Vowel Lengthening Effects in Natural Speech: 
Learning under sparse data and high variance

Rebecca L. Morley 
Ohio State University
morley.33@osu.edu

Vowel lengthening before voiced obstruents has been a much studied phenomenon for several decades. In spite of the large amount that has been written on the subject, however, there is neither consensus on the underlying cause of the phenomenon, nor complete consistency in experimental results. It has been argued that the effect is a part of universal phonetics (Chen 1970, Kluender et al. 1988, Crowther and Mann 1992), but that certain languages (such as English) show considerably larger differences than other languages (such as French). Vowel duration is often cited as the primary cue to word-final voicing in English (Chen 1970, Raphael 1972), and perceptual studies have consistently found robust categorical perception effects for the voiced/voiceless distinction, conditioned by preceding vowel length alone (Denes 1955, Lisker 1978, Port and Dalby 1982, Raphael 1972, Klatt 1976, Kluender et al. 1988, Luce and Charles-Luce 1985). However, production studies vary widely in the degree of vowel lengthening that they report (anywhere from 13 ms (Abdelli-Beruh 2004) to 150 ms (House, 1961), or in between (Peterson and Lehiste, 1960, Raphael, 1972, Chen, 1970, Crowther and Mann, 1992, Umeda, 1975)). Furthermore many studies find significant variability due to vowel, place of articulation of consonant, speaker, speaking rate, and prosodic position (Crowther and Mann 1992, Umeda 1975, House and Fairbanks 1953, Klatt 1976, Smith 2002, Mack 1982, Abdelli-Beruh 2004).

The results of the perceptual experiments suggest that the strength of the cue is enough to signal a contrastive relationship for speakers of American English. The variability of the production studies, however, raises the question of how such a robust relationship can be induced from the learner’s input. Furthermore, in the cited studies the stimuli were tokens from read speech recorded in a laboratory setting. An actual learner, of course, must learn phonetic detail from natural speech, which can be expected to contain more sources of variation – variation that could easily mask a relatively small difference in segment length.

Although it is generally assumed that learners must contend with considerable noise in the speech signal, few attempts have been made to quantify this variation. Modeling work usually assumes a symmetrically distributed random error term; experimental work attempts to carefully control the extent and type of variation; corpus studies typically report statistical significance, and only sometimes also provide effect sizes. The degree to which both linguistic and non-linguistic factors affect the distribution of phonetic variables has yet to be systematically studied. The current work uses measurements of vowel duration from the Buckeye Corpus of Conversational Speech (Pitt et al. 1997) to provide an estimate of the size of the vowel lengthening effect in natural speech; the findings are used to inform the search for a learning procedure capable of extracting the dependency from a background of competing cues.

Measured tokens were restricted to pre-pausal CVC words ending in voiced or voiceless stops; these words were expected to show the largest difference in vowel length based on coda consonant voicing. Taking all such tokens together, the duration distributions of pre-voiced and pre-voiceless vowels were statistically identical, indicating that without some sort of normalization, listeners would be unable to learn the lengthening cue. This result is not unexpected. There are a number of different factors that have been shown to affect vowel duration in a systematic way, most notably speaking rate, word frequency, intrinsic vowel length, and place of articulation of final stop. And it is possible to find a duration difference (56 ms) in the expected direction by restricting analysis to a sub-sample consisting of longer vowels in lower frequency words, at slower speaking rates.

However, data sparsity is a pervasive property of the entire sample. In the first place, there are few to no actual minimal pairs for final stop voicing, or even non-minimal word pairs sharing a vowel, or non-minimal pairs sharing a final place of articulation. For the above result, only 29 speakers out of 40 produce any tokens of the relevant type. Of these, there are only 14 speakers that have both
voiced and voiceless tokens. And of these, only 8 have distributions that are not completely overlapped. Additionally, no speaker produces tokens containing more than three different vowel types. For labial-final words across speakers, only the vowels /æ/ and /a/ have both voiced and voiceless tokens, and the /æ/ difference goes in the wrong direction. Similarly, for velar-final words, only /ɑ/ has both types. Comparable non-uniformity is found for other sub-groupings of the data.

Returning to the full sample, various trends can be observed when charting variation along a single dimension. As expected, vowel duration decreases as a function of increasing word frequency, and increasing speaking rate. Furthermore, as vowel duration decreases the difference between voiced and voiceless tokens also diminishes. This result can be attributed to a floor on absolute duration. As vowel durations get shorter they approach an absolute minimum. This is already happening for words in the second lowest quartile of frequencies (see Fig.1), and at the second slowest speech rate (also divided into quartiles) . It happens sooner for intrinsically short vowels (i,e,æ,a,o,u) than for intrinsically longer vowels (I,E,U).

Figure 1: Density Plots of vowel duration in pre-pausal CVStop words: Rows: Short vs Long vowels; Columns: log(frequency) quartiles.

These results are informative regarding the learnability of the vowel duration cue. Collapsing over all other factors, a 38% (70 ms) reduction in vowel duration can be attributed to frequency alone, and a 56% (131 ms) reduction to speaking rate. These are both larger than the voicing conditioned lengthening effect (less than or equal to 22% (56ms)). Another masking factor is the duration difference attributable to place of articulation (13% (23 ms) difference from velar to coronal to labial).

If this sample is representative of learners’ input generally, then these results have far-reaching implications for linguistic theory: from acquisition, to sociolinguistic variation, to language change. For a small phonetic dependency to be learned listeners must be able to effectively filter out the effects of speaking rate, word frequency, and place of articulation1. Due to the floor effect, however, the voicing-dependent lengthening cannot be extracted via decomposition into a series of summed Gaussians (cf. McMurray et al. 2011). Conversely, the statistical degeneracy of the input requires listeners to pool data in a way that potentially confounds results by factors like speaker, vowel, and word. On the one hand, the amount, and size, of confounding variation offer several opportunities for the type of ‘mis-learning’ posited to lead to sound change (cf.Ohala (1981), Blevins (2004)). On the other hand, the nature of the distributions suggests that the most likely outcome of mis-learning would be loss of the vowel length distinction altogether. Thus, theories that require generalization/learning to occur in particular ways potentially face a far more difficult problem than is revealed by models that assume normalized low-dimensionality input (cf. Sóskuthy (2013), Wedel (2006)).

---

1Children as young as two years old have been shown to produce a robust vowel length difference in monosyllabic CVC words, sometimes accompanied by deletion of final voiced stops ((Song et al., 2012, Ko, 2007)). This vowel length difference was also found in the parents’ child-directed speech (Song et al., 2012). If all learning and grammar change is assumed to be restricted to a critical development period with simplified, maximally informative input, then there may be no learning problem. This paper, however, adopts the hypothesis that phonetic and phonological properties of languages can change over the lifespan, and that shifts in distributional properties can lead to global grammar change.