The Relationship Between Phonetic Clarity and Surprisal

Anonymous (do NOT include your name)

December 2, 2024

1 Abstract

This study investigates the relationship between phonetic clarity approximated by word duration and surprisal, a metric of information content. A simple linear regression model was applied to investigate whether longer word duration correlates with higher surprisal. Results show a statistically significant but weak positive correlation, suggesting that other factors beyond phonetic clarity may also influence surprisal.

2 Introduction

Previous work has explored the idea that more informative speech is articulated more clearly in order to facilitate listener comprehension, based on the communicative efficiency hypothesis (Jaeger and Buz, 2017). This project reprises and tests the hypothesis that word duration is related to surprisal, measured using a bigram language model.

3 Related Work

Jaeger and Buz (2017) hypothesized that the higher the surprisal, the more clearly the speaker articulates to help the listener understand better, which is in line with the communicative efficiency hypothesis in linguistics. This paper continues their effort using speech data and statistical modeling to test this relationship empirically.

4 Data and Methods

4.1 Data Collection

Two key metrics were used:

- **Surprisal:** A bigram model was trained on Wikipedia text data (‘wiki.train.raw‘) to calculate word surprisal values. Sentences were selected to ensure all bigrams were present in the training corpus and consisted of at least 15 words.
- **Word Duration:** Fifteen sentences were recorded and pre-processed using the Munich Automatic Segmentation System (MAUS). Extract word durations in milliseconds.

4.2 Data Preprocessing

- Surprisal was computed for each word based on bigram probabilities and averaged for each sentence.
- Word durations were extracted from MAUS output, converted from samples to milliseconds, and averaged for each sentence.

5 Results

5.1 Visualization of Linear Regrssion

Figure 1 shows a scatter plot of the data points, with the linear regression line overlayed. The weak positive trend is evident.

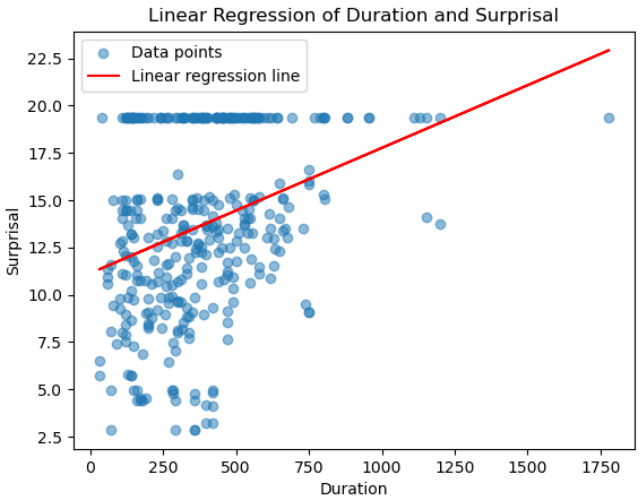

Figure 1: Linear Regression of Word Duration and Surprisal

5.2 Linear Regression Analysis

The regression analysis yielded the following results:

- **Intercept (β_0):** 11.14
- **Coefficient (β_1):** 0.0066
- **R-squared (R^2):** 0.1039
- **P-value:** 4.95×10^{-10}

The positive coefficient indicates a weak but statistically significant positive correlation between word duration and surprisal. This suggests that as word duration increases, surprisal tends to increase slightly.

5.3 Histogram of Word Durations

The histogram of word durations (Figure 2) revealed a positively skewed distribution, with most durations clustered at lower values.

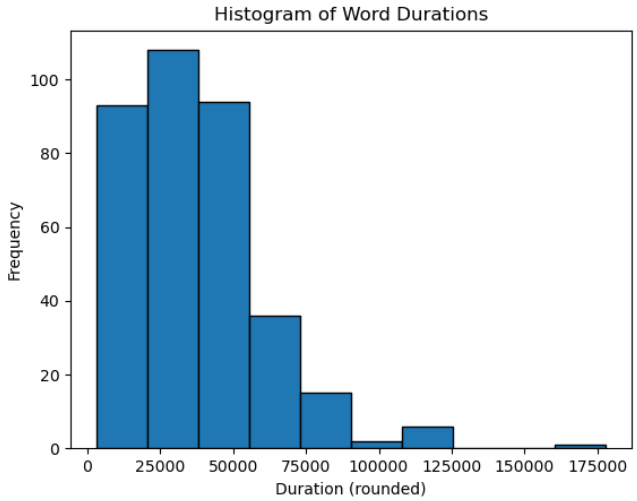

Figure 2: Histogram of Word Durations

6 Discussion

The results support the hypothesis that higher surprisal is associated with clearer articulation, reflected in longer word durations. However, the weak correlation, $R^2 = 0.1039$, suggests that factors other than phonetic clarity influence surprisal. Such factors may include contextual predictability, syntactic complexity, or speaker-specific differences.

The use of a bigram model to estimate surprisal is a potential limitation, as it may underestimate the true surprisal compared to more complex models. Additionally, word duration as a proxy for clarity may oversimplify articulatory differences.

7 Conclusion

This study finds a weak but statistically significant relationship between word duration and surprisal, consistent with the communicative efficiency hypothesis. However, the low R^2 value highlights the complexity of factors influencing surprisal. Future research should consider more sophisticated models of surprisal and phonetic clarity.

References

Jaeger, T. and Buz, E. (2017), 'Signal reduction and linguistic encoding', *Topics in Cognitive Science* **9**(3), 1–15.