An Explanation of the Voicing Effect: Limits on the expandability of lenis stops in English

1 Introduction


The voicing effect has been documented in many different languages and is considered by many to be a phonetic universal (e.g., Chen 1970, Kluender et al. 1988, Crowther and Mann 1992, Klatt 1976). Nevertheless, source of the phenomenon remains unknown. There is no consensus on what acoustic or articulatory properties give rise to the observed duration difference. Nor is the effect even consistently described as lengthening, but often as shortening before voiceless consonants, or “pre-fortis clipping” (Gimson 1970, Wells 1982). We hypothesize that this uncertainty persists due to an over-simplification of the empirical facts. In the first place, lengthening does not occur in all environments, or for all vowels (e.g., Peterson and Lehiste 1960, Umeda 1975, Hogan and Rozsypal 1980, Crystal and House 1982, De Jong 2004). Neither does preceding vowel duration robustly cue the voicing contrast at all speaking rates and for all types of speech (Umeda 1975, Port 1976, Crystal and House 1988, Smith 2002, Ko 2018). Other cues, such as voicing itself, or a period of aspiration following a stop release, may supersede vowel duration in perception (e.g., O’Kane, 1978, Raphael, 1981, Wardrip-Fruin, 1982, Revoile et al., 1982, Repp and Williams, 1985). Finally, lengthening can occur even when voiced obstruents are not actually voiced (i.e., in the absence of vocal fold vibration) (e.g., Sharf 1964, Walsh and Parker 1981, Keating 1984). This set of results shows that it is not accurate to treat vowel lengthening as a straightforward phonetic implementation rule.

In this paper we provide a more nuanced view of the voicing effect, highlighting less-cited works, and integrating results from other areas of speech perception and production. We will argue that the voicing effect is not the direct result of voicing at all (whether phonetic or phonological). Rather, vowel duration

\footnote{The apparent paradox involved in describing +voice segments as voiceless, or devoiced, has been traditionally solved by differentiating between a phonetic voicing feature, and a phonological voicing feature. The phonological feature is to be thought of as a completely abstract label which can be transformed through various rules to either a phonetically voiced or phonetically voiceless realization. The question of the best way to characterize laryngeal contrasts phonologically will be taken up in Section 7.}
differences are the result of a type of compensatory lengthening that is only appreciable at unusually long durations. This compensation occurs because voiced obstruents in English have an inherently shallow expansion curve, and a specified limit on how long they can become. External pressures to increase duration, such as phrase-final lengthening, must therefore shift to more expandable segments within the word, such as the preceding vowel. It should be noted that this type of compensatory expansion is different from what has been proposed, and subsequently rejected, by researchers such as Chen (1970), Port and Dalby (1982), Keating (1985). Our Expandability Hypothesis states 1) that lengthening processes apply to all segments of a word, but to inherently different degrees, 2) that segments may be underlyingly specified as short, such that they cannot lengthen past a certain maximum, and 3) that in the case that a given segment reaches its expansion maximum, other segments in the word will expand more in compensation.

We present evidence from a production experiment that further supports this view, showing that voiced obstruents in English lengthen less and less as speaking rate decreases, while voiceless obstruents continue to lengthen in a relatively linear fashion. Thus, as speaking rate slows, the difference in obstruent duration (across a set of minimal pairs) increases. The size of this difference is shown to be inversely correlated with difference in preceding vowel duration, exhibiting the predicted relationship.

The Expandability Hypothesis will also be used to account for the environments in which the voicing effect is not observed. Using a corpus of conversational speech, we corroborate the finding that the voicing effect is largely restricted to laboratory speech, and phrase-final contexts. This result is usually attributed to reductive forces that shorten and weaken segments that are produced rapidly or in non-prominent contexts. We posit that the directionality is actually reversed, i.e., that an effect is emergent only at unusually long absolute durations. Like vowel duration differences, differences in duration between voiceless and voiced obstruents only emerge in this range, and they are inversely correlated with vowel duration differences. Our explanation of the voicing effect turns out to have direct consequences for one of the most fundamental relationships in phonological theory, that between an underlying phoneme and its physically realized allophones.

The paper is organized as follows. In Section 2 we review a portion of the literature on the voicing effect. Section 3 contains the corpus study on American English, the results of which we model in Section 4 under the Expandability Hypothesis. The results of the production study are presented in Section 5. In Section 6 we elaborate on the Expandability Hypothesis, subjecting it to further tests, and providing an explanation for the perceptual side of the voicing effect. We summarize and conclude in Section 7, where we discuss the implications of the present work for theories of phonological contrast.

2 Background

It has been posited that the voicing effect is present in all languages to some degree, although that degree may vary considerably (Chen 1970, Kluender et al. 1988, Crowther and Mann 1992, Klatt 1976). The voicing effect may be minimal in at least Arabic (De Jong and Zawaydeh, 2002) and Catalan (Cuartero Torres, 2002), and possibly enhanced or exaggerated, in English, and possibly French, Spanish, Norwegian, Korean and Russian (House and Fairbanks 1953, Abdelli-Beruh 2004, Hillenbrand et al. 1984, Mack 1982). It is generally agreed that English (both British and American varieties) exhibits one of the strongest voicing effects.

Temporal compensation as an explanation for the voicing effect was proposed as early as Kozhevnikov and Chistovich (1965). Catford (1977) also suggested that vowel duration differences mirror the differences in consonant duration between voiced and voiceless obstruents. This temporal compensation was assumed to derive from a uniform rate of syllable production such that pre-voiced vowels would be longer than pre-voiceless by exactly the same amount as voiceless obstruents were longer than voiced. Because he found that syllable duration varied more than vowel duration in spoken CVC versus CVCC syllables, Chen ultimately rejected this explanation (Keating (1984) found a similar discrepancy in Polish). However, the current proposal does not require that all syllables be carefully matched for total duration. Compensation takes place within, not across, syllables, and is based on external lengthening pressures, such as changes in speech rate.
effects, with pre-voiceless vowels typically ranging between 60% to 70% of the length of their pre-voiced counterparts (Chen 1970, Harris and Umeda 1974, Mack 1982). English listeners also exhibit a remarkably consistent categorical perception effect based on vowel duration (e.g., Raphael, Crowther and Mann, Klatt, Hillenbrand et al., Denes). The fact that other cues to voicing have been shown to be unnecessary in such experiments, or, in fact, to be trumped by vowel duration has led to the claim that vowel duration is not only a sufficient cue to final obstruent voicing in English, but the primary cue (e.g., Raphael 1972, Luce and Charles-Luce 1985). Because stops are often unreleased in word-final position, it is plausible that a sound change has occurred (or is underway) in which the contrastive relationship between words like “bad” and “bat” has shifted away from the final obstruent itself, to be expressed in the duration of the preceding vowel.

While the above characterization of the voicing effect is largely undisputed, there is less consensus about the origin of the effect. The majority of the literature seems to assume an articulatory source without further explanation (Raphael 1975, House 1961, Ohala 1983). Belasco (1958), however, speculates that there is trade-off in the force of production between the vowel and coda consonant of a syllable. When the consonant requires more energy, or effort, the vowel is altered to require less, and vice versa. Thus, voiceless stops, involving forceful release and aspiration, condition shorter vowels, which require less energy. Similarly, Delattre (1962) argues that anticipation of a effortful articulation should shorten the preceding vowel. Moreton (2004) and Schwartz (2010) also offer proposals that are based on the more forceful articulation of voiceless stops which, they hypothesize, spreads to the preceding vowel. It has also been claimed that non-spontaneous voicing under reduced pressure requires careful movements of the vocal cords and thus proceeds slowly; that the transition from vowel to voiceless obstruent is more rapid than the transition to voiced (Chen 1970); and that the glottal opening gesture tends to occur a bit earlier for voiceless final consonants to ensure that there is no residual voicing on the consonant (Klatt 1976, Lisker 1974).

On the auditory side, Lisker (1957) suggests that longer vowels occurring with shorter voiced obstruents, and shorter vowels with longer voiceless obstruents, is an enhancement effect, reinforcing the length distinction (see also Jessen 2001, and Kluender et al. 1988). Javkin (1977) posits that vowels are consistently perceived as longer before voiced than voiceless consonants because listeners mis-attribute the glottal pulsing at the beginning of the consonant to the end of the vowel.

None of the foregoing proposals has emerged as a standard in the literature, and many are inconsistent with empirical data. Those of the articulation-based explanations that rely on actual vocal fold vibration cannot account for the fact that voicing effects occur even when voiced obstruents are phonetically devoiced (e.g., Walsh and Parker 1981, Chen 1970, Fox and Terbeek 1977). A universal basis for the effect is also called into question by the apparent absence of a lengthening effect in certain languages (Flege 1979, Hillenbrand et al. 1984, Keating 1979, 1985), or an effect in the opposite direction (Han 1994, Idemaru and Guion 2008).

3 A Corpus Study

Even in English, with one of the most robust voicing effects measured, durational differences are not found in all contexts. Production studies typically consist of either word lists or brief sentences read by participants in a laboratory setting. In sentence contexts, the target words are often in absolute final position. Such words also tend to be monosyllabic, which entails that the target vowel receives primary stress. When some or all of these factors are varied, the voicing effect can disappear.

Data from the Buckeye Corpus (Pitt et al. 2007) were analyzed in part to corroborate the finding that the voicing effect is reduced in conversational speech, as well as to investigate the reason for this reduction. The corpus consists of segmented, and transcribed sound files from different speakers, collected during individual interviews lasting approximately one hour. From this corpus we extracted all monosyllabic words of the form (C)onsonant-(V)owel-(C)onsonant ending with one of the following obstruents: voiced (d,b,g,z,f,v)
or voiceless (t,p,k,s,f,l). No nasalized or rhotacized vowels were included, to be sure that each word had only three underlying segments. Only tokens that were both phonemically and phonetically CVCs were included. For example, tokens of “past” realized as [pæs], and tokens of “allowed” realized as [ləud] were both excluded. Because there were no words ending in voiced dental fricatives, those ending in voiceless dental fricatives were also removed. The vowel /oI/ was also excluded for reasons of data sparsity. 20.3% of the stops in the remaining data were transcribed as glottalized (tq), which could represent a glottal stop or unreleased stop with glottalization on the vowel, but less than 1% were underlyingly voiced, so all such tokens were removed from analysis.

The analysis of the remaining data included the following factors, each of which is known to affect segment duration. Because the analysis was limited to CVC words, stress and word length are not included.

- **INHERENT VOWEL DURATION**: Tense and lax vowels in English are differentiated in part by duration. /I, e, ə, o/ all lax vowels, are reliably shorter than their tense counterparts (e.g., Peterson and Lehiste (1960), Klatt (1976), Stevens and House (1963)). /æ/, although technically lax, has much longer durations than any other lax vowel (e.g., Hillenbrand et al. 2000, Crystal and House 1988), and is actually a diphthong in some dialects, thus it is grouped with other inherently long (tense) vowels.

- **CONSONANT MANNER OF ARTICULATION**: Like vowels, consonants possess different inherent lengths. Furthermore, stops may be reduced in certain contexts to flaps, which consist of a very brief tongue tip gesture against the roof of the mouth. Because the durations of each have their own distributional characteristics, and may have different effects on the preceding vowel, each manner (fricative, stop, and flap) was analyzed in a separate statistical model.

- **SPEAKING RATE DEVIATION**: In order to estimate speaking rate for individual words, the average durations for each segment, across all word tokens, for each speaker, were used as the baseline. How much faster or slower a given word was than its mean value for a given speaker was used as a proxy for rate deviation. The z-scored difference between the mean and the observed duration for each of the three segments within a given word (onset, nucleus and coda) were averaged to obtain each word’s total deviation. Because this is actually a duration measure, a positive difference indicates that the individual segments within the word are generally longer than their average durations, and thus that the speaking rate is slower than average.

- **WORD FREQUENCY**: More frequently used words generally have shorter durations than less frequently used words, and both vowels and consonants within those words are affected (e.g., Jurafsky et al. (2001), Fidelholtz (1975), Fosler-Lussier and Morgan (1999), Hooper (1976), Pluymaekers et al. (2005)). Function words, generally the most frequent and the most contextually predictable words, tend to be inherently shorter than content words (Bell et al., 2009, Umeda, 1975). In order to keep both content and function words in the analysis, word frequencies were given as Zipf scores \( \log_{10}(\text{Frequency}) \), based on the frequency counts per million given in the SUBTLEX corpus (Van Heuven et al., 2014).

- **PHRASAL POSITION**: Prosodic boundaries have the effect of lengthening segments adjacent to those boundaries. The greater the number of nested phrases marked by the boundary, the greater the degree of lengthening, and the further its spread (Oller, 1973, Wightman et al., 1992, Fougeron and Keating, 1997, Byrd and Saltzman, 2003). Because the Buckeye Corpus is not annotated for syntactic boundaries, tokens were classified only as pre-pausal or non-pre-pausal, based on the end of a transcribed utterance.
3.1 A Graphical View of the Data

It is expected that speech collected from different individuals, and different conversations, without controlling for linguistic variables, will show a high degree of variability. Conversational styles of speech are also expected to exhibit considerable reduction in the realization of individual words, some component sounds of which may be entirely missing (e.g., (Harris and Umeda, 1974, Johnson, 2004, Jurafsky et al., 1998)). Nevertheless, for listeners to use a given feature in discriminating contrastive sound units, there must be some way for that feature to be extracted from the actual speech signal. To get a sense for the size of the voicing effect relative to other factors affecting vowel duration, we present here a number of different views of the raw data prior to analysis.

In Fig. 1 vowel durations for the entire set of CVC word tokens are plotted as a function of the voicing feature of the final obstruent. The density plot on the right suggests that there is a very small effect of voicing at the longest durations. However, the actual counts given in the left panel show that there are never more voiced than voiceless tokens at any duration. This is due to the fact that there are considerably more (phonemically) voiceless than voiced coda obstruents in the data (almost twice as many, although there are more voiced than voiceless tokens for fricatives. See Appendix). Vowels preceding voiceless obstruents have a slightly longer mode than pre-voiced, and at the very longest durations (>175 ms.), the relative proportion of the pre-voiced distribution is larger than the pre-voiceless, but the two distributions are completely overlapping.

However, when the data are grouped with respect to some of these other variables that affect duration, such as speaking rate, vowel class, and word frequency, a suggestive pattern starts to emerge. In Figure 2, it appears that the voicing effect actually goes in the opposite direction from what is predicted, especially at faster speaking rates. However, if we examine the distributions of vowels before fricatives and full stops as speaking rate slows (top to bottom), the voiced sub-distribution can be seen to be increase in duration to a much larger degree than the voiceless. The pre-voiced distribution moves closer and closer to the pre-voiceless, and begins to overtake it in the slowest 25% of the speaking rate distribution (Q4)). Thus, the longest (slowest) tokens are closer to the predicted effect than the shorter (fastest).

Figure 1: All CVC vowel durations
This trend is more pronounced, the longer the tokens. In Fig.3b only the very longest subset of tokens are plotted: content words of below median word frequency, containing inherently long vowels (i.e.,u,a,o,o,æ,ui,ao), produced at slower than average speaking rates, in pre-pausal contexts. Although still small, the expected effect of voicing appears to emerge. In comparing Figs.3a and 3b, the effect of absolute duration seems to be cumulative. This subset of data differs only in also containing tokens produced in non-pre-pausal contexts, resulting in a slightly shorter distribution, and what is apparently a slightly smaller voicing effect.
This set of plots suggests the following interpretation: the voicing effect is small, and masked by other factors, but increases with increased segment duration. Below, we provide the corresponding plots for obstruent durations in the same tokens, where we will see that a similar relationship holds. Voiced obstruents are typically shorter than voiceless obstruents, in the same way that preceding vowels are longer for the former than the latter (e.g., Klatt 1976, Umeda 1975, Miller and Volaitis 1989, Chen 1970, Luce and Charles-Luce 1985). In conversational speech, however, this duration difference is elusive.

Figure 4 is the counterpart of Figure 1 above, showing the raw consonant durations as both counts and probability densities. This figure shows that the asymmetry between the number of voiceless and voiced tokens is due entirely to the stop-final words as fricatives have more voiced than voiceless tokens. In both cases, however, the distributions have a great deal of overlap, with voiceless tokens only showing a slight tendency toward longer durations.
Grouping the data by speaking rate deviation quartiles reveals a similar trend as that observed for vowel duration. For both fricatives and stops, slower speaking rates result in more separation between the two distributions. This time it appears to be the voiced sub-distribution that is relatively invariant, while voiceless obstruents increase in duration.

Figure 4: Obstruent durations by fricative and stop (flaps excluded).
The difference in duration is more consistent for consonants than vowels, but the same dependence on duration is observable. The longest tokens show the largest effect, and that effect appears to be cumulative. The fact that duration differences between final obstruents mirror duration differences between the preceding vowel is suggestive of a co-varying relationship based on temporal compensation. Statistical tests confirm this hypothesis.
Figure 6: Consonant duration for the subset of longest tokens: content words below the median speaking rate, below the median word frequency, and containing the inherently longer vowels: (i,e,u,o,a,r,a,I,a,o).

3.2 Statistical Analysis

Vowel duration was modeled separately for flaps, fricatives and full stops. For each model, the factors listed above were included as main effects. In addition, the effect of voicing was directly tested, whether determined solely by the citation form of the word (phonemic voicing), or the actual presence of vocal fold vibration in the speech signal (phonetic voicing). All factors were sum-coded, so that the average of all levels were considered in the effects at each level.

- **PHONETIC VOICING**: transcription provided by corpus annotators (-1=voiceless; 1=voiced)
- **PHONEMIC VOICING**: voicing category of the phoneme in the citation form of the word (-1=voiceless; 1=voiced)
- **CONSONANT DURATION**: duration of the final consonant as measured by corpus annotators

Random intercepts for word and speaker were included in all models. Random slopes were added for all factors when doing so significantly increased model fit. Continuous numerical variables were log-transformed, where appropriate, and mean-centered to approximate a normal distribution with a mean of zero.

All statistics were performed using the lme4 package in R. Linear mixed effects models were run using the function lmer, fit by REML. T-tests used Satterthwaite’s method, and the lmerTest function was used to
obtain estimated p-values. Interactions that did not reach a significant estimated t-value were individually removed from the model, and goodness-of-fit model comparisons confirmed that including these factors did not improve model fit. Three-way interactions were avoided for reasons of interpretability as well as model convergence.

The results were similar for all three models. Full stops and fricatives showed the same significant effects, and similar relative effect sizes, except for the final interaction term. The model for flaps was simplified due to lack of effect for some factors, but the remaining factors patterned similarly to that of full stops. Because of these similarities, and for reasons of space, only the results for full stops are reported here.

Table 1: Vowel Duration Model for Full Stops

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>estimated df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.25E-02</td>
<td>4.06E-02</td>
<td>1.84E+02</td>
<td>-0.554</td>
<td>0.580</td>
</tr>
<tr>
<td>speaking rate deviation</td>
<td>1.10E+00</td>
<td>2.29E-02</td>
<td>1.30E+02</td>
<td>48.292</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>phonetic voicing</td>
<td>-1.58E-02</td>
<td>6.96E-03</td>
<td>1.17E+04</td>
<td>-2.267</td>
<td>0.023</td>
</tr>
<tr>
<td>word frequency</td>
<td>-6.32E-02</td>
<td>2.17E-02</td>
<td>2.47E+02</td>
<td>-2.916</td>
<td>&lt;.005</td>
</tr>
<tr>
<td>Vowel Class</td>
<td>-1.81E-01</td>
<td>2.04E-02</td>
<td>2.47E+02</td>
<td>-8.861</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>phonemic voicing</td>
<td>-1.09E-01</td>
<td>3.50E-02</td>
<td>1.95E+02</td>
<td>-3.122</td>
<td>&lt;.005</td>
</tr>
<tr>
<td>Consonant Duration</td>
<td>-2.91E-01</td>
<td>9.41E-03</td>
<td>1.18E+04</td>
<td>-30.946</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Phrasal Position</td>
<td>9.06E-02</td>
<td>3.86E-03</td>
<td>1.17E+04</td>
<td>23.434</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>speaking rate deviation:Vowel Class</td>
<td>9.18E-02</td>
<td>2.00E-02</td>
<td>8.87E+01</td>
<td>4.581</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>speaking rate deviation:phonemic voicing</td>
<td>1.78E-01</td>
<td>2.26E-02</td>
<td>1.28E+02</td>
<td>7.851</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>speaking rate deviation:Consonant Duration</td>
<td>6.48E-02</td>
<td>1.21E-02</td>
<td>9.44E+03</td>
<td>5.35</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>word frequency:phonemic voicing</td>
<td>-4.53E-02</td>
<td>2.01E-02</td>
<td>2.40E+02</td>
<td>-2.251</td>
<td>0.025</td>
</tr>
<tr>
<td>phonemic voicing:Consonant Duration</td>
<td>-6.57E-02</td>
<td>9.06E-03</td>
<td>1.15E+04</td>
<td>-7.248</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Many of the model factors are closely related, but reported correlations among main effects are surprisingly low. Speaking rate deviation, for example, is a duration-based measure which takes into account the durations of the final consonant and the vowel, but is only correlated at r=-0.28 with the intercept and r=0.24 with the consonant duration. Phonemic voicing and phonetic voicing are only correlated at r=0.097 in the analysis shown. Thus, each factor accounts for a sufficiently unique effect.

For each variable, the average value of its levels (if a factor), or of its range of values (if a continuous numerical variable) was the baseline for analysis. This allows us to conceptualize the results in a way that is similar to ANOVA, where each effect is an adjustment to the average value for the model. For example, the effect of Vowel Length is determined by whether short vowels are different from the average of long and short vowels, while all other effects are held constant at their average.

As expected, there was a significant main effect of speaking rate deviation. See Table 1. Longer vowel durations were found at slower speaking rates. However, the negative effect of consonant duration indicates that, for words of average speaking rate (and average frequency), vowels were longer when final consonants were shorter. Similarly, word frequency had a significant negative effect on vowel duration, such that more frequent words had shorter vowel durations. As predicted, lax vowels (minus ë) were also shorter than tense vowels, and vowels in pre-pausal words were considerably longer than in non-prepausal. However, both phonetic and phonemic voicing actually resulted in significantly shorter vowel durations, the opposite of what is predicted by the voicing effect.

However, the interaction of speaking rate deviation and phonemic voicing shows that the expected directionality emerges at slower speaking rates. A positive interaction of consonant duration and speaking rate indicates that at slower speaking rates, consonants increase in duration, and that consonant duration is less predictive of vowel duration at these slower rates. The interaction of frequency and voicing shows that the
negative effect of frequency (higher frequency = shorter V duration) is increased in voiced tokens, so that in lower frequency words, we get even greater increases in vowel duration. An additional interaction term of speaking rate and vowel length shows a positive adjustment to short vowels at slower speaking rates, which means that they increase in duration to be more similar to long vowels in slow speech. Finally, the significant interaction between phonemic voicing and consonant duration (for stops but not fricatives) indicates that shorter consonant durations are even more predictive of longer vowel durations when the consonant is the voiced member of the contrast. This may simply be because there is less variability in the duration of voiced stops.

The lack of a main effect for voicing in the predicted direction indicates that this cue is not extractable from normal speech in the aggregate, even when phrase-final position and other factors are partialled out. Part of the reason for this is the unbalanced nature of the data (to which we attribute a spurious negative correlation of voicing and vowel duration). To the degree that this type of speech data is typical, however, the listener/learner faces a similar analysis problem. However, interactions involving voicing indicate that the effect is not the same for all token durations. While the correlation actually goes in the wrong direction for the shorter word tokens, there appears to be a cross-over point in terms of absolute token duration where the correlation first disappears, on its way to a positive correlation in the longest subset of the data.

The results of analyzing only the subset of the longest tokens displayed in Figure 3b are given in Table 2. As before, linear mixed effects models of vowel duration for words ending in fricatives and stops were constructed (flaps were excluded because they should not occur in pre-pausal position), but again, due to highly similar results, only the stops are reported. The model is simplified because there were fewer data points, and less variability in the factors. Word frequency was no longer significant. The large effect of pre-pausal position on duration washes out differences in speaking rate deviation – all tokens are essentially slower than average. No interactions are included for similar reasons.

Table 2: Vowel Duration Model for Full Stops in longest tokens

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>estimated df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.66267</td>
<td>0.04125</td>
<td>105.06663</td>
<td>16.066</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Vocal Class</td>
<td>-0.0678</td>
<td>0.03412</td>
<td>121.06521</td>
<td>-1.987</td>
<td>0.049</td>
</tr>
<tr>
<td>phonemic voicing</td>
<td>0.09313</td>
<td>0.03376</td>
<td>109.00637</td>
<td>2.758</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Consonant Duration</td>
<td>0.18751</td>
<td>0.02447</td>
<td>1023.41315</td>
<td>7.664</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Because all word tokens are longer than average, the effect of consonant duration on vowel duration now goes in the opposite direction. Longer consonants predict longer vowels because both are subject to the same lengthening effect of pre-pausal position. However, there is now a significant main effect of phonemic voicing in the predicted direction. The emergence of a main effect at longer durations, we argue, is epiphenomenal of a difference in consonant durations. Shorter consonants predict longer vowels, as can be seen when the full set of data are analyzed, with a comprehensive range of vowel and consonant durations. Pre-pausal lengthening swamps the inverse correlation of consonant and vowel duration in the smaller set of data, but the voicing variable allows it to be uncovered by grouping shorter things (voiced obstruents) separately from longer things (voiceless obstruents). We hypothesize that voicing is positively correlated with vowel duration in this model because it is negatively correlated with consonant duration.

3We confirm that voicing is significantly negatively correlated with obstruent duration (p<.001) in a separate model with obstruent duration as the dependent variable, and fixed effects of speaking rate deviation, consonant voicing, vowel duration, and phrasal position. It is worth noting that, even at average speaking rates, consonant voicing is a significant predictor of consonant duration.
3.3 Summary & Discussion Of Results

These results show that, in conversational speech, we do not see the expected effect of voicing on vowel duration at average speaking rates, for average word frequencies, or phrase-internally. We only begin to see an effect emerge at slower speaking rates, in lower frequency words, for longer vowels, and in phrase-final position, as shown in interaction terms, and a separate analysis of phrase-final tokens. Tanner et al. (2019) also analyzed data from the Buckeye Corpus, although only utterance-final tokens, and reported that the voicing effect was smaller in faster words, at higher frequencies, and for inherently shorter vowels (high versus non-high). However, they interpreted this result as due to the reduction found in conversational speech. Our analysis, on the other hand, shows that consonant duration is a more consistent (negative) predictor of vowel duration. The contexts in which voicing emerged as significant (in the right direction) can all be explained as effects of consonant duration, a factor that is significantly correlated with voicing class. Thus these results support our hypothesis that the voicing effect is actually a duration effect, driven by compensation between the vowel and the coda obstruent.

4 The Expandability Hypothesis: Modeling the Corpus Data

Although we find little to no voicing effect in the corpus data, laboratory production studies typically report large durational differences, on the order of a 2:3 ratio, between pre-voiced and pre-voiceless vowel tokens. For example, Peterson and Lehiste (1960) report an average vowel duration of approximately 297 ms preceding voiced obstruents. Mack (1982) and House (1961) report similar values, while Luce and Charles-Luce (1985) find pre-voiced durations averaging 177 ms, and Umeda (1975) reports an average of 210 ms in pre-pausal contexts.

For the majority of the vowels in the Buckeye Corpus, on the other hand, durations this long are rare. Only the uppermost tail of the distribution reaches values over 200 ms, with over 93% of the distribution falling below this duration. Even among phonologically long vowels, only 9% of tokens reach 200 ms or longer. Median vowel duration for the subset of tokens plotted in Figure 3b is 200 ms, while median duration for all tokens is only 83 ms. The reported vowel durations seem to indicate that laboratory speech falls in the very upper range of conversational speech (cf. Gahl et al. 2012).

If differences are indeed only observed at unusually long vowel durations, a general vowel lengthening rule does not capture the phenomenon. However, if lengthening, of all kinds, is modeled as expansion of a word or syllable unit, then an explanation can be found in the fact that different segments appear to have different inherent elasticity, with voiced obstruents being among the least elastic (Cooper et al. 1985, Lehiste 1975, Oller 1973, Klatt 1974, Umeda 1975, Miller 1981).

Qualitatively, obstruent duration distributions are strikingly similar to vowel duration distributions. In the corpus data the voiced and voiceless tokens are distinguished only in the upper range, but in laboratory studies they are significantly different. Chen (1970), for example, finds that there is a 2:3 ratio of closure duration between voiced and voiceless stops when they are released. Luce and Charles-Luce (1985) also find that closure duration fails to reliably differentiate the voicing contrast on stops in most environments (e.g., across sentence position, place of articulation, and preceding vowel quality). However, in sentence-final position (averaging across the two tense vowels, and all places of articulation), they find a 3:4 ratio between voiced and voiceless stops.

These similarities are accounted for if vowel duration differences derive from consonant duration differences. To demonstrate that compensation between vowel and coda can capture these distributional properties, we created a simple model of speaking rate-based expansion. We take the rhyme (vowel nucleus and coda consonant) to be the linguistic unit that is expanded or compressed as a whole under changes in rate. An external expansion force causes all segments within the rhyme to lengthen, but affects the vowel and
consonant to different degrees (for the purposes of this model we ignore the potential contribution of the syllable onset).

In this model, as expansion force increases (speaking rate slows), and the rhyme grows longer, the coda obstruent also lengthens, at first relatively linearly, but asymptotically approaching a maximum duration. Voiced and voiceless stops are both modeled with logistic expansion functions, but with different parameter settings. See Fig. 7. Rhyme duration, for simplicity, is taken to be a linear function of the expansion force (dark solid line). To implement compensation, vowel duration is calculated by subtracting the coda duration from the target rhyme duration at a given speaking rate\(^4\). Thus, as coda lengthening diminishes, vowel lengthening increases. In reality, the vowel presumably cannot lengthen indefinitely, but over the range shown here it is assumed to fully compensate in expansion degree.

We generated a duration distribution for vowels and codas (half voiced and half voiceless) by sampling from a Normal distribution of expansion values shown graphically by the vertical lines superimposed on Fig. 7. The majority of the expansion distribution falls in the region where the durations of voiced and voiceless obstruents are very similar, resulting in little difference in vowel duration, as was found for most of the corpus data. Only the tail of the expansion distribution includes a more divergent region, accounting for the observed difference in longer word tokens. Figures 8a and 8b show the obstruent and vowel duration distributions, respectively, produced by the simulation.

\[ R = R_0 + bE, \]
\[ O = \frac{A}{1 + Be^{kE}}. \]
\[ V(E) = R(E) - O(E). \]

---

\(^4\)The rhyme is taken to expand as a linear function of the expansion force \((E): R = R_0 + bE\), where \(E\) is a Normally distributed random variable. The expansion function for voiced and voiceless obstruents has the same general form, but differing parameters \((A, B, k): O = \frac{A}{1 + Be^{kE}}\). Vowel duration at \(E\) is calculated by taking the difference between target rhyme duration and obstruent duration: \(V(E) = R(E) - O(E)\).
The distributions approximate what is observed in the corpus data: little differentiation between voiced and voiceless tokens except at the very upper duration range. The difference is also larger between obstruents than between vowels because the rhyme duration increases more for a given increase in expansion force, than the difference between the voiced and voiceless obstruent durations does. In other words, pre-voiceless vowels, although shorter than pre-voiced vowels produced at the same speaking rate, are longer than those pre-voiced vowels when the rate is slightly slower. This results in a difference only at the absolute minimum speaking rate.

5 A Production Study

We take the corpus results, in conjunction with the production literature as a whole, to provide strong preliminary support for the Expandability Hypothesis. In this section we report the results of a production experiment that confirms more directly the following prediction: that voiced obstruents will show diminished lengthening relative to voiceless obstruents as speaking rate decreases, and that the difference in preceding vowel duration will mirror the difference between the obstruent durations.

5.1 Procedure

Participants consisted of 34 undergraduate students at The Ohio State University who were given course credit for completing the experiment, which took about half an hour to run. Participants were seated in front of a computer monitor inside a sound-attenuated booth. Continuous audio was recorded from a desktop microphone using the sound editing software Audacity\(^5\). Participants were instructed that they would be asked to speak into the microphone in response to prompts on the computer screen.

The experiment began with a practice block to acclimate participants to the experimental task, and the different repetition rates involved. Prior to the start of the practice block, participants were given the

\(^5\)Available at http://audacity.sourceforge.net.
following instructions: “A + sign will appear on the screen. It will be black to begin with, then will change to red, and keep alternating. Your job is to repeat the word on the screen every time + changes color. Try to use the entire time that the + does NOT change color to say the word. Keep going until the flashing stops. Press any key when you are ready to practice with the word “lab”."

For the first trial, participants saw the following text: “Here’s the fastest speed”, followed, 1.5 seconds later by the appearance of the word “lab”. The word stayed on the screen as the “+” immediately appeared and began to change color. Color changes occurred 8 times. At the end of the 8 cycles, a new trial began. For each of the rate changes, participants were alerted to the change with the following text: “A little slower”, followed, after 2 seconds, by the appearance of the same word. There were 5 different rates, corresponding to the time it took for the plus sign to change from black to red: 350, 550, 750, 950, and 1150 ms. The slowest and fastest rates were chosen to be as extreme as possible while still being within the ability of participants to match6.

At the end of the practice session, after participants had gone through all 5 rates, they were told that they could begin the experiment whenever they were ready. The experimental trials were identical to the practice except that the rates went in order from slowest to fastest. Participants were presented with the following text: “You will begin with the SLOWEST speed, and the flashing will become faster”. Subsequently, each rate change was signaled with: “The speaking rate will now speed up a bit”. Participants saw a total of 16 words, each presented once. Trials were blocked by word, such that participants experienced all rates before beginning with a new word. At the end of a given block participants were alerted that “The next item will now appear on the screen”, followed by a pause of 2 seconds. Word order was randomized across participants. The 16 words corresponded to 8 minimal pairs: peas/peace, feet/feed, thief/thieve, lobe/lope, coat/code, doze/dose, half/have, tap/tab. Due to time constraints, only data for the pairs feet/feed, and thief/thieve were analyzed for this paper.

5.2 Data Selection and Annotation

Each participant produced approximately 8 tokens of each word at each rates. Of those 8, a single representative token from the center of the group was selected and measured. Because each token was surrounded by other tokens at the same repetition rate, it was possible to segment both the closure and the release interval for each stop. In the case of two adjacent stops at faster repetition rates, such as /t.k/ as in ’coat-coat’, the first stop was frequently unreleased, so only the closure was marked for the final stop. At slower repetition rates, however, even these had individual closures and releases. Occasionally the voiced stops and fricatives at the slower repetition rates were produced with a final epenthetic schwa. There were 29 such tokens out of a total of 680. Final schwa duration was not included in the total duration measurement, which was the sum of closure and aspiration duration for stops. In order to conduct a paired analysis, any missing values led to both members of the pair being discarded. There were only 4 such pairs.

The data were distributed among three undergraduate research assistants. One of the authors and two of the RAs then re-measured a subset of the data produced by the other two annotators. Discrepancies between any two raters were discussed as a group to establish shared criteria for ambiguous tokens. The two RAs then individually reviewed their previous measurements and made adjustments where their original segmentation did not meet the discussed criteria. The same two RAs each also re-measured half the data of the third RA who had left the lab at that point. Praat (Boersma and Weenink 2009) was used for segmentation and annotation.

Note that the fastest change time, 350 ms, is quite long in terms of vowel duration alone, as measured in the Buckeye Corpus. This presumably reflects the fact that coarticulation and reduction, along with prosodic organization, allow for individual segments to be much shorter in normal speech than in a laboratory word-repetition task.
5.3 Results

Figure 9: Closure duration, VOT and Total duration for final stops as a function of repetition rate (decreasing from left to right). Means and standard error bars.

The left panel of Fig. 9 confirms that closure duration for final voiced stops varied relatively little across repetition rates (given as a number between 1 and 5, where 5 was the fastest, and 1 the slowest). However, most stops were also produced with a period of aspiration (VOT). Taken together, voiced stops show a clear increase in duration that appears to plateau only at the slowest rate. For voiceless stops, closure duration shows a more linear, and steeper increase trajectory, patterning very closely with VOT. Because both duration measures show dependence on rate, total duration was used as the factor for testing the Expandability Hypothesis.

In Figure 10 vowel duration (filled circles) and total obstruent duration (filled triangles), are displayed together for each of the four words. Visual inspection shows that larger vowel durations were reached by the voiced member of each minimal pair (panels on right), while larger obstruent durations were reached by the voiceless member (panels on left). There is also a larger difference between obstruent and vowel durations for voiced-final tokens across all repetition rates, and that difference increases with decreasing repetition rate. These trends are highlighted in Figure 11, where the difference in duration between voiced and voiceless members of a given minimal pair, is plotted by rate. The negative correlation between the difference in duration of voiced and voiceless obstruents, and the difference in duration of their preceding vowels, is clearly demonstrated in this figure.
The statistical analysis confirms the significant correlation between the two difference measures. Because consonant duration difference is highly correlated with repetition rate, vowel duration difference shows a significant effect of both repetition rate and consonant duration difference. There was no significant
difference in fit between the two one-factor models. For the single-factor linear mixed-effects model of vowel duration difference, consonant duration difference was significantly negatively correlated (t-value = -4.72, df = 327, p < .001). For the single-factor linear mixed-effects model of consonant duration difference, repetition rate was significantly negatively correlated (t-value = -8.77, df = 298, p < .001). Random effects for participant and word type were included in both models, but random slopes did not improve model fit and were dropped.

These results strongly support the Expandability Hypothesis. Firstly, we confirm the predicted difference in lengthening between voiced and voiceless obstruents in coda position, which parallels what has been repeatedly found for obstruents in initial and medial position (Port 1976, 1981, Miller and Baer 1983, Miller and Volaitis 1989, Volaitis and Miller 1992). We also confirm that the difference in durations for both consonants and vowels is largest at the slowest repetition rate, replicating Ko (2018) who, in a similar study, tested three speaking rates (fast, normal and slow). However, the normal and fast speed conditions in that study were largely the same, and duration differences between voiced and voiceless pairs were not compared. Pairing consonant duration differences with vowel duration differences at each speaking rate, with rates controlled via metronome, confirms our prediction that the strength of the voicing effect increases with decreasing rate. There is little effect at the fastest rates, and only at the slowest rates does the difference approach the expected vowel lengthening effect for English, precisely where the difference in obstruent duration becomes appreciable.

6 Further Tests of The Expandability Hypothesis

We have argued thus far for a vowel lengthening effect that is driven by a failure of voiced stops to lengthen sufficiently. This model entails that segments, or at least vowels, cannot have fixed expansion parameters. However, even if expansion factors were constant, and voiced obstruents could lengthen indefinitely, vowels preceding voiceless stops would still be shorter due to the inherent difference in expandability between the three types of segment: vowels, voiceless obstruents, voiced. Furthermore, as duration increases over all, the difference between the durations of both obstruents, and the vowels preceding them, will also increase. This is essentially what Campbell (1992) demonstrated in his model of syllable-based timing.

For a given syllable, and a given target syllable duration, expansion or compression is distributed among the component segments. This model assumes a single expansion value for all segments, but modulates that value by the estimated segment variance, thereby effectively using smaller expansion constants for segments that lengthen less. Because the combination of individual segment durations comprise the target syllable duration, Campbell’s model also effectively implements compensation. Vowels expand less when inherently longer and more variable segments are also present in the syllable, and more, when inherently shorter and less variable segments are present. This mechanism is able to reproduce the duration differences between vowels in open versus closed syllables, and in syllables with complex versus simple codas (e.g., Elert 1965, Kristoffersen 2000, Kavitskaya 2002).

In fact, Campbell notes that his model “...appears to account quite simply, though in fact not completely, for the lengthening that has been observed in vowels of English before voiced consonants p. 217 (emphasis ours).” The piece missing is the non-linear portion of the lengthening function for voiced obstruents. Our analysis suggests that the degree of lengthening decreases as duration increases, compounding the compensation factor, and resulting in the characteristic voicing effect in English observed at the very longest word tokens.

Additional evidence for the view that large voicing effects may derive from duration ceilings comes from the literature on prosodic lengthening. Lengthening associated with phrasal boundaries has been found to affect the individual segments in a word as a function of their distance from that boundary (Turk and Shattuck-Hufnagel 2007, Cambier-Langeveld 1997, Berkovits 1993, Hofhuis et al. 1995, Campbell 1992, 19
Port and Cummins 1992). In phrase-final position, Berkovits (1993), measuring disyllabic Hebrew words, found that coda consonants lengthened the most proportionally, followed by the nucleus, and then the onset of the final syllable. Hofhuis et al. (1995) and Campbell (1992) also found more lengthening of coda and nucleus than onset in Dutch, and British English, respectively. The proximity effect is also observed intrasegmentally, with the latter part of the nucleus lengthening more than the initial.

However, Cambier-Langeveld (1997) and Cambier-Langeveld (2000) find that final lengthening in Dutch can sometimes also affects the penultimate syllable. This happens only when the final syllable is unstressed, or contains a schwa vowel. They explain this result in the following way: when inherent restrictions on segment duration prevent the required amount of lengthening from occurring in the final rhyme (insufficient expandability in our terms) lengthening is re-distributed over the preceding syllable. Prosodic lengthening has actually been modeled as a local change to articulation rate conditioned by, and occurring in the vicinity of, a prosodic boundary (e.g., Byrd and Saltzman 2003). A large final lengthening effect (e.g., Delattre 1966) that applies preferentially to coda segments increases the likelihood that voiced obstruents in coda will diverge sufficiently from their voiceless counterparts for an emergent voicing effect to be observed.

6.1 Perception of voicing in final position

The Expandability Hypothesis provides an explanation for the voicing effect as an epiphenomenon of the limited expandability of voiced obstruents in English. This hypothesis accounts for why the effect is only consistently observed for especially long duration tokens. However, in and of itself, this hypothesis does not explain the ability of listeners to reliably use vowel duration to predict post-vocalic obstruent voicing. In this section we will show that the Expandability Hypothesis is not only consistent with the perception literature, but corroborated by certain of the results.

For the remainder of the paper we will focus on word-final stops. It is primarily for stops that preceding vowel duration has been characterized as a contrastive cue. This is plausible because there are often very limited cues to stops in final position. In laboratory experiments such stops can often be literally absent, without voicing, aspiration, or release cues. Stimuli may consist only of a steady state vowel followed by a short period of formant transitions that are consistent with a following stop. Nevertheless, the following period of silence is interpretable as the closure of an unreleased stop.

Although the voicing contrast in final position is often described as being cued entirely, or primarily, by preceding vowel duration, there are actually conflicting results on this point. Raphael (1972) finds that preceding vowel duration cues are stronger than \( F_1 \) formant transitions, and the results of Crowther and Mann (1992) suggest that vowel duration is also a stronger cue than \( F_1 \) offset frequency. Most studies, however, do not directly test different cues against one another. Among those that do, the balance of evidence comes down against the effectiveness of the vowel duration cue. Wardrip-Fruin (1982) found that when preceding vowel duration conflicted with either formant transition cues, or actual vocal fold vibration, the latter dominated. Hogan and Rozszypal (1980) also reported that, for certain voiceless-final words, lengthening the vowel did not change the percept to voiced, but produced no effect, or resulted in stimuli that sounded unnatural. Revoile et al. (1982), using naturally produced stimuli, found that the identification of voiced stops was most strongly disrupted by the removal of vowel offset cues (also O’Kane 1978), while the identification of voiceless stops was most strongly disrupted by removing the release burst. Similarly, Repp and Williams (1985) found that the addition of a release burst to otherwise ambiguous stimuli reduced voiced responses. Changes to vowel duration, on the other hand, had little effect on voicing perception. Raphael (1981) concluded that vowel duration was only a weak cue to voicing for natural stimuli produced in carrier phrases, and that the effectiveness of various cues was strongly context-dependent.

Like word-final stops, medial stops are post-vocalic, and they are also subject to neutralization of both voicing and aspiration, resulting in productions that are essentially just short periods of silence (oral cavity closure). The duration of that silence, however, is known to affect voicing perception. In fact, it is standard in
the literature to describe the perceptual boundary in medial position in terms of the ratio of closure duration to preceding vowel duration (Lisker 1957, Port and Dalby 1982, Port 1979, 1981). For fixed consonant durations, increases in preceding vowel duration lead to increased perception of voicing, while increases in consonant duration (for fixed vowel durations), lead to decreased perception of voicing. In a production study of initially stressed bisyllabic words, Davis and Van Summers (1989) found both that vowel duration was longer before voiced stops than voiceless, and that voiceless closures were longer than voiced. In fact, there was a strong negative correlation between the two measures, suggesting that the differences resulted from temporal compensation.

Based on these results, it is reasonable to suppose that stop duration will affect perception in final position. Denes (1955) has shown that the perceptual boundary between the fricatives /s/ and /z/ in final position is dependent on both consonant and vowel duration. However, compared to the number of studies that test perception based on preceding vowel duration alone, there are relatively few that manipulate, or even report, the duration of final stops. These studies, for whatever reason, also tend to be cited less frequently. However, Raphael (1981) found that swapping voiceless closure durations for naturally produced “peg” and “peck” effectively switched the voicing percept for the two tokens. Repp and Williams (1985) similarly found an effect of overall closure duration on the perception of voicing on stop-final syllables followed by a stop-initial syllable (e.g., “labcoat” vs. “lapcoat”).

In fact, it would be surprising if the perception of stop voicing did not depend on the duration of the stop itself. The most predictive measure of voicing in word-initial position is V(oice)O(nset)T(time), an essentially temporal measure (Summerfield 1981, Miller et al. 1986, Miller and Volaitis 1989, Cole and Cooper 1975). And voiced stops tend to be shorter than voiceless in all contexts. Given that the voicing contrast is a property of the consonants themselves, we start with the assumption that vowel duration is a cue to consonant duration, and therefore only a secondary, or indirect, cue to voicing.

In categorical perception experiments participants often hear words in isolation, meaning that the cues to the identity of the segments are the same as the cues to the speaking rate. Because vowels generally show the largest changes under changes in speaking rate (e.g., Gay 1978), and vowel duration tends to be the strongest perceptual cue to speaking rate (Crystal and House 1982, Summerfield 1981, Port and Dalby 1982), vowel duration is often used as a proxy for local speaking rate. Thus, the perceptual dependence on vowel duration can be very simply explained as the effect of speaking rate.

The durational distributions for voiced and voiceless stops overlap considerably, meaning there is no absolute duration boundary that can be used to classify them with complete accuracy. In essence, the listener’s task is to decide whether what they are hearing is a voiced stop spoken slowly or a voiceless stop spoken quickly. We assume that listeners derive a hypothesis about speaking rate from the duration of the vowel. The hypothesized speaking rate then generates an expectation about the duration of the upcoming segment, under the assumption that listeners possesses implicit knowledge of the expansion functions of different segments. Shorter vowel durations, which comprise the majority of the corpus data, correspond to speaking rates at which voiced and voiceless stop durations are not significantly different from each other. In this range, vowel duration is ineffective as a cue to voicing. Only as speaking rate slows to the point where the two expansion trajectories begin to diverge does vowel duration become predictive.

At a slow speaking rate a voiceless stop gets quite long, a voiced stop, considerably less so. An ambigu-

---

7These words were produced phrase-medially and the absolute duration differences were only on the order of 10-20 ms

8Whether total stop duration, closure duration, or aspiration duration is the relevant durational measure may well be language-specific. Durvasula and Luo (2012) report longer vowel durations preceding voiced stops, as well as longer vowel durations preceding aspirated stops in Hindi (which has a four-way contrast system). Both effects can be attributed to compensation if the relevant duration measure is closure duration. Closure durations for aspirated stops were found to be shorter than unaspirated, and closure durations were also shorter for voiced than voiceless stops. Total stop durations were not reported. If aspiration duration is significantly shorter than closure duration and/or aspiration does not participate in lengthening, then these results are consistent with our hypothesis. A speaking rate study of the contrast in final position in Hindi is necessary to answer this question.
ous final stop stimulus will be chosen with a closure of middling duration. But this duration becomes less ambiguous as speaking rate decreases. At some point the token is far shorter than it would have been if it were actually a member of the “voiceless” category (cf. Massaro and Cohen 1983).

The compensation model of Section 4 is used to illustrate this effect. See Fig. 12. As before, speaking rate decreases to the right as expansion force increases. The duration of the voiceless stop (dotted line) gradually diverges from the duration of its voiced counterpart (solid line), as the rhyme is lengthened. This divergence is mirrored in the preceding vowel duration (dark solid line for vowel preceding voiced stop; dark dashed line for vowel preceding voiceless stop). If the listener is exposed to a relatively short vowel (Fig. 12a: horizontal burgundy line), their expectation for the duration of the upcoming stop will be roughly the same regardless of whether it is voiced or voiceless (vertical difference between the lower open circles). An observed stop duration (blue dot) that falls close enough to both expected values is assumed to be acceptable for either member of the pair, and will not be sufficient to distinguish between the two in the absence of other cues.

![Figure 12: Compensation Simulation](image)

For a very long vowel, on the other hand (Fig. 12b: horizontal burgundy line), there is a large difference in the expected durations of the voiced and voiceless stops. The same observed stop duration (blue dots) now falls well below both expected values. Assuming that a larger distance between the expected and observed values results in a lower probability, we hypothesize that this token will be perceived as voiced, especially in a task in which participants are told there are only two options.

In the classic perception experiment listeners are exposed to a continuum of uniformly spaced vowel durations, presented in a randomized order. For each presentation of each token, they must pick either the voiced or the voiceless alternative (e.g., “bad” or “bat”). Results consistently show a categorization function with a high, uniform rate of voiceless response at the shorter end of the continuum, followed by a brief transition period of mixed responses, and a high uniform rate of voiced responses at the longer end of

---

![Diagram](image)
the continuum. Given that we hypothesize that shorter vowels should not provide any cues to the voicing contrast, we would expect, all else being equal, that listeners would be at chance in identifying tokens at that end of the continuum. However, the nature of the actual experimental stimuli may bias perception strongly towards the voiceless member of the pair for two reasons. Ambiguous tokens are, by definition, phonetically voiceless. Depending on how exactly such tokens were created, stimuli may also retain other cues to the original speech token from which they were generated, such as an $F_1$ offset that is more consistent with a voiceless, rather than a voiced, stop. For example, the synthetic stimuli used in Denes (1955) were based on originally voiceless tokens. Raphael (1972) created both “voiced” and “voiceless” final synthetic stimuli, but the only difference was that the voiceless lacked any $F_1$ values at all during the transition period. All tokens in both experiments lacked a voicing bar, and contained no release bursts. In contrast, Repp and Williams (1985), using naturally produced stimuli, found a large perceptual difference between continua generated from an originally voiced stop (lab), versus from an originally voiceless stop (lap). Voiced responses were about 40% higher for the former across all but the two longest vowel durations.

The categorical perception results for word-final stops are therefore explainable as the result of, first, a default voiceless percept, due to residual cues that are more consistent with the voiceless member of the contrast, and second, unusually long vowel durations. At the longest vowel durations (vanishingly rare in actual speech), the expected duration of a voiceless stop is so long as to approach a probability of zero. The perception-production mismatch found in experiments testing the voicing contrast in initial position provides additional support for this account (e.g., Miller et al. 1986, Miller and Volaitis 1989, Volaitis and Miller 1992). The perceptual VOT boundary has been consistently found to be longer than the boundary estimated from production data. However, the two boundaries coincide when naturally produced, unedited stimuli are used in the perception task. Nagao and de Jong (2007) suggest that the mismatch may arise from the fact that the stimuli typically used in perception experiments are artificially impoverished. In other words, the edited tokens are so ambiguous that they can only be confidently classified at very long VOT, or very slow speaking rates – speaking rates that may not normally arise in conversational speech. Furthermore, the consistency in the reported perceptual cross-over point across experiments on word-final stops may be explained by the same artificiality. For voiceless closures with no audible release, the duration of the coda stop cannot be determined. Listeners may all infer some (more or less the same) plausible duration.

The foregoing shows that it is possible to explain the perception of voicing in final position without assuming that vowel duration is a primary, or even a sufficient, cue to the voicing contrast. Instead we argue that, in a laboratory task involving artificially impoverished stimuli, listeners use vowel duration to infer speaking rate, from which an expectation for the duration of the following stop arises. We assume that the role of vowel duration is the same for the perception of stops in word-medial position.

### 6.2 Preceding versus following vowels

Port and Dalby (1982) originally considered the possibility that the influence of vowel duration on voicing perception (in word-medial position) was tied to speaking rate, rather than an intrinsic cue to following voicing. However, they ended up rejecting the speaking rate hypothesis because they found a much stronger perceptual effect of preceding vowel duration than of following vowel duration (in word pairs such as “rapid”/“rabid”). Instead, they concluded that the C/V ratio was a quasi-invariant cue to voicing. Massaro and Cohen (1983) demonstrated that C/V was not, in fact, invariant under tempo changes, and that absolute durations also matter in classification. The effects of speaking rate are non-linear. Voiced stops expand less and less as speed continues to decrease, while their associated vowels expand even more, causing the C/V ratio to decrease in the upper range. Nevertheless, it does appear to be the case that vowels following stops in initial and medial position exert a significantly smaller perceptual effect than vowels preceding stops in medial and final position (Peterson and Lehiste 1960, Allen and Miller 1999). Stress, syllabic constituency, and final lengthening may all play a role in this result.
Unstressed vowels are likely to be shorter than their stressed counterparts, and to be lengthened less under slowed speaking rate. Thus, it is reasonable to suppose that unstressed vowels contribute less to the inference of speaking rate for stops in post-stress medial position (such as those used in Port and Dalby (1982)). This is also consistent with the general finding that lax and unstressed vowels show little to no voicing effect (Umeda 1975, Crystal and House 1982, De Jong 2004). While lengthening occurs preceding as well as following prosodic boundaries, the effect does not seem to be symmetric in English. While the consonant immediately following a prosodic boundary shows the effects of lengthening, typically the vowel following that consonant does not (e.g., Fougeron and Keating 1997, Cho and Keating 2009, Kim and Cho 2012). In fact, the post-boundary effect may be even more local than the first consonant\(^9\). In which case, following vowel duration may not correlate well with stops in phrase-initial position.

It may also matter whether the consonant and vowel share a syllable, a rhyme, or neither. Pre-vocalic and post-vocalic consonantal gestures are phased differently in English, with more overlap between the vowel and any preceding consonants, than between the vowel and any following consonants (Browman and Goldstein, 1988, 1990). However, coda and nucleus are described as more tightly linked than onset and nucleus (e.g., Selkirk 1982), and it may be that temporal compensation effects are more likely to be seen between the former.

In medial position the consonant’s syllabic affiliation is ambiguous. The onset maximization principle (e.g., Clements and Keyser 1983) places a single medial consonant in the onset of the following syllable. However, there is evidence that stress and sonority both affect syllabification, such that a medial consonant will be syllabified in the coda of a preceding syllable if it is stressed, and the following syllable is not (e.g., Treiman et al. 1994, Eddington et al. 2013)\(^{10}\). Thus, in a word pair such as rapid/rabid, the bilabial stop may be in coda position, and thus strongly affected by the duration of the preceding vowel. This interpretation is supported by the fact that in bisyllabic words with the opposite stress pattern (e.g., “adopt” versus “atop”), vowel duration differences between the vowels following the alveolar stop are either very small, or non-existent (as might be expected if the stop is in onset position), as were the closure duration differences (Davis and Van Summers 1989). Under this syllabification, the preceding vowel is in a different syllable and is therefore also predicted to affect perception of the following consonant voicing less than in rapid/rabid type pairs.

6.3 Predictions

There are many possible avenues for further tests of the Expandability Hypothesis. In this section we detail three general predictions, two of which have already been partially confirmed: 1) it should be possible to find “voicing effects” for segments in non-final position, 2) it should be possible to find “voicing effects” for segments other than obstruents, and 3) it should be possible to manipulate listener perception not just of voicing, but also of speaking rate.

If voiced stops in initial position lengthen less than voiceless, and tautosyllabic vowels compensate at very long durations, then we should be able to observe a voicing effect in onset position, where it is the

---

\(^9\)Byrd et al. (2005) find that coda consonants show a larger difference between medial and boundary position than onsets. Interestingly, the difference seems to lie in the fact that more of the coda gesture is lengthened. For both codas and onsets, the portion of the gesture closest to the boundary is lengthened the most – for codas the release portion, and for onsets, the constriction portion. However, the constriction portion for the coda consonant is also lengthened to a lesser degree, while the release for the onset is not (or at least, not consistently). If this effect is robust, it may be that stop consonants in absolute initial position (where constriction cannot be perceived) will show no lengthening effect.

\(^{10}\)In fact, it has been argued that all intervocalic singleton consonants are ambisyllabic, based on the fact that such segments are phonetically realized neither as an initial-position allophone nor a final-position allophone (Kahn 1976, Rudes 1977). The paragon case of this is the realization of coronal stops as flaps in post-stress medial position. At other places of articulation, however, medial stops may exhibit a closure duration and degree of aspiration that are intermediate between initial and final allophones (e.g., Davis and Van Summers 1989).
following vowel that appears to be lengthened. Although a number of studies report on the perception of VOT in initial position as a function of speaking rate (e.g., Miller and Baer 1983, Miller and Volaitis 1989, Volaitis and Miller 1992), stop duration is usually varied inversely with vowel duration, making it difficult to determine the relationship between consonant and vowel. In the handful of studies that vary vowel, rather than syllable, duration the results accord well with the Expandability Hypothesis. Differences are only observed for relatively long vowel durations, similar to those in final position (see Section 4). Viswanathan et al. (2019) report a significant difference in voicing perception only between vowels of 175 and 225 ms., no difference between longer vowels, at 225 and 275 ms., and no difference between shorter vowels, at 125, 150 and 175 ms. Toscano and McMurray (2015) find a significant difference in response rate across the same boundary, between vowels of 189 ms. and those of 377 ms. Nevertheless, as was noted in the previous section, effect size seems to be smaller in onset than in coda position. However, if it is possible to devise a scenario in which sufficient lengthening applied to the beginning, rather than the end, of the syllable, then it may be possible to induce a comparable “voicing effect”.

If the Expandability Hypothesis is correct, it should also be possible to find expansion compensation on segments other than vowels, as long as they are more expandable than voiced stops. Weismer (1979) reports a long-distance effect that supports this prediction wherein the VOTs of stops in onset position were longer when the coda stop was voiced. A difference in nasal duration preceding voiced versus voiceless stops has also been found both for monosyllabic words of the form “dens/dense” (Raphael et al. 1975, Port and Cummins 1992, Beddor 2009), as well as polysyllabic words of the form “cantor/candor” (Vatikiotis-Bateson 1984). Furthermore, Raphael et al. (1975) find that both vowel and nasal duration affect perception of voicing on final stops. Beddor et al. (2013) confirm that listeners are sensitive to the difference in nasal consonant duration. For auditory stimuli consisting of CÑC tokens (i.e., nasalized vowels but no nasal consonant), participants were overall more likely to fixate on the image corresponding to the CVNC word when C2 was voiceless (e.g., bent). They interpret this result as deriving from listener expectation that the nasal gesture will be coordinated differently in the two contexts: initiating earlier before a voiceless stop, and later before a voiced stop11. Thus, a nasal consonant of length zero, coupled with a fully nasalized vowel, is more consistent with “bent” than “bend”. However, no explanation is offered as to why the phasing relationship should be different in the two contexts. This difference, however, can be accounted for under the Expandability Hypothesis if expansion compensation at the word (or syllable) level acts to both lengthen individual gestures, as well as separate them (as is found for speaking rate effects (e.g., Stetson 1928, Hardcastle 1985), as well as prosodic lengthening effects (e.g., Byrd and Saltzman 1998, Byrd et al. 2000). At the longest durations, when voiced stops cannot expand further, nasal and vowel gestures should be pulled apart to a larger degree than in voiceless-final words, resulting in less vowel nasalization in the former.

Finally, it is predicted that a change in the perception of voicing should lead to a change in the perception of speaking rate. During the course of the vowel production, it is assumed that a hypothesis about both speaking rate and following segment duration is generated by the listener. For a vowel that is so long that it creates a strong expectation for a following voiced stop, a certain speaking rate is also inferred (given by the x-intercept of the vertical line on the left in Figure 12b). If listeners subsequently experience an unambiguously voiceless stop (strong aspiration), this should simultaneously cause a change in their experience of the speaking rate of the entire syllable (to the x-intercept of the vertical line on the right in Figure 12b). The voiceless stop should indicate that the speaking rate is actually slower than previously supposed, not (or not only) because the stop itself adds duration to the syllable, but because a vowel of the given duration, preceding a voiced stop, is expected to occur at a faster speaking rate than a vowel of the

11Pycha and Dahan (2016) find a difference in the ratio of nucleus to offglide for the diphthong í before voiced versus voiceless stops that they attribute to the same phenomenon. They also show in an eye-tracking study that listeners can use the nucleus duration in some cases to predict the voicing of the coda stop.
same duration, preceding a voiceless stop. To the best of our knowledge, there are no data yet that directly test this prediction.

7 Summary & Conclusions

In much modern work the voicing effect tends to be described in simplified terms, as a regular, universal, phonetically-driven process, often characterized as a sound change in progress. The frequently cited works tend to be those which demonstrate either strong perceptual effects of preceding vowel duration (e.g., Raphael 1972) or large vowel duration differences in production (e.g., Mack 1982). Such studies are primarily conducted using monosyllabic single-word stimuli in a laboratory setting, where speaking rate is much slower than for normally produced speech (production), and cues to stop identity are significantly impoverished, if not missing altogether (perception). In this paper we have synthesized a large literature on this phenomenon, demonstrating that vowel length differences correlated with coda stop voicing are dependent on a number of factors, that interact in a more complex way.

It has been known for some time that little to no difference in vowel duration is found when words are not phrase final or produced in isolation (e.g., Umeda 1975). Additionally, lax, unstressed, or otherwise inherently short vowels show little to no voicing effect even in such environments (e.g., Peterson and Lehiste 1960). We confirmed both these results using the Buckeye Speech Corpus, finding no over-all effect of final-obstruent voicing on vowel duration: vowels preceding voiced obstruents could be as short as the shortest vowel tokens, and vowels preceding voiceless obstruents could be as long as the longest vowel tokens. Similarly, voiced and voiceless obstruents themselves showed largely overlapping duration distributions. The data suggested that the expected difference might start to emerge at the very high end of the duration distribution, when analysis was restricted to low frequency words, low tense vowels, slow speaking rate, and/or pre-pausal context.

In production studies that manipulate speaking rate it has been shown that voiceless obstruents, in both word-initial pre-stressed (e.g., Miller and Volaitis 1989), and word-medial post-stress (e.g., Port 1976) position, are longer than voiced, with that difference increasing as speaking rate decreases. Extending the findings of Ko (2018) on coda obstruents in CVC words, we replicated the effect of decreasing speed on duration difference in coda position. Our data also demonstrated that the difference in obstruent duration, as well as in preceding vowel duration, was quite small at the fastest speaking rates. With data across 5 different speaking rates, we were able to show a strong negative correlation between the duration difference between voiced and voiceless coda obstruents, and the duration difference between their corresponding vowels.

It has been demonstrated in perceptual experiments that when aspiration or voicing are present, vowel duration can be over-ridden as a cue to “voicing” (Wardrip-Fruin 1982, Repp and Williams 1985). Furthermore, depending on the presence of other cues, as well as the absolute duration of the vowel, vowel duration may have no effect on phoneme identification (Revoie et al. 1982). Obstruent duration itself has also been shown to affect perception in the same way as vowel duration (Denes 1955, Raphael 1981, Repp and Williams 1985), just as it does in word-medial position (Port and Dalby 1982). This body of results argues against preceding vowel duration as a primary cue to the voiced/voiceless contrast in English. Indeed, it suggests that vowel duration may not even be a sufficient cue in the majority of contexts.

We have offered a proposal that accommodates this full set of results, as well as additional related findings. In the first place, the inverse correlation between obstruent duration and vowel duration, and its dependence on speaking rate, are attributed to a type of compensatory effect (see also Massaro and Cohen 1983, Campbell 1992). Because not all segments have the same degree of expandability, the segments within a word lengthen to different degrees. Additionally, for low-expandability segments, it is possible to reach a plateau in which essentially no further lengthening is possible. It is in this region that the non-linear effects of vowel lengthening can be observed. The model captures the fact that a certain length (or rate) is necessary
to observe the phenomenon, and explains why lengthening appears to be particularly robust in laboratory speech. The discrepancy that Campbell and Isard (1991) report in their model of syllable expansion is directly explained by this model, along with the finding that the C/V ratio boundary between voiced and voiceless obstruents in medial position is not actually invariant under speaking rate (Massaro and Cohen 1983).

The Expandability Hypothesis also predicts the observed interaction between vowel nasalization and stop voicing. Nasals, just like vowels, are longer before voiced than voiceless stops (e.g., Raphael et al., 1975). Because nasals, like vowels, and other sonorants, are highly expandable segments, they can also compensate for any failure of the voiced stop to expand sufficiently. Furthermore, the finding that listeners develop an expectation that there will be less vowel nasalization in words with voiced stop codas (“bend”), than voiceless (“bent”) (Beddor et al. 2013) is also explained. If the voicing effect is driven entirely by an external mechanism of stretching, or slowing, then both longer segment durations, as well as decreased overlap between neighboring segments, are to be expected. Decreased overlap results in less coarticulation, and thus less vowel nasalization.

Under our account, three things are necessary to create a strong voicing effect. The first is an inherent specification for short duration on the “voiced” member of the contrast. It must not be as expandable as its voiceless counterpart. Secondly, neighboring segments in the same word must be more expandable. Thirdly, there must exist sufficient contexts in which words can be slowed/lengthened to the extent necessary for the durational ceiling of the less expandable segment to be reached. Such contexts seem to arise infrequently in normal speech, but are most likely to occur phrase-finally, or pre-pausally, and English is known to demonstrate a large boundary-final lengthening effect, one that affects the segments closest to the boundary the most strongly (e.g., Turk and Shattuck-Hufnagel 2007).

An additional corollary of our account is that actual voicing, or any feature other than length, is not required for a “voicing” effect to arise. In fact, Sharf (1964) found that vowel duration differences were approximately the same whether words were produced normally or whispered, thus demonstrating that the effect was independent of actual vocal fold vibration. Ultimately, we predict that we should find a comparable vowel duration difference between any two sounds that can both appear in coda position and have significantly different expansion functions.

We have also argued that the perception of coda stops as voiced or voiceless is heavily dependent on the laboratory environment. In many such experiments, final stops consist entirely of a period of silence following a short transition from the end of the vowel, and participants make a two-alternative forced choice for a series of words heard in isolation. Investigation of the literature in this domain shows that it is only very long vowels (in the critical region where voiced stops have effectively stopped expanding) that give the strong impression of a following voiced stop, and only in the case where there is no conflicting evidence from the final stop itself. We have explained these results as driven by speaking rate rather than use of vowel duration as a cue to contrast. Listeners should expect a very long voiceless stop at the slow speaking rate implied by such a long vowel. Failing to observe evidence for a long final stop, we hypothesize, leads them to select the voiced alternative as the most likely option (cf. Repp et al. 1978).

The assumption that stops in coda position are often unreleased as well as devoiced, coupled with the

---

12 This result is interpreted as evidence for the “linguistically determined” (i.e., phonologized) nature of the vowel duration cue. This would also have to be the case if stops in present-day English are not reliably voiced.

13 Tests of this prediction may be difficult, however, due to other factors that affect vowel duration. For example, for English nonce bisyllabic words, Peterson and Lehiste (1960) find that the shortest vowel durations occur before voiceless stops, followed by voiceless fricatives, voiced stops, nasals, and then voiced fricatives. Umeda (1975) and Kavitskaya (2002) (for Russian) find a similar scale of durations, although the ordering may change somewhat depending on vowel quality. Yet, in terms of expandability alone, nasals might be expected to show the shortest vowel duration rather than one of the longest. And we have already seen that nasals are longer preceding voiced stops than preceding voiceless stops in coda position. The relationship between vowel and coda nasal, however, may be different in the case where the two gestures can overlap significantly. These relative timing relationships may also vary by vowel, accounting for some of the observed variability.
apparently strong perceptual effect of preceding vowel duration, has led many to hypothesize that the “voicing” contrast in this position has either changed, or is in the process of changing. Raphael (1972) argues that vowel duration should be the preferred cue due to its consistent availability, unlike voicing, which is variable, and release burst information, which may be missing. However, as we have already seen, vowel length differences are not necessary when other cues are available, nor is vowel length sufficient for determining contrast at most durations. Furthermore, unlike the effect of voicing on $F_0$, for example, vowel duration does not vary consistently as a function of the voicing contrast.

If pre-pausal stops were to be lost altogether, then absolute vowel duration could potentially shift from an indirect cue – via inference of stop duration –, to a direct cue to the “voicing” contrast. However, this would be possible only near the upper limit of the vowel length distribution. Furthermore, it is unlikely that stops would be lost from more perceptually salient contexts, where stop duration itself would remain a cue to contrast. In fact, we have seen that stop duration itself does not consistently differentiate the members of the contrast. The difference, just as with the difference in preceding vowel duration, only exists at especially long durations.

The distribution of lax versus tense vowels appears to be another example of this kind of pattern. Lax vowels are shorter on average than tense vowels, and maintain that difference under both stress and prosodic lengthening in laboratory settings (Peterson and Lehiste, 1960, Sharf, 1962, De Jong, 2004). However, data from the Buckeye Corpus show that the two types of vowels are not discriminable by duration at the durations typical of normal speech. See Figure 13. Tense vowels can be as short as lax vowels under the right conditions, and lax vowels can often be as long as tense ones. The data are consistent with a durational ceiling that leads to a robust length difference only at the extreme edge of the duration distribution. Finer-grained distinctions, such as that between high and non-high tense vowels, would likely require even more extreme values to be reliably differentiated by duration. If this kind of pattern is typical of duration-based differences then the so-called voicing effect may apply to many more contrasts, although effect size will likely vary. If this is correct then, even if vowel duration can be used reliably to distinguish stops in final position, and despite the fact that no single cue definitively expresses the voicing distinction on English obstruents, there seems little reason to assume that the contrast is unstable or particularly subject to change.

Figure 13: Distribution of vowel durations for CVC content words in the Buckeye Corpus. Vowels are divided into 3 groups based on predicted duration: lax (ε,ə,o,i), tense high (i,e,u), and tense non-high (a,o,o,æ,u,a,o).
7.1 The Right “Voicing” Features

Throughout this paper the relevant stop contrast in American English has been referred to as one of voicing. But, of course, it is precisely because of the fact that phonetic voicing is often absent from “voiced” stops that preceding vowel duration can be discussed as a possible cue to contrast. Furthermore, the presence or absence of vocal fold vibration is not always necessary or even sufficient for stop identification. Nevertheless, the traditional characterization of this contrast involves the binary feature $±$voice. In order to account for the surface realizations of the contrast it is necessary to treat the phonological feature as distinct from the phonetic feature of the same name. The first must be only an abstract label which is then transformed through a series of rules to the actual physical realization of the sounds. For example, in absolute initial position the /$−$voice/ stop is realized phonetically as voiceless and aspirated, while the /$+$voice/ member may be realized as voiceless unaspirated. In intervocalic stressed position /$−$voice/ is also realized as aspirated, but the /$+$voice/ segment is likely to be phonetically voiced in this environment. In intervocalic unstressed position both phonemes may be voiced, and/or flapped. In syllable-final position both stops may be realized as voiceless and unreleased. Following /s/ pre-vocically, both may be voiceless and unaspirated.

In order to discuss competing analyses of this contrast we will switch now to using the umbrella terms fortis and lenis. Fortis productions tend to be associated with aspiration and voicelessness, and lenis with voicing, but not exclusively. In the preceding analysis, fortis stops in English are specified as /$−$voice/, and lenis as /$+$voice/. Within the tradition of Laryngeal Realism, however, where a closer tie between phonetic and phonological features is sought, the proposed features are directly interpretable in articulatory terms as configurations of the vocal folds. Laryngeal contrasts can be specified over any of unary features [spread glottis], [voice], or [constricted glottis]. Double specifications are also possible with those features that are not articulatorily antagonistic.

There is general agreement that the unspecified category, or default status, of the laryngeal articulator (denoted as $\emptyset$) is equivalent to a plain voiceless, or short lag, stop. Following Kim (1970) and Iverson and Salmons (1995), the set of two-way contrasts involving one underspecified member, [voice]-$\emptyset$ and [spread glottis]-$\emptyset$, are taken to be realized phonetically (in stressed initial position) as [voice, voiceless unaspirated], and [voiceless aspirated, voiceless unaspirated], respectively. It is now fairly common, although not uncontroversial, to characterize English as a [spread glottis]-$\emptyset$ language: fortis stops are specified as [spread glottis], and lenis stops are completely unspecified.

However, strict underspecification of this kind, while it captures some phonetic and phonological effects, makes incorrect predictions for others. Firstly, both plain voiceless and voiceless aspirated stops are predicted to be targets of voicing assimilation, but never triggers, since both are unspecified for [voice]. In fact, assimilation to the voiceless member of a contrast should never occur with the hypothesized features. Conversely, stops unspecified for voicing should never resist assimilation, i.e., maintain voicelessness in a phonetic context antagonistic to it. Yet evidence for both unpredicted outcomes exists. Jamaican Creole appears to exhibit bidirectional assimilation to voiceless (Brown 2016). Jansen (2004) and Stevens et al. (1992) both report regressive voiceless assimilation in CC sequences across morpheme and word boundaries in English\textsuperscript{14}. There is also evidence that fortis stops in “true-voicing” languages (i.e., [voice]-$\emptyset$ languages, such as Hungarian and Russian) do not undergo passive voicing (Keating 1984, Kulikov 2012, Beckman et al.

\textsuperscript{14}Mester and Itô (1989) analyze apparent cases of [/voice] assimilation as simple delinking of the privative feature [voice]. Since the unspecified value is absence of voicing, the surface forms will be realized as voiceless. Thus such cases can be described as neutralization rather than assimilation. While this avoids the specification of [/voice], it introduces an unmotivated process of delinking without spread. It is not clear why voicing neutralization would occur just in those cases where the outcome of assimilation would otherwise be two voiceless segments, yet when the outcome would be two voiced segments assimilation is said to occur instead.
At the same time, it appears that not all aspirating languages (i.e., [spread glottis]-∅ languages) allow passive voicing of lenis stops intervocally (see Beckman et al. (2013) for Icelandic, Deterding and Nolan (2007) for Mandarin).

To deal with these discrepancies, Beckman et al. (2013) propose the following (Jansen (2004) offers a similar analysis). In “true voicing” languages that do not allow passive voicing, “unspecified” stops are actually specified with low voicing values such as [1voice] (while voiced are specified as [9voice]), which acts to prevent automatic (increases in) voicing. To explain the lack of passive voicing in certain aspirating languages, it is necessary to assume that “unspecified” stops are actually specified for a high enough [sg] value, such as [5sg], that would make passive voicing articulatorily difficult, if not impossible.

This move creates an ad hoc division between more “typical” systems that can be described with categorical features (0 or 10 values), and others that require intermediate values. Yet it is likely that well-behaved phonological or phonetic systems are only apparently so, because incompatible or more complex behavior can be placed in another part of the grammar. Although English is described in places as a straightforward [spread glottis]-∅ system, the original argument for this specification clearly shows the ambiguity that remains. One stated advantage of the autosegmental approach for this specification clearly shows the ambiguity that remains. One stated advantage of the autosegmental approach is the greater transparency across the phonetics/phonology divide, in other words, the minimization of allophonic rules. Even with the switch to [spread glottis]-∅, however, mismatches remain. Initial lenis stops are sometimes voiced, and fortis stops may be unaspirated or even voiced in some cases. Ultimately, because completely categorical divisions cannot be made, a weaker appeal to the apparent fact that aspiration distinguishes the contrast in more contexts than vocal fold vibration is necessary.

This problem only gets worse, as every degree of closer examination seems to yield finer and finer-grained distinctions. In this paper the focus has been on vowel duration which, regardless of the phonological features chosen, is likely a direct cue only to stop duration, not voicing or aspiration per se. The host of indirect, or ancillary, cues to the distinction, however, is potentially quite large. In addition to duration and intensity of voicing, aspiration, and F0 contour, length of vowel formant transitions with respect to steady state duration (Fitch 1981), F1 offset frequency (Crowther and Mann 1992), speed of jaw lowering, and jaw offset position Van Summers (1987), have all been shown to differ consistently in fortis and lenis stops in final position. In medial position following a stressed vowel, consistent differences are found between fortis and lenis stops in the timing of vocalic voice offset, and the signal decay time Lisker (1986), which should apply to final position as well. While not all of these potential cues may be sufficient to perceptually discriminate the contrast alone, the work of Bailey and Summerfield (1980) suggests that there are almost an unlimited number of fine-grained acoustic details that listeners are sensitive to in the determination of contrast. While many of these cues could be relegated to an acoustic level of representation, the division between levels is generally not well-defined.

Vowel duration is frequently characterized as a phonological “voicing” feature, but not closure duration or F0, even though the latter two cues have been shown to influence perception to the same, or an even greater, degree. Generally, an acoustic cue is assumed to be sufficient to express the associated contrast (read, contrastively specified) if it can be shown to support accurate phoneme identification in isolation. But it is well known that cues can be “traded off” with one another. That is, while a long enough closure duration can cue a fortis stop on its own, a shorter closure in tandem with a shortened vowel can also do so (e.g., Kohler 1979, 1984, Fitch 1981, Lisker 1986, Van Summers 1987, Bailey and Summerfield 1980, Klatt 1976, Malécot 1968).

It has been known for some time that the classical feed-forward model of both auditory and linguistic perception is incorrect. Perception does not occur in a vacuum, but is influenced by the perceiver’s hypotheses about what they expect to perceive, or are more likely to perceive. In the current scenario, the longer

---

15The latter is attributed to reduction in the amount of glottal opening in weak prosodic positions, leading to absence of aspiration, and even voicing if there is sufficient narrowing (Iverson and Salmons (1995), Honeybone (2005), Beckman et al. (2013)).
the segment duration, the less likely it is that the segment in question is underlyingly short (or lenis). This choice automatically entails the assumption of a particular speaking rate, faster if more of the duration is attributed to inherent stop length; slower, if more of the duration is attributed to the effects of speaking rate. However, the preceding vowel also provides information regarding speaking rate, and is likely to be the more salient of the two. As was modeled in Section 6.1, for a constant stop duration, as vowel duration increases, the probability that the stop is the fortis (or highly expandable) segment, correspondingly decreases. Thus, at a certain point, the probability becomes low enough that the stop is perceived as “short”, or lenis. Fundamentally, the analyses of speaking rate and of phonemic identity are not separate, or indeed, separable. Thus, the probability of fortis versus lenis is, in reality, the probability of a vowel of duration $d_V$, produced at speaking rate $v_i$, followed by a fortis stop of length $d_p$, as compared to the probability of a vowel of duration $d_V$, produced at speaking rate $v_j$, followed by a lenis stop of length $d_p$.

Thus, in a sense, all levels of analysis must happen simultaneously. This makes a straightforward acoustics-to-phonetics-to-phonology mapping unlikely. If the boundaries between these levels are eliminated, however, then so too are the numerous, potentially complex rules that would be necessary to connect them. At the same time, such a move would require abandoning the minimal specification assumption that drives much of current contrastive feature theory: that for a contrast involving $n$ phonemes, there should be at most $n-1$ contrastive features. Yet if nominally contrastive features ($\pm$ voice, spread glottis, etc.) are not actually realized in all contexts, even when complex allophonic rules can be deployed, then the better alternative may be to allow many weighted features that directly subserve phoneme identification, features that may be redundant, or only weakly discriminative, but when taken together allow highly accurate classification.

### 7.2 Further Implications for Feature Theory

We have argued in this paper that “voiced” obstruents in English, whatever other features define them, must possess the property of shortness. The chain of reasoning for that conclusion, however, turns out to be diametrically opposed to the dominant view of the role of prominence in phonological theory. While the phonetic realization of underlyingly contrastive features will differ by context, the most prominent environment, usually initial stressed position, is assumed to most faithfully reflect those features. Features are said to be enhanced, or more strongly signaled, in such contexts (e.g., Kingston and Diehl (1994)). Conversely, observed enhancement is taken to indicate features that are “controlled”, or underlyingly specified, as opposed to being supplied by context-sensitive rules (e.g., Ohala (1981)). Enhancement can be realized as increased acoustic amplitude, increased size of articulatory gestures, and increased gestural, and thus, segmental, duration. In addition to making individual features more salient, enhancement is also assumed to be a mechanism for distinguishing the members of a phonemic contrast by increasing their phonetic distance from one another (e.g., De Jong 1995, Cho 2016, Cho and Jun 2000). For the above reasons, slower than normal speaking rate is taken to be an enhancing mechanism that is predicted to lead to lengthening, but only of contrastively specified features (e.g., Solé 2007).

Within the enhancement framework, failure (of segments in initial position) to lengthen with decreased speaking rate implies the lack of a contrastive features. Thus, fortis stops in French and Thai are unspecified for laryngeal features, while lenis stops in English are the unspecified member of the contrast (Kessinger and Blumstein (1997), Beckman et al. (2013)). Additionally, an observed interaction between a given phonetic cue and an additional variable that affects duration is taken to indicate that the cue is underlyingly specified as inherent to the contrast. All of the following results have been taken as evidence that preceding vowel duration is purposefully manipulated by speakers to mark the fortis/lenis distinction. De Jong (2004) compared the effect of stress on vowels in syllables closed by a voiceless stop, versus those closed by a voiced stop, and found that while stress acted to lengthen each syllable relative to its unstressed counterpart, pre-voiceless stressed vowels were lengthened less than pre-voiced stressed vowels. Braunschweiler (1997)
reported that the voicing effect in pre-voiced position in German did not seem to lengthen vowels to the degree that would obscure an existing long versus short vowel distinction. Similarly, Peterson and Lehiste (1960) showed that vowel duration differences preceding voiced and voiceless segments was greater for long vowels than short vowels in English. In the opposite direction, Klatt (1973) found that the difference in duration between the stressed vowel in a monosyllabic word and the same vowel in a bisyllabic word was larger (by percentage) for syllables closed by voiced stops than those closed by voiceless stops (see also Van Summers (1987), De Jong (1991), Crowther and Mann (1992), Raphael (1975), Smith (2002)).

In this paper, however, we have conceptualized stress, prosodic boundary marking and speaking rate simply as external forces which, among other effects, can act to lengthen segments. We have argued that an emergent account of the voicing effect, as a type of temporal compensation, achieves both greater parsimony and greater explanatory coverage of the empirical data. Under our account, it is the failure of lenis stops to increase in duration as other segments in the word lengthen that indicates an underlying specification: for a segment that is shorter than some maximum length. The asymmetric effects on lenis versus fortis stops of factors such as stress, inherent vowel length, and word length, are expected. They follow directly from the fact that more lengthening leads to more compensation, as the gap between the target duration and the actual stop duration increases.

If one accepts these conclusions then it follows that lengthened vowels are not allophones of underlyingly shorter vowels, and there is no linguistic rule that generates them. And if this is true of vowel lengthening, what other phenomena might it also be true of? Like minimal specification, the prominence assumption requires potentially extensive transformations of underlying forms: to the more frequent, non-prominent contexts of normal speech. However, if allophonic rules are only limited proxies for more complex perceptual integration, then, as we suggested in the previous section, we may be better off replacing them altogether. We have explained the apparent categorical perception of vowel duration with these more general perceptual mechanisms. Very slow, hyper-articulated speech is the unusual context in which vowel duration is predictive of voicing. Properties such as intense aspiration and especially long duration associated mostly with this speech style should, under our view, be the least central to the fortis/lenis contrast, not the most. This flipped view of contrast offers an intriguing avenue for future research.

References


---

16The correlation between voicing and length is likely to have arisen due to the difficulty of maintaining the necessary conditions for voicing over extended closure periods (e.g. Ohala 1983, 2011). Nevertheless, it is possible, by virtue of greater articulatory effort, to maintain voicing if desirable. Lengthening with partial or total devoicing is also an alternative. That lenis stops in English are now frequently devoiced rules out a phonetically motivated synchronic account of duration differences. Furthermore, the evidence that such segments are not allowed to lengthen past a certain point is an indication that duration itself has become an independent cue to the contrast.


36


Appendix

Word List (CVStop)

Voiced (4106 tokens; 96 unique words):
bad, bag, bed, big, bob, cab, cad, cod, code, could, cub, dab, dad, dead, did, died, dig, Dodd, dog, dude, fed, feed, fog, food, gig, god, good, guide, had, he’d, head, hid, hide, hood, how’d, hub, hug, hyde, jed, jedd, job, kid, knob, lab, lag, laid, lead, league, led, leg, lid, lied, life, load, loud, mad, made, maid, med, meg, mid, mud, need, paid, pig, read, red, rhode, rid, ride, road, rob, rode, rub, sad, said, she’d, shed, should, showed, side, sued, tag, ted, they’d, tied, todd, tub, tube, wade, we’d, web, weed, wide, would, you’d

Voiceless (14796 tokens; 183 unique words):
back, bat, beat, beep, bet, bike, bit, bite, boat, book, bought, buck, but, butt, cake, cap, cape, cat, caught, chalk, cheap, check, chick, chip, coke, cook, cop, cope, cup, cut, date, debt, deck, deep, dip, dot, doubt, duck, duke, fake, fat, feet, fight, fit, folk, foot, fought, fuck, gap, gate, get, got, gut, hat, hate, heat, heck, height, hick, hip, hit, hook, hop, hope, hot, hype, jack, jeep, jet, jock, joke, kat, keep, kick, knit, know, lack, lake, lap, late, let, light, like, lock, look, lot, luck, luke, mac, make, map, mat, meet, met, might, mike, mock, nap, neat, neck, net, night, nope, nose, not, note, nut, pack, pat, peek, pet, pere, pick, pipe, poke, pop, pope, pot, psych, puck, put, rat, rate, rec, right, rock, rope, route, sake, sat, seat, set, shake, shape, sheet, ship, shit, shock, shoot, shop, shot, shut, sick, sight, sit, site, soap, soup, suit, take, talk, tap, tape, taught, tech, that, thick, this, thought, tight, tip, took, top, type, vet, vote, wait, wake, week, weight, wet, whack, what, whip, white, wick, woke, wreck, wright, write, wrote, yet, zip

Word list CV Fricative

Voiced (5666 tokens; 77 unique words)
b’s, boys, c’s, cahs, cause, cave, cheese, choose, chose, cows, d’s, days, dies, does, dos, faze, five, gave, gays, give, goes, guys, has, have, hayes, haze, he’s, his, how’s, jazz, joe’s, keys, knees, knows, laws, leave, live, lose, love, move, news, noise, p’s, pays, phase, raise, rave, rise, rose, save, says, seas, sees, shave, she’s, shoes, shows, size, so’s, t’s, Taj, these, they’ve, those, ties, toes, toys, twos, use, was, wave, ways, we’ve, who’s, whose, wise, you’ve

Voiceless (2813 tokens; 80 unique words)
base, bash, bass, bath, beef, biff, boss, both, bus, bush, calf, case, cash, chess, chief, choice, cuff, cuss, dose, face, faith, fish, fuss, gas, geese, goose, gosh, guess, half, hash, house, joyce, juice, kiss, knife, las, laugh, lease, less, loose, los, mass, math, mess, mice, miss, moss, mouth, nice, niche, niece, pace, path, peace, piece, piss, push, race, rash, Reese, rice, rough, rush, safe, south, teeth, this, thought, tiff, tooth, toss, tough, vice, voice, wash, wife, wish, with, yes, youth