THE EFFECT OF LANGUAGE MODEL PROBABILITY ON PRONUNCIATION REDUCTION

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond

Linguistics Department
University of Colorado, Boulder

ABSTRACT

We investigate how the probability of a word affects its pronunciation. We examined 5618 tokens of the 10 most frequent (function) words in Switchboard: *I*, *and*, *the*, *that*, *a*, *you*, *to*, *of*, *it*, and *in*, and 2042 tokens of content words whose lexical form ends in a t or d. Our observations were drawn from the phonetically hand-transcribed subset [1] of the Switchboard corpus [2], enabling us to code each word with its pronunciation and duration. Using linear and logistic regression to control for contextual factors, we show that words which have a high unigram, bigram, or reverse bigram (given the following word) probability are shorter, more likely to have a reduced vowel, and more likely to have a deleted final t or d. These results suggest that pronunciation models in speech recognition and synthesis should take into account word probability given both the previous and following words, for both content and function words.

1. INTRODUCTION

Many factors are known to play a role in whether a word's pronunciation is reduced, including the phonetic context, rate of speech, and neighboring disfluencies. In recent work, we and others have shown that the trigram probability of a word $P(w_i|w_{i-2}w_{i-1})$ also predicts pronunciation reduction [3, 4, 5] and can play a role in a dynamic LM [6, 7].

This paper extends our work by comparing different kinds of LM probabilities including the unigram and bigram probabilities of the target word, as well as the 'reverse' bigram probability given the following word. We also study whether probability affects reduction in both content and function words, by examining 5618 tokens of the function words *I*, *and*, *the*, *that*, *a*, *you*, *to*, *of*, *it*, and *in*, as well as 2042 tokens of content words whose lexical form ends in a t or d. Both datasets are drawn from 38,000 phonetically hand-transcribed words [1] from the Switchboard corpus [2]. We coded each word with its duration and pronunciation and then used linear and logistic regression to control for contextual factors like phonetic context and rate of speech. We then studied whether various LM probabilities predicted whether words would be reduced (have a reduced vowel, a deleted final t or d, or simply be shorter).

Thanks to the National Science Foundation for partial support of this work via NSF IIS-9733067, to Eric Fosler-Lussier for supplying us with the N-gram probability distributions and collaborating on many earlier related projects, and to Cynthia Girand.

2. LANGUAGE MODEL PROBABILITY

We estimated the following LM probabilities using a backoff trigram grammar with Good-Turing discounting trained over the entire ($\approx$ 3 million word) Switchboard corpus:

Probability Measure	
Unigram	$P(w_i)$
Bigram	$P(w_i\|w_{i-1})$
'Reverse' Bigram	$P(w_i\|w_{i+1})$
Centered Trigram	$P(w_i\|w_{i-1}, w_{i+1})$

3. STUDY 1: FUNCTION WORDS

Our first experiment studied the 10 most frequent English function words in the Switchboard corpus. (These are also the ten most frequent words in the corpus.)

3.1. The Function Word Dataset

The function word dataset consists of all instances in the phonetically transcribed corpus of the 10 most frequent English function words *I*, *and*, *the*, *that*, *a*, *you*, *to*, *of*, *it* , and *in*. We analyzed 5,618 tokens (out of 9,000 total after excluding various non-comparable items;(see §3.3). Each observation was coded for two dependent reduction factors:

vowel reduction: We coded the vowel of each word as *full* or *reduced*. The full vowels included basic citation pronunciations, e.g. [dh iy] for *the*, as well as other non-reduced vowels. The reduced vowels that occurred in the words were [ax] and [ix].[1]

duration: the duration of the word in milliseconds.

[1] In general we relied on Berkeley transcriptions for our coding, although we did do some data cleanup, including eliminating some observations we judged likely to be in error; see [3] for details.

3.2. The Regression Analysis

We used multiple regression to evaluate the effects of our predictability factors on reduction. A regression analysis is a statistical model that predicts a *response variable* (in this case, the word duration, or the frequency of vowel reduction) based on contributions from a number of other *explanatory factors*. Thus when we report that an effect was significant, it means that after accounting for the effects of the other explanatory variables, adding the explanatory variable in question produced a significantly better account of the variation in the response variable. For duration we used ordinary linear regression to model the log duration of the word. For vowel quality, which is a categorical variable, we used logistic regression.

3.3. Control Factors

The reduction processes are each influenced by multiple factors that must be controlled to assess the effect of probability on reduction. We briefly review these factors here and our method of controlling for them. First, we excluded tokens of function words based on the following three factors:

Planning problems: We removed words which are immediately followed by disfluencies indicative of 'planning problems' (pauses, filled pauses *uh* and *um*, or repetitions), since they tend to have less-reduced pronunciations [8, 3, 4, 9]. We also removed words that were preceded by filled pauses.

Phrase boundaries: Words are known to be lengthened at intonational phrase boundaries. As an attempt to control for this fact, we removed words which are initial or final in our pseudo-utterances, since pseudo-utterances of our datasets are bounded by turns or long pauses.

Special forms We removed cliticized words (e.g., *you've*, *I've*, *it's*) and the variant *an* of the indefinite article *a*.

We then controlled other variables known or suspected to affect reduction by entering them first in the regression model. Thus the base model for an analysis was a regression on the following set of control factors:

Rate of Speech: Words are shorter when spoken faster (see, e.g., [5]). We measured rate of speech at a given word by taking the number of syllables per second in the smallest pause-bounded region containing the word. Our regression models included both log rate and log squared rate.

Segmental Context: The form of a word is influenced by the segmental context—for example, vowels are less like to be reduced when they are followed by another vowel. We controlled for the class (consonant or vowel) of the following segment.

Syllable type of target: We coded the target word for syllable type (open or closed) (e.g., *it* vs. *a*). This variable interacts closely with segmental context.

Reduction of following vowel: The accentual pattern of the utterance plays a crucial role in reduction. We partially controlled for accent by coding whether the vowel in the syllable following the target word was reduced or full. [2]

We also included a number of terms for the interactions between these variables. [3] Because the 10 words in this dataset were all very frequent we did not include unigram probability in this analysis.

3.4. Results

3.4.1. Vowel Reduction in Function Words

$P(w_i|w_{i-1})$ was a significant predictor of reduction ($p < .0001$), even after controlling for the various contextual factors discussed above. The higher the bigram probability, the greater the expected likelihood of vowel reduction.

The predicted likelihood of a reduced vowel in words which were highly predictable from the preceding word (at the 95th percentile of probability) was 48 percent, whereas the likelihood of a reduced vowel in low predictability words (at the 5th percentile) was 24 percent.

Higher conditional probabilities of the target word given the following word $P(w_i|w_{i+1})$ were also a predictor of a greater likelihood of reduction ($p = .002$). The predicted likelihood of a reduced vowel in words which were highly predictable from the following word (at the 95th percentile of conditional probability) was 42 percent, whereas the likelihood of a reduced vowel in low predictability words (at the 5th percentile) was 35 percent. Note that the magnitude of the effect was a good deal weaker than that with the previous word.

Even after accounting for the individual effects of the conditional probability given preceding and following words, there is a small additional significant effect of the preceding and following words together, as measured by the 'centered' trigram probability given the two surrounding words ($P(w_i|w_{i-1} \cdots w_{i+1})$ ($p < .02$).

[2] This partially controls for stress since the reduction of the following vowel should correlate with its stress level, and hence the stress level of the target word.

[3] We were unable to control for additional aspects of the preceding segment environment (e.g., vowel identity and coda identity), some aspects of prosodic structure (especially metrical prominence) and social variables (register, age, gender, race, social class, etc.). We did control for some of these variables in earlier work [4] and still found robust effects of the LM predictability measures.

3.4.2. *Duration of Function Words*

Probability also affected duration. $P(w_i|w_{i-1})$ was a significant predictor of durational shortening ($p < .0001$). The higher the conditional probability, the shorter the target word. High probability tokens (at the 95th percentile of the probability) have a predicted duration of 92 ms; low probability tokens (at the 5th percentile) have a predicted duration of 118 ms.

$P(w_i|w_{i+1})$ was also a strong predictor of shortening; the higher the probability the shorter the target ($p < .0001$). Tokens which were highly probable (at the 95th percentile of the probability) have a predicted duration of 99 ms; tokens with low probability (at the 5th percentile) have a predicted duration of 123 ms.

As with vowel reduction, there is a small additional significant effect of the preceding and following words together, as measured by the centered trigram probability ($p < .0001$).

3.5. The Function Word Dataset: Discussion

Function words with a higher LM probability are shorter and more likely to have reduced vowels. The probability given the preceding word and given the following one both play a role, on both duration and deletion. The magnitudes of the duration effects are fairly substantial, in the order of 20 ms or more, or about 20 percent, over the range of the conditional probabilities (excluding the highest and lowest five percent of the items).

4. STUDY 2: FINAL-T/D CONTENT WORDS

Our previous results show that function words are reduced when predictable. But function words have extremely high frequencies and might be more likely to cliticize or collocate with neighboring words than content words. This next study, then, studies whether our results extend to content words. Because content words have a much wider range of frequencies than function words, they also allow us to investigate the role of unigram probability.

4.1. The Final-t/d Content Word Dataset

The Final-t/d Content Word dataset is again drawn from the phonetically-transcribed Switchboard database. We examined 2042 word tokens (out of 3000 total, after excluding various non-comparable items discussed below). Each observation was coded for two dependent reduction factors:[4]

deletion of final consonant: Final t-d deletion is defined as the absence in the transcription of a phone corresponding to a pronounced oral stop segment corresponding to a final t or d in words. For example, the phrase 'but the' was often pronounced [b ax dh ax], with no segment corresponding to the t in *but*.

duration: The duration of the word in milliseconds.

4.2. Control Factors

As with the function word analyses, we excluded tokens of words which occurred in disfluent contexts, or initially or finally in pseudo-utterances. We also excluded polysyllabic words from the duration analyses to make the items more comparable. Other factors were controlled by including them in the regression model before considering the predictability factors. They included variables already discussed: rate of speech, rate of speech squared, whether the next vowel was reduced or not, following segment type (consonant or vowel), coda type of the syllable, as well as the following additional factors:

Inflectional status: [11] and others have noted that a final t or d which functions as a past tense morpheme (e.g., *missed* or *kept*) is less likely to be deleted than a t or d which is not (e.g. *mist*).

Identity of the underlying segment: We coded the identity of the underlying final segment (t or d).

Number of phones: The number of phones in the word is of course correlated with both word frequency and word duration.

Number of syllables: For the deletion analysis only, since the duration analysis was limited to monosyllabic words.

4.3. Results

Using multiple regression, the LM measures (including the unigram probability) were tested on the two shortening variables of deletion and duration by adding them to each of the regression models after the base model.

4.4. Duration

The duration analysis was performed on 1412 tokens of the final-t/d content words.

We found a strong effect of the unigram probability of the target word ($p < .0001$). Overall, high frequency words (at the 95th percentile of frequency) were 18% shorter than low frequency words (at the 5th percentile).

The probability of the target given the next word significantly affected duration: more predictable words were shorter ($p < .0001$). Words with high probability (at the 95th percentile of the probability given the next word) were 12% shorter than low probability words (at the 5th percentile).

The probability of the target given the previous word $P(w_i|w_{i-1})$ also significantly affected duration ($p = .0009$).

[4] Our earlier work also considered other reduction factors including other kinds of deletion [3] and tapping [10].

4.5. Deletion

The deletion analysis was performed on 2042 tokens of t/d-final content words.

Again, we found a strong effect of unigram probability ($p < .0001$). High frequency words (at the 95th percentile) were 2.0 times more likely to have deleted final t or d than low frequency words (at the 5th percentile).

The probability of the target given the previous word did not significantly affect deletion. The only previous word variable that affected deletion in target words was the unigram probability of the previous word. More frequent previous words lead to less deletion in the target word ($p = .007$).

We also confirmed our result from earlier work [10] that deletion is not sensitive to predictability effects from the following word. Neither the probability of the target word given the next word nor the unigram probability of the next word predicted deletion of final t or d.

4.6. Final-t/d Content Word Dataset: Discussion

Content words with higher unigram probabilities are shorter and are more likely to have deleted final t or d than content words with lower unigram probabilities. As is the case with all of our results, this is true even after controlling for rate of speech, word length, and contextual and discourse factors. The effect of unigram probability was the strongest overall factor affecting reduction of content words.

We also found an effect of conditional probability. Content words which have a higher probability given the following word are shorter, although not more likely to undergo final segment deletion. Overall, however, the effects of conditional probability on reduction are much weaker in content words than in function words. Neither $P(w_i|w_{i-1})$ nor $P(w_i|w_{i+1})$ had any effect on deletion, and $P(w_i|w_{i-1})$ had no effect on duration. Failure to find effects may be due to the smaller number of observations in the content word dataset or the general lower unigram and conditional probabilities of content words.

The only effect of the previous word was an effect of previous-word unigram probability. High-unigram previous words led to *longer* target forms and *less* final-t/d deletion. One possible explanation for this unexpected result is based on the fact that the previous-word unigram is in the denominator of the equation defining $P(w_i|w_{i-1})$. Perhaps the effect of previous word unigram is really a consequence of conditional probability, but the size of our content-word dataset is too small to see the effects of the numerator. Another possibility is that the lengthening of content words after frequent previous words is a prosodic effect. For example, if the previous word is frequent, it is less likely to be accented, which might raise the probability that the current word is accented, and hence that it is less likely to be reduced.

5. CONCLUSION

More probable words are reduced, whether they are content or function words, and whether the probability is computed from previous or following words. Function words were strongly affected by conditional probability, while content words showed weaker effects of surrounding context, but strong effects of unigram probability. We are currently incorporating these factors into ASR lexicons. Some of the conditional probability effect can be modeled via reduced pronunciations for multiwords [12]. For others, we are investigating dynamic lexicons, in which pronunciations change as a result of context.

6. REFERENCES

[1] Steven Greenberg, Dan Ellis, and Joy Hollenback, “Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus,” in *ICSLP-96*, Philadelphia, PA, 1996, pp. S24–27.

[2] J. Godfrey, E. Holliman, and J. McDaniel, “SWITCHBOARD: Telephone speech corpus for research and development,” in *IEEE ICASSP-92*, San Francisco, 1992, IEEE, pp. 517–520.

[3] Daniel Jurafsky, Alan Bell, Eric Fosler-Lussier, Cynthia Girand, and William D. Raymond, “Reduction of English function words in Switchboard,” in *ICSLP-98*, Sydney, 1998, vol. 7, pp. 3111–3114.

[4] Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, and Daniel Gildea, “Forms of English function words – Effects of disfluencies, turn position, age and sex, and predictability,” in *Proceedings of ICPhS-99*, 1999, pp. 395–398.

[5] Eric Fosler-Lussier and Nelson Morgan, “Effects of speaking rate and word frequency on conversational pronunciations,” *Speech Communication*, vol. 29, pp. 137–158, 1999.

[6] Eric Fosler-Lussier, “Contextual word and syllable pronunciation models,” in *Proceedings of the 1999 IEEE ASRU Workshop*, Keystone, Colorado, 1999.

[7] Eric Fosler-Lussier, *Dynamic Pronunciation Models for Automatic Speech Recognition*, Ph.D. thesis, University of California, Berkeley, 1999, Reprinted as ICSI technical report TR-99-015.

[8] Jean E. Fox Tree and Herbert H. Clark, “Pronouncing “the” as “thee” to signal problems in speaking,” *Cognition*, vol. 62, pp. 151–167, 1997.

[9] Elizabeth Shriberg, “Phonetic consequences of speech disfluency,” in *Proceedings of the International Congress of Phonetic Sciences (ICPhS-99)*, San Francisco, 1999, vol. I, pp. 619–622.

[10] Michelle L. Gregory, William D. Raymond, Alan Bell, Eric Fosler-Lussier, and Daniel Jurafsky, “The effects of collocational strength and contextual predictability in lexical production,” in *CLS-99*. University of Chicago, Chicago, 1999.

[11] William Labov, “The internal evolution of linguistic rules,” in *Linguistic Change and Generative Theory*, Robert P. Stockwell and Ronald K. S. Macaulay, Eds., pp. 101–171. Indiana University Press, Bloomington, 1972.

[12] Michael Finke and Alex Waibel, “Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition,” in *EUROSPEECH-97*, 1997.