Applying Multiple Regression Models for Predicting Word Duration in a Corpus of Spontaneous Speech

Na’im R. Tyson

Department of Linguistics
The Ohio State University
ntyson@ling.ohio-state.edu

Abstract

Using word duration as a representative of pronunciation variation, the objective of this research was to delineate a set of variables known to affect word duration and determine the total amount of variation in duration accounted for by them in a multiple linear regression model. More importantly, computing the amount of variation each variable contributes (independently of the others) is crucial in proving its predictive power. Authors such as [1] claim that probabilistic measures such as unigram probability greatly affect whether a word is likely to be reduced in its pronunciation (i.e. the more likely a word is to appear, the greater the chance of it being reduced). However, after performing a regression analysis on word durations from the Variation in Conversation (ViC) corpus of spontaneous speech, and computing partial correlation coefficients of each factor, the results showed that probabilistic measures such as unigram and bigram probability account for less than 1% of the variation in word duration. This finding suggests that the predictive power of certain variables is dependent on the type of corpus being examined — in the case of the spontaneous speech studies in [1], the examined corpus consisted of phone conversations, while the ViC corpus contains monologues.

1. Introduction

Examining read speech, i.e. speech produced under constrained laboratory conditions, could serve as a basis for validating presumptions about speech production, but there are some inherent limitations of using this type of speech to make predictions in spoken language. Read speech often fails to illustrate the effects of reduction in speech within a discourse; for example, in earlier studies of medial /t,d/ deletion in conversational speech, over 10% of medial alveolar stops were deleted, but at the same time, a study of read speech showed that deletion in word-medial contexts was rare [2].

However, if subjects simply spoke without restriction, then it might be possible to measure the influence of both word frequency and discourse effects on an individual’s speech production. The hypothesis tested here is that more frequent words require less effort to produce because of previous mentions. Also, repeated mentions of words contributing to the current information within the topic of the conversation might be produced in some reduced form. To test this hypothesis, we used word duration as a measure of both fluency and reduction, since highly practiced articulatory movements (which are similar to reduced movement) should be shorter (faster) than less practiced ones, and used a regression model to control for other factors that might affect duration.

2. Factors Affecting Word Duration

There are three major types of factors that might lead to the reduction of word duration: localized, invariant and contextual. Localized factors are those properties that are inherent to a word type such as number of syllables, number of phones, etc. Invariant factors include those properties that are external to the word but have some relation to the corpus as a whole such as bigram probability. Contextual factors, on the other hand, take into account only the context of what the speaker has said during the discourse, up to the word token of interest.

2.1. Speech Rate

Prior research findings indicate that measuring speech rate using syllables is advantageous over using phones. Within the Switchboard corpus of impromptu phone conversations, the phone deletion rate is approximately 13%; on the other hand, the rate of syllable deletion is only 2.5% [3]. Noticing that words produced in fast speech tend not to follow their ancillary canonical pronunciations, [4] reports that 25% of the words in the ViC corpus had segments (phones) deleted, while only 10.5% of the total words had at least one syllable omitted. Although [3] and [4] define syllable deletion differently (in the sense that the former research is based on phonetic evidence and the latter is based on the absence of vocalic elements to verify syllable deletion), the fact remains that phone deletion rates are much higher, and more unstable, compared to syllable deletion rates. As a result, computing speech rate required the number of syllables as part of the calculation.

2.2. Word Repetitions

Once a word has been uttered in the current discourse, thereupon becoming part of the focal point of the conversation, subsequent mentions of the same word might be redundant and prone to reduction. In a replication of an earlier experiment done by [5], we used durations of words in the ViC corpus as a rough measure of pronunciation variation, and plotted the average word duration as a function of the number of syllables in the word and the number of repetitions of the word. The relationship between average word duration and the number of syllables and repetitions is shown in Fig. 1. Five- and six-syllable words are not presented in the graph because there were no samples repeated more than twice. In general, as words were repeated, the average duration continued to decline and the effect was approximately linear in nature.
The Probabilistic Reduction Hypothesis of information content in individual words. In spontaneous speech [3], there has yet to be a way of estimating the amount of information other applications in studying pronunciation on conversational contexts, whereas highly informative words will experience little or none. Words carrying low amounts of information have shorter durations, and measuring their information content should indicate that conveying information in a discourse involves the use or reuse of words carrying less redundancy. For example, the same principle holds true for an individual word in the sense that words with high degrees of co-occurrence with other words are very predictable, and are more likely to experience reduction if they are likely to be next in a sequence [3].

2.3. Predictability

Given listeners’ expectations of what they will hear in a discourse, speakers do not have to fully articulate their productions of words [6]. Percentages of correct identification for redundant words after hearing a nonredundant word were higher in contrast to percentages of identification of nonredundant words after hearing redundant words [6]. Redundant words vary from nonredundant words in duration, volume and amplitude. Likewise, listeners do not require a good signal to hear words because, having heard a word before, they can identify subsequent repetitions based on the context of old words [7]. The problem with this notion is that if talkers use less effort for repeated mentions of words [8], and this reduction is constrained somehow by discourse context, [8] does not state a means by which to quantify predictability.

2.4. Probabilistic Measures of Context

The Probabilistic Reduction Hypothesis claims that words undergo more reduction when they are more predictable or probable in discourse context [1]. In short, words with high degrees of co-occurrence with other words are very predictable, and are associated with more reduction from their citation forms. The same principle holds true for an individual word in the sense that it is more likely to undergo reduction if it is likely to be next in a sequence [3]. Findings such as these validate the intuition of previous research in pronunciation variation but do not address discourse-related, contextual factors.

2.5. Information Content

Conveying information in a discourse involves the use or reuse of words, and measuring their information content should indicate if a word is subject to variation. The presumption here is that words carrying low amounts of information have shorter durations, whereas highly informative words will experience little or no reduction. Although computing information content has had other applications in studying pronunciation on conversational speech [3], there has yet to be a way of estimating the amount of information content in individual words.

In Johnson’s studies of massive reduction of words (i.e. words losing two or more syllables in spontaneous speech), function words undergo massive deletion considerably more often than content words [4]. Given his observations, content and non-content words received separate treatment in this analysis since content words might bear more information and be less probable to experience duration reduction.

3. Methods

3.1. Corpus Description

The Variation in Conversation (ViC) consists of the conversational speech of Midwestern, Caucasian Americans from Central Ohio (for a full description of the ViC database, see [9]); moreover, this corpus captures phonological variability via detailed phonetic descriptions of words in their citation and uttered forms [10]. Speech was collected from forty speakers, where each speaker spoke for approximately one hour. Factors such as the age and gender of the interviewer and interviewee were controlled such that measurements are available from sessions with male interviewers, with either male or female interviewees, and female interviewers, with either male or female interviewees. Speakers expressed their views on a variety of topics, but the intent of the data gathering is for the interviewee to produce a monologue (with relatively little interruption from the interviewer).

With the amount of detail encoded in these transcriptions, both at the word and discourse levels, we removed words based on a selection criterion. Each word token had a citation and corrected form (produced by the speaker), but word tokens that co-occurred with background noise were difficult to measure and had to be excluded with the aid of SGML tags[11] — leaving only those word tokens that had concomitant pronunciations.

3.2. Localized Properties

Localized properties of words are necessary to measure in this analysis given the aforementioned observations on the reduction behavior in content and non-content words [1, 4]. Each word in the corpus can be characterized in terms of several inherent properties: 1) whether the word is a content or function word, 2) the number of syllables in the actual form of the word produced by the speaker, 3) the number of phones in the speaker’s production of the word, and 4) the duration (in seconds) of the first occurrence of the word in the discourse. Recording the first duration of a given word token by a given speaker was important in order to establish a constant point of reference for later repetitions of the same word. Johnson’s studies on massive reduction in the ViC corpus illustrate that both syllable deletion and phone deviation rates are in upwards of 10% [4], which means that it is not always the case that the number of syllables and the number of phones are strongly predictive of the duration in subsequent repetitions of the first token.

Standard information retrieval applications use stopwords to broadly correspond to linguists’ notions of function (non-content) words in order to ignore them while performing searches on document collections. However, since most stop-word lists are intended for written language, we expanded a standard list of stopwords to include hesitation words (uh, um, ah, etc.). In the regression model presented here, a binary variable received a ’1’ if the word is a stopword, and ’0’ otherwise.
3.3. Invariant Factors

Invariant factors are calculated once and have the purpose of measuring co-occurring relationships between a word and the context in which it appears. Moreover, calculating these invariant factors gives an estimate of the predictability a word has in the context of other words. The metrics used here are forward bigram probability, otherwise known as the conditional probability of a word given the prior one and backward bigram probability (once the original textual transcriptions are reversed). Both measures may reveal important information about co-occurring words. If two words always occur in the same context such as you know, of the, etc., then these two probabilities will be the same.

3.4. Contextual Factors

Unlike invariant factors, which do not vary from token to token of a word, contextual factors vary with the immediate context of the user’s discourse. One such factor is the speaking rate of a word (uttered by a speaker) which is:

\[
\text{speaking rate} = \frac{\text{no. of syllables in phonetic transcription}}{\text{total time of five preceding words}}
\] (1)

Note that the preceding five words do not include the target word.

In contrast to bigram probability, which is computed from the corpus, unigram probability deals with the number of words encountered thus far. Considering words to be a sequence of random events, the unigram probability of the target word \(w_i\) in a particular speaker’s discourse is:

\[
p(w_i) = \frac{C(w_i)}{n_j}
\] (2)

Equation (2) states that the probability of the target word \(w_i\) for speaker \(j\) is the quotient of the number of times the target word \(w_i\) appeared and \(n_j\), which is the number of words uttered by speaker \(j\) thus far.

To address earlier concerns regarding information content, the entropy, or amount of uncertainty, is obtained from (2):

\[-\sum p(w_{ij}) \log_2(p(w_{ij}))\] (3)

3.5. Regression Analysis

For this research, each factor in the regression analysis served as an explanatory variable to predict a dependent (response) variable. The duration of the first mention of a word was an explanatory variable used as a proxy for the unavailable canonical duration of the word. However, each of the second and subsequent durations of words were response variables.

4. Regression Results

Among 27,905 word tokens containing 2,674 word types, 22,381 random events, the unigram probability of the target word encountered thus far. Considering words to be a sequence of the corpus, unigram probability deals with the number of words duration of the word. However, each of the second and subsequent variables used as a proxy for the unavailable canonical

4.1. Counting Repetitions

Because it was not obvious how to count repetitions, given the structure of the corpus, we decided to implement four different methods to determine which one accounted for the most variation word duration. Method 1 counted repetitions solely within interviewer turns. An interviewer turn here is a span of a speaker’s speech that appears after an <IV E R> SGML tag, but before the following one. Method 2 counted repetitions within interviewer turns; however, repetitions were only counted if the word appeared within a window of fifteen prior words. In other words, counting inside the interviewer turn is based on the number of times the token had been mentioned already in the preceding fifteen words [5]. Method 3 involved the same sort of counting within a window of fifteen words, but this time counting was done throughout the entire discourse rather than within interviewer turns; this test mimicked a test previously performed by [5]. Method 4 simply measured repetitions across the entire discourse without regard to interviewer turns or the previous context. Instead of using all the variables needed for a single, large-scale regression, only the first duration, speaking rate, number of syllables and whether the word is a stopword were used as predictive variables, with the number of repetitions being the variable that varied with each method.

With \(r^2 = 0.5042\), Method 4 accounted for the most variation in word duration amongst the eight speakers. The other methods produced exactly the same r-squared value \((r^2 = 0.4938)\) because these methods produced the same counts of repetitions. Also, interviewer turns were relatively short throughout the entire corpus.

4.2. Full Regression Model Results

All explanatory variables can now be put together to compute the overall regression model. It is important to note, though, that the objective of this research was to determine the value of each variable’s contribution in accounting for word duration. Therefore, partial correlation coefficients should be reported also and to calculate them required the following formula [12]:

\[pr^2 = \left(\frac{t_B}{\sqrt{r_B^2 + n - k - 1}}\right)^2\] (4)

Equation (4) measures partial correlation as the quotient of the t-value of variable \(B\) and the square root of the t-value squared plus the number of degrees of freedom \(n\) subtracted by the total number of coefficients \(k\) in the model. Subtraction of one is necessary because the number of coefficients includes the intercept in conjunction with the rest of the explanatory variables. Equally important, summing partial correlation coefficients will not give a value equivalent to the correlation coefficient; this type of correlation measures a variable’s distinct contribution ignoring the correlation with other variables in the model.

5. Discussion

According to the results obtained from the multiple regression model in Table 1, localized factors, i.e. those factors whose values are local to the words of the speaker, appear to account for the most variation. Knowing whether the word is a stopword gave indication to whether words were likely to be shorter; this was due to the fact that almost two-thirds of the subset of the corpus were one-syllable, non-content words as illustrated in Fig. 2. In spite of the prevalence of the localized variables,
Table 1: Full model regression results where \( F(10, 22370) = 3009, p < 2.2e-16 \). Diamonds represent localized factors, circles show contextual factors and asterisks constitute invariant factors.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( pr^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>◊ first duration</td>
<td>7.11%</td>
</tr>
<tr>
<td>◊ stopword</td>
<td>2.67%</td>
</tr>
<tr>
<td>◊ number of syllables</td>
<td>0.040%</td>
</tr>
<tr>
<td>◊ number of phones</td>
<td>12.47%</td>
</tr>
<tr>
<td>◊ unigram probability</td>
<td>0.059%</td>
</tr>
<tr>
<td>◊ entropy</td>
<td>0.23%</td>
</tr>
<tr>
<td>◊ rate</td>
<td>4.19%</td>
</tr>
<tr>
<td>◊ repetition number</td>
<td>0.24%</td>
</tr>
<tr>
<td>* bigram probability</td>
<td>0.068%</td>
</tr>
<tr>
<td>* reverse bigram probability</td>
<td>0.000002%</td>
</tr>
</tbody>
</table>

| r-squared                       | 57.34%       |

The number of syllables made a negligible contribution to the analysis, thereby coinciding with the empirical observation that syllables tend not to delete as much as phones.

Of the contextual factors, rate of speech accounted for the most variation. Entropy made a small but noteworthy contribution, which lends credence to the observation that information content and pronunciation variation have some association. Counting the number of repetitions failed to account for much variation simply because words can only become so short after consecutive mentions — as shown in Fig. 1.

Invariant factors explained infinitesimal amounts of variation in the corpus. In fact, the reverse bigram probability proved to be statistically insignificant in the regression analysis and can be thrown out accordingly.

6. Conclusion

Results showed that localized factors work best in terms of their individual contributions, but contextual and invariant factors should be expanded in number to see whether there are improvements. For contextual factors, adding whether the word is phrase-initial or phrase-final might prove the assumption that phrase-initial words are longer in duration, but the present state of the corpus is not amenable to such an approach. Probabilistic measures hardly accounted for any variation; however, this fact could change as more speakers’ transcriptions become available.

7. Acknowledgements

The author would like to thank Dr. Mary Beckman for her useful comments on an earlier draft of this paper.

8. References


