Speech Fluency Measurement for Aphasia Patients

Authors: Steffi Chern, Zihan Geng, Mason Kim
Advisor: Dr. Joel Greenhouse
Client: Dr. Davida Fromm

I. Introduction

According to the National Stroke Association, around 80,000 new cases of aphasia occur each year. Aphasia is a communication disorder that results from damage to parts of the brain that control language and speech. It is most commonly caused by a sudden stroke or head injury, but it can also develop gradually from a brain tumor. The type of aphasia depends on which part of the brain is injured, ranging from mild to severe, with Wernicke's aphasia being one of the most severe types. The level of severity is determined by the WAB Aphasia Quotient (AQ) score. The lower the AQ score, the more serious the condition is. A common symptom among people with aphasia is difficulty producing and understanding language, such as using made-up words, repeating certain words or phrases many times, making frequent pauses in a sentence, or failing to put words together into a coherent sentence. These difficulties can make it challenging for individuals with aphasia to perform daily activities, participate in social events, or maintain relationships with friends and family. Therefore, to improve the quality of life for those affected by this condition, we aim to investigate the differences in fluency between people with and without aphasia, and improve the reliability and validity of **fluency measurement** by quantifying the behaviors of people with aphasia.

II. Methods

1. Data

There were initially five datasets that we were given and each dataset contains about 202-559 patients. The datasets can be categorized as one of three specific tasks: procedural, expositional, and narrative. The procedural task is when the patient was asked to recreate the procedures to a certain activity. The expositional task is when the patient was asked to view a picture and utilize the context clues of the picture to formulate a story. Finally, the narrative task is when the patient was asked to retell a popular, fictional story. For our analyses in this report, we focused on the Cinderella (narrative task) dataset. The Cinderella task is performed when the patient was asked to retell the story of Cinderella. The data were collected through FLUCALC, which is an automated analysis tool that deciphers transcripts and records the syntax of the patient's speech. The transcripts were originally collected through numerous formal aphasia testing on patients and non-patients.

The datasets were collected from various Aphasia labs across the United States. Each subject is identified by their Aphasia type or as a control (no Aphasia). There were a few instances with missing or incorrect labels, thus we have discarded them to avoid errors or biases that may occur as a result.

2. Variables

There are a total of 69 variables in each dataset. However, our research specifically focuses on the several variables that are directly related to fluency measurement. Here is a list of the important variables we selected in this study:

- Age: age of participant in the form of (year; month)
- Sex: sex of participant
- Group: Aphasia type or control
- Aphasia Quotient score: the score that rates the overall severity of Aphasia symptom of the participant (empty for the control group)
- Number of utterances in the sample: number of total utterances in the sample
- Number of words in the sample: number of total words in the sample
- Words per minute (speech rate): number of total words in the sample divided by number of minutes of the sample
- Number of whole word repetitions: number of whole word repetitions in the sample
- Percentage of whole word repetitions: number of whole word repetitions in the sample divided by the number of total words in the sample
- Number of phonological fragments: number of phonological fragments in the sample
- Percentage of phonological fragments: number of phonological fragments in the sample divided by the number of total words in the sample
- Number of phrase repetitions: number of phrase repetitions in the sample
- Percentage of phrase repetitions: number of phrase repetitions in the sample divided by the number of total words in the sample
- Number of word revisions: number of word revisions in the sample
- Percentage of word revisions: number of word revisions in the sample divided by the number of total words in the sample
- Number of phrase revisions: number of phrase revisions in the sample
- Percentage of phrase revisions: number of phrase revisions in the sample divided by the number of total words in the sample
- Number of filled pauses: number of filled pauses in the sample
- Percentage of filled pauses: number of filled pauses in the sample divided by the number of total words in the sample
- Between utterance pause duration: total # of msec from the last word in an utterance to the first word in the next consecutive utterance for that speaker divided by the total number of consecutive utterances for that speaker

- Internal utterance duration: total duration of pauses between words or phrases with in an utterance divided by the total duration of utterances (time from first to last word of utterance)

3. Statistical Methods

Throughout our research, we employed various techniques to conduct statistical analysis. As part of the exploratory data analysis, we first examined the histograms of quantitative variables individually to gain insights into the distribution and nature of the data. Next, we analyzed the variables by visualizing boxplots to explore the distribution across different Aphasia groups for each variable. To determine whether the means of these Aphasia groups were significantly different for each feature, we conducted an analysis of variance (ANOVA) test at the 95% confidence level. However, this only informed us whether or not at least one Aphasia group had a significantly different mean than the other Aphasia groups. In conjunction with the ANOVA test, we conducted the Tukey's Honest Significant Difference (HSD) test to identify specifically which pairs of Aphasia groups had significantly different means in this multiple comparison setting.

To gain a deeper understanding of the differences between each Aphasia group and how each feature contributed to these differences, we performed Principal Component Analysis (PCA) on the important variables we pinpointed in the previous section. Since PCA is a method for dimensionality reduction, we can use it to visualize how each variable contributes to the data simultaneously.

Finally, we used Gaussian Mixture Models (GMM) based on the results from the PCA to better visualize the separation of the Aphasia groups. GMM identifies clusters of subjects, assigns each patient to a cluster, and provides us with a probability estimate of each cluster that a patient belongs to. Specifically, we used a soft clustering approach in which each patient is assigned a probability of belonging to each of the groups. This is in contrast to hard clustering, which assigns each patient to a single group. One challenge when using GMM is determining the optimal number of clusters to fit the data. To address this, we used the Bayesian Information Criterion (BIC), which is calculated by balancing the fit of the model with the complexity of the model. By performing GMM and choosing the optimal number of clusters, we can gain a deeper understanding of the structure of the data and identify underlying patterns that may not be apparent from the PCA results alone.

III. Results

1. Exploratory Data Analysis

We start by plotting the histograms for all continuous variables from the Cinderella dataset except Age and Aphasia Quotient score as part of the univariate data analysis. Note that all the analysis below all only pertain to the Cinderella task.

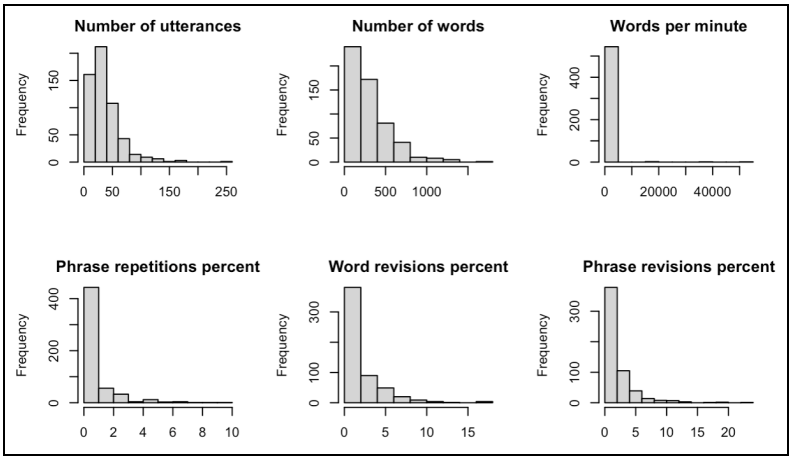

Figure 1: Histograms of variables

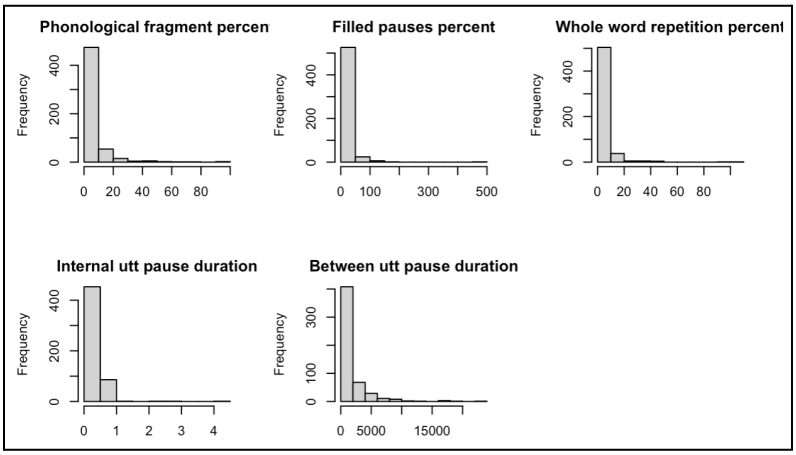

Figure 2: Histograms of variables

We observed that the distribution of all selected variables except for Age, Sex, Group and Aphasia Quotient scores are strongly right skewed, with a few large outliers, as shown in Figure 1 and Figure 2. A log transformation is performed on all right skewed variables. The distribution of those variables are closer to a normal distribution after log transformation, as shown in Figure 3 and Figure 4. We continue to use the log transformed variables in our further analysis and modeling.

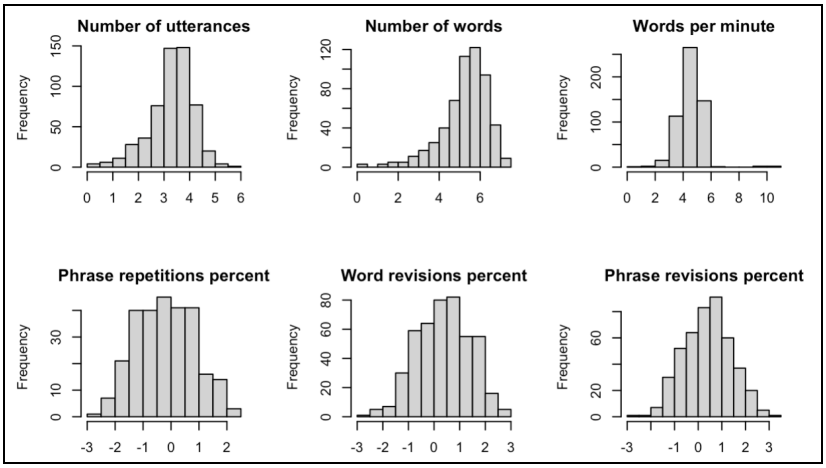

Figure 3: Histograms of variables (log transformed)

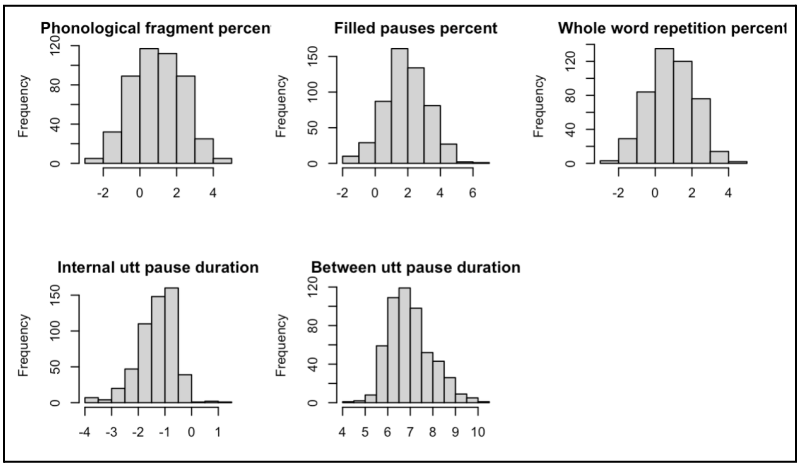

Figure 4: Histograms of variables (log transformed)

Next, we used side by side boxplots to explore the distribution of each continuous variable across different groups (including the Aphasia types and the control group). From the boxplots, we observe that for the variables number of utterances, number of words and words per minute, the control group has a higher median and overall distribution than the other Aphasia groups, as in figure 5, 6 and 7. The Broca Aphasia group has the lowest median and distribution across all groups for these three variables.

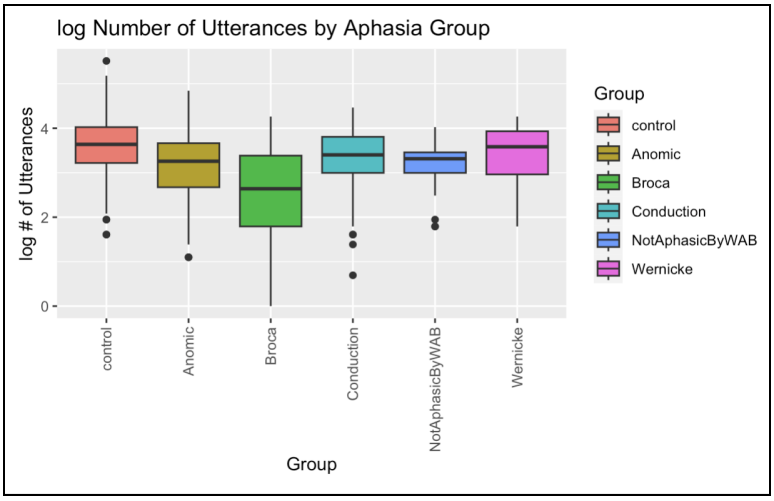

Figure 5: Log number of Utterances by Aphasia Group

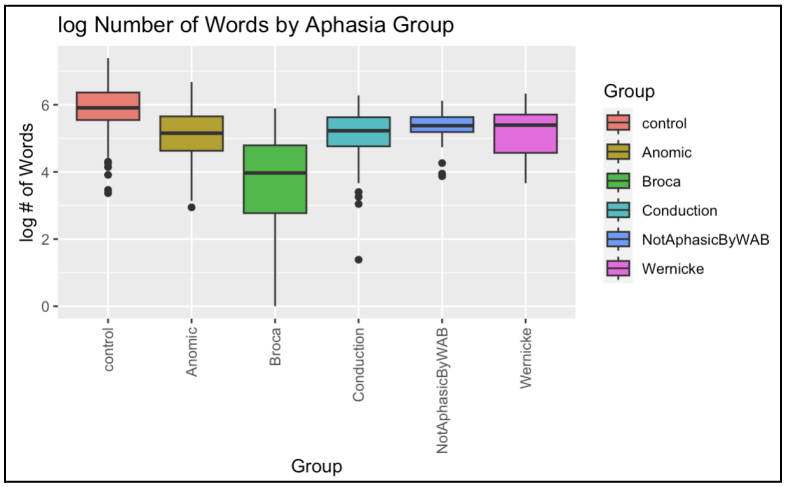

Figure 6: Log number of Words by Aphasia Group

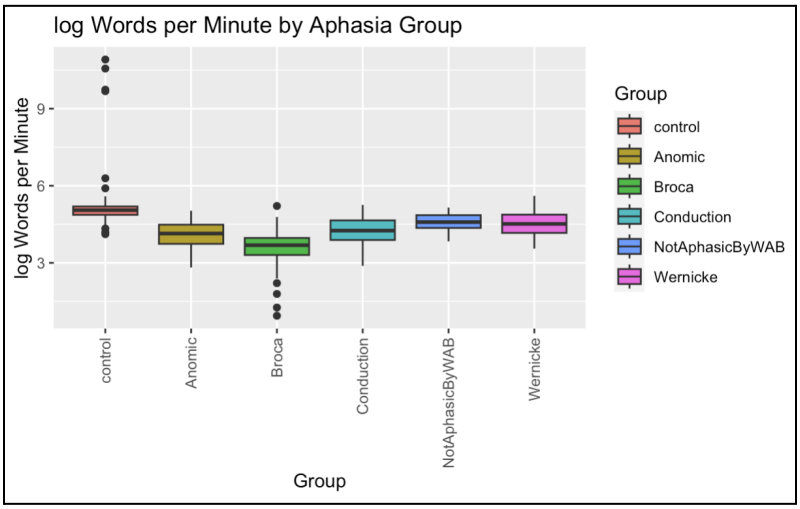

Figure 7: Log number of Words per Minute by Aphasia Group

For the variables percentage of whole word repetitions, percentage of phonological fragments, percentage of phrase repetitions, percentage of word revisions, percentage of phrase revisions, percentage of filled pauses, between utterance pause duration and internal utterance duration, the control group have a lower median and distribution than the other Aphasia groups, as in figure 8, 9, 10, 11, 12, 13, 14 and 15. In figure 8 and 10, the Broca Aphasia group has a low median and 25% quartile in its distribution for the variables percentage of Phrase Revisions and percentage of Phrase Revisions. We observed a large proportion of these variables for the Broca group are 0, which may result in a lower median for the Broca group. For the other variables, the Broca group has the highest median, followed by the Anomic and Conduction group. Some interesting patterns we have noticed is that the distribution of the important variables for the NonAphasicByWAB group is more similar to that of the other Aphasia groups and different than the control group.

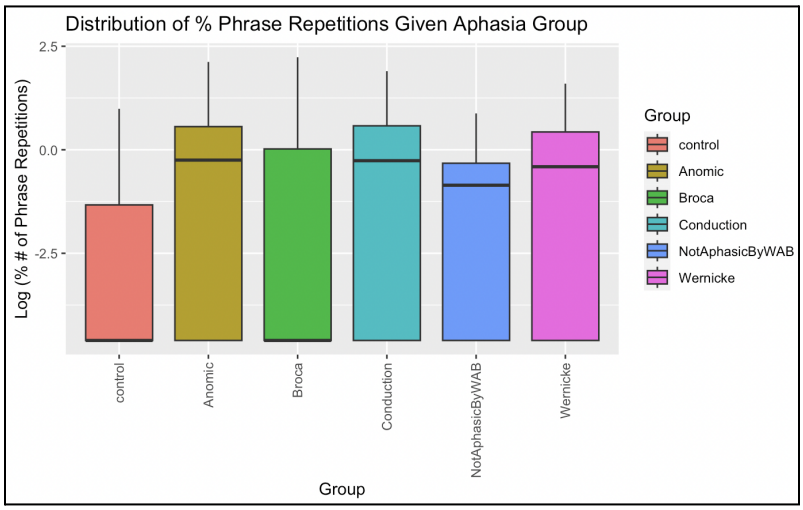

Figure 8: Log percent of Phrase Repetitions by Aphasia Group

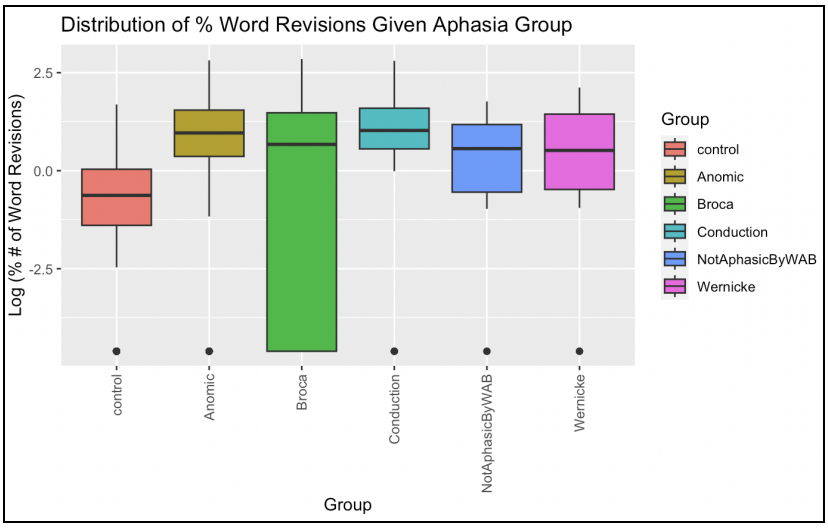

Figure 9: Log percent of Word Revisions by Aphasia Group

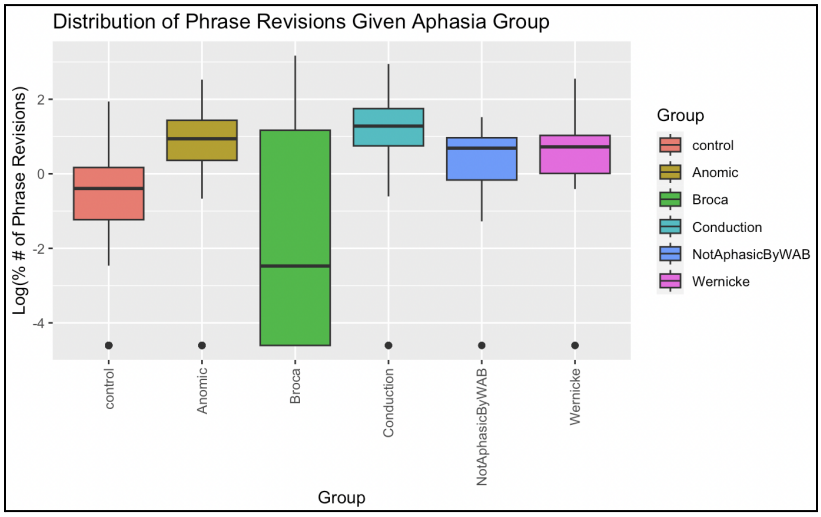

Figure 10: Log percent of Phrase Revisions by Aphasia Group

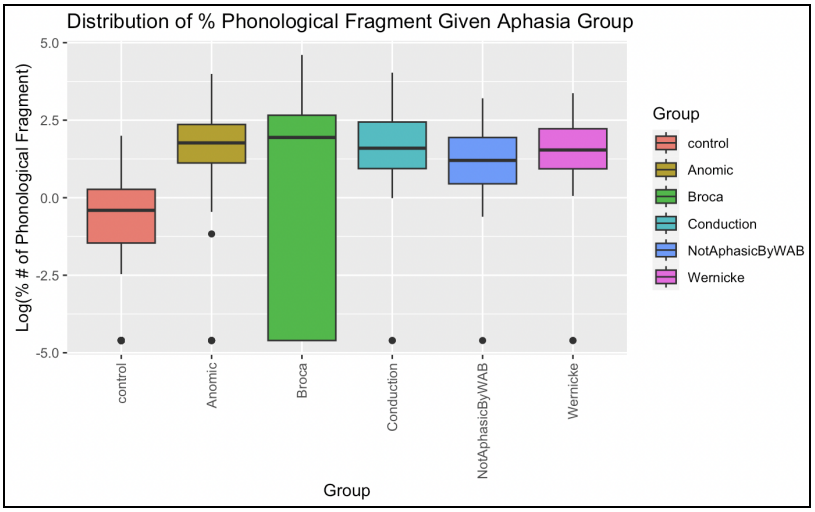

Figure 11: Log percent of Phonological Fragments by Aphasia Group

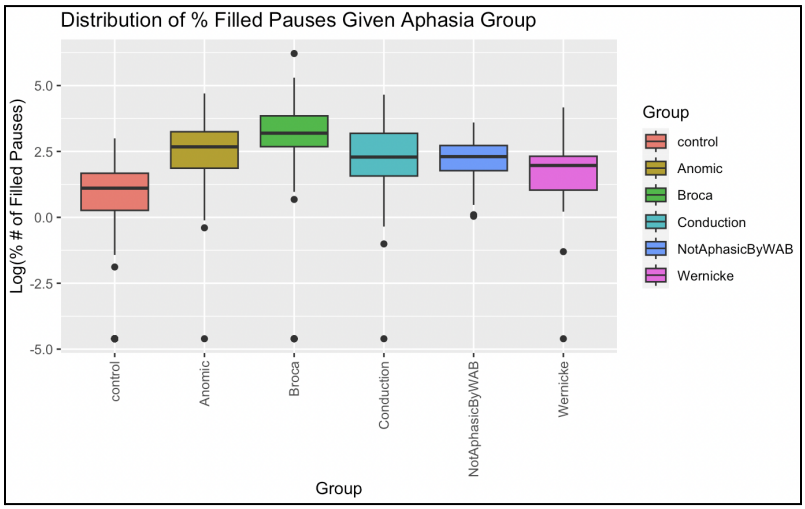

Figure 12: Log percent of Filled Pauses by Aphasia Group

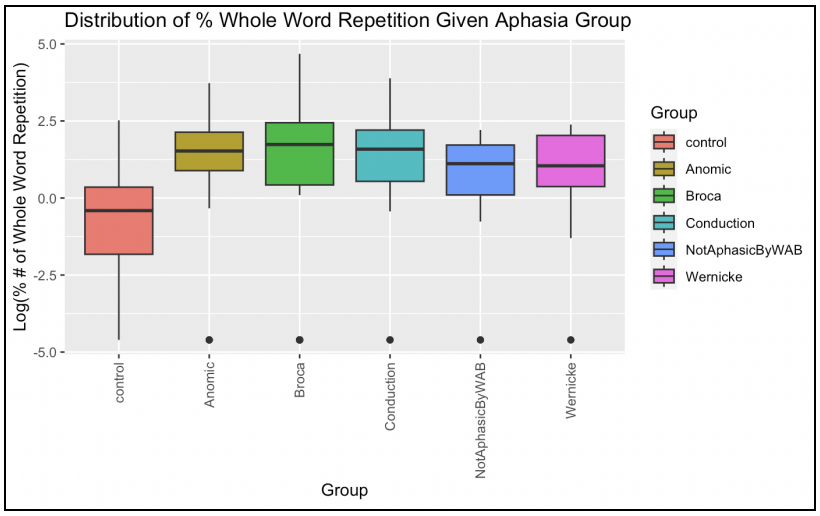

Figure 13: Log percent of Whole Word Repetition by Aphasia Group

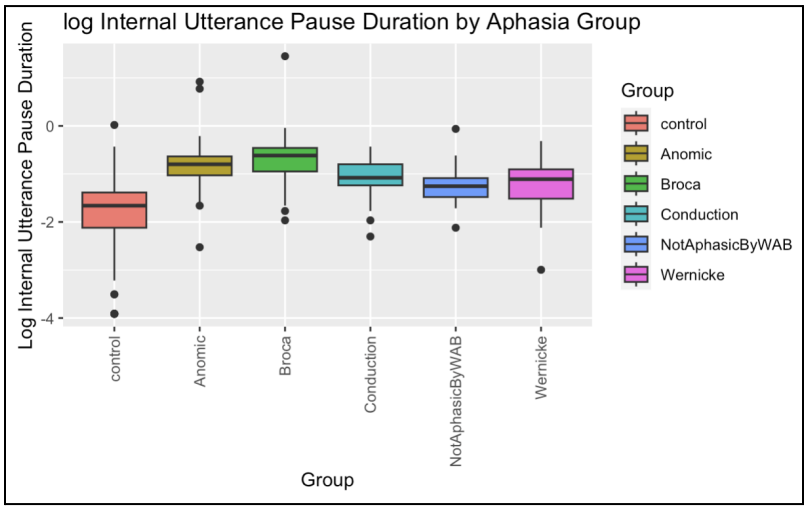

Figure 14: Log Internal Utterance Pause Duration by Aphasia Group

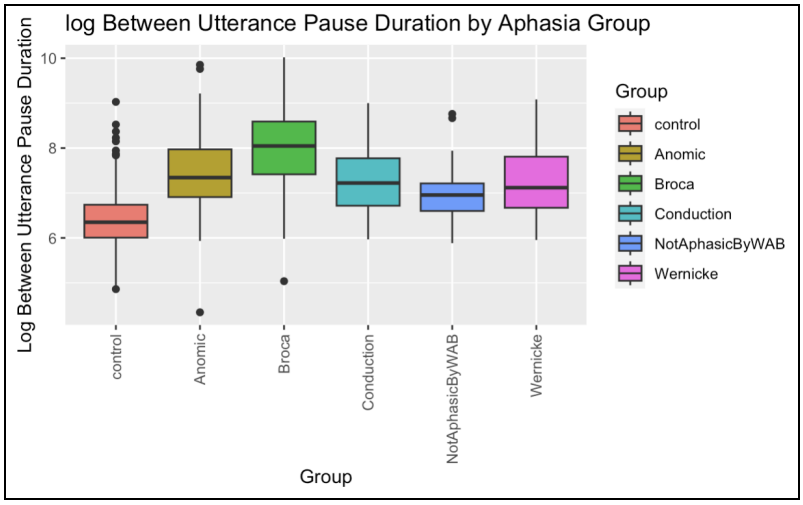

Figure 15: Log Between Utterance Pause Duration by Aphasia Group

2. ANOVA and HSD Tests

We performed the ANOVA tests on all the log transformed percentage variables and the pause duration variables. Since we obtained a p-value of < 0.05 after running the ANOVA test on all the selected variables, we conclude that at least one Aphasia group has a significantly different mean than that of the other Aphasia groups for all the log transformed selected variables.

We computed the 95% confidence intervals of the difference in the mean of each pair of Aphasia groups (including the control group) for these variables with the HSD test.

Figure HSD-1: 95% family-wise confidence interval of log percent of phrase repetitions

For the variable log percentage of phrase repetitions in figure HSD-1, the control group has significantly different mean than all Aphasia groups. The Conduction and

Broca group have significantly different means than the NotAphasicbyWAB group. The control group has significantly lower mean than all groups. The NotAphasicbyWAB group has significantly lower mean than the Broca group and the Conduction group.

Figure HSD-2: 95% family-wise confidence interval of log percent of word revisions

For the variable log percentage of word revisions in figure HSD-2, the control group has significantly different mean than all Aphasia groups. The control group has a significant lower mean than all groups.

Figure HSD-3: 95% family-wise confidence interval of log percent of phrase revisions

For the variable log percentage of phrase revisions in figure HSD-3, the control group has significantly different mean than all Aphasia groups. The Conduction group

has a significantly different mean than the NotAphasicbyWAB group. The control group has a significant lower mean than all groups. The NotAphasicbyWAB group has a significantly lower mean than the Conduction group.

Figure HSD-4: 95% family-wise confidence interval of log percent of phonological fragments

For the variable log percentage of phonological fragments in figure HSD-4, the control group has a significantly different mean than all Aphasia groups. The control group has a significant lower mean than all groups.

Figure HSD-5: 95% family-wise confidence interval of log percent of filled pauses

For the variable log percentage of filled pauses in figure HSD-5, the control group has significantly different mean than all Aphasia groups except for the Wernicke group. The control group has a significant lower mean than all groups.

Figure HSD-6: 95% family-wise confidence interval of log percent of whole word repetitions

For the variable log percentage of whole word repetitions in figure HSD-6, the control group has significantly different mean than all Aphasia groups. The control group has a significant lower mean than all groups.

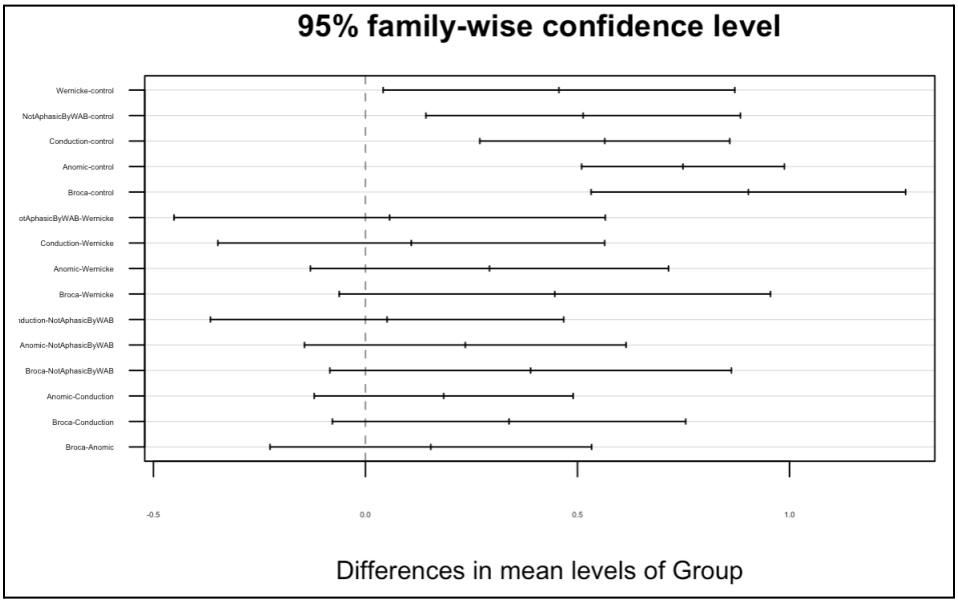

Figure HSD-7: 95% family-wise confidence interval of log internal utterance pause duration

For the variable log internal utterance pause duration in figure HSD-7, the control group has significantly different mean than all Aphasia groups. The control group has a significant lower mean than all groups.

Figure HSD-8: 95% family-wise confidence interval of log between utterance pause duration

For the variable log between utterance pause duration in figure HSD-8, the control group has significantly different mean than all Aphasia groups, and the Broca group has significantly different mean than all the other groups. The control group has significantly lower mean than all groups and the Broca group has significantly higher mean than all groups.

From the ANOVA tests, we could observe that at least one group has a significantly different mean of each percentage variable and the pause duration variables across different Aphasia groups. From the HSD tests, for all variables, the control group has a significantly different mean. For the variable between utterance pause duration, both the control group and the Broca group have significantly different mean than other groups, with the control group having lower mean and the Broca group having higher mean.

3. Principal Component Analysis (PCA)

We performed principal component analysis to help us determine which variables contribute most to explaining the variability in fluency. By examining the loadings of the variables on the principal components, we can identify the variables that have the highest weight in explaining the variability in the data, as shown in figure 16. Positive loadings indicate that the variable increased with the associated principal component, while negative loadings indicate that the variable decreased with the principal component. The

magnitude of the loadings indicate the strength of the relationship between the variable and the principal component. PC1 captures around 34.76% while PC2 captures around 25.29% of the total variance in the data (refer to Figure 17 for more details).

Based on figure 16, we observed that the log-transformed variables Words Per Minute, Number of Words, and Number of Utterances had the highest positive loadings on the first principal component, while Internal and Between Utterance Pause Duration had the highest negative loadings on the first principal component.

The log-transformed variables that contributed the most to the second principal component were Percentage Phrase Revisions, Percentage Word Revisions, and Number of Utterances.

	PC1	PC2
Log Number of Utterances	0.27	0.40
Log Number of Words	0.36	0.37
Log Words Per Minute	0.42	0.09
Log % Whole Word Repetition	-0.26	0.36
Log % Phonological Fragment	-0.28	0.33
Log % Phrase Repetitions	-0.16	0.34
Log % Word Revisions	-0.16	0.38
Log % Phrase Revisions	-0.07	0.41
Log % Filled Pauses	-0.35	0.07
Log Internal Utterance Pause Duration	-0.40	-0.04
Log Between Utterance Pause Duration	-0.37	-0.16

Figure 16: Correlation Between Each Variable and Principal Components

Comparing multiple groups of Aphasia patients based on various features can make it difficult to visualize all the features at once. However, by performing PCA on the features and projecting the data onto the first two principal components, we can create a scatter plot where each point represents an individual patient. This allows us to visualize how each Aphasia group is clustered based on the features and can help identify patterns and differences among the groups, as shown in figure 17.

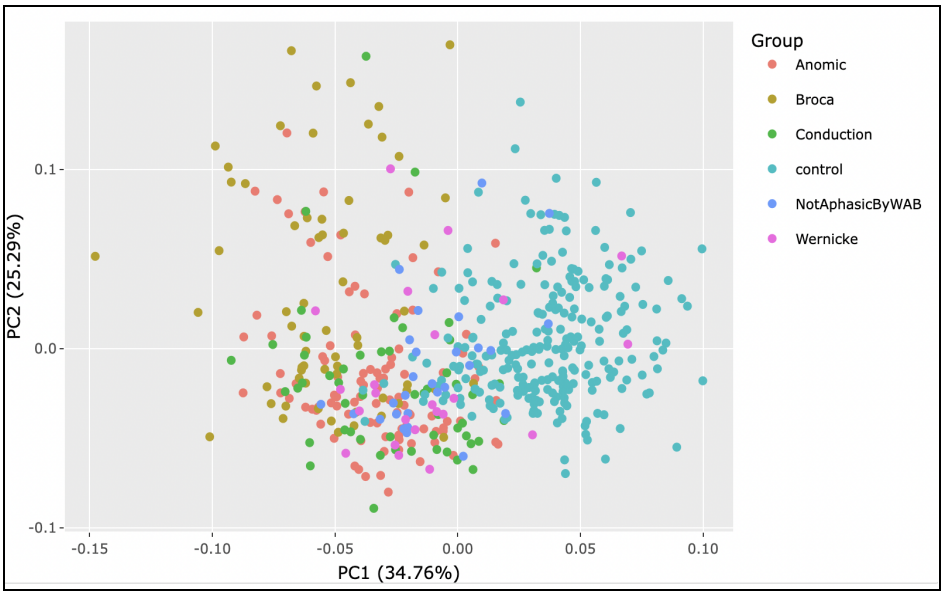

Figure 17: PCA Scatter Plot by Aphasia Groups

From Figure 17, we observed that the first two principal components captured around 60% of the total variance in the Cinderella data. The scatter plot showed relatively clear clusters for the control and Broca's Aphasia groups. Some patients who were labeled as having either NotAphasicByWAB or Wernicke Aphasia showed similar characteristics to patients in the control group. Most of the other Aphasia groups, namely Anomic, Conduction, NotAphasicByWAB, and Wernicke, were clumped together in the lower middle part of the plot.

Patients labeled as belonging to the control group have positive scores on the first principal component (PC1), indicating that they differ from the Aphasia groups in terms of this component. However, their scores on the second principal component (PC2) vary widely along the y-axis. Patients labeled as belonging to the Broca's Aphasia group have negative scores on PC1, indicating that they differ from the control group in terms of this component. However, their scores on PC2 also vary widely. In the next section, we investigate the grouping of patients based on the PC scores more formally.

4. Gaussian Mixture Models (GMM)

By fitting a GMM to the data, we can assign each patient based on their PC1 and PC2 scores to the Aphasia group that he/she most likely belongs to. The algorithm assumes that the data is generated by a mixture of Gaussian distributions, where each distribution corresponds to a different Aphasia group in the data. As shown in figure 18, if we let GMM fit 6 clusters and assign to each one the most prevalent Aphasia type in that cluster, we see three major clusterings: the control group, the Broca's group, and the rest of the Aphasia groups.

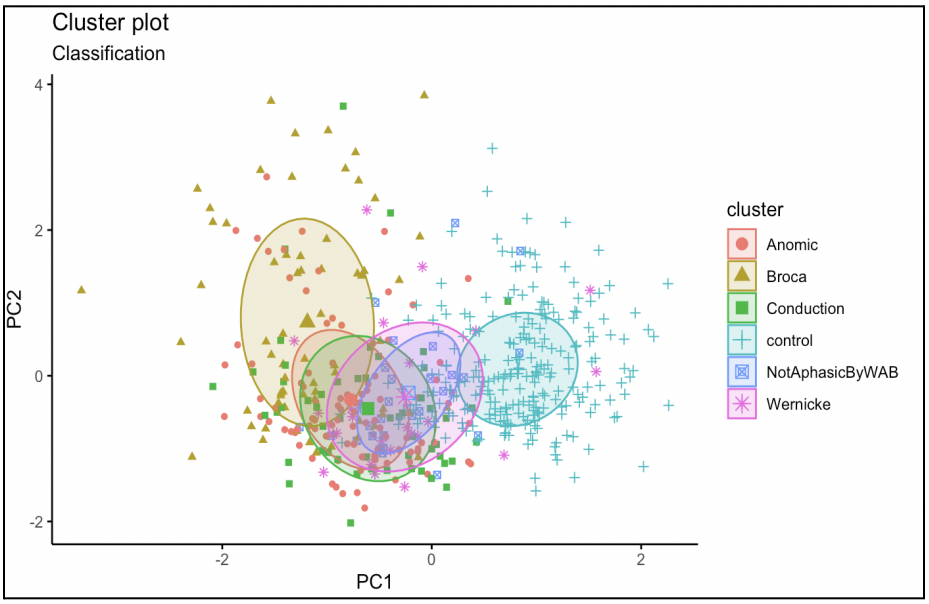

Figure 18: Clustering with Actual Aphasia Group Labels

We also used GMM to determine the optimal number of clusters to fit to the data without knowing what the actual Aphasia group labels are. This is chosen based on the BIC, which turned out to be three clusters, as seen in figure 19. This aligns with our PCA results and the previous graph, which also suggest that the data can be roughly clustered into three groups corresponding to the control group, the Broca's group, and the remaining Aphasia groups.

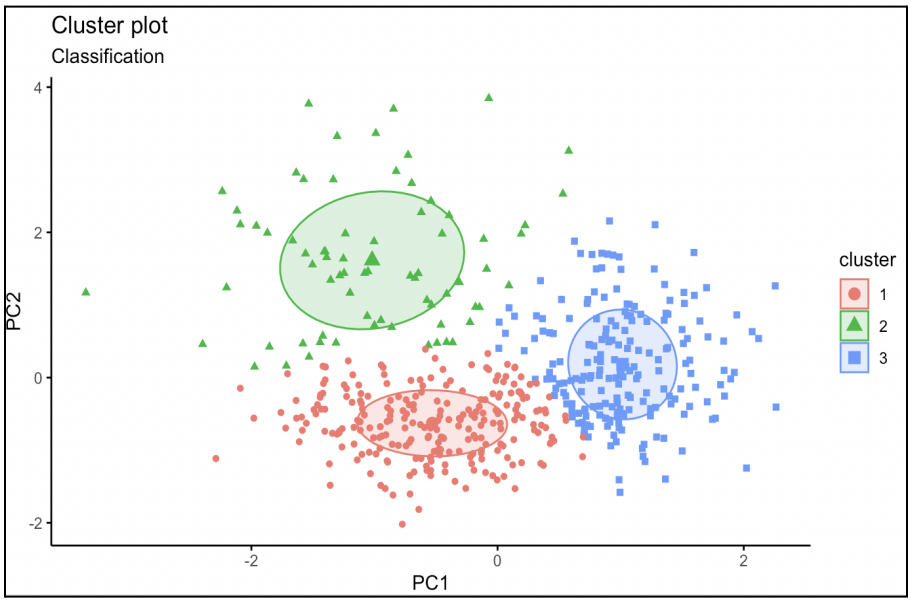

Figure 19: GMM Clustering Without Actual Aphasia Group Labels (based on BIC)

To evaluate the performance of the GMM in grouping Aphasia patients, we can use a confusion matrix, which compares the predicted cluster assignments from the GMMs with the actual Aphasia types of the patients. The rows of the matrix in Figure 20 correspond to the true Aphasia types, and the columns correspond to the predicted cluster assignments from the GMM in Figure 18. The entries in the matrix represent the number of patients that are correctly or incorrectly classified. We obtained an absolute classification error of 22.81%, which is calculated by dividing the total number of misclassified subjects by the total number of subjects. We noticed that the GMMs have a relatively high proportion of correctly classified patients as control or Broca's Aphasia, which is likely due to the distinct fluency speech characteristics of these groups compared to the other Aphasia groups. This suggests that the GMMs have a higher probability of correctly distinguishing these groups from the other groups.

	Predicted					
Class	Anomic	Broca	Conduction	control	NotAphasicByWAB	Wernicke
Anomic	76	8	4	6	1	4
Broca	7	50	5	0	2	2
Conduction	22	4	20	3	1	3
control	6	0	1	232	3	2
NotAphasicByWAB	7	0	1	12	7	1
Wernicke	6	1	3	2	0	11

Figure 20: Confusion Matrix: Number of Correct/Incorrect Predictions

IV. Discussion

In this study, we observed that the different Aphasia types and the control group have significantly different means for the fluency predictors of interest. From the HSD plots, we found that the control group has a significantly different mean than that of most of the other Aphasia groups, for all the predictors we investigated.

Based on the results from the PCA, we observed that the first principal component is related to differences in the speed and quantity of speech, whereas the second principal component is related to differences in the quality and accuracy of speech. We also found that some patients who were labeled as having either NotAphasicByWAB or Wernicke Aphasia showed similar speech characteristics to patients in the control group, suggesting that their symptoms could be relatively mild with respect to speech fluency. Our findings from the PCA provide insights into which variables are most important in explaining the variability in fluency.

From the GMM results, we found that the Aphasia patients could be classified into three major groups. This clustering corresponds to the severity of each Aphasia group with regards to speech fluency. The three groups are the control group, the Aphasia groups with better fluency, and the Aphasia groups with worse fluency. The clustering

also provides an explanation for why we observed the Broca's group in a cluster on its own, since this group represents the worst speech fluency group. As we shift our attention to the confusion matrix, it appears that there may be underlying patterns in speech fluency that are not fully accounted for by the current WAB Aphasia Quotient (AQ) score. The confusion matrix reveals that some patients who were classified as having a particular Aphasia type based on their AQ score actually exhibit speech characteristics that are more similar to a different Aphasia group or control group. This may indicate that the current system of classifying Aphasia based on the AQ score is not capturing all of the nuances and complexities of the condition. Further research is needed to investigate these potential patterns and to develop more accurate and comprehensive methods for classifying and characterizing Aphasia.

There are some limitations in our study. Firstly, the data used in our analysis were collected from multiple Aphasia labs, which may introduce biases and inconsistencies in the dataset and therefore limit the generalizability of our findings. Additionally, the lack of data for some of the Aphasia groups, including the Transmotor and Transensory groups, may have limited the ability to detect clear patterns in the clustering. We also observed that the Broca's Aphasia and Transmotor Aphasia groups correspond to the groups with worse fluency, but due to the limited number of observations for the Transmotor group, we excluded them from the analysis, leaving the Broca group as the only group with the worst fluency. It is also important to note that the clustering may not be perfect since GMM makes certain assumptions about the data that may not be accurate. These limitations suggest that future studies with larger sample sizes and more consistent data collection methods may be needed to further investigate the clustering of Aphasia patients and to develop more accurate methods for characterizing the condition.

For future steps, we plan to examine datasets from other tasks, such as Cat, Sandwich, Stroke, and Illness, which assess different speech skills of the participants. Our current results are based on the Cinderella dataset, and performing the analysis on other datasets may provide more insight into the relationship between different speech skills and Aphasia types. Moreover, we used unsupervised learning techniques in our analysis, but we also plan to explore supervised learning methods to build models that can accurately classify Aphasia patients based on their speech characteristics. These future steps will help us expand our understanding of Aphasia and develop more comprehensive methods for diagnosing these patients.

V. Acknowledgements

The authors express their special thanks to the faculties in Carnegie Mellon University -- Dr. Joel Greenhouse, Dr. Davida Fromm, and Dr. Zach Branson for their valuable feedback and suggestions throughout the research process.

VI. References

Brian MacWhinney, “AphasiaBank”

VII. Appendix

1. Contingency Table

During our initial research on the dataset, we observed from the boxplots that the Broca's Aphasia group had very different distributions compared to the other groups for each variable. We investigated further and discovered that this was due to many patients in the Broca's group having variable values equal to zero. Figure 21 provides more information about the percentage of patients in each group for each variable that had values of zero.

	Anomic <dbl>	Broca <dbl>	Conduction <dbl>	control <dbl>	NotAphasicByWAB <dbl>	Wernicke <dbl>
mor_Utts	0.000	0.000	0.000	0.000	0.000	0.000
mor_Words	0.000	0.000	0.000	0.000	0.000	0.000
words_min	0.000	0.000	0.000	0.000	0.000	0.000
Phonological_fragment_percent	0.051	0.258	0.038	0.209	0.036	0.087
Phrase_repetitions_percent	0.313	0.652	0.377	0.594	0.393	0.304
Word_revisions_percent	0.131	0.288	0.094	0.180	0.107	0.130
Phrase_revisions_percent	0.121	0.470	0.094	0.148	0.107	0.174
Filled_pauses_percent	0.010	0.030	0.019	0.074	0.000	0.043
IW_Dur_Utt_Dur	0.000	0.000	0.000	0.000	0.000	0.000
No_Switch_Dur_Num_No_Switch	0.000	0.000	0.000	0.000	0.000	0.000

Figure 21: Contingency Table for Cinderella Dataset

2. PCA

To clarify what PCA does, it is a technique used to reduce the number of variables in a dataset while preserving as much of the original information as possible. The result is a new set of variables, called principal components, that are linear combinations of the original variables. These principal components explain the majority of the variation in the data and allow us to visualize the data in a lower-dimensional space, making it easier to interpret and analyze. Here, we display an additional plot regarding the contribution of each variable to the principal components using the “factoextra” R package. Variables that are close to each other are positively correlated, while variables that are on opposite sides are negatively correlated.

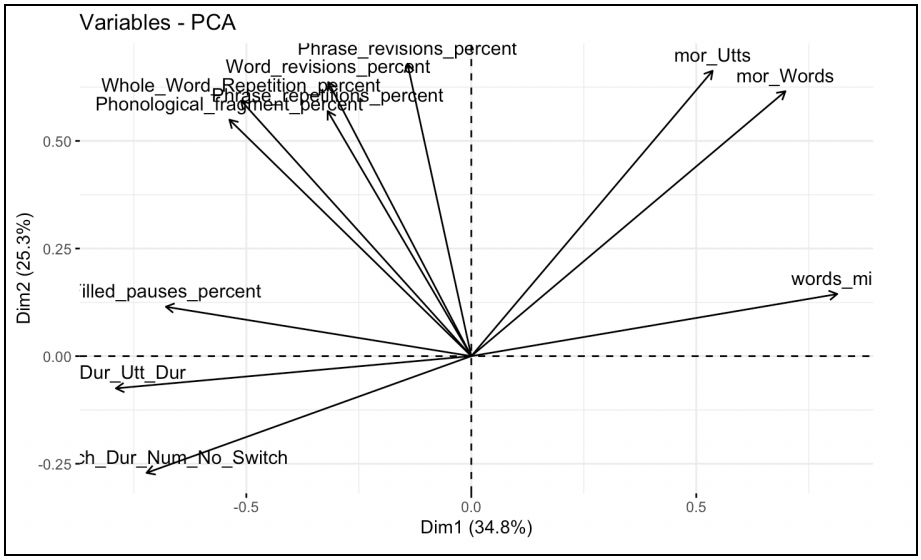

Figure 22: Contribution of Each Variable to Each Principal Component (PC)

3. GMM and BIC

GMM is a statistical model used mainly for clustering. It assumes that the data is generated from a mixture of several Gaussian distributions, where each distribution represents a cluster, and that all features are independent. GMM can identify clusters in data.

BIC is a statistical measure used for model selection. It balances the goodness of fit of a model with its complexity, allowing us to determine the optimal number of clusters in a GMM model. BIC takes into account the likelihood of the data given the model and a penalty term for the number of parameters (number of clusters in this case) in the model. This encourages selection of models that are simpler but accurate. Based on the “mclust” R package we used to calculate, the model with the highest BIC value is considered the best model.