Modelling the dynamics of speaking fluency in second language learners

Khalajzadeh, Niloofar

Modelling the Dynamics of Speaking Fluency in Second Language Learners

Niloofar Khalajzadeh

Daily Supervisor: **Dr. Jelle Goeman**, Biomedical Data Sciences, Leiden University Medical Centre
Second Supervisor: **Dr. Nivja de Jong**, Centre for Linguistics, Leiden University

6 June , 2024

MASTER THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

The present study investigates the dynamics of speaking fluency in second language learners, emphasizing the variability of pause duration throughout speaking tasks. Recognizing that traditional static measures of fluency may overlook the dynamic nature of spoken language, this research employs Generalized Additive Mixed Models (GAMMs) to capture the non-linear relationships and individual variability among participants. The dataset comprises speech recordings analyzed for silent and filled pauses, previous and next word frequencies, and speech rate.

The analysis reveals significant patterns in pause durations linked to cognitive load and linguistic complexity. Notably, the study identifies that pause durations are influenced by the complexity of words preceding and following the pauses, with filled pauses being particularly sensitive to upcoming complex words. Additionally, the study highlights the challenges posed by varying task lengths and the difficulty in distinguishing between different types of pauses.

Hierarchical clustering method is employed to group participants based on the predicted pause duration, revealing distinct clusters that reflect adaptive strategies in managing speech flow and cognitive load.

The findings contribute to a deeper understanding of L2 fluency, offering insights that can enhance language assessment and educational practices. Future research should also consider incorporating speaker cognition perspectives such as pronunciation, lexical choice, and syntactic complexity to further elucidate the cognitive processes underlying speech production in L2 learners.

Keywords — S peaking Fluency, Second Language Learners, GAMM, Pause Durations, Hierarchical Clustering

Contents

Chapter 1

Introduction

Speaking fluency, a key component in second language (L2)[1] proficiency, has traditionally been characterized by its static nature in language assessment [1]. It is usually quantified through measures like pause frequency, speech rate, and lexical complexity during performance snapshots. However, the static approach overlooks the inherent dynamicity of spoken language. The cognitive processes involved in language production *conceptualization*, *formulation*, and *articulation* are not constant but vary throughout a speaking task [2]. This study aims to advance the assessment of speaking fluency by modeling its dynamics, thus reflecting the true complexity of language performance in L2 speakers.

Speaking fluency includes the ease and efficiency of the psycholinguistic processes involved in speaking [3]. This project adopts the view that fluency extends beyond smooth speech production to include dynamic changes in speaking patterns as responses to varying cognitive demands over time [4][5]. For instance, fluency dips when the speaker encounters linguistic challenges or when generating new ideas, leading to pauses or slowed speech [6].

This research underscores the dynamic nature of fluency. For example, native speakers demonstrate fluctuations in speaking rates and pause patterns within a single performance, likely reflecting changes in cognitive load [7]. Less is known about such fluency dynamics in L2 speakers, who face additional challenges due to the reduced automatization of language processing [6].

The project's primary objective is to identify and model time trends in L2 fluency measures specifically, whether speakers show tendencies for increasing or decreasing pauses as a speaking task progresses. It is aimed to extend the understanding of L2 fluency beyond static descriptions by accounting for its variability over time.

Utilizing a corpus of L2 speech performances, Generalized Additive Models (GAMs) will be employed to explore time-dependent patterns in fluency. Key variables, such as

[1]Refers to any language learned in addition to a person's first language

the duration of silent[2] and filled[3] pauses, will be examined for dynamic trends. Furthermore, it be will performed a cluster analysis to investigate whether incorporating these dynamics into statistical models improves the prediction of pause duration.

This research is poised to make contributions to the field of second language acquisition (SLA) and language testing. It follows dynamic nature of fluency, which suggest that fluency varies with the syntactic and lexical challenges faced by speakers [8][9]. By recognizing fluency as a dynamic construct, this project will pave the way for more nuanced and effective assessment methodologies in L2 speaking tests.

The pursuit of modelling the dynamics of speaking fluency in L2 learners is not just an academic exercise; it holds practical implications for language education and assessment. As fluency is a salient feature of language proficiency, providing a more detailed and dynamic analysis will enhance the validity of speaking assessments and contribute to a deeper understanding of L2 speaking processes. This research, therefore, represents an essential step forward in the intersection of SLA theory, psycholinguistics, and language assessment practices.

[2]A break in conversation or speech during which there is no talking or noise

[3]A non-silent pause in an otherwise fluent speech, where instead of a silent pause there is a filler. The filler can be non-lexical or semiarticulate utterances such as *huh*, *uh*, *erm*, *um*, or *hmm*.

Chapter 2

Related Work & Prerequisites

In this section, some necessary linguistic and statistical concepts are introduced.

2.1 Linguistic Context

Communication through speech is fundamental to human interaction and is remarkably complex, both in acquisition and execution. Mastering a language, let alone a second or third, is a significant cognitive feat expected in many facets of modern society, be it for professional advancement or social integration [10].

Language production involves a series of connected stages, beginning with conceptualizing the ideas to be expressed, followed by formulating the linguistic structure of these ideas, and culminating in the articulation of sounds [2]. During this process, speakers often face challenges that can disrupt their speech flow, leading to various forms of disfluencies such as silent pauses, where speech halts entirely, and filled pauses, marked by placeholders. These interruptions are particularly prevalent in L2 speech, where less automated language processing can increase their occurrence [6].

The study of fluency in L2 speech revealed that it is not merely the presence but the distribution of such disfluencies that may be indicative of a speaker's proficiency. Native speakers typically produce language in chunks, which affords them the cognitive space to plan successive segments of speech. This ability results in pauses predominantly occurring at clause boundaries. In contrast, L2 speakers, who often lack an extensive repertoire of pre-fabricated language chunks, may experience pauses within clauses as they grapple with real-time language construction [11].

Additionally, it is posited that nouns, which often introduce new information, may slow down speech more than verbs, suggesting a universal pattern in how language is processed and produced across different linguistic and cultural backgrounds. This is evidenced by research showing slower speech articulation before nouns and a higher likelihood of pauses in the proximity of noun onset [12].

This research has also explored the moments preceding the articulation of lexical items, with a focus on nouns and verbs. Studies suggest a significant planning phase

of around 500 milliseconds before the onset of such words, aligning with the idea that conceptualizing a single content word might require up to 600 milliseconds. This phase of preparation often results in a measurable deceleration of speech rate, alongside a higher likelihood of encountering pauses, defined as intervals exceeding 150 milliseconds. These observations contribute to the broader understanding of fluency dynamics in language processing[12].

This focus on the 600 ms window preceding word onset also played a pivotal role in this thesis's predictive models. The duration of pauses was postulated to be influenced by the complexity of upcoming words, considering the cognitive demands they impose on the speaker. The chosen modelling approach is designed to capture these subtleties, using this time frame as a predictor for pause duration in statistical analyses.

In light of these findings, this thesis sets that cognitive fluency is a dynamic construct influenced by the fluctuating cognitive demands placed upon the speaker. The goal is to investigate the presence of time trends in L2 fluency, such as the propensity for pauses to increase or decrease during speech tasks, which reflect the challenges faced in the conceptualization and formulation of the message [13][14].

2.1.1 PRAAT

For the empirical analysis, the thesis utilizes data obtained via PRAAT.

PRAAT is a powerful software tool developed by Paul Boersma and David Weenink at the University of Amsterdam, designed for the analysis of speech in phonetics. It is widely used by linguists, speech scientists, and those involved in language learning research to analyze and manipulate speech and sound recordings. PRAAT allows users to perform a broad range of acoustic analyses, synthesize speech, and create high quality visual representations of sound, such as spectrograms, pitch tracks, and intensity curves.

In the context of measuring fluency in second language learners, PRAAT is utilized to develop scripts that can automatically analyze aspects of speech fluency. This includes measuring silent pauses, filled pauses, and the speed of speaking. The use of PRAAT for these purposes eliminates the need for transcribing speech or manually annotating speech data to measure these fluency aspects.

PRAAT scripts can be programmed to automatically detect silent pauses and filled pauses within speech. The detection algorithms can use various acoustic features such as duration, pitch, and vowel quality. For filled pauses, characteristics like long duration, low pitch, and vowel qualities similar to a schwa [1] or a back mid-open vowel indicate a higher likelihood of a segment being a filled pause. Additionally, the "pronounced lazily" criterion, where less effort in vowel pronunciation, also helps in identifying filled pauses [6].

The speed of speaking is another critical aspect of fluency that PRAAT can measure. This is typically done by calculating the articulation rate, which involves counting the number of syllables or words spoken within a certain time frame. By automating this

[1] A schwa is a vowel sound in English that is typically unstressed and sounds like a quick, relaxed "uh"

process, the script can efficiently provide metrics related to the speed of speech without manual counting.

The ultimate goal of these measurements is to assess fluency for language testing purposes. By relating the automated measures of fluency (including silent and filled pauses, and speaking speed) to human judgments of fluency, researchers can evaluate the validity of these automatic measures for assessing language proficiency. This involves statistical analyses to see how well the automatically derived fluency metrics predict human ratings of fluency.

2.2 Generalised Additive Models (GAM)

In statistics, a generalized additive model is a generalized linear model in which the linear response variable depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions.

GAMs were originally developed by Trevor Hastie and Robert Tibshirani to blend properties of generalised linear models with additive models. They can be interpreted as the discriminative generalization of the naive Bayes. [15]

The model relates a univariate response variable, Y, to some predictor variables, x_i. An exponential family distribution is specified for Y (for example normal,binomial or Poisson along with a link function g(for example the identity or log functions) relating the expected value of Y to the predictor variables via a structure such as

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p) \tag{2.1}$$

The functions f_i may be functions with a specified parametric form (for example a polynomial, or an unpenalized regression spline of a variable) or may be specified non-parametrically, or semi-parametrically, simply as "smooth functions", to be estimated by non-parametric means. So a typical GAM might use a scatterplot smoothing function, such as a locally weighted mean, for $f_1(x_1)$, and then use a factor model for $f_2(x_2)$. This flexibility to allow non-parametric fits with relaxed assumptions on the actual relationship between response and predictor, provides the potential for better fits to data than purely parametric models, but arguably with some loss of interpretability.

Generalized additive models are a valuable tool in linguistic research as they allow for the modeling of complex, non-linear relationships between variables, which is common in language data.

In linguistics, the fluidity and complexity of human language present unique challenges that GAMs are particularly well-suited to address. For example, the relationship between word complexity and pause duration in speech may not be straightforward; certain words might cause longer pauses due to their complexity or unfamiliarity to the speaker. A GAM can flexibly model such intricate relationships without the need for prespecifying the form of the relationship, using smooth splines to adapt to the observed data.

Moreover, for researchers interested in exploring linguistic patterns, GAMs can provide intuitive visualizations of the results, allowing for an accessible interpretation of the effects of different predictors on language related responses.

2.3 Generalised Additive Mixed Models (GAMM)

Generalized additive mixed effect models are a type of statistical model that combines the flexibility of GAMs with the ability to account for random effects in mixed effect models.

Like GAMs, GAMMs allow for non-linear relationships between predictors and the response variable by fitting smooth functions to each predictor. However, GAMMs also allow for the inclusion of random effects, which capture the variability of observations within groups or clusters.

A generalized additive mixed effects model (GAMM) can be written as:

$$Y_i = f_1(X_{1,i}) + f_2(X_{2,i}) + ... + f_p(X_{p,i}) + Z_i b + \epsilon_i \tag{2.2}$$

In this formula, Y_i represents the response variable for the ith observation, and $X_{1,i}$ to $X_{p,i}$ represent the values of the p predictor variables for that observation. The functions f_1 to f_p represent the relationships between each predictor variable and the response variable.

The term $Z_i b$ represents the random effects in the model. Z_i is a matrix that specifies the random effects design for the ith observation, and b is a vector of random effects coefficients.

Finally, ϵ_i represents the error term for the ith observation.

Like GAMs, GAMMs is particularly useful when dealing with linguistic observations that have hierarchical or grouped structures, such as multiple measurements from the same participants or linguistic items nested within languages. This makes GAMMs powerful for linguistic studies that require the accommodation of individual differences, such as variability in language learners' performance across different tasks.

Overall, GAMs and GAMMs provide a way to explore the dynamics of language without making rigid assumptions about the form of the relationships between variables, which is often the case in linguistic research that deals with complex, nuanced phenomena like language fluency or word recognition processes.

Chapter 3

Methods

This chapter outlines the methodology employed to construct a statistical model focused on the dynamics of speaking fluency in second language learners, aiming to determine if there are significant trends in fluency measures, such as the frequency and duration of pauses, during speaking tasks. By exploring the steps from data collection to analysis, this research seeks to answer whether the propensity to pause within L2 speaking shows a pattern of increase, decrease, or stability over time within speech tasks, thus reflecting the dynamic nature of fluency.

3.1 Data Collection

The dataset for this thesis was compiled by Master's students enrolled in the Second Language Acquisition course at Leiden University. It comprises a structured collection of speech analysis data from 34 participants, who each undertook two distinct tasks in both their native language (Dutch for the majority) and their second language (English), resulting in a total of four tasks per participant.

The collection process was carefully designed to standardize the experimental conditions and ensure consistency across all participants. They were presented with a situation description, which they were asked to read. This description served as a prompt for the task and was designed to simulate real-life scenarios that would elicit spontaneous speech. One illustrative scenario involved the description of a crime scene, where participants were positioned as eyewitnesses to an accident and subsequently required to report their observations to a police officer as precisely as possible.

Each participant had a speaking window of two minutes, within which they were expected to articulate their response. However, participants had the flexibility to conclude their speech by pressing a "stop" button earlier than allocated time.

Upon completing the experiment, the audio recordings of the participants' responses were systematically collected and uploaded into PRAAT software for further analysis. This software enabled the detailed parsing and manipulation of the speech data, facilitating the subsequent analysis of fluency metrics and other linguistic features inherent

in the participants' spoken responses [16].

3.2 Dataset Properties

The dataset for this study is structured around several key pieces of information tied to each speech event recorded. At the beginning of each data entry, there's a **SoundfileID** that uniquely combines the participant ID with the language used in the task (L1 or L2) and the specific task number (Task1 or Task2).

Time intervals for speech events are delineated with **tmin** and **tmax** markers. For events with a duration, such as phrases or filled pauses, *tmin* indicates when the event starts, and *tmax* signals its conclusion. In cases where the event is a singular point, like a syllable nucleus, *tmin* and *tmax* will hold the same value, pinpointing the exact moment of occurrence.

The dataset also identifies the type of variable being measured or observed, designated as **Tier**. This includes various speech elements such as *phrases* (differentiating speech or silent pauses), *DFauto* (automated detection of filled pauses), and *nuclei* (which counts syllables). Another **Tier** of records is the *Lg10WF* which represents the log frequency of word occurrence drawn from an external corpus, providing a numeric value that reflects word usage frequency in broader language use.

Additionally, there's information labeled as **Text** that corresponds to the values for the specified variables. This could be a binary system (0 or 1, used for *phrases* and *DFauto* to indicate presence or absence of a pause) or an absolute count (as with nuclei to denote syllable numbers).

3.3 Data Preprocessing

To prepare the dataset for analysis, the `.txt` files are initially combined into a single dataset encompassing all recordings.

To simplify data handling, *Participant*, *Language*, and *Task* details from each recording are derived. The *SoundfileID* format **ParticipantID_Language_Task.wav** across all recordings enabled us to split this identifier into three new features, enhancing the ability to sift through the data efficiently.

The evaluation of the data necessitated the exclusion of three participants due to inaudible recordings. Additionally, gaps in the dataset were observed, attributed to some participants not completing all tasks. In line with the study's exploratory aim of discerning speech fluency patterns rather than exact quantitative analysis, a decision was made to forego imputation for the absent data. This choice was driven by the intent to maintain the integrity of the data and to avoid the introduction of potential bias, thereby acknowledging and accepting the inherent variability and partial completeness of the dataset.

Furthermore, a preliminary observation indicated that the initial seconds of recordings predominantly consisted of pauses. Presuming this duration represented the time

taken by participants to commence their tasks, it is decided to exclude the first pause from all recordings. Similarly, to ensure the precision of pause analysis, the closing seconds of each task is omitted, acknowledging these moments typically involved silence as participants concluded their tasks and signalled completion by pressing the stop button. The exclusion pause intervals vary for each participant but on average, each initial and ending pause lasts approximately 2 seconds.

3.4 Data Exploration

In this part, an examination of a specific participant's data is conducted to provide an overview and identify general trends within the dataset.[1] This participant is chosen because she/he had the longest speaking performance in each task and exhibited sufficient silent and filled pauses, making her/his data suitable for fitting a model to predict pause duration. This analysis allows for a more detailed understanding of the key features relevant to utterance fluency measurements.

Reviewing the participant `JAESPP1`'s speech patterns in Figure 3.1, the frequency of speech occurrences and pauses between two tasks performed in their first and second language can be compared.

The silent pauses, are interspersed throughout both tasks; however, their occurrence seems slightly more frequent and longer in first task especially in L2, hinting at more hesitation. This pattern could be attributed to the participant' initial adjustment to the task environment and requirements. As Task1 is the first opportunity for participant to speak within the set framework, she/he may experience challenges in managing her/his time effectively and figuring out the most efficient way to deliver her/his responses in second language. The filled pauses are less frequent than silent pauses but present a similar distribution across both tasks. Overall, this participant preferred to use more silent pauses rather than filled ones in both languages.

Here, each utterance fluency measurement is explained.

Speech rate: Defined as the total number of syllables divided by the overall time from the beginning to the end of the recorded task, this metric offers insight into the general pace of speech.

The analysis in Figure 3.2 segmented the speech rate from each task into equal intervals of five seconds to observe fluctuations throughout the performance. This division reveals a notable stability in the middle seconds of each task, showing that the participant has become more accustomed to the speaking context as she/he progressed. This also reflects her/his increasing comfort and preparation for continued discourse, illustrating dynamic changes in speech rate from the beginning to the end of the tasks.

Articulation rate: Calculated as the total number of syllables divided by the time spent speaking, this rate excludes any silent intervals, offering a perspective on the flow of speech.

[1] Figures for all participants can be found in the Appendices.

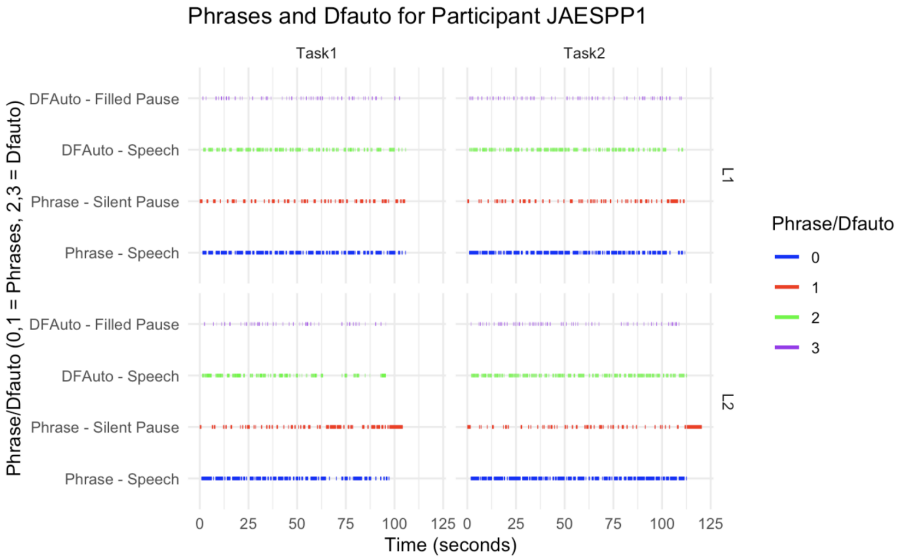


Figure 3.1: *Analysis of Speech and Pause for participant JAESPP1*

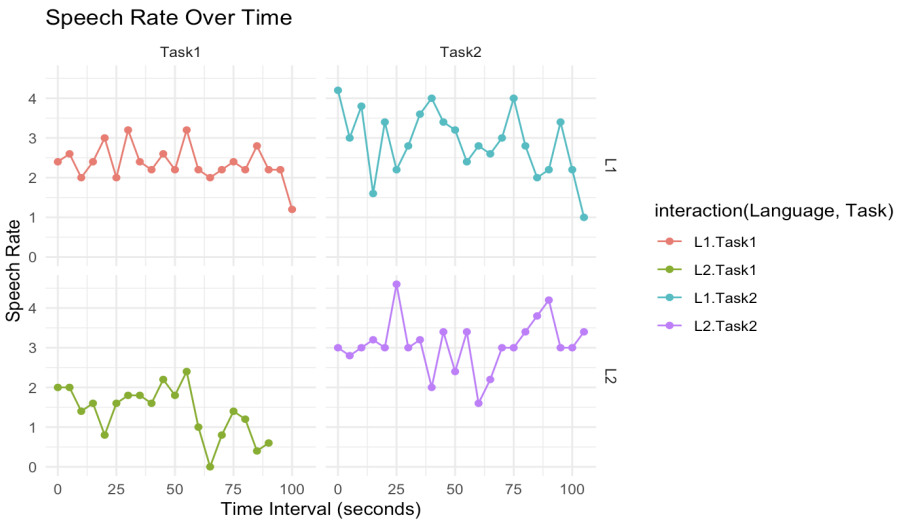


Figure 3.2: *Speech Rate in different languages and tasks for participant JAESPP1*

The formulas for speech rate and articulation rate are quite similar, but Figure 3.3 reveals a distinct difference in their patterns towards the end of each task. The plot indicates a brief pause just before the end of performances in her/his second language, during which the articulation rate drops for both tasks. In contrast, in her/his first language, the participant maintains a consistent speaking rate, similar to the middle of the performance. Overall, it is evident that the participant experienced more difficulties performing the first task in their second language.

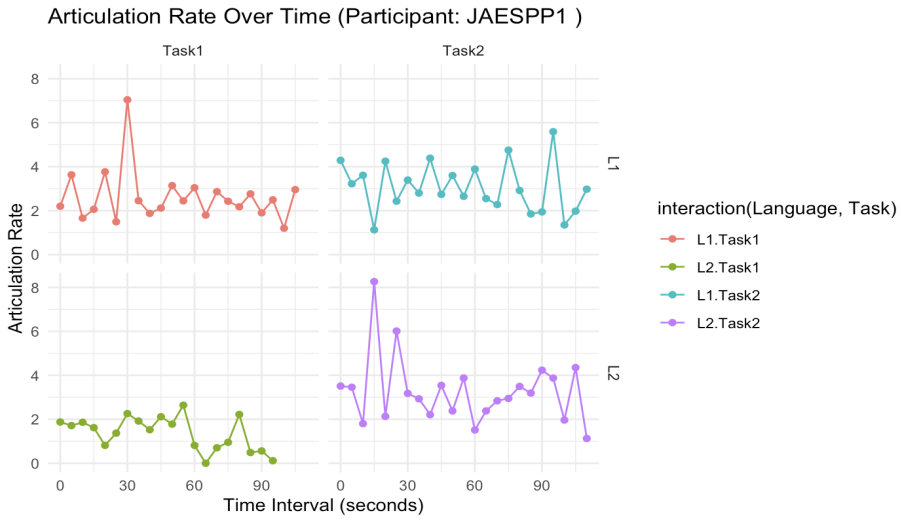


Figure 3.3: *Articulation Rate in different languages and tasks for participant JAESPP1*

Number of silent pauses: The frequency of silent pauses is measured by dividing the total number of silent pauses by the total time or speaking time. This measure can shed light on the occurrence and distribution of pauses.

Figure 3.4 the number of silent pauses varies across tasks and languages. In the second language, there's a fluctuating pattern with pauses in both tasks and also it seems that she/he used more number of silent pauses in L2, but in the first language, the number of pauses appears repeatedly which can show the smooth dynamics of her/his performance as a native speaker. This suggests that while speaking in L1, the participant may pause frequently, possibly for shorter durations, whereas in L2, the participant takes even more but potentially longer pauses, which could be indicative of searching for words or processing language structures.

It should be mentioned that in the second task for both languages, a notable challenge is observed especially during the initial 40 seconds, as evidenced by the high number of silent pauses present. These findings are consistent with the observations presented in Figure 3.1.

Number of filled pauses: Similarly, the count of filled pauses is normalized by the total time or speaking time, which can inform on the speaker's fluency and potential moments of cognitive processing.

In Figure 3.5, in the first language, filled pauses are relatively consistent across both tasks, showing a stable pattern in the use of filled pauses in the native language. In contrast, filled pauses in the second language show more variability especially around 20 to 30 seconds of both tasks which have the highest number of filled pauses. This pattern indicates moments of increased cognitive load and difficulty in language processing, which is common in second language production. It shows a strategy to gain time to think

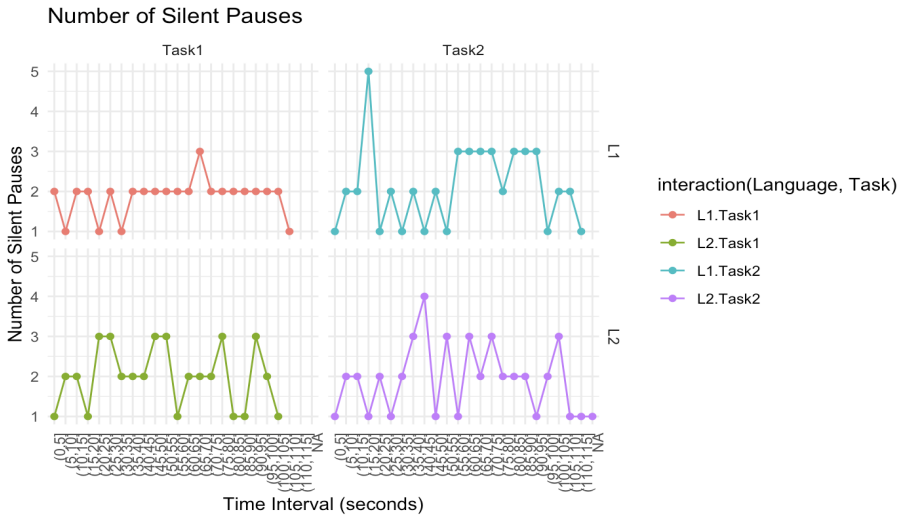


Figure 3.4: *Number of Silent Pauses in different languages and tasks for participant JAESPP1*

and plan speech during more challenging parts of the tasks in L2 and also reflect a lower level of comfort and automaticity when speaking in a non-native language.

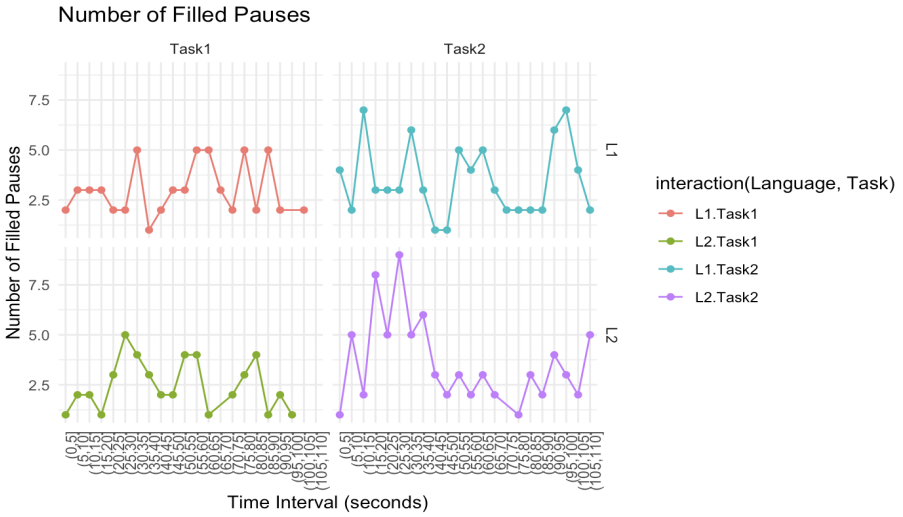


Figure 3.5: *Number of Filled Pauses in different languages and tasks for participant JAESPP1*

3.5 Choice of Statistical Model

The study initially employed Random Forest models to analyze the dynamics of fluency, capitalizing on their strengths in handling nonlinear relationships and complex interactions between multiple variables. Random Forest is particularly adept at managing high-dimensional data and preventing overfitting, making it suitable for incorporating a wide range of linguistic and temporal features that influence pause duration. Additionally, the feature importance scores from Random Forest models offer valuable insights into which predictors most significantly affect pause durations.

However, despite the initial use of Random Forest models, they are not considered the optimal choice for all aspects of linguistic data analysis. The primary limitation of these models is their "black box" nature, which obscures how decisions are made and complicates the interpretation of how variables interact within the model. This lack of transparency can be problematic in linguistics, where understanding the specific influence of each variable is crucial. Furthermore, Random Forest models do not allow for the explicit modeling of various forms of functional relationships and interactions as flexibly as some other statistical approaches, such as **Generalized Additive Models**.

Linguistic data often exhibit complex, non-linear relationships that traditional linear models may not capture effectively. The relationship between word complexity and pause duration, for example, is unlikely to be linear; instead, it might have a nuanced pattern where extremely complex or simple words have disproportionately large or small effects on pause duration compared to words of average complexity. GAMs use smooth spline functions to model non-linear relationships flexibly and that makes them an ideal choice in these cases. Additionally, there is inherent variability in speech data; different speakers or even the same speaker under different circumstances might show different patterns of pausing. GAMs can accommodate this variability by allowing for the inclusion of random effects or by constructing splines that can interact with other variables like speaker proficiency or speech rate. This feature is essential in linguistics analysis, where individual differences and context-dependent variations are the norm rather than the exception.

Moreover, when it comes to predictive performance, GAMs can offer superior results by providing a better fit to the actual data. Since language data can have underlying structures that are far from simple linear trends, the flexibility of GAMs in capturing such structures can lead to more accurate predictions. This is particularly relevant for predicting pause duration, as the research involves understanding how pause duration might vary over time and with different linguistic factors at play.

Lastly, while GAMs might be complex due to their non-linear nature, the resulting models can be quite interpretable. The smooth functions fitted by a GAM can be visualized, giving clear, intuitive insights into the relationship between predictors like word complexity and response variables such as pause duration. This can reveal valuable linguistic insights, such as points where an increase in complexity leads to significant changes in pause duration, which can then be related to underlying language processing mechanisms.

3.5.1 Model Specification

In exploring the dynamics of speaking fluency, understanding the factors that influence pause duration is pivotal. Pause duration can be indicative of cognitive processing during language production, such as the search for words or the planning of speech. Predicting pause duration, therefore, can shed light on the underlying linguistic and cognitive mechanisms at play.

To predict pause duration effectively, several predictors are developed, each capturing different aspects of speech and cognitive processing:

1. **Length of previous pause:** Captures the immediate history of pausing in seconds, reflecting possible continuity in speech rhythm.
2. **Length of previous speech:** Provides context on the fluency before the current pause, potentially influencing pause initiation.
3. **Previous 600 milliseconds of word frequency:** Reflects the frequency of occurrence of least frequent(most complex) word in past 600 milliseconds, which could affect the cognitive load and pause duration.
4. **Previous speech syllables:** Indicates the amount of speech production before the current pause and may relate to speaker fatigue or information processing.
5. **Type of the previous pause:** Differentiates between filled and silent pauses, giving insight into different pausing strategies or hesitation phenomena.
6. **Logarithm of previous speech rate:** This is calculated by dividing the number of syllables from the beginning of the performance by the total time elapsed prior to the current pause.
7. **Logarithm of previous articulation rate:** This is obtained by dividing the number of syllables from the beginning of the performance by the speaking time (without silent pauses) elapsed prior to the current pause.
8. **Absolute time:** Takes into account the position of the pause within the overall speech, which may reveal trends over the course of speaking.
9. **Language:** Considers the effect of whether the task is in L1 or L2, as language proficiency can significantly influence pausing.
10. **Task:** Reflects the nature of the speaking task, as different tasks may elicit varying pausing behaviors.
11. **Type of current pause:** Identifies whether the current pause is filled or silent, integral for differentiating between types of disfluency.
12. **First 600 milliseconds of upcoming word frequency:** Reflects the frequency of occurrence of least frequent(most complex) upcoming word in the first interval of 600 milliseconds following the pause

13. **Second 600 milliseconds of upcoming word frequency:** Provides insight into the frequency of occurrence of least frequent(most complex) upcoming word in the second interval of 600 milliseconds after the pause

14. **Third 600 milliseconds of upcoming word frequency:** Extends the lookahead to the complexity of word even further beyond the pause, potentially influencing longer-term planning processes.

For the response variable, pause durations in speech data typically show a right-skewed distribution with many short pauses and fewer long ones, which can challenge linear modeling techniques that assume normally distributed residuals. By transforming these durations to a logarithmic scale, their distribution is normalized, enhancing the reliability of linear models. This transformation is particularly beneficial as it compresses the scale of measurement, minimizing the influence of extreme values and improving homoscedasticity. Additionally, using a log scale shifts the focus to proportional rather than absolute changes in pause durations, offering a more accurate understanding of speech dynamics, where the relative increase or decrease in pause duration is more significant than the exact time.

3.5.2 GAMs Components

In GAMs, fixed effects, smooth terms, interactions, and random effects play distinct roles in modeling complex relationships in data. Here, it should be explained what the roles of each predictor are.

- **Fixed Effects:** These are variables for which data are collected on all possible levels or categories. They represent parameters that have a constant effect across different observations in the model. In the model, Language, Task, Type of Previous Pause, and Type of Current Pause are considered fixed effects because their existence in the model is the same for each participant, regardless of the variability among them.

- **Smooth Terms:** They are used in GAMs to model non-linear relationships between the predictors and the response variable. These terms are fitted using spline functions, which allow for flexibility in the shape of the function, adapting to the structure of the data. Length of Previous Pause and Length of Previous Speech and all other predictors are modeled as smooth terms to capture their potentially complex and non-linear effects on the response variable. These effects might vary by language, indicating that the relationship isn't strictly linear across different levels of these predictors.

- **Interactions:** Using interactions in GAMs assess whether the effect of one predictor on the response variable changes at different levels of another predictor. Interactions can be specified between fixed effects, between smooth terms, or between fixed and smooth terms, allowing for detailed exploration of how combined

factors influence the response. It is hypothesized that the effect of one of predictors varies by language or task in order to find the meaningful interactions which improve the reliability of the model.

- **Random Effects:** These effects account for variation at a group level that is not explained by the fixed effects alone. These are used when data is collected from hierarchical or nested structures, such as measurements from multiple participants. Participants are included as random effects to handle individual variability among participants that could affect the outcome independently of the fixed or smoothed predictors.

3.5.3 Model Diagnostics

In this part, a comprehensive assessment of the GAM model is introduced. This involves checking for any violations of the model assumptions, as well as identifying potential issues that could affect the validity and reliability of the results.

To ensure robust analysis with Generalized Additive Models, it is necessary first to assess the smoothness of each predictor. This involves using diagnostic plots such as partial residuals plots and examining the estimated degrees of freedom for each smooth term. Smooth terms with degrees of freedom close to 1 indicate that a linear term is adequate.

The model fit can then be evaluated by examining the explained deviance or using a metric analogous to the R-squared statistic, commonly referred to as R^2_{adj} in the context of GAMs. This step assesses how well the model captures the variance in the data.

Residual analysis, which is essential for identifying potential issues in the models. By plotting residuals against fitted values, one can check for heteroscedasticity, ensuring that residuals do not show any systematic patterns. Additionally, normality of residuals should be verified using Q-Q plots, confirming the assumptions about the error distribution. The `gam.check` function from the `mgcv` package in `R` is typically used for these diagnostics.

Regarding multicollinearity, although GAMs are generally robust due to their focus on non-linear relationships, it is still prudent to examine the correlations between predictors using the `concurvity` function in `R`. This check helps ensure that interpretations of the model components remain valid and unaffected by high inter-predictor correlations.

3.5.4 Model Validation

The study utilized the `mgcv` package to implement cross-validation, training the model on multiple data folds and testing on the remaining folds in a rotational manner. This approach effectively evaluates the model's capacity to generalize across new datasets. Additionally, model fit is assessed using the Akaike Information Criterion (AIC). AIC gauges model quality by balancing model complexity against goodness of fit, with lower AIC values indicating a more optimal balance. Further, residual analysis can still be con-

ducted to detect any patterns or systematic deviations in the model residuals, providing essential diagnostics for model validation.

3.6 GAMMs in detail

As discussed in Section 2.3, GAMMs are useful for modeling complex and nonlinear relationships between dependent and independent variables using smooth basis functions. These functions can capture various effects, including random effects that account for variability within groups or clusters [17]. In analyzing the dynamics of fluency in language processing, GAMMs are employed after fitting GAMs, incorporating participants as random effects to account for individual variability.

There are three main types of random effects in GAMMs:

1. **Random intercept:** Adjust the height of other model terms with a constant value
2. **Random slope:** Adjust the slope of the trend of a numeric predictor
3. **Random smooth:** Adjust the trend of a numeric predictor in a nonlinear way.

A fourth type combines random intercept and slope. After conducting the analysis with all four types for the GAMM, the models will be compared using the AIC test to identify the best fit. Subsequently, cluster analysis will group participants with similar performance based on the selected GAMM model.

3.6.1 Modeling the residuals

In this part of the analysis, an additional component can be included to account for autocorrelation by modeling residuals. This can provide insights into the time dependency of the residuals.

The covariance structure of the residuals should be addressed after removing fixed and random effects from the data to capture any remaining patterns. Autocorrelation aims to represent temporal or spatial dependence within the model. For instance, when taking measurements at specific time intervals, the objective is to model the relationship between y_t and $y_{t\pm n}$ for some n. The same principles can be applied to capture spatial autocorrelation.

Generally, time series data is processed by first removing trends through fixed-effect predictors, ensuring that the data is stationary (meaning that its statistical properties do not change over time), and then modeling the residual noise that remains.

Autocorrelation in the residuals is typically assessed using the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (pACF). These functions measure the correlation of residuals over different time lags. The ACF shows the correlation of the model's residuals with themselves at various lags. A spike at lag 0 is always observed because it represents the correlation of the residuals with themselves, which is always

perfect. The pACF, on the other hand, reveals the extent of correlation between the residuals at each lag after removing the effects of correlations at shorter lags.

It's important to note that accounting for autocorrelation can be challenging, especially when time is also modeled as a smooth term.

3.6.2 Cluster Analysis

As previously mentioned, it will be conducted a cluster analysis on the dynamic parameter of each participants. Hierarchical clustering will be used to determine the optimal number of clusters and identify the fluency patterns of each participant.

This method organizes data into a tree-like structure by progressively splitting clusters based on similarity measures, providing a visual representation known as a dendrogram. Hierarchical clustering is particularly useful in linguistic analysis as it helps determine the optimal number of clusters that group participants based on their fluency patterns. By analyzing the dendrogram, it becomes possible to identify distinct clusters that represent varying fluency patterns among participants. Once the clusters are identified, this method can highlight how speaking fluency dynamics differ within each cluster for both the first and second languages, offering insights into variations in speech performance across participants.

Chapter 4

Results

In this chapter, the focus will be on presenting the results from the analysis of fluency dynamics among second language learners.

4.1 GAM

The initial step involves determining which predictors are essential for capturing the full spectrum of the dynamics, including past, present, and future aspects of speaking fluency. This entails a careful evaluation of the available predictors to ensure that the model comprehensively represents the varied dimensions of language fluency.

A forward selection technique is employed to determine the most effective predictors for the model. This method involves incrementally adding each predictor to the model and observing the impact on model fit, using AIC for evaluation. Through this approach, it is determined that eight predictors significantly enhance the model's performance. These predictors include the previous 600 milliseconds of word frequency, the logarithm of previous speech rate, absolute time, the next first 600 milliseconds of word frequency, the next second 600 milliseconds of word frequency, and three other fixed effects except Type of Previous Pause which are mentioned in Chapter 3. The effectiveness of these predictors is also supported by analysis conducted with a correlation matrix, details of which are provided in the Appendices.

All these analyses involved only one specific participant, `JAESPP1`. This approach is chosen to initially identify which predictors significantly impact pause duration with one participant before extending the analysis to all participants. These predictors are also tested on several other participants to ensure their significance. `JAESPP1` is selected by reviewing the speech and pause plots like Figure 3.1 for each participant, focusing on those with the longest task performances in both languages and a sufficient number of silent and filled pauses.

The results of GAM analysis for participant `JAESPP1` in Table 4.1 show how various factors affect pause duration in speech. The `edf`s shown in the table are the **effective degrees of freedom** of the the smooth terms. Essentially, more `edf` imply more com-

plex, wiggly splines. When a term has an `edf` value that is close to 1, it is close to being a linear term. The intercept is estimated at -0.68, with a highly significant t-value

Table 4.1: *Results of generalised additive model predicting pause duration without interactions for participant JAESPP1*

Fixed Effects	**Estimate (SE)**	**t-value**	**p-value**
Intercept	-0.68 (0.02)	-27.52	< 0.001*
(PauseType)Silent	0.51 (0.02)	21.83	<0.001*
(Language)L2	-0.07 (0.030)	-2.72	0.006*
(Task)Task2	-0.05 (0.35)	-1.30	0.19
Smooth Terms	**edf**	**F**	**p-value**
previous 600msecs word frequency	1.00	28.21	< 0.001*
log(previous speech rate)	1.00	2.02	0.15
absolute time	1.72	1.75	0.16
first 600msecs upcoming word frequency	1.00	9.20	0.002*
second 600msecs upcoming word frequency	1.00	0.86	0.35
R-sq.(adj)	**Deviance explained**	**RSME**	R^2
0.71	72%	0.22	0.66

of -27.52, showing that the base log pause duration is significantly less than zero when all other predictors are held at their reference levels. This suggests that without the influence of other factors, the expected pause duration is around 500 milliseconds.

The coefficient for silent pauses is positive at 0.51, with a significant t-value of 21.83, indicating that silent pauses significantly increase the pause duration. On a logarithmic scale, this means that silent pauses are expected to be 66% longer than filled pauses.

The coefficient for speaking in a second language is -0.07, with a t-value of -2.72, suggesting a small but significant reduction in pause duration when speaking in L2 compared to L1. This implies that pauses in L2 are about 7% shorter than pauses in L1.

The effect of performing in Task2 as compared to Task1 is not significant, with a p-value = 0.19, indicating no substantial difference in the duration of pauses between Task2 and Task1.

Word frequency prior to a pause, measured over 600 milliseconds, significantly influences pause duration, supporting the hypothesis that complex words necessitate longer pauses, likely due to heightened cognitive load. In contrast, neither the logarithm of previous speech rate nor absolute time significantly affects pause duration. However, the frequency of occurrence of upcoming complex word in the first interval of 600 milliseconds following a pause also significantly prolongs the duration, corroborating the cognitive load theory which asserts that preparation for complex words entails extended pauses.

Overall, the model explains 72% of the deviance in pause duration, demonstrating a good fit. The explained deviance and the Root Mean Square Error (RMSE) are

highlighting the model's precision and reliability in predicting pause duration based on analyzed contextual factors.

Furthermore, incorporating interactions in the model can significantly alter the results, indicating that their influence should not be overlooked. For instance, the finding that absolute time does not affect pause duration may seem counterintuitive. However, interactions, such as those between absolute time and other variables like language or task, might reveal differential effects across conditions, potentially explaining the initially surprising results.

4.1.1 Adding Interactions

In the results presented in Table 4.2 below, significant interactions are identified between absolute time and language, next first 600 milliseconds word frequency, and the type of current pause. These interactions were the only ones among all potential interactions that were found to be statistically significant.

Table 4.2: *Results of generalised additive model predicting pause duration with interactions for participant JAESPP1*

Fixed Effects	**Estimate (SE)**	**t-value**	**p-value**
Intercept	-0.68 (0.02)	-27.45	< 0.001*
(PauseType)Silent	0.51 (0.02)	21.62	<0.001*
(Language)L2	-0.07 (0.030)	-2.72	0.007*
(Task)Task2	-0.05 (0.35)	-1.34	0.19
Smooth Terms	**edf**	**F**	**p-value**
previous 600ms word frequency	1.00	28.19	< 0.001*
log(previous speech rate)	1.00	1.89	0.17
absolute time:L1	1.95	1.14	0.34
absolute time:L2	1.00	4.03	0.04*
next first 600ms word frequency:Filled pause	1.00	8.02	0.004*
next first 600ms word frequency:Silent pause	1.43	1.90	0.26
next second 600ms word frequency	1.00	0.94	0.34
R-sq.(adj)	**Deviance explained**	**RSME**	R^2
0.71	72.7%	0.21	0.67

The interaction of absolute time with language, specifically for L2, shows a significant effect with a p-value of 0.04. This indicates that the evolution of pause duration over time is statistically significant for second language speakers, showing that the dynamics of pause duration in second language differ from those in the first language. The significance also implies that as time progresses within a speaking task, the manner in which pauses lengthen or shorten is distinct for L2 speakers compared to L1 speakers.

Moreover, the model reveals that the frequency of upcoming words in the first 600 milliseconds after a pause significantly interacts with the type of pause, particularly

when the pause is filled. This result suggests that when filled pauses are followed by less frequent words, the pause duration tends to increase significantly. This is likely because preparing to speak complex words requires more cognitive effort, thus extending the duration of the filled pause.

Visualisation

This section will examine the influence of each predictor on pause duration, focusing on the effects of smooth terms and other variables illustrated in the plots in Figure 4.1 and Figure 4.2.

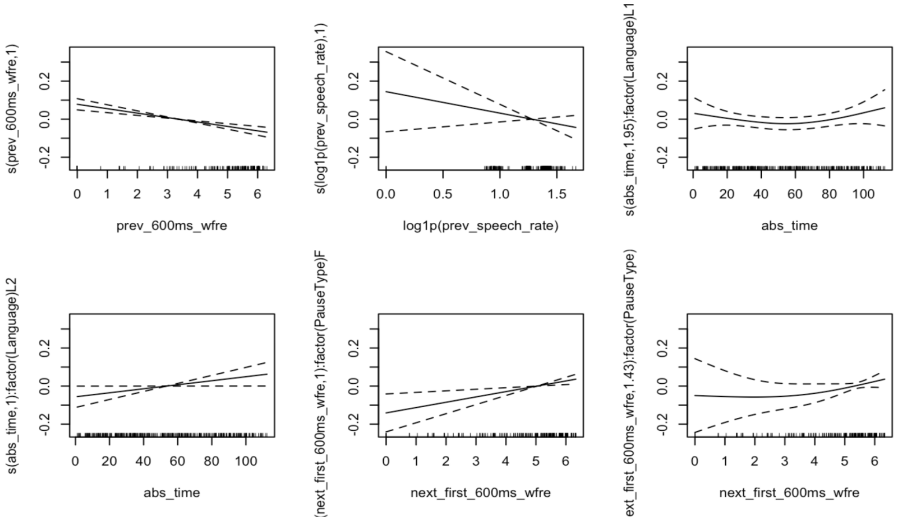


Figure 4.1: *Effective predictors on pause duration for participant JAESPP1*

Starting with the upper left plot, the effect of word frequency in the previous 600 milliseconds shows a downward trend, indicating around 18.13% decrease in pause duration as a word becomes more frequent (less complex).

The subsequent plot examines the relationship between the logarithm of the previous speech rate and pause duration, showing a decrease in pause duration by approximately 20% as speech rate increases. This shows that faster speech leads to shorter pauses, indicating that the speaker have experienced less difficulty in speech retrieval or planning. Due to the logarithmic scale, there is no data at higher speech rate and the widening confidence intervals suggest increasing uncertainty about this effect at those levels.

In the plot examining the effect of absolute time on pause duration within first language, a quadratic trend is observed. Pause duration decreases initially, reaches a minimum around 50 seconds, then begins to increase, suggesting an adjustment period where speaker stabilizes her/his speech pace before possibly needing more pauses as they approach task completion. The change in pause duration from 20 seconds to 50 seconds is a decrease of 10%, and from 50 seconds to 80 seconds, there is an increase of 5%.

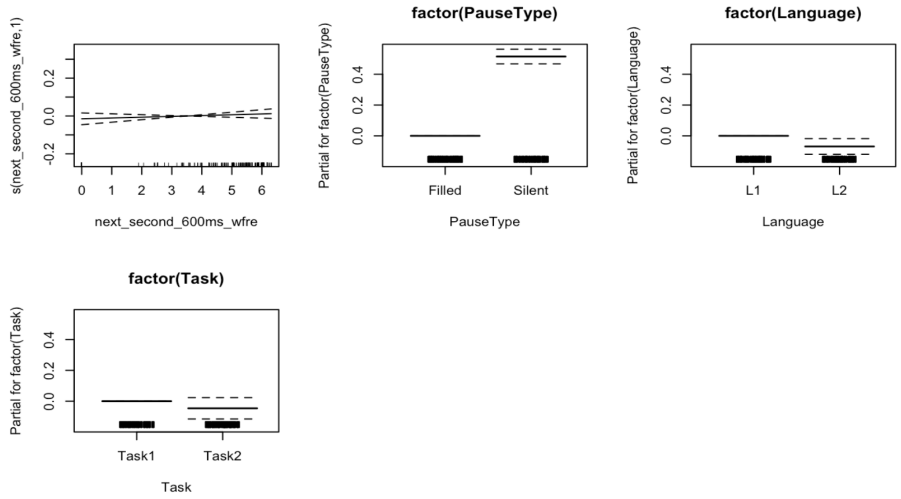


Figure 4.2: *Effective predictors on pause duration for participant JAESPP1*

The corresponding plot for second language shows a gradual increase in pause duration over time, indicating a consistent extension of pauses as speakers progress through the task, possibly adapting to or fatiguing from the task demands.

The interaction plot between the frequency of upcoming words and filled pauses shows an increase in pause duration as word frequency increases, particularly when the pause is filled.

Conversely, for silent pauses, the relationship between upcoming word frequency and pause duration follows a quadratic trend. Initially, pause duration decreases with increasing word frequency, but then it increases as the frequency of less complex words rises. This is unusual, as people generally pause more for complex words and less for simpler ones.

The effect of word frequency in the second interval of 600 milliseconds shows a horizontal trend across the frequency range, indicating no significant impact on pause duration.

In the factor plot for pause type, the solid and dashed horizontal lines likely represent summary statistics such as the mean (solid line) and confidence intervals(dashed lines) for each type of pause. Silent pauses are approximately 65% longer than filled pauses on average, indicating that silent pauses might incorporate more profound cognitive processing and hesitation compared to more fluidly used filled pauses.

For language effect, pause duration in second language are shorter by about 5% compared to first language, suggesting more frequent but shorter pauses in a second language, possibly due to different cognitive processing or comfort levels with the language.

Lastly, the task effect plot indicates that Task2 involves slightly shorter pauses than Task1, showing that Task2's demands might allow for or necessitate quicker transitions between speech segments compared to Task1, which might require longer pauses due to

its nature or complexity.

Figure 4.3 illustrates the comparison between actual and predicted pause durations across different types of pauses over time. The vertical difference between points represents the discrepancy between the model's predictions and the actual pause durations; where red lines indicate the model predicted shorter pauses than actually occurred, and green lines indicate the model predicted longer pauses.

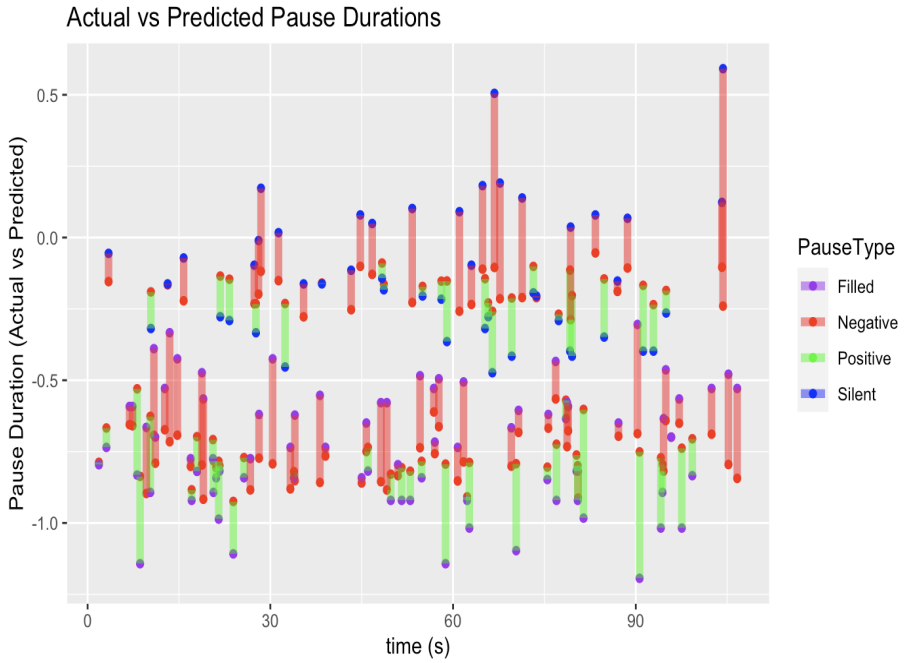


Figure 4.3: *Actual and Predicted Pause Durations for participant JAESPP1*

Around the middle of the performance (approximately 60 seconds), there is a noticeable concentration of green lines. This shows that the model overestimated the duration of pauses during this period. This can be due to several factors such as mid-performance cognitive load where participant might be processing more complex or new information, causing her/him to pause longer than at the start. The model's overestimation reflect its sensitivity to these complexities.

Towards the end of the performance, a prevalence of red lines suggests that the model underestimated pause durations. This underestimation could be linked to fatigue or culmination of thought processes towards the end of speaking tasks, where participant might take unexpectedly longer pauses to gather thoughts or conclude her/his speech. The model's underprediction here might indicate a lack of adjustment for end-of-task cognitive or psychological dynamics, which tend to elongate pauses more than expected.

4.1.2 Goodness of fit in GAM model

In Figure 4.4, the diagnostic plots are used to assess the fit of this Generalized Additive Model. Each plot provides unique insights into the model's behavior and assumptions.

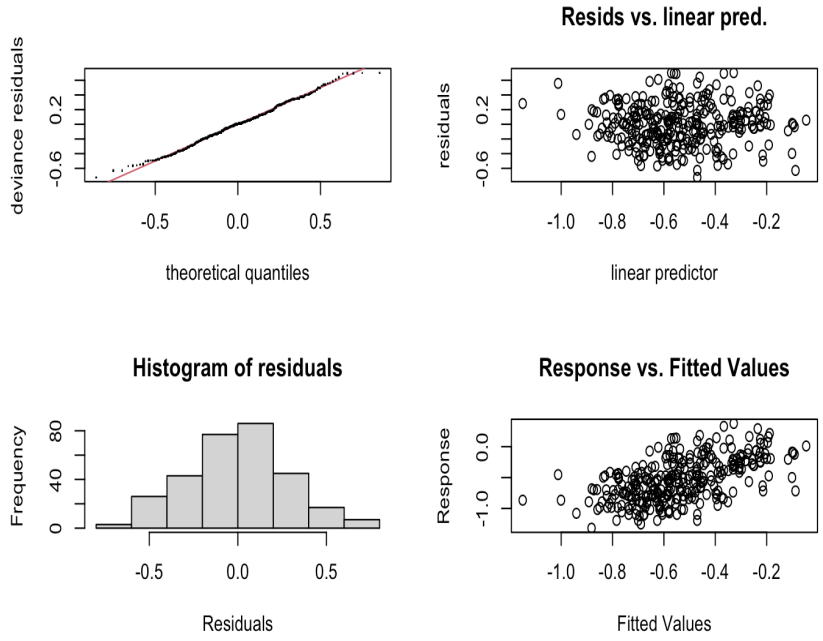


Figure 4.4: *Goodness of fit in GAM model*

1. Quantile-Quantile (Q-Q) Plot of Deviance Residuals (Upper Left): This plot assesses whether the residuals of the model are normally distributed by comparing their distribution to the theoretical quantiles of a normal distribution. Points that lie close to the red diagonal line indicate adherence to normality. In this plot, the residuals generally align well with the expected line, except for minor deviations at the extremes, suggesting that the normality assumption is reasonably satisfied.

2. Residuals vs. Linear Predictor (Upper Right): Used to detect non-linear patterns, heteroscedasticity, or other systematic deviations from a model's assumptions. The plot shows a random scatter of residuals around the horizontal axis with no obvious patterns, indicating that the model captures the relationship in the data effectively without systematic errors.

3. Histogram of Residuals (Lower Left): Provides a visual representation of the distribution of residuals to further assess normality. The histogram reveals a somewhat symmetric distribution but not perfectly normal, with a slight skewness visible. This slight bias shows minor imperfections in model fit but generally indicates adequate model performance.

4. Response vs. Fitted Values (Lower Right): Evaluates the model's predic-

tive accuracy by comparing the fitted values against the actual responses. The concentration of data points around the zero line in a horizontal band indicates a good fit of the model. There are no visible trends or patterns, suggesting that the model is free from issues of non-linearity and heteroscedasticity, and provides a reliable prediction across the observed value range.

4.2 GAMM

In Chapter 3, four distinct GAMM models were explored. This modeling approach was adopted to account for the variability among participants, which can influence the pause duration. Incorporating this factor as random effect allows for a more nuanced understanding of the data, acknowledging that individual participant differences can have significant impacts on the results.

Based on the evaluation of various GAMMs models incorporating participants as random effects, the AIC test is used as it is employed in GAM model to compare models differing in complexity: models with a random intercept, a random slope, both a random intercept and slope, and a random smooth term. As presented in Table 4.3, the GAMM featuring only a random intercept yielded the lowest AIC value, indicating a preferable balance of model fit. Consequently, this model is selected for further analysis.

Table 4.3: *Result of AIC test using different GAMM models*

	df	**AIC value**
GAMM with a Random Intercept	39.01	**-773.56**
GAMM with a Random Slope	38.74	-706.60
GAMM with a Random Intercept and Slope	41.57	-770.86
GAMM with a Random Smooth	47.47	-765.94

4.2.1 Residual Autocorrelation

Before proceeding with the GAMM analysis, it is essential to evaluate the correlation among the residuals, which represent the differences between the observed and predicted values. This assessment helps to determine if there exists any time-dependent patterns or relationships in the residuals, referred to as autocorrelation.

In Figure 4.5, the correlations for non-zero lags remain within the blue dashed confidence bounds, suggesting minimal autocorrelation. This indicates that the residuals do not carry significant time-dependent patterns beyond what would be expected by chance.

Here, the pACF values also predominantly stay within the confidence bounds, indicating no significant partial autocorrelation at various lags. This further confirms that the residuals from the model are not significantly autocorrelated, affirming the independence of errors over time. This independence supports the model's adequacy in capturing

the essential patterns in the data without leaving behind time-dependent structures in the residuals.

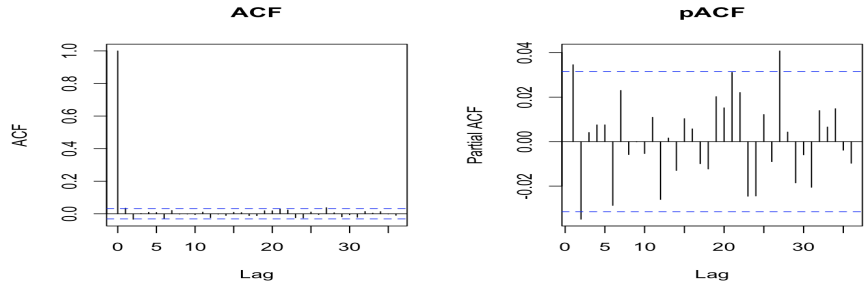


Figure 4.5: *ACF-pACF Residuals*

The outcomes derived from the chosen model are detailed in Table 4.4, facilitating subsequent examinations of fluency dynamics. The model aimes to understand the effects of various predictors on pause duration and included interactions to observe differences between the first and second language settings.

Interestingly, the outcomes reveal no important global discrepancy between L1 and L2 speech as indicated by the pause durations, implying individuals retain their fluency in talking across both languages. The shorter pauses noticed in Task2 show that participants may have adapted to the performance requirements, based on what they learnt from their performance during the first task. Additionally, the influence of absolute time on pause duration for L2 is more complex indicating that there are subtle changes throughout the course of speech. Moreover, considering subjects as random term indicates a considerable individual difference in how speakers control their pause lengths – due to personal or stylistic variations in speech patterns.

Figure 4.6 depicts the effects of word frequency in the previous 600 milliseconds on the logarithm of pause duration for both first language (left plot) and second language (right plot), demonstrating how word complexity influences pause durations differently across languages.

In both L1 and L2, as speakers navigate from moderate to high word frequency, the pause duration tends to decrease. However, in L2, there is increased uncertainty around moderate word frequencies, indicating that the model struggles to predict whether there will be longer pauses or not in this range.

Table 4.4: *Results of generalised additive mixed model with a random intercept predicting pause duration*

Fixe Effects	**Estimate (SE)**	**t-value**	**p-value**
Intercept	-0.71 (0.01)	-48.14	< 0.001*
(PauseType)Silent	0.56 (0.007)	72.11	<0.001*
(Language)L2	-0.01 (0.007)	-1.71	0.08
(Task)Task2	-0.02(0.007)	-3.24	0.001**
Smooth Terms	**edf**	**F**	**p-value**
previous 600ms word frequency:L1	1.00	65.75	< 0.001*
previous 600ms word frequency:L2	1.71	52.70	< 0.001*
log(previous speech rate):L1	1.69	10.26	< 0.001*
log(previous speech rate):L2	1.91	4.24	0.01*
absolute time:L1	1.83	5.84	0.003**
absolute time:L2	1.32	5.10	0.02*
next first 600ms word frequency:L1	1.97	0.004	0.50
next first 600ms word frequency:L2	1.90	8.70	0.006**
next first 600ms word frequency:Filled pause	1.00	14.65	0.00***
next first 600ms word frequency:Silent pause	1.94	7.56	0.00***
next second 600ms word frequency:L1	1.00	2.67	0.10
next second 600ms word frequency:L2	1.00	5.04	0.02*
Random Effect	**edf**	**F**	**p-value**
Participant	17.79	12.67	< 0.001*
R-sq.(adj)	**Deviance explained**		
0.66	66.7%		

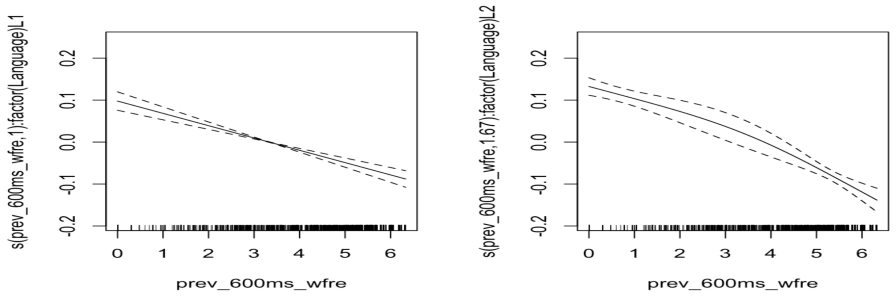


Figure 4.6: *Effect of word frequency prior to a pause on the pause duration*

In Figure 4.7, both curves demonstrate how linguistic and cognitive processing demands influence pause duration. In L1, there is a more gradual adaptation, reflecting inherent fluency and comfort with the language. In contrast, L2 speakers might overcompensate at lower speeds and struggle more as speeds increase, showing a critical threshold where cognitive load overtakes fluency. The point around log(1) for both languages likely represents a critical fluency threshold. Below this speech rate, increasing speed aids fluency up to a point by reducing pauses.

Figure 4.8 depicts the effect of absolute time on pause duration differentiated by L1 and L2.

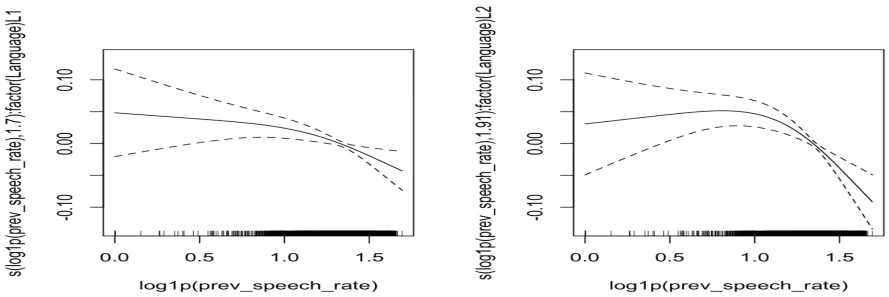


Figure 4.7: *Effect of speech rate prior to a pause on the pause duration*

The initial increase in pause duration for L1 could be tied to the speakers dealing with the introductory complexity of the task or finding their pace in speech delivery. Once adapted, the stabilization of pause durations shows that speakers manage to maintain a steady state of cognitive load and speech planning, balancing fluency with processing demands effectively. On the other hand, the consistent increase in pause duration in L2 illustrates the relentless challenge faced by second language speakers.

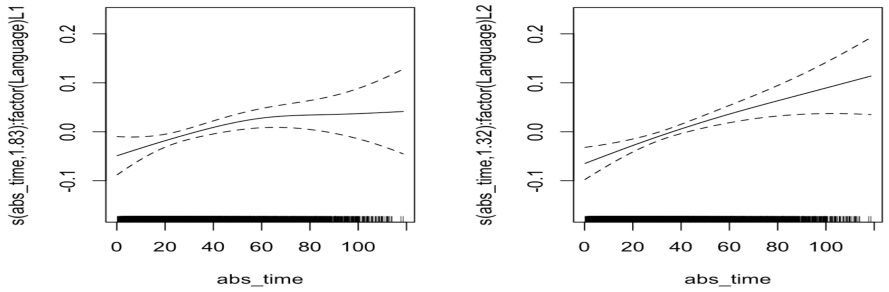


Figure 4.8: *Effect of absolute time on the pause duration*

The effect of word frequency on pause duration shows a U-shaped curve for L1 in Figure 4.9. Initially, as frequency increases from 0 to around 4, the pause duration slightly decreases, showing that moderate number of complex words do not require additional pause time. This reflects a familiarity or efficiency with handling moderately complex constructs in the native language. As complexity continues to decrease beyond 4, the pause duration begins to increase, implying that simple word constructs require more processing time.

Figure 4.10 shows the effects of word frequency in the next 600 milliseconds on pause duration, modulated by type of pause. The increasing trend in filled pause duration suggests that speakers use filled pauses strategically to manage the cognitive load of upcoming complex speech, possibly to maintain fluency without going completely silent. On the other hand, as word frequency increases from a low to a moderate level, the duration of silent pauses decreases, suggesting that initial complexities in the upcoming speech do not significantly disrupt the speech flow enough to necessitate longer silent

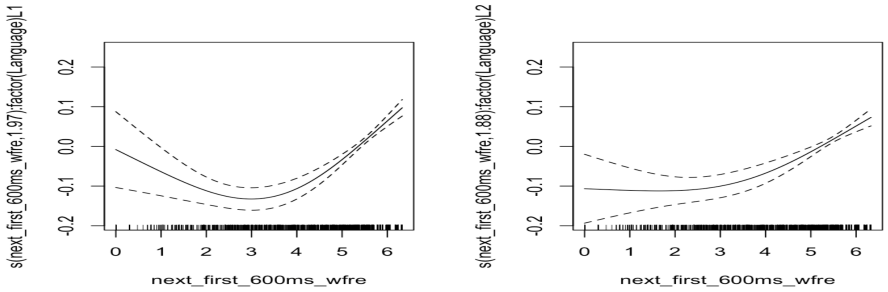


Figure 4.9: *Effect of first 600milliseconds upcoming frequent word on the pause duration in first and second language*

pauses. However, as frequency continues to increase, the trend reverses, and silent pauses become longer.

The contrasting trends between filled and silent pauses illustrate different adaptive strategies in managing speech flow and cognitive load. Filled pauses serve as a buffer, allowing continuous speech flow with minor interruptions, while silent pauses are employed more drastically once a complexity threshold is exceeded.

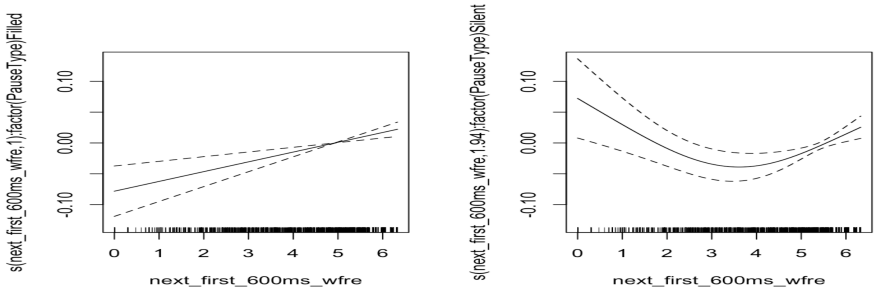


Figure 4.10: *Effect of first 600milliseconds upcoming frequent word on the pause duration in filled and silent pause*

The plots in Figure 4.11 depict the relationship between word frequency in the next second interval of 600milliseconds and pause duration. Specifically, they show that as the frequency of less complex words increases in this interval, there is a corresponding decrease in the length of pauses. Conversely, as the frequency of occurrence of more complex word is low, the duration of pauses tends to increase.

From Figure 4.12, most data points closely follow the diagonal line, indicating that the random effects for the participants are approximately normally distributed. Verifying that the random effects are normally distributed supports the reliability and accuracy of the model's predictions.

The displayed plots in Figure 4.13 compare actual and predicted pause durations for participant JAESPP1 across two tasks and two languages. In each plot, the red line

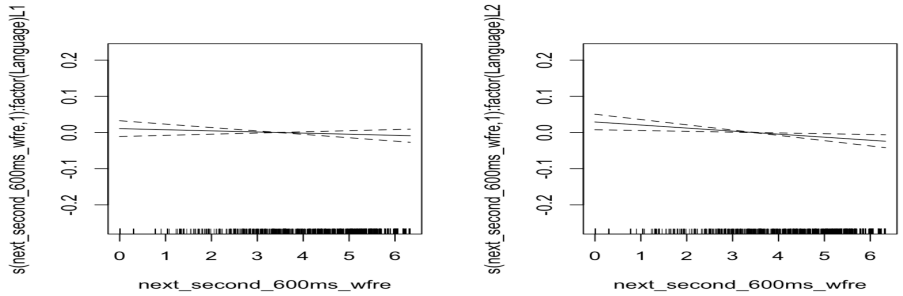


Figure 4.11: *Effect of second 600milliseconds upcoming frequent words on the pause duration*

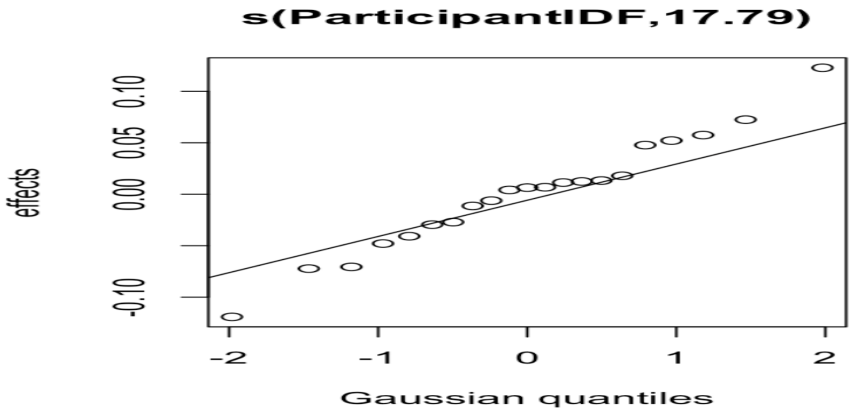


Figure 4.12: *Q-Q plot of the random effect for participants*

represents the predicted pause durations, while the blue line represents the actual pause durations.

The top two plots correspond to the first language, with the left plot showing data from Task1 and the right plot showing data from Task2. Similarly, the bottom two plots correspond to the second language.

These plots illustrate the fluctuations in pause duration over time. The blue lines display the variability observed in the participant's speech, while the red lines show the model's estimations based on the given predictors.

While the predicted pause durations generally follow the trend of the actual pause durations, there are instances where the model does not perfectly capture the peaks and troughs of the actual data. Several factors can contribute to this discrepancy.

First, human speech is highly complex and influenced by numerous factors, both linguistic (e.g., word choice, sentence structure) and non-linguistic (e.g., speaker's emotional state, cognitive load). A model may not be able to account for all these nuances, leading to differences between predicted and actual values.

Second, there may not be enough data to accurately estimate the model parameters, particularly for rare events or extreme values. This can lead to less reliable predictions in those areas.

Moreover, the actual pause durations might include noise due to measurement errors or external influences not accounted for by the model. This noise can create variability that the model cannot predict.

These factors highlight the inherent challenges in modeling complex human behaviors such as speech.

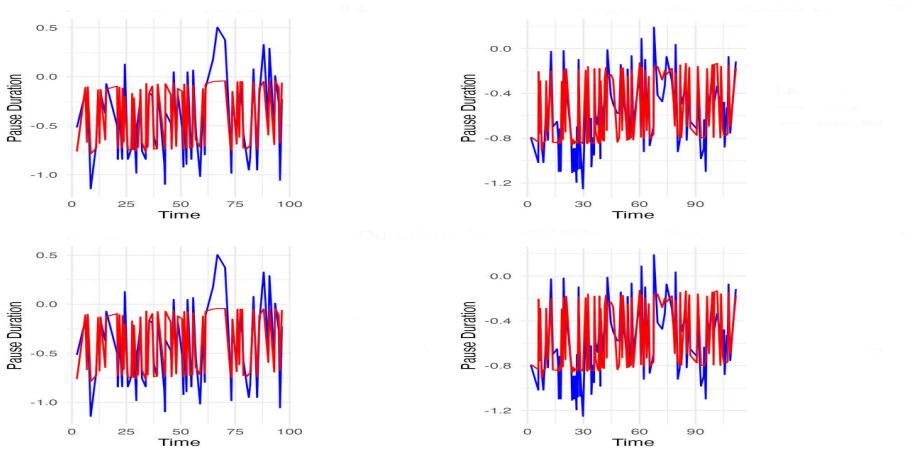


Figure 4.13: *Actual and Predicted Pause Durations for participant JAESPP1*

Based on the provided information about diagnostic plots for GAM model, the goodness of fit of the GAMM model can be evaluated from Figure 4.14. These plots suggest that the model is a reasonable fit for the data. The Q-Q plot and the histogram indicate a reasonable approximation of normality, though there are potentially some slight skewness in the distribution of residuals. The plots of residuals vs. linear predictors and response vs. fitted values do not exhibit clear or systematic deviations, which supports the adequacy of the model fit.

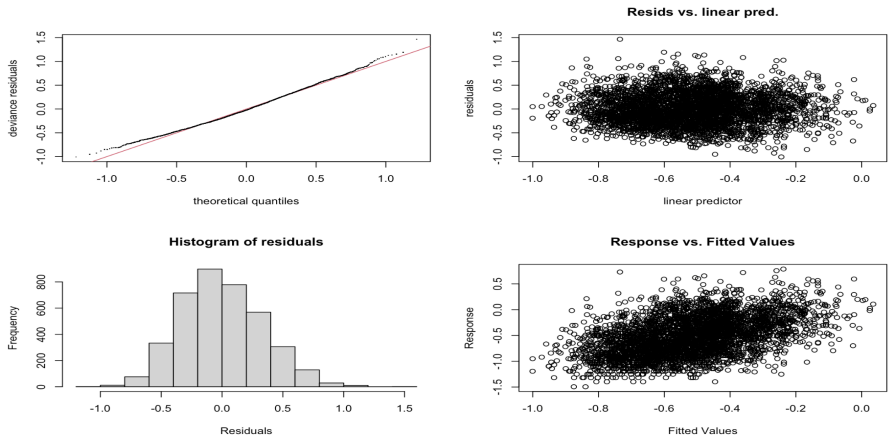


Figure 4.14: *Goodness of fit of the GAMM model*

4.2.2 Cluster Analysis

The initial plan for clustering aimed to differentiate between participants' speaking performances by analyzing their dynamic speech patterns such as "absolute time" which is considered the most important predictor in this analysis. However, the best-fit model for this data includes only an intercept and lacks a random slope for this predictor, making it unsuitable for capturing individual differences in speech performance over time. Therefore, clustering based on these parameters is not feasible.

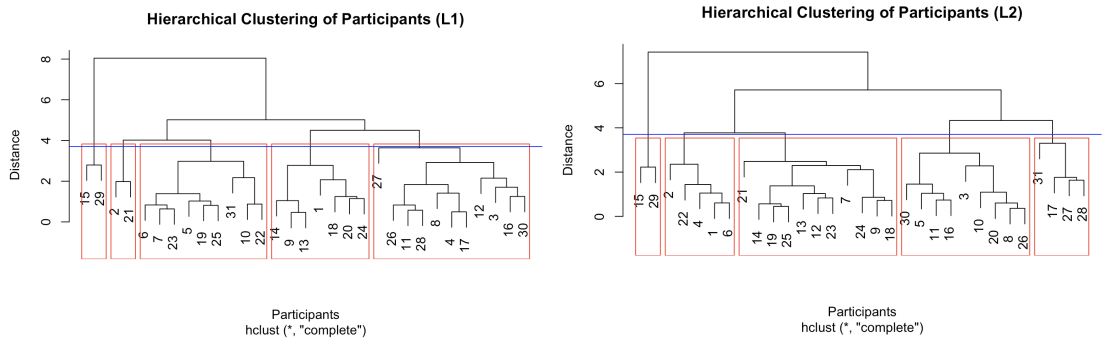
A model with only an intercept does not account for variations in how different participants' speech performances evolve over time. It provides a single average estimate, which is insufficient to reveal the dynamic, individual-specific patterns necessary for effective clustering.

As an alternative, participants can be clustered based on the predicted pause durations derived from the predictors used in the selected model. This involves first determining the optimal number of clusters through hierarchical clustering, followed by grouping participants based on their predicted pause durations. By calculating the mean predicted pause duration in 5-second intervals for each cluster, separately for both languages, it can be effectively identifed and grouped participants with similar speech performance patterns.

This approach enables clustering of participants with similar predicted pause duration, providing insights into differences in fluency and pause patterns across the participant pool.

As mentioned, the optimal number of clusters is determined using hierarchical clustering, with an emphasis on the inspection of height differences between successive linkage points in the dendrogram. In Figure 4.15, significant jumps in height between linkage points indicate natural divisions within the data, suggesting that five clusters is the optimal number for both languages.

Based on the bar charts provided for the mean pause duration by cluster in both



(a) **(b)**

Figure 4.15: *Hierarchical Clustering of Participants in first and second language*

first and second language of the participants in Figure 4.16, except for Clusters 4, other clusters show longer mean pause durations in the second language compared to the first.

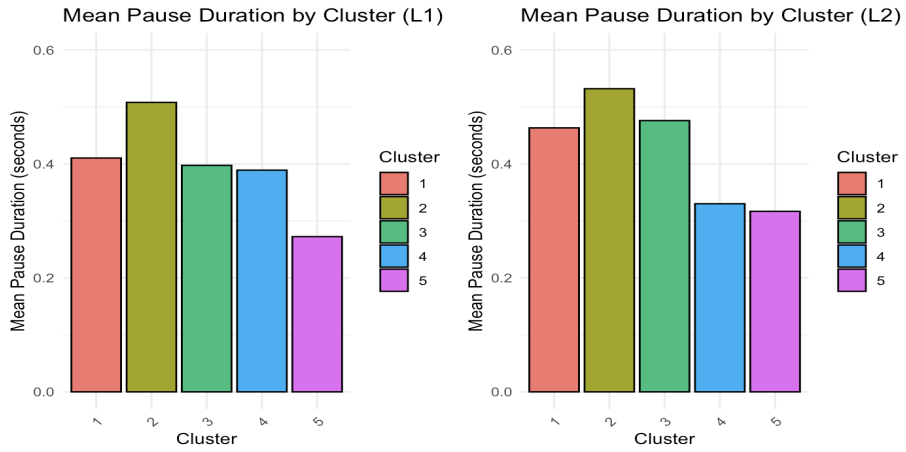


Figure 4.16: *Histogram of mean predicted pause duration in first and second langauge in each cluster*

According to the cluster analysis, as shown in Figure 4.17, the majority of participants are grouped in the first cluster, which exhibits the longest performance durations in both languages. This observation is further illustrated in Figure 4.18. Notably, there are only three participants whose performances in the second language are clustered differently from their performances in the first language; these participants are highlighted in bold in the tables. Aside from these exceptions, the data indicates that most participants had similar performance patterns in both their first and second languages, suggesting a high level of proficiency in their second language comparable to their native language.

In the analysis of mean pause durations over time in both the first and second

Cluster 1	AMGOPP1, EKATPP1, ELISPP1, ELISPP2, EVGEPP2, JAESPP1, JAESPP2, KIKIPP2, KISTPP1, MYHIPP1, MYHIPP2, MYROPP1
Cluster 2	AMGOPP2, LOESPP1
Cluster 3	DAHAPP1, HRHHPP2, JAWOPP2, KIKIPP1, KISTPP2, SHRIPP1, SUCAPP1, SUCAPP2, TABLPP1
Cluster 4	HRHHPP1, MAPOPP1, **TABLPP2**, EKATPP2, JADEPP1, **EVGEPP1**, **KISTPP1**
Cluster 5	JAWOPP1, SVENPP1

(a) L1

Cluster 1	AMGOPP1, EKATPP1, ELISPP1, ELISPP2, EVGEPP2, JAESPP1, JAESPP2, KIKIPP2, MYHIPP1, MYHIPP2, MYROPP1
Cluster 2	AMGOPP2, **EKATPP2**, LOESPP1, **EVGEPP1**, **KISTPP1**
Cluster 3	DAHAPP1, HRHHPP2, JAWOPP2, KIKIPP1, KISTPP2, SHRIPP1, SUCAPP1, SUCAPP2, TABLPP1
Cluster 4	HRHHPP1, MAPOPP1, TABLPP2, JADEPP1
Cluster 5	JAWOPP1, SVENPP1

(b) L2

Figure 4.17: *Clustering participants based on mean predicted pause duration*

language in Figure 4.18, it is evident that participants generally exhibit longer pause durations in L2 across almost all clusters, reflecting increased difficulties associated with processing a less familiar language. The variability of pause durations is more pronounced in L1, where fluctuations show that participants are more comfortable and thus vary their pacing more freely when using their native language. In contrast, L2 shows a pattern of fewer but higher peaks in pause durations. It is worth noting that while most participants are in the first cluster, which exhibits the longest performance durations, only 12 participants in the first language and 11 in the second language completed their tasks within the predetermined time. Out of 31 participants, many chose to press the stop button before the allocated time had passed.

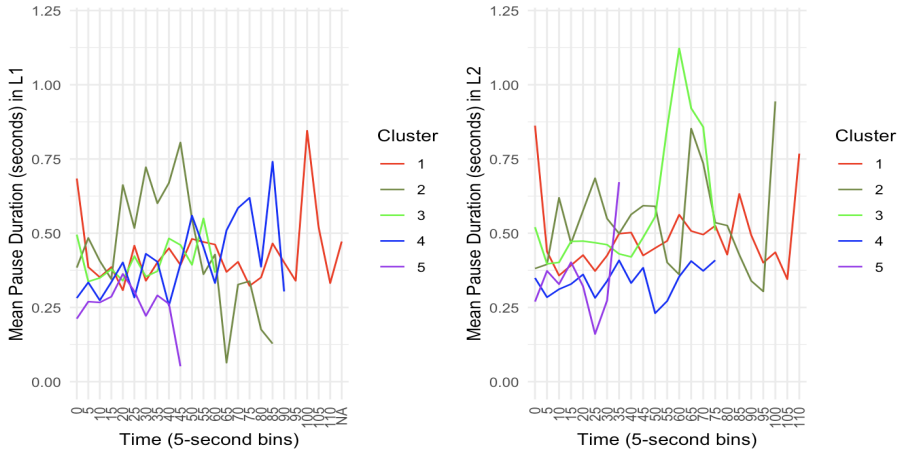


Figure 4.18: *Mean predicted pause durations over time for participants grouped by cluster*

Chapter 5

Discussion

In the discussion section of this research, several notable challenges were encountered.

The first challenge was the variation in the length of tasks across different participants. This variability posed a problem as the inconsistency in task length could affect the reliability of the results. To achieve more reliable outcomes, future research should ensure that all participants complete tasks of the same duration. This standardization would help in obtaining consistent and comparable data across participants.

Another significant challenge was distinguishing between filled pauses and silent pauses within longer pauses, such as those lasting six seconds. This distinction is particularly important when analyzing predictors like the frequency of words in the previous 600 milliseconds or the next first interval of 600 milliseconds. The analysis revealed an unexpected trend where an increase in word frequency coincided with longer pause durations. Initially, this seemed counterintuitive. However, it was considered that the complexity might arise from the possibility of subsequent or preceding pauses influencing the current pause. This complexity is further compounded by the lack of detailed information on whether silent pauses contain filled pauses. Therefore, a more refined method for classifying and analyzing pauses is necessary. Future studies should aim to gather more detailed data to accurately differentiate between types of pauses and better understand their impact on pause duration.

The current study focused on aspects of fluency over time. However, it did not extensively examine other linguistic factors such as pronunciation accuracy, lexical choices, grammatical correctness, the use of cohesive devices, and syntactic complexity, all of which may also fluctuate over time. Future research should address these aspects to provide a more comprehensive understanding of linguistic performance. Another limitation of this study is its exclusive focus on speaking performances. To gain deeper insights into the fluctuations of linguistic aspects over time, future research should also consider the speakers' perspectives. Investigating speaker cognition through methods such as stimulated recall and idiodynamic procedures could provide valuable information on how speakers experience and manage their language production in real-time.

In conclusion, for future research, it is essential to incorporate models that can estimate slopes for each participant. This could involve using mixed-effects models with

random slopes or other advanced statistical techniques capable of capturing individual temporal dynamics. By including these slopes, future studies could better differentiate between participants based on how their speech patterns change over time, leading to more insightful cluster analysis results.

Appendices

Reference to online repository for the code and all figures: GitHub

Table 1: *Results of generalised additive model predicting pause duration with interactions for participant TABLPP2*

Fixed Effects	**Estimate (SE)**	**t-value**	**p-value**
Intercept	-0.71 (0.06)	-12.94	< 0.001*
(PauseType)Silent	0.63 (0.06)	21.62	<0.001*
(Language)L2	-0.11 (0.06)	1.70	0.09
(Task)Task2	-0.005 (0.06)	-0.08	0.93
Smooth Terms	**edf**	**F**	**p-value**
previous 600ms word frequency	1.86	4.76	0.01*
log(previous speech rate)	1.00	0.19	0.66
absolute time:Language1	1.00	0.25	0.61
absolute time:Language2	2.60	1.25	0.34
next first 600ms word frequency:Filled pause	1.00	0.97	0.32
next first 600ms word frequency:Silent pause	1.00	1.27	0.26
next second 600ms word frequency	1.00	1.41	0.24
R-sq.(adj)	**Deviance explained**	**RSME**	R^2
0.76	81.7%	0.16	0.37

Table 2: *Results of generalised additive mixed model with a random slope predicting pause duration*

Fixed Effects	**Estimate (SE)**	**t-value**	**p-value**
Intercept	-0.48 (0.02)	-21.55	< 0.001*
(PauseType)Silent	0.56 (0.007)	72.27	<0.001*
(Language)L2	-0.01 (0.007)	-1.54	0.12
(Task)Task2	-0.02(0.007)	-2.05	0.03*
Smooth Terms	**edf**	**F**	**p-value**
previous 600ms word frequency:L1	1.001	81.85	< 0.001*
previous 600ms word frequency:L2	1.67	100.92	< 0.001*
log(previous speech rate):L1	1.80	6.14	0.001**
log(previous speech rate):L2	1.00	16.25	< 0.001*
absolute time:L1	1.64	5.42	0.01*
absolute time:L2	1.54	6.51	0.01*
next first 600ms word frequency:L1	1.83	22.22	< 0.001*
next first 600ms word frequency:L2	1.50	8.70	0.006**
next first 600ms word frequency:Filled pause	1.00	8.70	< 0.001*
next first 600ms word frequency:Silent pause	1.94	7.67	0.00**
next second 600ms word frequency:L1	1.00	0.56	0.45
next second 600ms word frequency:L2	1.00	9.68	0.001**
Random Effect	**edf**	**F**	p-value
Participant	16.82	45.39	< 0.001*
R-sq.(adj)	**Deviance explained**		
0.65	66.1%		

Table 3: *Results of generalised additive mixed model with a random intercept and slope predicting pause duration*

Fixed Effects	**Estimate (SE)**	**t-value**	**p-value**
Intercept	-0.48 (0.02)	-18.31	< 0.001*
(PauseType)Silent	0.56 (0.007)	72.11	<0.001*
(Language)L2	-0.01 (0.01)	-1.73	0.08
(Task)Task2	-0.03(0.01)	-3.12	0.001**
Smooth Terms	**edf**	**F**	**p-value**
previous 600ms word frequency:L1	1.00	79.16	< 0.001*
previous 600ms word frequency:L2	1.67	93.56	< 0.001*
log(previous speech rate):L1	1.71	3.35	0.02*
log(previous speech rate):L2	1.00	8.31	0.003**
absolute time:L1	1.64	5.66	0.009**
absolute time:L2	1.32	13.57	< 0.001*
next first 600ms word frequency:L1	1.96	43.55	<0.001*
next first 600ms word frequency:L2	1.87	23.00	<0.001*
next first 600ms word frequency:Filled pause	1.00	14.63	0.00
next first 600ms word frequency:Silent pause	1.94	7.54	0.00
next second 600ms word frequency:L1	1.00	1.00	0.31
next second 600ms word frequency:L2	1.00	7.11	0.007**
Random Effect	**edf**	**F**	**p-value**
Participant	18.46	15.53	< 0.001*
R-sq.(adj)	**Deviance explained**		
0.66	66.7%		

Table 4: *Results of generalised additive mixed model with a random smooth predicting pause duration*

Fixed Effects	**Estimate (SE)**	**t-value**	**p-value**
Intercept	-0.48 (0.02)	-18.31	< 0.001*
(PauseType)Silent	0.56 (0.007)	72.60	<0.001*
(Language)L2	-0.01 (0.007)	-1.78	0.07
(Task)Task2	-0.03(0.007)	-3.12	0.001**
Smooth Terms	**edf**	**F**	**p-value**
previous 600ms word frequency:L1	1.00	79.17	< 0.001*
previous 600ms word frequency:L2	1.67	93.56	< 0.001*
log(previous speech rate):L1	1.71	3.35	0.02*
log(previous speech rate):L2	1.00	8.31	0.003**
absolute time:L1	1.64	5.66	0.009**
absolute time:L2	1.32	13.57	< 0.001*
next first 600ms word frequency:L1	1.96	43.55	< 0.001*
next first 600ms word frequency:L2	1.87	23.00	<0.001*
next first 600ms word frequency:Filled pause	1.00	8.27	0.004**
next first 600ms word frequency:Silent pause	1.94	7.59	0.00**
next second 600ms word frequency:L1	1.00	1.00	0.31
next second 600ms word frequency:L2	1.00	7.11	0.007**
Random Effect	**edf**	**F**	**p-value**
Participant	18.48	1.68	< 0.001*
R-sq.(adj)	**Deviance explained**		
0.66	66.8%		

Hierarchical clustering algorithm

- Begin with n observations and a (distance/(dis)similarity) measure (e.g., Euclidean distance) of all pairwise dissimilarities. Treat each observation as its own cluster.
- For $i = n, n-1, ..., 2$:
    1. Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are the least dissimilar (= the most similar). Fuse these two clusters. The dissimilariy of these two clusters indicates the height in the dendrogram at which the fusion should be placed.
    2. Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters

Bibliography

[1] N. H. De Jong, "Fluency in second language testing: Insights from different disciplines," *Language Assessment Quarterly*, vol. 15, no. 3, pp. 237–254, 2018.

[2] W. J. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behavioral and brain sciences*, vol. 22, no. 1, pp. 1–38, 1999.

[3] P. Lennon, "Investigating fluency in efl: A quantitative approach," *Language learning*, vol. 40, no. 3, pp. 387–417, 1990.

[4] C. J. Fillmore, D. Kempler, and W. S. Wang, *Individual differences in language ability and language behavior*. Academic Press, 2014.

[5] P. Lennon, "The lexical element in spoken second language fluency," in *Perspectives on fluency*, University of Michigan, 2000, pp. 25–42.

[6] N. H. de Jong, J. Pacilly, and W. Heeren, "Praat scripts to measure speed fluency and breakdown fluency in speech automatically," *Assessment in education: Principles, policy & practice*, vol. 28, no. 4, pp. 456–476, 2021.

[7] B. Roberts and K. Kirsner, "Temporal cycles in speech production," *Language and Cognitive Processes*, vol. 15, no. 2, pp. 129–157, 2000.

[8] R. J. Hartsuiker and L. Notebaert, "Lexical access problems lead to disfluencies in speech," *Experimental psychology*, 2009.

[9] T. T. Kircher, M. J. Brammer, W. Levelt, M. Bartels, and P. K. McGuire, "Pausing for thought: Engagement of left temporal cortex during pauses in speech," *NeuroImage*, vol. 21, no. 1, pp. 84–90, 2004.

[10] N. H. de Jong, "Fluency in speaking as a dynamic construct.," *Language Teaching Research Quarterly*, vol. 37, pp. 179–187, 2023.

[11] A. Davies, *The native speaker: Myth and reality*. Multilingual Matters, 2003, vol. 38.

[12] F. Seifart, J. Strunk, S. Danielsen, *et al.*, "Nouns slow down speech across structurally and culturally diverse languages," *Proceedings of the National Academy of Sciences*, vol. 115, no. 22, pp. 5720–5725, 2018.

[13] F. Goldman-Eisler, "Psycholinguistics: Experiments in spontaneous speech," 1968.

[14] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and speech*, vol. 44, no. 2, pp. 123–147, 2001.

[15] T. J. Hastie, "Generalized additive models," in *Statistical models in S*, Routledge, 2017, pp. 249–307.

[16] G. B. Elena van Zuilen, "Analyzing stationarity in language learning: A comprehensive procedure."

[17] E. J. Pedersen, D. L. Miller, G. L. Simpson, and N. Ross, "Hierarchical generalized additive models in ecology: An introduction with mgcv," *PeerJ*, vol. 7, e6876, 2019.