Research Question & R Code with Outputs

Anushri Kartik-Narayan

2023-06-30

## Research Question

Vowels may either be produced entirely using the oral cavity or partially using the nasal cavity. Vowels are thus classified as either oral or nasal respectively. Nasal vowels typically present on a spectrogram with anti-formants, which are lighter (lower intensity) bands of their component frequencies (harmonics; formants) than the formants of their oral vowel counterparts. The question can then be asked, is there a quantifiable difference in intensity between the formants of oral vowels and the anti-formants of nasal vowels?

This analysis hypothesizes that there is a difference in harmonic frequency amplitudes between oral and nasal vowels. To test this hypothesis, the Phoneme Dataset containing harmonic frequency amplitudes for instances of oral and nasal vowels was evaluated. The dataset was retrieved from: <https://datahub.io/machine-learning/phoneme#readme>

In this dataset, the variables V1-V5 refer to the amplitude values of the five component frequencies of each vowel instance. The “Class” variable refers to the vowel’s classification, where “Class 1” refers to nasal vowels and “Class 2” refers to oral vowels.

## Data Analysis

Importing necessary libraries.

library(tidyverse) # For data analysis.

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggpubr) # For data visualization.

Importing and cleaning the dataset of vowel instances.

# Read in dataset from downloaded csv file.  
phoneme\_data = read.csv("phoneme\_csv.csv")  
  
# Convert all negative values to absolute values, resulting in all positive amplitudes.  
clean\_data = abs(phoneme\_data)  
  
# Print preview of cleaned dataframe.  
head(clean\_data)

## V1 V2 V3 V4 V5 Class  
## 1 0.489927 0.451528 1.047990 0.598693 0.020418 1  
## 2 0.641265 0.109245 0.292130 0.916804 0.240223 1  
## 3 0.870593 0.459862 0.578159 0.806634 0.835248 1  
## 4 0.628439 0.316284 1.934295 1.427099 0.136583 1  
## 5 0.596399 0.015938 2.043206 1.688448 0.948127 1  
## 6 0.164735 0.642728 0.980619 0.386415 0.242046 1

Generating the average harmonic amplitude for each vowel instance.

# Add new column "Mean.Harmonic.Amplitude" containing arithmetic means of V1, V2, V3, V4, and V5 for each vowel instance.  
data.means = mutate(clean\_data, Mean.Harmonic.Amplitude = rowMeans(clean\_data))  
  
# Print preview of new dataframe.  
head(data.means)

## V1 V2 V3 V4 V5 Class Mean.Harmonic.Amplitude  
## 1 0.489927 0.451528 1.047990 0.598693 0.020418 1 0.6014260  
## 2 0.641265 0.109245 0.292130 0.916804 0.240223 1 0.5332778  
## 3 0.870593 0.459862 0.578159 0.806634 0.835248 1 0.7584160  
## 4 0.628439 0.316284 1.934295 1.427099 0.136583 1 0.9071167  
## 5 0.596399 0.015938 2.043206 1.688448 0.948127 1 1.0486863  
## 6 0.164735 0.642728 0.980619 0.386415 0.242046 1 0.5694238

Generating summary statistics for the average harmonic amplitudes, grouping vowel instances by vowel class. Class 1 refers to nasal vowels; Class 2 refers to oral vowels.

# Calculate summary statistics of oral and nasal vowels, including mean, standard error, and median.  
summary.stats = data.means %>%   
 group\_by(Class) %>% # Group data by vowel class variable.  
 summarize(  
 Class.Mean = mean(Mean.Harmonic.Amplitude), # Calculate mean within vowel class.  
 std.err = sd(Mean.Harmonic.Amplitude)/n(), # Calculate standard error within vowel class.  
 median = median(Mean.Harmonic.Amplitude) # Calculate median within vowel class.  
 )  
  
# Print summary statistics.  
summary.stats

## # A tibble: 2 × 4  
## Class Class.Mean std.err median  
## <int> <dbl> <dbl> <dbl>  
## 1 1 0.812 0.0000540 0.809  
## 2 2 1.03 0.000129 1.01

Running an Independent Two-Sample T-Test on the average harmonic amplitudes of the two vowel classes.

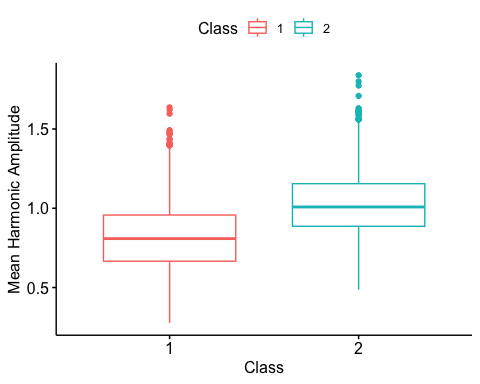
# Run Independent Two-Sample T-Test.  
tested = t.test(Mean.Harmonic.Amplitude ~ Class, data = data.means)  
  
# Print results of Independent Two-Sample T-Test.  
tested

##   
## Welch Two Sample t-test  
##   
## data: Mean.Harmonic.Amplitude by Class  
## t = -34.89, df = 2987, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0  
## 95 percent confidence interval:  
## -0.2257027 -0.2016841  
## sample estimates:  
## mean in group 1 mean in group 2   
## 0.8122418 1.0259351

## Visualization

Generating a box and whisker plot of the average harmonic amplitudes for vowel instances of each class.

# Plot Mean.Harmonic.Amplitude of each vowel instance, grouping by vowel class.  
plot = ggboxplot(data.means, x = "Class", y = "Mean.Harmonic.Amplitude", color = "Class", xlab = "Class", ylab = "Mean Harmonic Amplitude")  
  
# Print plot.  
plot



## Conclusion

Based on the above analysis, the hypothesis is supported. There is a statistically significant difference between the average harmonic amplitudes of nasal vowels and oral vowels. Thus, there is a difference in intensity between anti-formants and formants, suggesting acoustic divergences that are potentially used to audibly distinguish these sounds.