coda-dependent durational difference is not restricted to syllabic nuclei, but includes sonorant consonants, while Slater and Coleman (1996) showed that, for a British English speaker, the differences tended to be greatest in a confined region of the syllable, the specific location being determined by the syllable's segmental structure.

In a companion study to the present paper (Nguyen & Hawkins, 1998; Hawkins & Nguyen, submitted), we confirmed the existence of the durational difference and showed that it is accompanied by systematic spectral differences in four accents of British English (one speaker per accent). For three speakers/accents, F_2 frequency and the spectral centre of gravity (COG) in the /l/ were lower before voiced compared with voiceless codas, as illustrated in Figure X.1. (The fourth speaker, not discussed further here, had a different pattern, consistent with the fact that his accent realises the /l/-/r/ contrast differently.) Since F_1 frequency in onset /l/s did not differ due to coda voicing, whereas both F_2 frequency and the COG did, we tentatively concluded that our measured spectral differences reflect degree of velarisation, consistent with impressionistic observations. Thus the general pattern is that onset /l/ is relatively long and dark when the coda of the same syllable is voiced, and relatively short and light when the coda is voiceless.

Do these differences in the acoustic shape of onset /l/ affect whether the syllable coda is heard as voiced or voiceless? If they do, the contribution of the onset is likely to be small and subtle, because the measured acoustic differences are small (mean 4.3 ms, 11 Hz COG, 16 Hz F_2 over three speakers). However, though small, the durational differences are completely consistent and strongly statistically significant. Spectral differences are more variable but also statistically significant. Moreover, at least some can be heard. Even if only the more extreme variants provide listeners with early perceptual information about coda voicing, there are far-reaching implications for how we model syllable- and word-recognition, because the acoustic-phonetic properties we are concerned with are in *non*adjacent segments and, for the most part, seem to be articulatorily and acoustically independent of one another. So, by testing whether these acoustic properties of onset /l/ affect the identification of coda voicing, we are coming closer to testing the standard assumption that lexical items are represented as sequences of discrete phonemic or allophonic units, for in standard phonological theory, longer duration and

* We are grateful to Dennis Norris for use of his TSCOP equipment in the perceptual experiment, and to a number of other people for various types of help, especially John Coleman, Eric Fixmer, Ruth Kearns, John Local, and Richard Ogden.

greater darkness of a sonorant are not considered to be fundamental attributes of a voiced stop at the other end of a syllable.

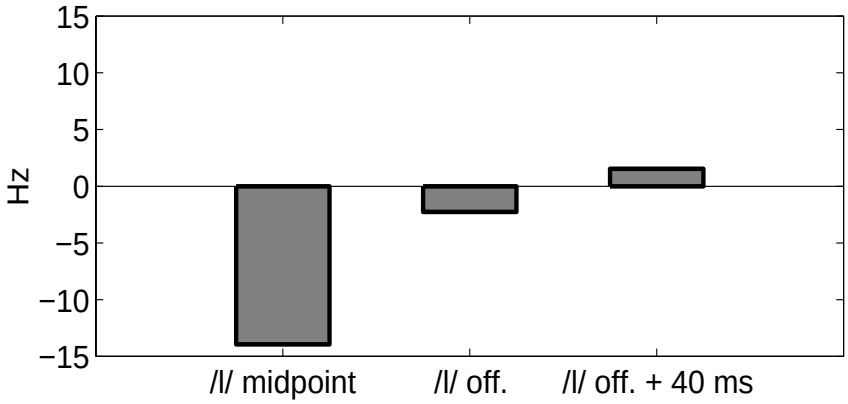

Figure X.1. Difference in mean spectral centre of gravity (COG) at three points during the onset /l/ and vowel of all syllables produced by one speaker. The COG was computed from a DFT spectrum (49-ms Hanning window) converted to an auditory excitation pattern with decibel magnitude (Moore and Glasberg, 1987) in the range 50-3500 Hz. /l/ midpoint = COG centred at the midpoint of the periodic portion of /l/; /l/ off. = COG centred at the offset of the /l/; /l/ off. + 40 ms = COG centred 40 ms after vowel onset.

We argue, then, that whereas the durational pattern simply extends the well-established distinction in vowel length due to coda voicing into another sonorant part of the syllable, the association of darkness in the onset with voicing in the coda potentially adds a different dimension. There is good evidence that decisions about segmental and lexical identity are heavily context-dependent. But to the best of our knowledge, there is no documented evidence that information in nonadjacent phonetic segments influences perception, except when the information either changes the number of lexical candidates or directly reflects a property fundamental to the articulation of the upcoming segment. Thus vowel-to-vowel coarticulation reflects tongue movements essential to the two vowels (cf. Öhman, 1966; Fowler & Smith, 1986), just as long-domain spectral influences due to /r/ or /l/ reflect the spread of attributes essential to the liquids (Hawkins & Slater, 1994; Tunley, 1999; West, 1999). Similarly, although a lip-rounded vowel can affect the identity of an (ambiguous) fricative across an intervening stop (Mann & Repp, 1980), the acoustic properties that spread to the fricative are fundamental to the identity of the conditioning vowel. The current case also differs from documented instances of word identification being delayed until after the acoustic offset of a syllable (*cf.* Grosjean & Gee, 1987): our effects are tautosyllabic, and do not depend on the identification of one versus two words. Perhaps the case most like our own is work by Raphael & Dorman, cited by Fowler (1983), in which the type of syllable onset caused shifts in /d-t/ boundaries in the coda, but these data were interpreted as due to changes in vowel duration consequent on different onsets. The properties we are discussing are different in two crucial ways: they involve changes in the *same* onset segment dependent on coda voicing, and, as noted above, in standard phonetic theory, longer duration and greater darkness are not fundamental attributes of a voiced obstruent articulation, nor even an associated secondary articulation; hence they cannot “spread” to a sonorant elsewhere in the utterance. We argue below that longer and darker onset sonorants could in fact contribute fundamentally to perception of coda voicing, but the argument cannot be couched in terms of standard phonetic theory.

If information about coda voicing is available to listeners in the syllabic onset, then coda voicing could reasonably be modelled as a property of the entire syllable. Such an approach has a number of consequences, one of them being that speech might be analysed and matched in a nonsegmental way to words in the lexicon, which might itself be represented nonsegmentally. The investigation described here was intended to provide data that could justify our developing these arguments.

We cross-spliced /l/s in pairs of monosyllables differing in both coda voicing and lexical status (e.g. /l/ of *larg* was cross-spliced with that of *lark*) and presented original and cross-spliced syllables in a speeded lexical decision task. We expected longer reaction times (RTs) for cross-spliced syllables than originals if acoustic properties of /l/ provide information about the voicing of the coda (cf. studies of coarticulatory cues in word recognition: Marslen-Wilson & Warren, 1994; Streeter & Nigro, 1979; Whalen, 1991).

X.2 Experiment 1: Method

X.2.1 Material

Stimuli were isolated monosyllables spoken by S1 in Nguyen & Hawkins (1998) and Hawkins & Nguyen (submitted), which may be consulted for more details. S1, the first author, has lived in several regions of England and the USA; she maintained an Educated Northern British accent, with clear initial /l/. The data were part of a larger dataset recorded directly into a Silicon Graphics computer (16 kHz SR) in a sound-treated room with high quality equipment. S1 spoke as naturally as possible while keeping pitch, intonation, and rate as constant as possible. Although her average spectral differences were smaller than those of Hawkins and Nguyen's S2 and S3, durational differences were greater, and stress and F_0 were especially consistent.

X.2.1.1 Test stimuli

The test stimuli were 39 pairs of (C)lVC monosyllables differing in the voicing of the final stop (voiced for one member of the pair and voiceless for the other). For 14 pairs, the voiced member was an English word, e.g. *load*, and the voiceless member a nonword, *loat*, while the reverse was true for the other pairs. All words had frequencies of less than 50 per million (mean 4.41 in the Brown Corpus). We controlled as far as possible the frequencies of the lexical competitors for each pair, defined as other English monosyllables with the same onset and nucleus: 33 pairs had only low-frequency lexical competitors (<100 per million); three had one high-frequency competitor, and three had two. Syllable onsets were /l bl pl gl kl fl sl/); vowels were /ɪ ɪ eɪ eə ɛ æ aɪ ɜ ɑ ʌ ɒ əʊ u/).

Tokens were selected from the dataset on grounds of greater-than-average differences in /l/ duration and F_2 frequency, yet small differences in F_0 at the cross-splicing point. The /l/ was longer (mean +23 ms), and had a lower F_2 (mean −17 Hz) in voiced-coda syllables.

X.2.1.2 Control stimuli

There were 39 pairs of FVC stimuli, constructed identically to the test stimuli except that F is a fricative. Thus the FVC pairs differed in the voicing of the final stop, and were matched with the (C)lVCs for phonemic structure, lexical status, word frequency and frequency of lexical competitors. These FVCs were intended as a control set for which cross-splicing was not expected to have any effect.

X.2.1.3 Filler and practice stimuli

There were 89 pairs of filler items; and 38 other monosyllables to be used as practice items. The fillers fell into two categories. (i) 32 word-nonword pairs of (C)lVC and FVC monosyllables. Unlike the test and control stimuli, their codas were either both voiced or both voiceless, but differed along some other phonetic dimension, e.g. *chafe-chayshe*. (ii) 57 word-nonword monosyllabic pairs whose onset contained neither /l/ nor a fricative, and whose codas were both voiced or both voiceless e.g. *noon-noove*. All filler words were low frequency. The test, control and filler sets all contained the same proportions of voiced and voiceless words. The 38 practice pairs fitted some of the above patterns, but were otherwise chosen at random.

X.2.1.4 Cross-splicing of test and control stimuli

To avoid cutting right at the acoustic discontinuity at /l/ release, each (C)lVC pair was cross-spliced at the zero-crossing 4 periods after the end of /l/ (defined as an abrupt rise in formant amplitudes and, usually, frequencies). Thus cross-spliced *lark* had the /l/ + 4 periods from *larg*, while cross-spliced *larg* had /l/ + 4 periods of *lark*. The average duration between the offset of /l/ and the splicing point was 18.6 ms (sd = 3.6 ms, n = 78). The FVC stimuli were cross-spliced at the boundary between aperiodic and periodic excitation. Four stimuli were derived from each pair, the two original syllables, and two cross-spliced ones. Both members of each pair came from the same repetition of the production study. (C)lVC pairs were rejected and replaced by others if the cross-splicing resulted in audible discontinuities in pitch, vowel quality, or loudness. At the cross-splicing point, mean F_0 was 181 Hz; the mean absolute difference in F_0 between the members of each pair was 5.7 Hz (σ 3.8 Hz, n = 39); mean absolute difference in intensity was 1.09 dB (σ 0.89 dB). For comparison, measurements made at the same location in the unspliced stimuli (that is, on either side of the putative cross-splicing point) were as follows: mean absolute difference in F_0 was 1.4 Hz (σ 1.3 Hz, n = 78); mean absolute difference in intensity was 0.34 dB (σ 0.32 dB). Thus, although cross-splicing produced bigger differences in F_0 and intensity than in the original stimuli at the same location, these differences seem to be limited within a reasonable range, given the nature of the experiment. As noted above, pairs with noticeable discontinuities were rejected.

X.2.2 Tapes

The stimuli were recorded onto two DAT tapes, each containing either the original or the cross-spliced version of each test and control item. Each tape began with the 38 practice stimuli, followed by two further blocks, with one member of each pair in each of these two blocks. The four blocks (2 × 2 tapes) were balanced for proportion of words *vs.* nonwords, voiced *vs.* voiceless codas, and original *vs.* cross-spliced syllables. Stimuli were randomised within blocks. ISI was 4 s. A warning beep sounded 1 s before the onset of each stimulus. For each test and control item, a 5 kHz pulse coinciding with the cross-splicing point was recorded onto the other track of the tape. This pulse (inaudible to the subject) triggered the millisecond clock used to measure RTs.

X.2.3 Procedure

Each subject heard one tape over headphones in a sound-treated room, so each person heard only the original or the cross-spliced version of each pair of test and control items, counterbalanced and randomised with the other stimuli. Subjects indicated whether each item was a real word by pressing the appropriate one of two buttons; they were asked to respond as quickly and accurately as possible. Right/left positions of the "yes" and "no" buttons were counterbalanced. RT was measured from the cross-splicing point.

X.2.4 Subjects

18 paid subjects (9 per group) completed the experiment, all native speakers of English aged 18-31 years with no known hearing impairment. Six further subjects (three per group) were rejected for low percentage of correct fast responses (< 1200 ms).

X.3 Results of Experiment 1

Of the 2808 total possible responses to the test and control items, 6 were missing, 259 (9.2%) were discarded as incorrect, and 98 (3.5%) with RTs longer than 1200 ms were discarded after inspecting the RT distribution. Unless otherwise specified, statistical significance was assessed by ANOVAs with splice type (original/cross-spliced), lexical status (word/nonword) and coda voicing (voiced/voiceless) as fully-crossed within-group IVs, and subjects as a between-groups random factor. A preliminary ANOVA showed that cross-splicing did not affect the percentage of correct responses.

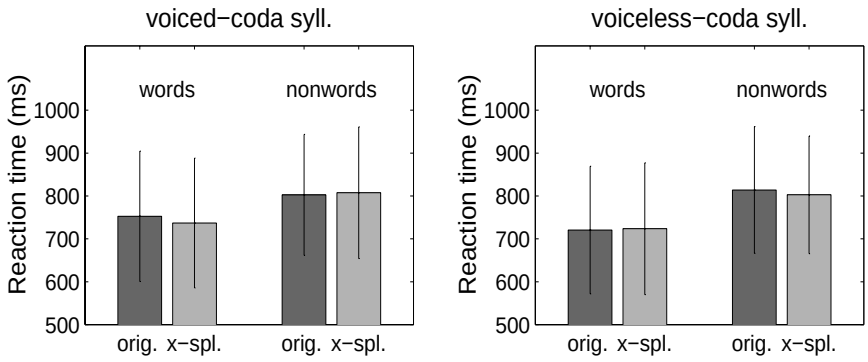

Figure X.2. Mean reaction times and standard deviations for (C)lVC syllables. Left panel = voiced codas; right panel = voiceless codas; orig. = original syllables; x-spl. = cross-spliced syllables.

Figure X.2 shows RTs for (C)lVC syllables. As expected, RTs were significantly faster for words than nonwords: mean 734 ms *vs.* 808 ms, $F(1,17) = 30.61$, $p < 0.001$. However, cross-splicing did not affect RTs: original and cross-spliced syllables had mean RTs of 774 ms and 769 ms respectively. Although this testifies to the success of our splicing technique, it does not support our main prediction. Because there was no difference between the cross-spliced and original test stimuli, responses to these stimuli were not compared with those to the controls (FVCs), which are not discussed further. No other differences between conditions were significant.

Since RTs to cross-spliced syllables were not longer overall, we tested whether they were sensitive to the *magnitude* of changes resulting from cross-splicing, such that responses might be slower when cross-splicing entailed greater modification of the acoustic shape of the syllable onset. If so, then RTs should increase in

proportion to the acoustic difference in duration and F_2 frequency between the two /l/s in a cross-spliced pair. (The COG could be substituted for F2.) In a standard linear statistical framework, hereafter the voicing-cue model, this prediction can be stated as follows:

(1) $RT(\text{x-spliced}) - RT(\text{original}) = 0.5(a_1[\text{dur(voiced)} - \text{dur(voiceless)}] + a_2[F_2(\text{voiceless}) - F_2(\text{voiced})] + b)$
where $a_1 > 0$ and $a_2 > 0$

In (1), b is the intercept in a standard linear regression equation; a_1 and a_2 are greater than 0 in order to satisfy the requirement that the relationship is positively correlated; and the constant 0.5 is used because two single-parameter models are combined. Fitting the model to the empirical data involves finding the set of coefficients, a_1, a_2 and b, that minimises the squared deviations between the observed differences in RT, and those predicted from the durations and F_2 frequencies for /l/. These coefficients were estimated in four multiple linear regressions, for voiced-coda words, voiceless-coda words, voiced-coda nonwords and voiceless-coda nonwords, as shown in Figure X.3.

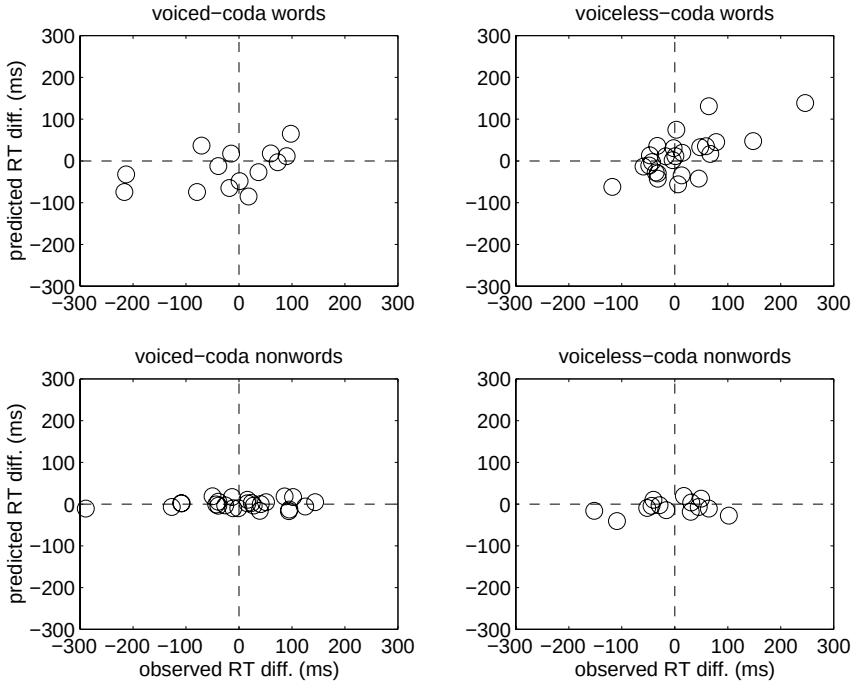

Figure X.3. Predicted difference in RT between cross-spliced and original versions of stimuli, as a function of observed differences. Predicted differences were based on the difference between the original and cross-spliced /l/s in duration and F_2 frequency.

Listeners' RTs were positively correlated with the magnitude of the acoustic change that cross-splicing introduced for real words, whereas they were completely unrelated for nonwords. That is, whereas the data points for words rise towards the right, those for nonwords lie along a horizontal line. Amongst real words, listeners' RTs were only strongly correlated when the coda was voiceless (multiple $R = 0.689$, $F(2,22) = 9.95$, $p < 0.001$, Spearman $rho(25) = 0.59$, $p < 0.01$, one-tailed). In other words, when a real word ended with a voiceless coda e.g. *lark*, the longer and darker its cross-spliced /l/, then the slower were listeners' responses. The correlation for voiced-coda real words is weaker. The parametric test does not achieve statistical significance at the 5% level (multiple $R = 0.465$, $F(2,11) = 1.52$, $p = 0.261$), but the rank-order Spearman correlation does ($rho(14) = 0.49$ $p << 0.05$). Responses to real words with voiceless codas thus appear to be most affected by cross-splicing, but it seems reasonable to conclude that both voiced- and voiceless-coda real words were processed in a qualitatively different way from nonwords in this experiment.

A number of factors might account for the pattern of these data. First, the difference between words and non-words (positive correlations for words but zero correlations for nonwords), suggests that these phonetic mismatches were processed at the lexical level, which is in keeping with a nonsegmental view of speech perception. Second, the long vowel of voiced-coda words may have been long enough to override the perceptual salience of small changes in onset /l/. Some support for this point lies in the slight (though nonsignificant) tendency for RTs to be longer when the coda was voiced (777 ms *vs.* 768 ms, $F(1,17) = 1.61$, $p = 0.222$). On this

account, the voiceless-coda words offered the highest probability of finding a correlation between the acoustical measures and RTs.

Alternatively, the longer RTs (mean +75 ms) for nonwords than words may have prevented fine acoustic details of /l/ from measurably influencing RT to nonwords, analogous to a ceiling effect. This explanation is not very convincing since, at around 800 ms, the RTs to nonwords were not especially long. A more convincing possibility is that responses to some stimuli were delayed because the cross-spliced stimuli sounded unnatural. This "general disruption" hypothesis differs from the voicing-cue hypothesis on one point. If cross-splicing itself interferes with processing the syllable, then it should interfere regardless of whether the coda is voiced or voiceless. Only the size of the acoustic differences between the two /l/s should matter. For example, cross-splicing should increase RT even when onset /l/ is shorter in the voiced-coda syllable, provided that the durational difference is large enough. In keeping with the linear statistical framework adopted above, the general-disruption hypothesis can be stated thus:

(2) $RT(\text{x-spliced}) - RT(\text{original}) = 0.5(a_1 \times |\text{dur(voiced)} - \text{dur(voiceless)}| + a_2 \times |F_2(\text{voiced}) - F_2(\text{voiceless})| + b)$
where $a_1 > 0$ and $a_2 > 0$

The general-disruption hypothesis cannot be properly tested because /l/ was always longer in the voiced-coda syllables. However, since F_2 frequency in /l/ was higher in the voiced-coda than the voiceless-coda syllable in a few pairs, the general-disruption model could be fitted to the RT data as described above, though only the influence of F_2 is really being assessed. As with the voicing-cue model, predicted and observed RTs were only significantly correlated for the voiceless-coda words, using these parametric tests. However, the multiple correlation coefficient was lower than with the voicing-cue model ($R = 0.50$), and achieves a much lower level of significance ($F(2,22) = 3.58$, $p < 0.05$). Thus, the voicing-cue model accounts more reliably than the general-disruption model for the reaction-time data.

X.4 Discussion of Experiment 1

These data show that onset /l/ by itself does not provide a strong indication of coda voicing: cross-splicing affected neither the percentage of items correctly identified nor average RT. On the other hand, the regression analyses suggest that the acoustic properties of onset /l/ can disrupt lexical decisions in the right circumstances, namely when the onset of a real word indicates a voiced coda but the syllable rhyme is consistent with a voiceless one. Though these results show only a small perceptual influence, they deserve attention.

First, they are unlikely to be artifacts of the cross-splicing itself. Although splicing did introduce some differences in F_0 and intensity (see Section X.2.1.4), the fact that its effect was not significant in the ANOVAs testifies to the acceptability of the experimental manipulation. Likewise, the general disruption model does not account for the data as reliably as the voicing-cue model. Moreover, had splicing introduced significant artifacts, one would expect differences in RT for the nonwords. Instead, the restriction of positive correlations to real words is consistent with Whalen, Best & Irwin's (1997) finding that people are sensitive to allophonic errors in real words but not nonwords.

Second, work by Wolf (1978) and Summers (1988) demonstrates differences in vowel onset spectra that provide weak perceptual information about the voicing of the coda obstruent that conditions them. If vowel onsets can cue coda voicing, then syllable onsets could be expected to as well, at least if they are sonorant.

Third, if acoustic properties of onset /l/ provided strong cues to coda voicing, then they would be expected to have been well known for years, as have durational properties of the preceding vowel. Since they are not well known, it seems reasonable to assume that perceptual influences from onset /l/ should be small. But if they exist at all, there could be important implications for theories of how we understand words.

We thus interpret these results as promising, but we need more supportive data before developing the theoretical implications. A second experiment was thus carried out to directly test whether acoustic properties of onset /l/ can indicate the voicing of the coda. This experiment is briefly described in the next section.

X.5 Experiment 2

The purpose of Experiment 2 was to check in a more controlled way than Experiment 1 allowed that the duration and F_2 frequency of onset /l/ can cue the voicing of a coda obstruent. Pilot studies suggested that F_0 might also be influential, so, since this has theoretical interest (cf. Kingston & Diehl, 1995), F_0 was a third variable. (Recall that in Experiment 1, F_0 differences were minimised in order to avoid discontinuities at the cross-splicing point.) The word *led* was synthesised using Sensimetrics' Sensyn (similar to KLSYN88) from a token spoken by S1. The impressionistic quality of synthesis was good. From this stimulus, five more were made such that the duration of /l/ varied in 20-ms steps over the range 70-170 ms. At each of the six durations a total of 4 stimuli

were made by using different combinations of F_2 frequency and F_0 during the /l/. F_2 was constant during the /l/, at either 1850 Hz or 1740 Hz; in each case it jumped to 1860 Hz at vowel onset, from where it rose to 2080 Hz over the next 110 ms. Likewise, F_0 began at either 1800 Hz or 1680 Hz, and fell linearly to 1620 Hz at vowel onset, from which point on it fell over identical values in all stimuli. Thus there were 24 (6 durations x 2 F_2 x 2 F_0) stimuli, identical except for properties of the initial /l/.

Each stimulus was truncated 80 ms after vowel onset, and 300 ms of white noise (rise time 20 ms, decay time 50 ms) was added to it, starting 70 ms from vowel onset. Thus the end of each stimulus seemed to be obliterated by noise: only its initial /l/ and 70-80 ms of the vowel could be heard. The stimuli were randomised in ten blocks of 24, preceded by 16 practice items, and played one at a time to 12 listeners using a Tucker-Davis DD1 D-to-A system controlled by a PC computer. Other conditions were as in Experiment 1. Listeners pressed one response button when they thought the word heard was *let*, and another if they thought it was *led*. ISI was 3 s.

Both duration and F_0 had the predicted effect on responses, but F_2 frequency did not: shorter /l/s and higher F_0 produced more *let* responses. Duration had by far the most influence, with a mean of 73% *let* responses when /l/ was shortest, falling roughly linearly to 36% when /l/ was longest ($F(5,55) = 16.21$, $p < 0.0001$). F_0 alone had a much smaller effect—only about 8% difference: 59% *vs*. 51% *let* to stimuli with high *vs.* low F_0 respectively ($F(1,11) = 12.12$, $p = 0.005$). There were no significant interactions, although F_0 tended to exert less influence as /l/ lengthened: responses to stimuli with high and low F_0 differed by 13% when /l/ was shortest (70 ms: 80% *vs.* 67% *let*) but only 4% when /l/ was longest (170 ms: 38% *vs.* 34%). The possibility that non-durational differences may exert most influence when /l/ is short is worth pursuing, since F_2 frequency, whose influence was far from significant, shows the same pattern as F_0 in this respect. That F_2 frequency did not significantly affect responses bears more study in any case, in view of our measurements of natural speech, and other perceptual data (Newton, 1996). For example, the influence of large durational and F_0 differences may outweigh that of F_2 frequency in this paradigm, or there may be interactions with vowel height.

In summary, listeners knew about and could use acoustic properties of onset /l/ to help predict the voicing of the coda of the same syllable. Although the properties of onset /l/ did not unequivocally signal the voicing of the coda, they were salient enough to swing decisions by as much as 17-35% on either side of chance.

X.6 General discussion

X.6.1 Preliminary comments

Experiment 1, using natural speech, showed that lexical decisions could be disrupted when an onset appropriate for a voiced coda was followed by the nucleus and coda appropriate for a voiceless obstruent. There is some evidence for a similar but much weaker effect in the opposite case, when the onset was appropriate for a voiceless coda but the rhyme was appropriate for a voiced coda. Experiment 2, using variants on a good-quality synthetic *led*, confirmed that, in the absence of other cues, both the duration of the /l/, and its F_0, influence listeners' responses about the voicing of the coda. F_2 frequency, however, did not. More experiments are planned to further explore influences of F_0 and F_2 frequency.

It seems, then, that the acoustic shape of onset /l/ indicates the voicing of the coda and affects the speed of lexical access. These acoustic properties of onset /l/ are subtle and have only a limited influence on perception when pitted against strong perceptual cues. But it is precisely because they are subtle that we find them interesting. In the cross-spliced syllables, we would not expect /l/ to always play a decisive role in perception of the coda as voiced or voiceless, given that differences in the vowel are still associated with the appropriate coda. In their original contexts, however, there seems no question that they could significantly contribute to the percept of coda voicing. Acoustic evidence in the onset increases the total duration in which coda voicing is consistently indicated, and its presence in more than one phonological unit might enhance the percept further.

X.6.2 Implications for perception of low-level phonetic contrasts

Our findings fit well with assumptions of Auditory Enhancement Theory (AET: Kingston & Diehl, 1994, 1995). In AET, many acoustic properties combine to form complex Intermediate Perceptual Properties (IPPs), which in turn combine to contribute in varying proportions to the identification of a distinctive feature. Longer onset /l/s before voiced codas could contribute to the (slightly redefined) IPP of the C:V duration ratio. The darkness of the /l/ could contribute to the IPP called the low-frequency (LF) property, since that is associated with a lower spectral centre of gravity. In AET, the C:V duration ratio and LF property are the two IPPs that feed into the phonological feature [voice]. Their physical parameters combine to form a unitary percept, defined as an abstract feature.

Both these IPPs are defined in terms of adjacent acoustic segments. It is trivial to extend the definition of the C:V duration ratio to include durations of sonorant syllable onsets, and the LF property to include F_0 in the onset. But Kingston and Diehl would be justified in objecting to our lumping darkness with the LF property, which is currently defined as a psychoacoustic blending of high-amplitude energy in F_1 and F_0 in the vicinity of a vowel-obstruent boundary: F_1 frequency at vowel offset, and periodicity in the obstruent closure. Our "dark" property is different: darkness is a property of the entire frequency spectrum, or perhaps the mid-frequencies, and it is distributed across the syllable onset or the onset and nucleus. It remains to be seen whether we are looking at one IPP or two in the LF property and onset darkness, but at some level of analysis we are almost certainly looking at a property of the whole syllable. Nevertheless, our durational and spectral findings seem compatible with AET's principles, and thus with an interesting model of the perception of low-level phonetic contrasts. Since our findings are also compatible with principles of psycholinguistic models of word recognition, we have a potential link between two disciplines that traditionally pursue separate paths.

X.6.3 Implications for lexical access

We have shown that lexical decisions were most likely to be disrupted when an onset appropriate for a voiced coda was followed by the nucleus and coda appropriate for a voiceless obstruent. This pattern might arise because the perceptual system is, metaphorically, taken by surprise when the vowel is short. One might characterise this process in terms of a Weber ratio, involving the contrast between a long onset and short rhyme. A short rhyme may allow properties of the onset /l/ to be discriminated better.

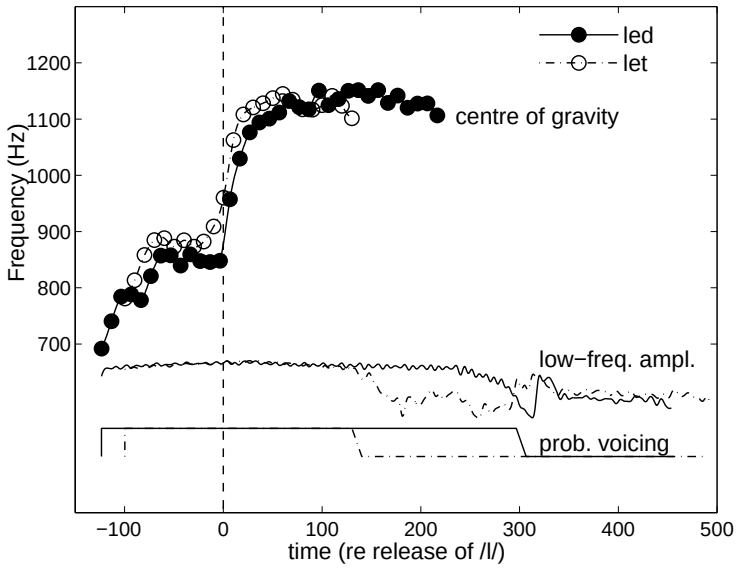

Figure X.4. Time-course of centre-of-gravity, low-frequency amplitude, and probability of voicing for *led* and *let*, aligned at /l/ offset. See text for explanation.

Figure X.4 illustrates our argument. It shows actual parameter values for one unspliced token of *led* and one of *let*, spoken in isolation by S1. Consider for now only the spectral centre of gravity, calculated as described for Figure X.1 from the beginning of /l/ to the end of the vowel. The COG is higher throughout the /l/ and the first half of the vowel in *let* (open circles). In the cross-spliced version of *let*, the syllable-onset would be taken from the other syllable, *led*, so the onset would be relatively long and dark, and the rhyme would have a short vowel paired appropriately with its voiceless coda. The long dark /l/ would presumably lead listeners to initially expect a voiced coda. When they hear the end of the vowel sooner than expected, listeners discover that their initial interpretation was wrong: RT increases while the interpretation changes as the vowel suddenly ends.

In contrast, when the cross-spliced onset is short and followed by a long vocalic nucleus, the likelihood that the coda is voiced will rise as more of the vowel is heard beyond a certain minimum duration. By the end of the word, the relatively weak initial cues to coda voicelessness lose out to stronger contradictory evidence from the syllable nucleus and the coda itself. There is no need for sudden late readjustment, because of the length and consistency of the cues for coda voicing in the syllable rhyme; decisions are thus not significantly slower. Of the explanations we have looked at, this one fits the data best.

There are at least three ways of conceptualising this process. (1) The listener makes early decisions about coda voicing but backtracks when she or he discovers an error. (2) The listener continually predicts what is about to occur, based on past and current information, and makes late decisions. (3) The listener continually predicts what is about to occur, makes early decisions, and corrects them later if necessary—in other words, early decisions are continually updated. Our work cannot speak to the first possibility, but one way to distinguish between the second and third possibilities could be to pit these acoustic-phonetic differences against a context effect that is generally thought to take place late in the decision process (cf. Frauenfelder & Tyler 1987: 8-17). If the acoustic-phonetic properties are used in perception, they should presumably be used early, and so should reduce the magnitude of post-lexical effects. These issues must be addressed in future work. Here, we focus on the type of model that these data seem to support, although we point out where our views have implications for the above questions.

Our general approach is uncontroversial: from the input signal, the system identifies properties that are linguistically relevant, combines them into higher-order categories, and identifies their linguistic status in a matching process based on activation of and competition between representational units. This process repeats as long as the signal continues. What distinguishes our model from others is the emphasis on the acoustic signal's fine structure, our consequent conceptualisation of how words are first recognised, and the implications for lexical and phonological structure.

In our model, the fine structure of the sensory signal contributes crucially to spoken word recognition. Consider Figure X.4 again. In addition to the COG data, there are measures for low-frequency amplitude and the probability of voicing. These parameters extend through the ends of the stops and show the expected differences between the two words. They are just some of many parameters that can be readily calculated from the acoustic signal. In reality there would presumably be many channels, whose nature and initial structuring into auditory streams could be subject to psychoacoustic principles (*cf.* Bregman, 1990). Such fluctuating, multi-channel streams of information may resemble what the auditory system transmits to the brain's word recognition system. There, they can be matched against similar representations stored in memory, which might comprise or include abstract templates, or prototypes, developed from the statistical distribution of items that have been heard over many years. (Such abstract representations presumably include multimodal information derived from non-auditory as well as auditory input channels (*cf.* Remez, Fellowes, Pisoni, Goh, & Rubin, 1998), but for simplicity the present description refers only to the acoustic/auditory modality.) In other words, we are suggesting that the first contact stage of spoken word recognition involves matching incoming auditory shapes with memories of other auditory shapes. And that in at least the initial stages of word recognition, lexical representations are simply time-varying auditory shapes.

We propose that distributional evidence accumulated over repeated hearings allows additional classes of specification to develop from the unsegmented lexical representation that the auditory signal initially contacts (*cf.* Figure X.4). One such class is abstract features whose time-varying, phonetically unsegmented values vary continuously between 0 and 1. These can be thought of as fuzzy truth values (Massaro, 1998) or as probabilities (cf. McClelland & Elman, 1986; Norris, 1994). Unlike feature values in those models, which roughly speaking represent the expected degree to which that feature is or should be present, feature values in our model represent a range of acceptable values depending on speech style. Other information derived from statistical evidence includes syllabic and prosodic structure, and links to grammatical and semantic systems. From the spoken signal, the mature listener can identify elements at all levels of the structure simultaneously. For example, an onset /l/ whose acoustic properties were indicative of a voiced coda would raise activation associated with coda voicing throughout the syllable while having no effect on features associated with manner or place of coda articulation. Likewise, the presence of a stressed syllable could be identified before the place of articulation of its initial consonant, or *vice versa*, depending on context and clarity of the signal.

The incoming speech signal is continuously monitored for information on each unit, giving rise to continuous modulation of feature values and structural relationships, and the best match with lexical items is sought. The model allows for the system to track time functions and hence vocal-tract dynamics and their acoustic consequences. It also allows for information that is relatively weak to provide an unambiguous percept, as long as it is consistent over time (*cf.* Warren & Marslen-Wilson, 1987; Hawkins & Warren, 1994). Disruption of this consistency delays perceptual decisions, as demonstrated in Experiment 1.

If the first stage of word recognition is based on relatively nonsegmental principles, how then do we get to segments, and to other discrete higher-order linguistic units? Discrete units presumably have some sort of mental representation, since they play a part in people's linguistic lives—consider speech errors, writing systems, dog Latin, poetry, and so on. We envisage that each person gradually builds up a system of discrete categories, corresponding to the familiar levels of linguistic analysis, from the distribution of sensory patterns stored in memory. So, for example, features which regularly co-occur become bound into segments as phonology develops. How exactly this might happen is still largely a matter of opinion, with the two main schools of

thought involving either exemplar-based or idealised prototype-based systems. In our view, we cannot yet expect to distinguish between the two, and indeed the distinction may not be clearcut, but it seems undeniable that the process of category formation depends initially on identifying repeated patterns over many instances of like items (*cf.* Jusczyk, 1993).

In our view, processes dependent on episodic memory should be important in the early contact stages, when information-rich sensory patterns are matched against similar time-varying, streamed, but otherwise unsegmented mental representations. Research in cognition suggests that episodic memory contributes significantly to how humans form categories in nonlinguistic systems. It may similarly contribute to speech perception (Pisoni, Lively, & Logan, 1994; Lively, Pisoni & Goldinger, 1994). The perceptual learning and attunement approach of direct realists, described for speech by Best (1994a, 1994b, 1995), likewise ascribes the formation of abstract cognitive categories to the individual's observation of repeated co-occurrences between events. Eventually, the individual learns to attend to certain types of information and to ignore others. This approach is compatible with the principle that a wide variety of higher-order classes can be formed from a smaller number of individual streams in the sensory input, by weighting and binding them together in different combinations.

The episodic memory and perceptual learning/attunement approaches could be compatible. They may describe the same processes from different viewpoints, though we tend to favour the idea that attunement may begin to operate on more abstract categories, when episodic memory exerts less influence on lexical access. Conceivably, they have somewhat complementary roles in different types of memory. For example, episodic memory of instances could be a primary aspect of memories for speakers and for vocally-expressed emotions, while processes better described in terms of attunement might have more influence in representations of lexical items and accents. The present data on /l/ would presumably be relevant to all four types of memory. Amongst the advantages of both approaches are, first, that access to exemplars in memory easily allows access to knowledge of acceptable ranges of variation in feature values for particular lexical items, and second, that different people may develop slightly different phonologies and grammars due to exposure to different input, just as they develop different lexicons and word associations, and as bilinguals' experiences affect how they organise their languages. Although these processes are presumably most plastic in an individual's early years, we envisage them as continuing throughout life, thus allowing for accommodation to new accents, for example.

This type of model can easily accommodate our perceptual results. At the end of the cross-spliced *voiceless* word in Experiment 1, the voiced coda is still activated to a certain degree, for the vowel is short enough that coda activation caused by the onset has not yet decayed. Hence the delay in the response, because the cross-spliced stimulus is more ambiguous than the original voiceless-coda word. In contrast, at the end of the cross-spliced *voiced* word, the activation of the voiceless coda (produced by the cross-spliced onset) has already been reduced to (or towards) 0 because of the relatively long vowel duration. In Experiment 2, the nature of the task presumably activates both words, but the acoustic properties of the /l/, relative to those of the vowel, provide enough evidence in favour of just one word for responses to differ significantly from chance.

Although there are other ways to explain these data, the proposed model has several advantages. Most specifically, because it maps fine acoustic detail onto nonsegmental whole-word memories, it economically accounts for properties of the syllable onset being able to cue the phonological status of the coda. Moreover, in Experiment 1, nonwords did not show the positive correlations found for words (Fig. X.3), which suggests that the mismatches were processed at the lexical level. This in turn supports a nonsegmental perceptual model in which fine phonetic detail distributed across the entire syllable is encoded in the lexicon. More generally, the proposed explanation is relatively simple: it invokes only the notions of activation, decay, and some sort of competition. The questions raised at the beginning of this section—at what point in the decision process the increase in RTs occur, and whether the increases reflect a slowing in the decision time itself, or a backtracking and explicit error correction—lose their importance, for this model does not require concepts like "early decision strategies" and "backtracking" to correct errors induced by early decisions. A "lexical decision" can be made at whatever time circumstances demand, for it always reflects the same thing: the current best match to a lexical representation, based on all the available evidence.

X.6.4 Some connections

Although we advocate a nonsegmental and phonetically-detailed approach to modelling, our data could be described within a segmental framework. However, that is less appealing because it makes more assumptions, e.g. that coarticulatory effects must be "undone" through perceptual compensation for coarticulation (*cf.* Mann & Repp, 1980). Our perceptual findings may not be incompatible with a TRACE-like model, with the addition of a set of connections between detectors for syllable-initial phonemes and detectors for syllable-final ones. In a sense, however, such connections would make the model a little less segmental, given that they would have to represent syllable constituents explicitly. More generally, accounting for how distributed acoustic cues are

perceived in the framework of a phonemic model such as TRACE seems counterintuitive to us, for we see no convincing reason why distributed acoustic information should be fed into short-domain detectors (phonemes), only to be redistributed again as these detectors in turn activate long-domain units (words).

Our basic ideas are implicit or explicit in most statistically-based continuous-information models of speech perception and word recognition, as indicated above. However, our approach is novel in its focus on the role of fine phonetic detail. By contrast, most models in practice assume a phoneme input, so they ignore the perceptual salience of systematic variation in acoustic-phonetic fine detail and thus begin at a level considerably higher than our conceptualisation of initial lexical contact. An exception amongst models of adults' spoken word recognition is the revised cohort model (Marslen-Wilson, 1993; Marslen-Wilson & Warren, 1994), but even this work has not examined how particular acoustic-phonetic properties correlate with perceptual responses. Grossberg's adaptive resonance theory (Grossberg, Boardman & Cohen, 1997), although it uses phonemes as obligatory units, likewise tackles a number of issues relevant to our concerns, including word recognition at different speech rates (see also Protopapas, 1999).

Superficially, our model resembles LAFS (Klatt, 1979), but there are crucial differences. The most obvious is that phoneme strings are central to LAFS, whereas they are non-essential in ours, usually a by-product of literacy in an alphabetic writing system, and only loosely associated with lexical representations. Second, our input and initial contact representation is nonsegmental, but not unanalysed: sensory information is analysed into streams, which is attended to and further processed in different ways. This is true also for LAFS, but the emphasis is different. In particular, prosodic phonological structure seems to be secondary in LAFS, whereas by its nature it is fundamental to systems like ours, which must take account of syllabic structure, for example. Finally, to reduce errors, LAFS prefers lexical decisions to be made late, but not too late, whereas our hierarchical, nonsegmental model should be less affected by these constraints. Undoubtedly however, some differences between the two models are due to the twenty-year gap between them.

Our position is compatible with work in a number of other disciplines. Early antecedents in phonological acquisition include the nonsegmental approaches of Waterson (1971) and Ferguson & Farwell (1975); see also MacNeilage & Davis (1990). More recently, Jusczyk's (1993) WRAPSA model of how infants learn to recognise words is especially compatible, as are aspects of Suomi's (1993) DAPHO, especially in the status it accords phonemes. By showing that perceptual cues to segmental phonological contrasts can be present in non-adjacent phonological segments, the particular contribution of the present study is to provide supportive data for models such as these, and thereby to underline the likelihood that adults may also recognise spoken words by reference to holistic rather than segment-by-segment analyses. The usual processes that adults use to recognise spoken words may be much the same as the processes young children use, not only because young children begin that way because they have no phonology, but also because speech itself is structured to facilitate or even demand the use of those processes.

There are implications of the present work for increasing the naturalness and robustness of speech synthesis, inasmuch as properties that influence speech perception by humans are presumably broadly similar to those that produce natural-sounding synthetic speech that is robust in adverse listening conditions. However, although our approach would seem to be compatible with statistically-based techniques in machine speech recognition, current systems do not use long- as well as short-term dependencies, and it is not currently clear how they could be made to do so (S. Young, T. Robinson, pers. comm.; Nguyen & Hawkins, 1999).

X.6.5 Concluding remarks

This demonstration of the influence of fine acoustic detail in a syllable onset on perception of the phonological status of the coda is unusual in combining psycholinguistic methods with phonetic materials and principles. We have shown that tampering with fine acoustic detail in nonadjacent phonetic segments can disrupt lexical access, and supported these observations with more direct evidence of their perceptual cueing power. We argued from this evidence that the initial contact stage of word recognition involves matching relatively unanalysed auditory streams with memories of such streams, or with closely-related representations constructed from such streams. Discrete units, and abstract representation, could be built up from these auditory representations (*cf.* Beckman and Pierrehumbert, this volume), **[eds: you asked us to draw your attention to this cross-ref, requested by Reviewer 2]** but if they are, then they function later in the process of lexical access. This viewpoint lends itself naturally to models of speech perception based on statistical (distributional) principles and identification via processes of excitation and decay of perceptual units.

X.7 References

Best, C.T. 1994a. Learning to perceive the sound pattern of English. In C. Rovee-Collier & L.P Lipsitt (eds.), *Advances in Infancy Research.* Norwood, NJ: Ablex Publishing Corporation, Vol. 9, 217-304.

Best, C.T. 1994b. The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J.C. Goodman & H.C. Nusbaum (eds.), *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words.* Cambridge, MA: MIT Press, 167-224.

Best, C.T. 1995. A direct realist view of cross-language speech perception. In W. Strange (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Speech Research.* Baltimore: York Press, 171-206.

Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Brown Corpus. http://wapsy.psy.uwa.edu.au/uwa_mrc.htm

Chen, M. 1970. Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* 22: 129-159.

Denes, P. 1955. Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America* 27: 761-764.

Ferguson, C.A., & Farwell, C.B. 1975. Words and sounds in early language acquisition: English initial consonants in the first fifty words. *Language* 51: 419-439.

Fowler, C.A. 1983. Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General*, 112: 386-412.

Fowler, C.A., & Smith, M.R. 1986. Speech perception as 'vector analysis': an approach to the problem of invariance and segmentation. In J.S. Perkell & D.H. Klatt (eds.), *Invariance and Variability in Speech Processes*. NJ: Lawrence Erlbaum, 123-136.

Frauenfelder, U.H., & Tyler, L.K. 1987. *Spoken Word Recognition.* Cambridge, MA: MIT Press. (Reprinted from *Cognition: International Journal of Cognitive Science* 25, 1987.)

Grossberg, S., Boardman, I., & Cohen, M. 1997. Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance* 23: 481-503.

Grosjean, F., & Gee, J.P. 1987. Prosodic structure and spoken word recognition. In U.H. Frauenfelder & L.K. Tyler (eds.), *Spoken Word Recognition.* Cambridge, MA: MIT Press, 135-155.

Hawkins, S., & Nguyen, N. (submitted) Acoustic properties of syllable-onset /l/ dependent on syllable-coda voicing. Submitted to *Journal of Phonetics.*

Hawkins, S., & Slater, A. 1994. Spread of CV and V-to-V coarticulation in British English: Implications for the intelligibility of synthetic speech. *Proceedings of the 3rd International Conference on Spoken Language Processing* 1: 57-60.

Hawkins, S., & Warren, P. 1994. Implications for lexical access of phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics* 22: 493-511.

House, A.S., & Fairbanks, G. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America* 25: 105-113.

Jones, D. 1972. *An Outline of English Phonetics.* Cambridge, U.K.: Cambridge University Press.

Jusczyk, P. (1993). From general to language-specific capacities: the WRAPSA Model of how speech perception develops. *Journal of Phonetics* 21: 3-28.

Keating, P.A. 1985. Universal phonetics and the organisation of grammars. In V.A. Fromkin (ed.) *Phonetic Linguistics: Essays in Honor of Peter Ladefoged.* Orlando, FL: Academic Press, 115-132.

Kingston, J., & Diehl, R.L. 1994. Phonetic knowledge. *Language* 70: 419-454.

Kingston, J., & Diehl, R.L. 1995. Intermediate properties in the perception of distinctive feature values. In B. Connell & A. Arvaniti (eds.), *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*. Cambridge: Cambridge University Press, 7-27.

Klatt, D.H. 1979. Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics* 7: 279-312.

Lively, S.E., Pisoni, D.B., & Goldinger, S.D. 1994. Spoken word recognition: Research and theory. In M.A. Gernsbacher (ed.), *Handbook of Psycholinguistics*. Academic Press, 265-301.

MacNeilage, P.F., & Davis, B.L. 1990. Acquisition of speech production: The achievement of segmental independence. In W.J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modelling.* Dordrecht: Kluwer, 55-68.

Mann, V.A., & Repp, B.H. 1980. Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception and Psychophysics* 28: 213-228.

Marslen-Wilson, W. 1993. Issues of process and representation in lexical access. In G.T.M. Altmann & R. Shillcock (eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting.* Hillsdale, N.J.: Lawrence Erlbaum.

Marslen-Wilson, W., & Warren, P. 1994. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review* 101: 653-675.

Massaro, D.W. 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* Cambridge, Mass.: MIT Press.

McClelland, J.L., & Elman, J.L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18: 1-86.

Moore, B.C.J., & Glasberg, B. 1987. Formulae describing frequency selectivity as a function of frequency and level and their use in calculating excitation patterns. *Hearing Research* 28: 209-225.

Newton, D.E. 1996. The nature of resonance in English: an investigation into lateral articulations. *York Papers in Linguisics*

17: 167-190.
Nguyen, N, & Hawkins, S. 1998. Syllable-onset acoustic properties associated with syllable-coda voicing. *Proceedings of the 5th International Conference on Spoken Language Processing.* Available on CD-ROM: ICSLP-98. Paper #539.
Nguyen, N., & Hawkins, S. 1999. Implications for word recognition of phonetic dependencies between syllable onsets and codas. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A.C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* 1: 647-650.
Norris, D. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52: 189-234.
Öhman, S. 1966. Coarticulation in VCV utterances: spectrographic measurement. *Journal of the Acoustical Society of America* 39: 151-168.
Peterson, G.E., & Lehiste, I. 1960. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* 32: 693-703.
Pisoni, D.B., Lively, S.E., & Logan, J.S. 1994. Perceptual learning of nonnative speech contrasts: Implications for theories of speech perception. In J.C. Goodman & H.C. Nusbaum (eds.), *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words.* Cambridge, MA: MIT Press, 121-166.
Protopapas, A. 1999. Connectionist modeling of speech perception. *Psychological Bulletin* 125: 410-436.
Raphael, L. 1972. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America* 51: 1296-1303.
Remez, R. E., Fellowes, J. M., Pisoni, D. B., Goh, W. D., & Rubin, P. E. 1998. Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Communication* 26: 65-73.
Slater, A., & Coleman, J. 1996. Non-segmental analysis and synthesis based on a speech database. *Proceedings of the IVth International Conference on Spoken Language Processing* 4: 2379-2382.
Streeter, L.A., & Nigro, G.N. 1979. The role of medial consonant transitions in word perception. *Journal of the Acoustical Society of America* 65: 1533-1541.
Summers, W.V. 1988. *F*1 structure provides information for final-consonant voicing. *Journal of the Acoustical Society of America* 84: 485-492.
Suomi, K. (1993). An outline of a developmental model of adult phonological organization and behaviour. *Journal of Phonetics* 21, 29-60.
Tunley, A. 1999. *Coarticulatory influences of liquids on vowels in English.* Unpublished PhD dissertation, University of Cambridge, U.K.
van Santen, J.P.H., Coleman, J.S., & Randolph, M.A. 1992. Effects of postvocalic voicing on the time course of vowels and diphthongs. *Journal of the Acoustical Society of America* 92(4/2): 2444.
Warren, P., & Marslen-Wilson, W. 1987. Continuous uptake of acoustic cues in spoken word recognition, *Perception and Psychophysics* 41: 262-275.
Waterson, N. 1971. Child Phonology: A prosodic view. *Journal of Linguistics* 7: 179-211.
West, P. 1999. The extent of coarticulation of English liquids: An acoustic and articulatory study. In J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A.C. Bailey (eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* 3: 1901-1904.
Whalen, D.H. 1991. Subcategorical phonetic mismatches and lexical access. *Perception and Psychophysics* 50: 351-360.
Whalen, D.H., Best, C.T., & Irwin, J.R. 1997. Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics* 25: 501-528.
Wolf, C.G. 1978. Voicing cues in English final stops. *Journal of Phonetics* 6: 299-309.