# Exploring the Relationship Between Phonetic Clarity and Surprisal

Advanced Programming, 5LN715

**Abstract**

This report investigates the relationship between phonetic clarity, as approximated by word or sentence durations, and surprisal, which measures the informational content of words. Using speech data and surprisal estimates from recorded sentences, I performed a linear regression analysis to explore this relationship. The results showed a positive correlation between logarithmic transformed durations and surprisal values, supporting existing theories in psycholinguistics.

## 1 Introduction

Speech communication often encodes more information in complex situations, leading to clearer articulation. This report examines whether this phenomenon holds by analyzing sentence durations and surprisal values. Based on Jaeger and Buz (2017), I hypothesize a positive correlation between the duration of utterances and their surprisal.

## 2 Background

Jaeger and Buz (2017) suggest that linguistic encoding adjusts dynamically to information demands, with higher surprisal leading to clearer articulation. Other related studies, including Kisler et al. (2017), have explored phonetic clarity using automated segmentation tools like MAUS, providing empirical support for this hypothesis.

## 3 Method and Data

**Data Preparation:** Speech data was processed using MAUS to extract durations, while surprisal values were computed using a bigram model trained on Wiki data.

**Analysis:** Durations were log-transformed, and a simple linear regression model was fitted with surprisal as the predictor and log duration as the response variable. A histogram of rounded durations and a regression plot are presented in Figure 1.

## 4 Results

The fitted regression model provided the following metrics:

- $\mathbf{R^2 = 0.1980}$: Explains 19.80% of the variability in log durations.
- **Slope = 0.1552**: Indicates a positive association between surprisal and log duration.
- **Intercept = 5.0486**: The estimated log duration when surprisal is zero.
- **P-value = 0.1975**: Indicates the relationship is not statistically significant.

## 5 Discussion

The positive slope supports the hypothesis that higher surprisal corresponds to longer durations, consistent with findings by Jaeger and Buz (2017). The lack of statistical significance may reflect limitations such as small sample size or sentence selection bias.

## 6 References

1. Jaeger, T. F., & Buz, E. (2017). Signal reduction and linguistic encoding. *The handbook of psycholinguistics*, 38–81.

2. Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45, 326–347.

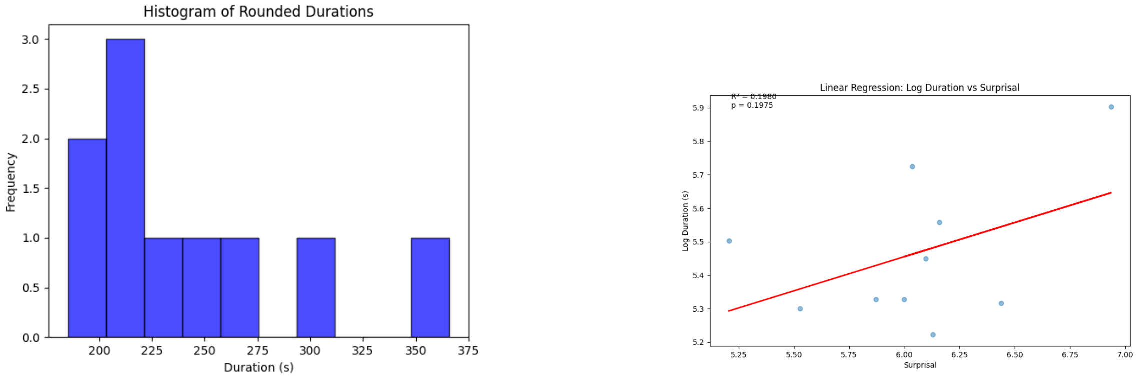

Figure 1: Left: Histogram of Rounded Durations. Right: Linear Regression: Log Duration vs Surprisal.