Lengthened Vowels and Filled Pauses

An Acoustic Analysis of Two Fluency Features Uttered by Dutch and English Second Language Learners

Katarina Stankovic

Thesis submitted in partial fulfillment

for the degree of Master of Arts in Linguistics

(Modern Languages)

Supervisor: Dr. N.H. de Jong

Second reader: Dr. W.F.L. Heeren

Acknowledgements

I am glad to have had the opportunity to work on this project and I would like to thank my supervisor for providing guidance and support. I would also like to thank the second reader of my thesis in advance for her attention. Additionally, I would like to thank to Jos Pacilly for encouraging me to code.

Thank you to Xomara Vliese for driving me forward. Also, a big thank you to my father and Tanja Motina for hosting me during this COVID-19 lockdown.

Abstract

This thesis investigated whether the duration and standard deviations of F0, F1, F2 and F3 frequencies were similar in disfluency types (filled pauses and lengthened vowels) uttered by speakers of either sex (female and male). The analyses were done on second language Dutch and English speech materials that were collected for fluency evaluative purposes. The results of this study showed that intervals of vocalic articulation in filled pauses and lengthened vowels have a similar duration. The results for the standard deviation of F0-3 frequencies in Dutch disfluencies showed similarities for filled pauses and lengthened vowels uttered by females and males. Whereas the results from the English materials showed less uniformity for the effects of the disfluency types and sexes.

CONTENTS

Introduction

Hesitation and disfluencies might be heard in the speech of even the most eloquent speakers. Commonly, when native Dutch or English speakers are hesitant or disfluent, their speech may slow down or stop for a filled pause (e.g. *eh, uh, um, mm*) (Stouten, Duchateau, Martens, & Wambacq, 2006; Wieling et al., 2016). Language learners are prone to being disfluent or hesitant due to a partial knowledge of grammar and pronunciation (Segalowitz, 2010).

Filled pauses or slowed down speech can be observed and applied for objectively defining scales of speech fluency for language testing (*Common European framework of reference for languages: Learning, teaching, assessment*, 2001; De Jong, 2018; Tavakoli, Nakatsuhara, & Hunter, 2017). Previously, the speed of speech delivery and the number of filled pauses have been measured to gauge speech fluency (Bosker, Pinget, Quené, Sanders, & De Jong, 2013; Tavakoli et al., 2017). Often, filled pauses were manually annotated or orthographically transcribed for fluency research and applications, but listening for this feature can be costly work (De Jong, 2018). Promisingly, researchers develop formulas and algorithms for automatically measuring features from within utterances that can be applied to estimating fluency. For example, one script was written to detect syllables and measure speech rates (De Jong & Wempe, 2009) and another detected filled pauses too (De Jong, Pacilly, & Heeren, 2020). However, little is known about the acoustics of filled pauses uttered by speakers who are not native Dutch or English speakers and who have diverse accents. This thesis will study the acoustics of filled pauses uttered by speakers who were recorded speaking Dutch and English as their second language.

It was found that measuring the duration and calculating the standard deviation of formant frequencies and the fundamental frequency for vocalic phonation could be indicative of filled pauses in natively spoken English (Audhkhasi, Kandhway, Deshmukh, & Verma, 2009; Krikke & Truong, 2013; E. Shriberg, 2001). Experimental results showed that some lengthened vowels were falsely detected as filled pauses (Audhkhasi, Kandhway, et al., 2009; Kaushik, Trinkle, & Hashemi-Sakhtsari, 2010; Krikke & Truong, 2013; Stouten et al., 2006). Although the acoustic measurements have been applied to automatic fluency evaluation (Audhkhasi, Deshmukh, Kandhway, & Verma, 2009), there are no reports of the acoustic features in natively pronounced filled pauses also being present in filled pauses uttered by people speaking their second language. Previously, findings indicated that dynamic formant measurements from the vocalic parts of filled pauses can be speaker specific (Hughes, Wood, & Foulkes, 2016), and other formant frequency measures could remain the same in a second language, e.g. when females who spoke Dutch as their first language uttered filled pauses in English spoken as their second language (De Boer & Heeren, 2019). The aim of this research is to test the hypothesis that filled pauses and lengthened vowels could have similar acoustic features when spoken in Dutch or English as a second language.

In this thesis, I will compare acoustic measures, like duration (in seconds) and stability of pitches (standard deviation of F0 frequencies in Hertz) and formants (standard deviation of F1, F2 and F3 frequencies in Hertz), of the vocalic intervals in filled pauses and of lengthened vowels. The acoustic data analysed come from two corpora of Dutch and English speech spoken as a second language. Mixed effects models will be used to factor for random differences per speaker, and to test for fixed effects of the type of disfluency uttered and the speaker's sex. Firstly, in Chapter 1, this thesis will provide an overview of the premises drawn from previous acoustic analyses, research on disfluencies in Dutch and English and filled pause detection studies. In Chapter 2, the methodology of the research will be described. In Chapter 3, the results of the tests will be reported. The results and the studies will be discussed in Chapter 4. Finally, Chapter 5 will provide a critical evaluation of this thesis.

Chapter 1 Literature Review

1.1 Acoustic Analyses of Filled Pauses and Lengthened Vowels

Whilst studies of speech fluency were language bound, Dutch and English hesitation and disfluencies were reported as sounding similar. For instance, filled pauses were found to be pronounced as neutral vowels within open syllables or as neutral vowels in syllables that end with a labial nasal articulation (De Leeuw, 2007; Wieling et al., 2016). Researchers put forward that filled pauses could be detected using acoustic measurements taken from within vowel phoneme segments in spontaneous Dutch and English speech recordings (Audhkhasi, Kandhway, et al., 2009; Stouten et al., 2006). Moreover, lengthened vowels in these languages could share acoustic qualities with filled pauses, because these were often falsely automatically detected (Audhkhasi, Kandhway, et al., 2009; Krikke & Truong, 2013; Stouten et al., 2006). Hence, acoustic features of lengthened vowels posed a problem for detecting filled pauses in Dutch or English.

The resonances of speech sounds are modulated by a speaker’s vocal tract, and in good quality recordings these can be automatically tracked as peaks in spectra (Ladefoged & Johnson, 2015). Spectrograms show the peaks of spectra over time, and the peak tracks can be used to view the frequencies of a pitch and of formants (Ladefoged & Johnson, 2015). Although formant tracks are evidently most clearly present in spectrograms when oral vowels are perceived, some formant frequencies may change over time of pronunciation (Boersma, 2014).

Generally, the first two formant frequencies are used to describe vowels in most languages. The F1 frequency indicates the height of a speaker’s tongue and F2 frequency denotes whether the tongue gathers in the front or back of the oral cavity (Ladefoged & Johnson, 2015). Typically, acoustic analyses of vowels are done separately for female and male speakers when measuring F1 and F2 frequencies as peaks in spectra because these are presumed to result from resonances from within a speaker’s vocal tract which would differ in size (Boersma, 2014; Whiteside, 2001; Whiteside, 1996). Even so, measuring dynamic formant frequencies of vowels was reported as being a variable process because this involves material, equipment and procedures that could affect the resulting measurement (Kent & Vorperian, 2018). For example, one review of measurement methods showed that there is no consensus for at what points formant frequencies are representative for vowels, and they demonstrated with an image taken from a spectrogram that some monophthongs in English can exhibit formant raising or dipping, like in the vowel /u/ (transcribed with International Phonetic Association convention) which is present in the word “too” (Kent & Vorperian, 2018). Additionally, the review reported that vowels spaces ranged differently for adult female and male speakers (Kent & Vorperian, 2018). The researchers also explained that a formant frequency peak could appear near another spectrum peak depending on the speaker’s voice and the vowel quality (Kent & Vorperian, 2018).

Moreover, the VT has two openings, the oral cavity and the nasal cavity, and these both could impact the acoustic variation of voice (Boersma, 2014). Nasal sounds could be articulated when a speaker allows their body to filter resonances through the nasal cavity (Ladefoged & Johnson, 2015). Vowels pronounced adjacent to nasal closures could be coarticulated and might exhibit nasalization (Boersma, 2014; Kaiser, 1997). Boersma (2018) pointed out that the algorithms which track formants in automated formant analyses should be used with caution and explained that automatic formant analysis should be done per resonance filter because the measurements could violate the assumption that the formant resonances come from the same articulatory gestures.

The results of experimental disfluency detection showed that the two filled pause variants and some lengthened vowels could be detected using the same acoustic features (Audhkhasi, Kandhway, et al., 2009; Stouten et al., 2006). This suggests there is no change in pronunciation of vowels throughout filled pauses (Audhkhasi, Kandhway, et al., 2009). The filled pause detection studies did not indicate separate methodological approaches for tracking the frequencies of female and male voices (Audhkhasi, Kandhway, et al., 2009; Stouten et al., 2006). Whilst, algorithms for measuring fundamental and formant frequencies require adjustments be made for females and males (Boersma, 2014). This thesis will examine whether vocalic parts of filled pauses and lengthened vowels share acoustic features and whether those acoustic features are similar for female and male Dutch and English second language speakers.

1.2 Prosody of Filled Pauses and Lengthened vowels

Research of prosody has aimed to uncover how speakers' intonations map to phrases and how speakers' pitches vary regionally. For instance, one study compared how pitch contours, measured as the F0 frequency, were realized in different semantic focus structures in sentences uttered by speakers in and around the Netherlands (Peters, Hanssen, & Gussenhoven, 2014). When people speak, the prosody of what is being said can be measured by prominences within units like the syllable or an intonational phrase (Frota, Arvaniti, & D'Imperio, 2012). Standard Dutch or English words can be differentiated by the duration of syllables and not by prominent tones of voice within syllables, hence these were termed stress accented languages (Jun, 2006). However, prosodic cues, like intonation, amplitude, duration and silent pauses, are present in Dutch and English speech and these might disambiguate meaning and communicate affect (Gussenhoven, 2016; Turk & Shattuck-Hufnagel, 1996). The prosodic hierarchies of Dutch and English contain small units such as phonological phrases, i.e. where feet and syllables lend prominence to word stress, and they contain larger prosodic units known as intonational phrases (Jun, 2006). Additionally, English has intermediate phrase between the previously mentioned prosodic units (Gussenhoven, 2006).

Two notable acoustic phenomena of fluent English intonational phrases were prolonged syllables at prosodic phrase boundaries, followed by silent pauses (Ferriera, 1993; Shriberg, 2001). Shriberg (2001) reported that the prolongation of syllables and the interruption of speech could also be prevalent in disfluent regions of speech which involved hesitation, however the degree of lengthening in syllables was longer in disfluency than at fluent phrase boundaries. The prolongation of speech due to a speaker hesitating was also reported as having an intonational difference from pitch movements in the lengthening of syllables before fluent phrase boundaries (Shriberg, 2001). Shriberg (2001) wrote that tones in the region of disfluency, prior to interruption of speech, would have a flat or slightly falling pitch contour, instead of the pitch movements that otherwise would prevail in fluent English (Shriberg, 2001). Additionally, Shriberg (2001) stated that this was not the case where the disfluency involved the speaker detecting error because usually there was no lengthening involved prior to an interruption of speech. Shriberg (2001) asserted that the intonation of filled pauses is like lengthened speech when a speaker hesitates, these have a flat or slightly descending pitch contour.

Clark and Fox Tree (2002) claimed that filled pauses could also be classified as interjections, meaning that these did not have syntactic constituents and that these will be dependent on intonation or intermediate phrases. However, Clark and Fox Tree (2002) also reported that filled pauses could make up a single prosodic unit. Another study of the intonation in filled pauses done in respect to the intonation of surrounding speech illustrated that tones were not necessarily flat for this disfluency (Shriberg & Lickley, 1993). Graphical linear representations of four speakers' pitches at the start of, and end of filled pauses showed that the tone height of "uh" and "uhm" in native English speech would often get lower towards the end of the syllable. Still, some lines representing the pitch of individual filled pauses were positively sloped, indicating a rising intonation (Shriberg & Lickley, 1993). This study concluded that filled pauses which occurred within clauses were predictable based on the height of the speakers' pitch in surrounding speech, because tones of filled pauses were relatively centered within the range of tones in surrounding speech (Shriberg & Lickley, 1993).

Correspondingly, researchers looked at whether speakers' pitches are useful prosodic cues for improving filled pause detection devices and many of these studies proposed that speakers' intonations would be flatter in filled pauses than in speech (Audhkhasi, Kandhway, et al., 2009; Kaushik et al., 2010; Krikke & Truong, 2013; Stouten, 2008; Stouten et al., 2006). However, Stouten et al. (2006) excluded pitch as a viable feature to use with their filled pause detection device because they found no significant differences in the mean pitch frequencies, the variation of pitch or the relative pitch frequencies for "uh" or "uhm" and other phoneme-like segments of speech. Audhkhasi et al. (2009) also compared how filled pauses were detected by the standard deviation of F0 frequencies because they presumed there were no tonal contours during a filled pause. They found that this approach had the lowest precision and suggested that filled pauses do not always have a flat pitch (Audhkhasi, Kandhway, et al., 2009). Admittedly, the results of previous filled pause detection studies showed that pitch might not provide an accurate indication of disfluent native speech.

Yet, the experimental filled pause detection studies showed that durational features could improve searching for a spectral stability. Similarly, Shriberg (2001) reported that the duration of the vocalic parts of filled pauses were longer than vowels which resembled the vocalic sounds heard in filled pauses of the ATIS corpus, like /ə/ and /ʌ/ transcribed in IPA convention. However, the average speed of speech and articulation could be different for language learners and native speakers (Bosker, Quené, Sanders, & De Jong, 2014; Trofimovich & Baker, 2006). In the study of what made Dutch speech sound fluent, by Bosker et al. (2014), native speech was manipulated to be slower, in order to match the average speech and articulation rate for Dutch language learners. The result was that the slowed down native speech was considered less fluent than the originally recorded speech (Bosker et al., 2014). In contrast, recorded learner speech was sped up to the average native speakers’ speech and the results indicated that raters considered the manipulated speed more fluent than the original speech record (Bosker et al., 2014). The temporal measure used for the speech rate manipulation was the number of syllables per second and for the articulation rate manipulation was the number of syllables per second excluding silent pauses (Bosker et al., 2014). Such findings showed that the articulation rate was slower for second language learners (Bosker et al., 2014; Trofimovich & Baker, 2006).

Hence, the duration previously attributed to filled pauses and lengthened vowels in experimental detection studies will be compared in this thesis. Additionally, this thesis will investigate whether the standard deviations of F0 frequencies are similar in filled pauses and lengthened vowels of females and males speaking Dutch and English as a second language, because it is not clear whether these share a flat intonation.

1.3 Disfluencies in Dutch and English

People learn first languages (L1s) from birth, whereas second languages (L2s) are learned later on in life (Meisel, 2009). Some individuals might acquire the L2 through formal education and these people can be tested for language proficiency longitudinally (Geeslin & Long, 2014). The pronunciation of words and production of sentences in a L2 could change with the amount of exposure a learner has to the L2 and the learner’s L1 could also give rise to a particular accent in their L2 pronunciation of words or utterances (Geeslin & Long, 2014).

People acquiring L2s gain sociolinguistic competences that allow them to distinguish dialects and accents, learners would learn markers of "social class, regional provenance, national origin, ethnicity, occupational group", such as phonology, vocal rhythm or loudness and paralinguistics (*Common European Framework of Reference for Languages: Learning, teaching, assessment*, 2001, p.121). The second language learner could choose to speak a L2 dialect or with a particular vocal register or accent and this might also add variation to the pronunciation of L2 words or utterances (*Common European framework of reference for languages: Learning, teaching, assessment*, 2001).

Variants of filled pauses (previously transcribed as *uh, uhm, er, em, mmm)* in Dutch and English have been considered socio-phonetic speech tokens that speakers use preferentially when hesitating (Braun & Rosin, 2015; De Leeuw, 2007; Mcdougall & Duckworth, 2018; McDougall & Duckworth, 2017; Segalowitz, 2010; Tottie, 2011; Wieling et al., 2016). Three filled pause variants of Dutch and English have been described single syllables, consisting of either: (i) only a neutrally articulated vocalic sound, (ii) a neutrally articulated vocalic sound that transitions into a bilabial nasal consonant or (iii) solely a bilabial nasal consonant (De Leeuw, 2007). Analyses from different research fields have provided insights to how specific filled pauses variants are for individual speakers in Dutch and English. Section 1.3.1 outlines issues relating to how distinguished prolonged syllables and filled pauses are in spontaneous fluent speech. Section 1.3.2 describes how filled pauses could change during SLA. Section 1.3.3 contains findings from forensic speaker comparisons done in English.

1.3.1 Lengthened Syllables and Filled Pauses in Spontaneous Speech

Shriberg (2001) found that the duration of speech within regions of disfluent English could be similar using the structure of repair first described by Levelt (1983). The structural aspect of speech repair allowed the disfluency to be defined by regions of speech where there was an interruption of lexical fluency. The structure of repair consisted of three regions, firstly there is a reparandum (i.e. the fluent speech that would be interrupted due to a speaker detecting error), then an editing phase (i.e. a region in the structure of repair that precedes the commencement of fluent speech, where the filled pause could be found) and finally a repair (i.e. the resumption of fluency). Shriberg (2001) explained that when the reparandum and the editing phase were removed, the speech would be lexically fluent. The acoustic analyses of disfluency within the structure of repair shed light on some acoustic patterns in regions of the disfluent English speech, like that the reparandum and filled pauses within the editing phase could both be prolonged and also share an intonation (Shriberg, 2001). Lengthened vowels were not studied as a discrete speech disfluency type in acoustic studies, although the speed of Dutch and English speech was commonly measured as a feature of fluency (De Jong, 2018). Shriberg (2001) stated that English speech may be prolonged and coarticulated with filled pauses, and also clarified that unclear phoneme boundaries could make it difficult to differentiate filled pauses from lengthened word endings, in cases where word-final syllables are articulated similarly to filled pauses.

1.3.2 Filled Pause Variability during Second Language Acquisition

Few studies of acoustic variation between L1 and L2 filled pauses were done with aims of understanding whether speakers showed phonetic integrity during second language acquisition (here forth SLA) (De Boer & Heeren, 2019; Gósy, Gyarmathy, & Beke, 2017). For instance, De Boer and Heeren (2019) explored whether females acquiring English as a L2 whose L1s were Dutch had adapted their acoustic productions of filled pauses over periods of time. They were interested in how measures of duration, pitch and formant frequencies could be predicted in a second language by the L1, and if filled pauses changed with exposure to the L2 (De Boer & Heeren, 2019). De Boer and Heeren (2019) have shown that vocalic filled pause interval midpoint F3 frequencies remained similar in the L1 and L2. Whereas, midpoint measures of F1 and F2 frequencies and proportions of “uh” and “um” changed with exposure to the L2 (De Boer & Heeren, 2019). Similarly, Gósy et al. (2017) looked at whether these phonetic properties changed for vocalic filled pauses uttered by people acquiring English as a L2 whose L1s were Hungarian. Gósy et al. (2017) suggested that speakers produced filled pauses with shorter durations at advanced SLA proficiency levels, and they showed that the formant frequencies did not change with the level of proficiency, i.e., the formant frequencies were not different between the L1 and the L2 (Gósy et al., 2017).

Only De Boer and Heeren (2019) had looked at how F0 frequencies changed in filled pauses uttered by female speakers acquiring a L2. There are no other studies that looked at how F0 frequencies of filled pauses changed during language acquisition, or whether filled pauses could be detected using this prosodic feature in L2 speech samples. They reported that measured F0 frequencies at the midpoints of filled pauses did not change for Dutch students acquiring English (De Boer & Heeren, 2019).

Due to the ambiguity of whether the duration, fundamental frequency and formant frequency dynamics are stable in filled pauses during SLA, this study will look at whether these measures are similar in filled pauses and lengthened vowels uttered by females and males in their L2.

1.3.3 Specificities of "uh" and "um" from Forensic Speaker Comparisons

In some countries, forensic linguistic studies of speech were done for purposes of identifying or profiling speakers (Foulkes & French, 2012). Recent studies have shown that filled pauses are effective for determining speaker specific speech traits (Hughes et al., 2016). For example, Hughes et al. (2016) had verified that the duration of nasal articulation in filled pauses within English speech could be useful for forensic speaker comparisons. Hughes et al. (2016) had shown filled pauses that consisted of both vocalic and nasal articulation were best compared with other filled pauses using dynamic measures of the first three formant frequencies and the duration of the nasal sound. Whereas, filled pauses containing a single vocalically articulated syllable had been best compared with other filled pauses using formant measures which were taken at midpoint of the duration of the filled pause (Hughes et al., 2016). These findings showed that the formant frequencies of filled pauses were speaker specific, and that formant frequencies could dynamically differ in the vowel-like realizations of "uh" and "um" (Hughes et al., 2016). Although filled pauses were shown to be speaker specific amongst same sex individuals who spoke the same regional dialects, not many studies researched whether static and dynamic formant frequency measures from filled pauses in a second language are also speaker specific[1].

[1] A Master thesis submitted to Leiden University researched this topic. It was found that acoustic measures, like the mean duration and static F3 frequencies of filled pauses were dependent on the speaker and the mean duration and static F1 frequency could be independent of the language (Dutch or English) being spoken (see Sleebos, 2018).

1.4 The Detection of Filled Pauses in Spoken Dutch and English

Sometimes speech disfluencies are misrecognized by automatic speech recognition (ASR) devices because these components of speech fluency are often not considered within language models based on a grammar, lexicon and pronunciation (Stouten & Martens, 2004). Filled pauses were included into Dutch and English ASR language models, by researchers who sought to reduce word recognition error rates, however it was found that this would not significantly improve the ASR (Stouten et al., 2006). Hence, these researchers looked at further reducing word recognition error rates by detecting filled pauses externally from the ASR using acoustic features and this showed better results. The external detection of filled pauses was done using multiple spectral features. Yet, further efforts to detect filled pauses in English used fewer acoustic features during intervals of vocalic articulation as indication of disfluency.

Audhkhasi et al. (2009) compared three filled pause detection techniques and found that the most accurate one was based on a stability of formant frequencies. The proposed formant-based technique for automatically detecting filled pauses was calculated as a log likelihood ratio of the standard deviation for F1 frequencies within 11 analysis frames, the same was done for F2 frequencies (Audhkhasi, Kandhway, et al., 2009). Audhkhasi et al. (2009) explained that the standard deviation of F1 and F2 frequencies measured within filled pauses would increase with larger windows, and they stated that the analysis frame rate was 10 milliseconds. By measuring formant frequencies automatically with this setting and calculating the standard deviation over 11 frames, they found that 78.7% of the standard deviation of F1 frequencies in filled pauses was under 40 Hertz (Hz). For normal speech, only 19.5 % of the standard deviation for F1 frequencies was under 40 Hz.

Other experimental research on filled pause detection showed that formant frequencies were more stable in filled pauses than in other vowels too; the filled pause detection studies reported less standard deviation of formant frequencies in the vocalically articulated sounds of filled pauses than that was observed in other vowels (Audhkhasi, Kandhway, et al., 2009; Krikke & Truong, 2013). For instance, Krikke and Truong (2013) stated the stability found within filled pauses was measured as the standard deviation of F1 and F2 frequencies for 9 analysis frames at a frame rate of 10 milliseconds. Yet, in these studies of filled pause detection, some lengthened vowels affected the degree of accuracy of filled pause detection, e.g. when words were articulated with tense or prolonged and steady vowels, like in the words “too” or “no” (respectively, Krikke & Troung, 2013, p. (respectively, in Audhkhasi et al., 2009, p. 4; Krikke & Truong, 2013, p. 4). Although formant tracks are evidently most clearly present in spectrograms when oral vowels are perceived, some formant frequencies may change over time of pronunciation (Boersma, 2014).

Notwithstanding, Audhkhasi et al. (2009) wrote that they referred to lengthened vowels as filled pauses too, because these were perceived as disfluencies and some would be detected with their technique. Namely, Audhkhasi et al. (2009) had 484 filled pauses and lengthened vowels in their test data. However, when reporting on the accuracy of the tested techniques, they were most concerned with 192 filled pauses that were labeled as prominent due to acoustic features, like energy, duration and proximity to silent pauses, by one listener. All previously reviewed filled pause detection studies implemented a durational threshold to define whether an observed spectral stability was significantly different to that observed in other vowels, e.g. Stouten et al. (2006) claimed that genuine filled pauses were at least 0.15 s long and Audhkhasi et al. (2009) wrote that the optimal threshold duration for detection would be 12 frames.

Additionally, Audhkhasi, Kandhway, et al. (2009) listed that the recall of filled pause detection was affected by filled pauses that had a low volume, that were coarticulated with surrounding speech or that had a short duration. Hence, filled pauses went undetected when these were preceded and followed by words (Audhkhasi, Kandhway, et al., 2009). Yet, they had given little attention to the phonetic varieties of filled pauses within the corpora. They impressionistically had described filled pauses as “ahh” or “umm” (Audhkhasi et al., 2009). Despite acknowledging the difference of articulation, by spelling out their aural impressions of filled pauses, Audhkhasi et al. (2009) had not included a description for a separate treatment of filled pause variants to the methodology of their proposed detection technique. Moreover, the filled pause detector proposed by Audhkhasi et al. (2009) was tested on speech from what might be assumed a homogenous group of people and the paper does not state whether the speakers were all of a single sex or whether separate approaches are necessary for speech of females and males.

1.5 Research Questions and Hypotheses

The filled pauses in English speech were found to have acoustic features that could also be present in lengthened vowels, i.e. a duration of vocalic articulation that is prolonged and stable F0 and F1-3 frequencies contours (Audhkhasi, Kandhway, et al., 2009). Filled pauses and lengthened vowels uttered by second language learners of Dutch and English might also share these acoustic features. Hence, the following research questions will be explored:

1. Are the vocalic parts of filled pauses and prolonged vowels similar in duration? Does this differ for female and male speakers of L2 Dutch or English?
2. Do second language speakers have a stable fundamental frequency in filled pauses and in lengthened vowels? How does this compare for female and male speakers of L2 Dutch or English?
3. Is the stability of formant frequencies similar in filled pauses and prolonged vowels? Are there differences for formant stabilities in these disfluencies for female and male speakers of L2 Dutch or English?

The following list of hypotheses come from the literature reviewed:

1. The duration of filled pauses and lengthened vowels will be similar, regardless of the speaker’s sex;
2. Speakers will have an equally stable fundamental frequency in lengthened vowels and in filled pauses, regardless of their sex;
3. The stability of formant frequencies in filled pauses will be more stable than in lengthened vowels, regardless of the sexes of speakers.

Chapter 2 Methodology

This chapter will define how the research questions listed in the previous section will be answered. What follows are descriptions of the materials, the data, the data collection procedures and the statistical analyses.

2.1 Methodological Approach

The Dutch and English second language speech material used in the following studies were not recorded in the same circumstances. Recording procedures can affect acoustic analyses (Zsiga & Podesva, 2014). Therefore, the acoustic analyses of filled pauses and lengthened vowels will be conducted separately for the two sets of materials.

The first aim of this study is to compare the duration of vocalic parts of filled pauses and lengthened vowels perceived in the materials. Previously, phoneme-like segments of speech were found to have a shorter duration than those within filled pauses, this was discovered using a segmentation algorithm created and described by Stouten et al. (2006). Similarly, Shriberg (2001) showed that the duration within the vocalic articulation in filled pauses is longer than vowels with a similar pronunciation quality. However, when Audhkhasi et al. (2009) tested their filled pause detector, they implied that vowels which were prolonged in native speech could share spectral features which included a durational threshold used for detecting filled pauses. The previous methodologies for measuring the duration for the vocalic parts of filled pauses were not described by Shriberg (2001), and those used by Stouten et al. (2006) will not be replicated. Instead, the duration will be measured in seconds from the start until the end of vocalic articulation where filled pauses and lengthened vowels are perceived (more details follow in Section 2.3).

The second aim is to examine the stability of female and male speakers’ fundamental frequency during a perceived filled pause or lengthened vowel disfluency. Previously, Stouten et al. (2006) wrote they examined pitch variation, but they had not reported the methodology for how this was measured. The other acoustic features they examined were described in relation to phoneme-like speech segments (Stouten et al., 2006). Presumably, they looked at pitch variation within phoneme-like segments marked by their speech segmentation algorithm. Audhkhasi et al. (2009) recorded measuring pitch stability as the standard deviation of F0 frequencies where energy thresholds did not indicate silences or fricatives. Whilst, Shriberg and Lickley (1993) measured F0 frequencies of filled pauses at the start and end of “uh” and “um”, because they aimed to find the relation of these syllables and the intonation of surrounding utterances. For this thesis, the fundamental frequencies will be measured within vocalic part of filled pauses and in lengthened vowels, pertaining to what was examined in previous detection studies. However, Audhkhasi et al. (2009) had not reported different methods for measuring the stability of pitch for female and male speakers. Whilst, practically analyses of pitch must be done separately for the dimorphic sexes dues to physiological differences in the vocal tract that affect pitch tracking algorithms (Boersma, 2014). Hence, pitch will be automatically measured using the “To Pitch (ac)” functionality in PRAAT (Boersma & Weenink, 2018), and adjustments will be made to the automatic F0 frequency tracker to comply with what Boersma (2014) advised for measuring speech of females and males using this algorithm (details in Section 2.4).

The third aim is to measure the amount of stability in formant frequencies when filled pauses and lengthened vowels are perceived in the speech materials. Previously, Audhkhasi et al. (2009) examined formant frequencies within recorded articulation that was distinguished by energy thresholds which would not indicate silences or fricatives. Krikke and Truong (2013) also examined how filled pauses were detected using formant measures, along with other acoustic measures, but did not report the methods of taking these measurements prior to training and testing the detection device. In acoustic analysis, measuring formant frequencies was reported most reliable during the articulation of oral vowels (Boersma, 2014). Moreover, according to Boersma (2014), the formant tracker used for the following studies must be adjusted for female and male speech in order to improve the accuracy of the algorithm (details in Section 2.4). Hence, the tracking of the formant frequencies within filled pauses and lengthened vowels will be modified for female and male speakers, and will only be measured where articulation is perceived as vocalic. The formant frequencies will be automatically measured using the “To Formant (burg)” in PRAAT (Boersma & Weenink, 2018). Additionally, in the raw material processing protocol for collecting F1-3 data, 5% of the lower and higher quantile measures of formant frequencies were used to exclude extreme measurement values per disfluency.

2.2 Materials

The second language speech materials used in this study were created for prior research done on aspects of second language speech fluency and second language testing. Initially, digital audio files containing the materials were distributed via a secure online platform. The speech materials were transferred as WAV-files. Additionally, blank (.TextGrid) files were also provided for the purpose of annotation, and these were processed using PRAAT (Boersma & Weenink, 2018). A PRAAT script written by Pacilly (2018) was provided too to ease the annotation procedure. What follows in Section 2.2.1 and 2.2.2 are detailed descriptions of the origin of these recordings.

2.2.1 The Dutch Speech Materials of Second Language Learners.

There are 114 Dutch recordings with a total duration of 2267 seconds, i.e. approximately 38 minutes. Bosker, Pinget, Quené, Sanders and De Jong (2013) previously collected the speech samples from the "What is Speaking Proficiency" (WISP) corpus to examine the relationship between aspects of utterance fluency and perceived fluency ratings. The WISP corpus contains speech elicited with tasks designed for testing speech adequacy and proficiency in a study done by De Jong, Steinel, Florijn, Schoonen and Hulstijn (2012). The speech samples had been preselected for the study done by Bosker et al. (2013) so that all recordings were adjusted to start at a phrase boundary and end at a silent pause of 250 milliseconds. Each speech sample is approximately 20 seconds in length (Bosker et al., 2013). This sample from the WISP corpus includes 38 individual speakers, of which eight are Dutch first language speakers (Bosker et al., 2013). Those eight first language Dutch speakers were excluded from this study. The thirty second language speakers of this sample were first language speakers of either English and Turkish (Bosker et al., 2013). The remaining samples of second language speech contained 1795 seconds for this analysis, i.e., approx. 30 minutes. There are 19 female and 11 male second language Dutch speakers. The sampling frequency of these recordings is 44100 Hz.

2.2.2 The English Speech Materials of Second Language Learners.

There are 120 files containing second language English speech, with a total duration of approximately 13905 seconds, i.e., approx. 232 minutes. The recordings were collected for a study conducted by Tavakoli et al. (2017) that had tested whether analytic fluency measures could provide evidence for characterizing fluency descriptors for “Aptis Speaking tests”. The recordings contained speech of 32 speakers with varying first languages (Tavakoli et al., 2017). However, due to the noisy quality of some recordings and a prior instruction for sampling for the study of De Jong et al. (2020, preprint), 63 files were not used. A sample of approximately 120 minutes remains annotated for this study. The sample contains speech of 6 female and 12 male speakers. The sampling frequency of these recordings is 11025 Hz.

2.3 Data

Each recording was manually annotated in PRAAT. Firstly, boundaries of filled pauses and lengthening were marked where I perceived these to begin and end. Secondly, upon further inspection, the articulation of a vocalic sound was manually annotated where I perceived it to begin and end. Visually inspecting spectrograms allowed me to distinguish when vocalic sounds ended, and possibly when the /m/ sound (transcribed with IPA) of a filled pause variant like “um” began. See Figure 1 below, it depicts the user-interface of PRAAT with an exemplary annotated TextGrid where a filled pause was perceived.

The materials which were annotated as vocalic sounds were processed with a script (for details see Section 2.4). The script extracted the following data into a table:

- A grouping identifier (for Dutch speaking females,“NL_females”; for Dutch speaking males, “NL_males”; for English speaking females, “UK_females”; for English speaking males “UK_males”);
- A unique speaker identifier and a unique annotated disfluency identifier;
- The type of disfluency uttered (lengthened vowel or filled pause) when the vocalic part is measured. Additionally, the filled pause variant (“v” or “vn”) perceived when the vocalic part is measured is printed;
- The duration of the total disfluency and the vocalic articulation (in seconds);

- The computed standard deviation of F0-3 frequencies for the annotated interval containing vocalic articulation (in Hertz);

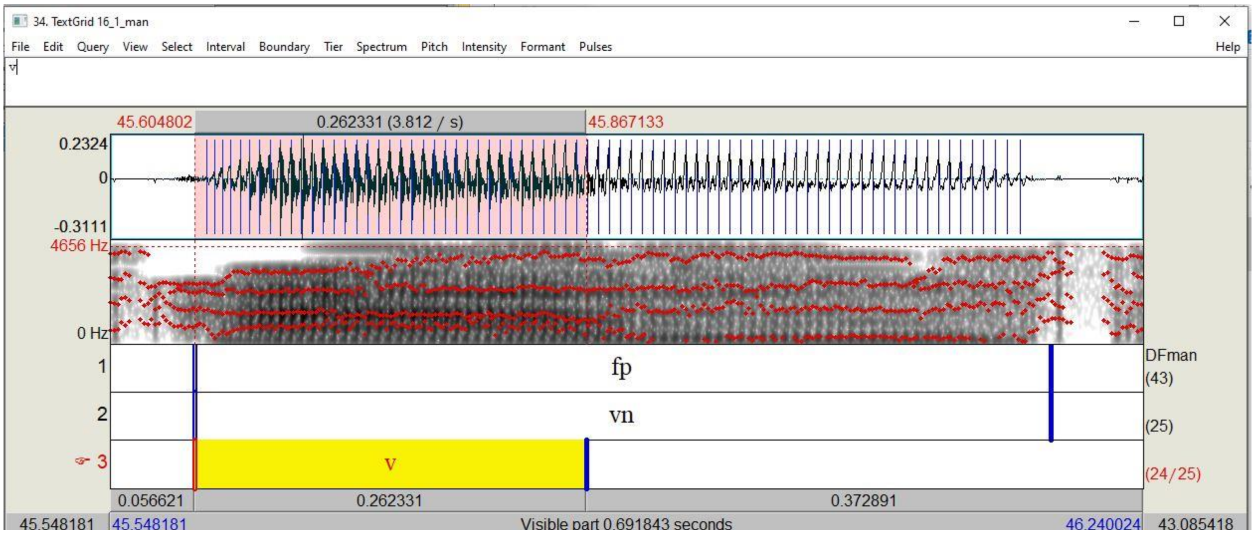


Figure 1 A Sound object and TextGrid object opened simultaneously in PRAAT, exhibiting how a disfluency (filled pause) was annotated in three tiers

2.4 Data collection procedure

Throughout this thesis, acoustic measures were taken using speech processing software PRAAT (Boersma & Weenink, 2018). Following the annotation procedure described in Section 2.3, a (.praat) script was used to extract the acoustic data. The order of the procedures within the script were:

1. The recordings and annotations are sorted by the language spoken and the perceived sex of the speaker. The script sorts the materials by opening the (.wav) audio files listed in one of four (.txt) files. I compiled the lists manually by noting file names that shared language spoken and speaker sex. Annotated (.TextGrid) files corresponding to the (.wav) files name are also opened. The opened files are saved in four separate directories.
2. The audio file and annotated (.TextGrid) file with the same name are opened from within the grouping directories. The annotated file containing three tiers. The first tier contained information on when filled pauses and lengthening was heard. The second tier is an annotation of the manner of articulation perceived, i.e. vocalic (v), vocalic transitioning to a bilabial nasal (vn), or nasal (n). The

third tier annotates the start and end of perceived vocalic articulation within the filled pauses and lengthened vowels. Every tier also contains empty intervals.

3. The duration is calculated based on the intervals in the third tier, labelled "v". The duration is calculated as the time minus the start time of the vocalic interval.
4. The standard deviations of the (F1, F2, F3) formant frequencies are calculated based on the intervals in the third tier, labelled "v", using the steps listed below. The start and end times calculated for each interval in procedure number 3 are used in "a.", hereunder. Empty arrays are filled, using a loop, with the standard deviations of F1, F2 and F3 frequencies for each interval within a recording, with the following procedures:
   a. Each interval labelled "v" in the third tier is extracted, 10 milliseconds before and after the interval boundaries;
   b. The "To formant (Burg method)" algorithm is used to automatically measure the formant frequencies. For female speakers, the maximum frequency was 5500 Hz and the algorithm determined 5 formants, and for male speakers, the maximum frequency was 5000 Hz and the algorithm determined 5 formants (as proposed by Boersma, 2014).;
   c. Every formant frequency value, if not undefined, is compared to the 5th and 95th quantile value found for the formant frequencies within the disfluency interval. The formant frequencies between these values are stored in a new empty vector. Values that were smaller or larger than the quantiles were replaced by the mean formant frequencies of the interval to exclude possible formant tracking errors and these are also stored in the new vectors. The standard deviations for the formants are computed with a PRAAT function for vectors.
   d. The calculated standard deviation of F1, F2 and F3 frequencies per interval labelled "v" are printed in the outcome table.
5. The standard deviation of the (F0) fundamental frequency is calculated based on the intervals in the third tier, labelled "v", using the procedures listed below. The start and end times calculated for each

interval in step 3 are used. An empty vector is filled with the standard deviation of F0 frequencies for each interval within a recording, with the following procedures:

a. Each interval in the third tier is extracted, 10 milliseconds before and after the interval boundaries;
b. The “To pitch (ac)” algorithm is used to automatically measure the formant frequencies. For female speakers, the pitch floor was 100 Hz and pitch ceiling was 500 Hz, and for male speakers, the pitch floor was 75 Hz and pitch ceiling was 300 Hz (as proposed by Boersma, 2014).;
c. The standard deviation of the F0 frequencies was calculated using all pitch estimates from the start to the end of the disfluent interval by querying the standard deviation using PRAAT’s “Get standard deviation” function for Pitch objects.
d. The calculated standard deviation of F0 frequencies are placed in the vector.

6. As the script opens the files and processes the measurements (duration (s), standard deviation of F0, F1, F2, and F3 frequencies (Hz)), the script also adds metadata (unique speaker code, unique disfluency code, the annotations from tiers 1 and 2 directly above the measured vowel and the speaker’s group) to the outcome table.
7. The table is saved.

2.5 Statistical Analysis

The sample from WISP corpus sample contains 395 annotated intervals, i.e. 362 filled pauses and 33 lengthened vowels. The sample from the Aptis Speaking test corpus holds 1632 annotated intervals, i.e. 1328 filled pauses and 304 lengthened vowels. The annotated filled pauses and lengthened vowels were examined using the statistical software R (R Core Team, 2019) in RStudio (RStudio Team, 2018).

To address the research questions of this study, mixed effects models were used to test whether the outcome acoustic variables, i.e. durations (s) and the standard deviations of F0-3 frequencies (hertz), could be factored by an interaction of the perceived disfluency type (lengthened vowel or filled pause) and the speaker’s sex (female or male). Considering that all speakers produced lengthened vowels and filled pauses repeatedly, the random factor in the following studies will be the speaker (Winter, 2019).

Initially, the data was filtered with R package “dplyr” into two data frames for either set of materials (Wickham, François, Henry, & Müller, 2019). The outcome variables of this study were tested statistically using the R package “LmerTest” (Kuznetsova, Brockhoff, & Christensen, 2017). The LmerTest package fits mixed effects models and does t-tests providing p-values using Satterthwaite's method (Kuznetsova et al., 2017). The fixed effects for each following model were dummy coded using the default settings in R.

To counteract possible family-wise error, a Bonferroni correction is used where multiple testing is done per sample. Hence, the previous alpha level ($\alpha = 0.05$) is divided by the number of tests per sample (4), making the Bonferroni corrected alpha level ($\alpha = 0.0125$).

The data was examined for missing values and in order to meet the assumptions for linear regression, and it was further filtered into data frames used for hypothesis testing. The assumptions of each model were assessed visually (Appendix A contains histograms of residuals, qq-plots of residuals and residual plots for each model). Additionally, the homogeneity of variance for the residuals per model was tested with Levene’s test ($\alpha = 0.01$) using the R package “car” (Weisberg, 2019).

Within the Dutch materials 23 data points (5.8%), and 85 data points within the English materials (5.2%), for the standard deviation of F0 frequencies were “undefined” in the initial data frames because those disfluent intervals did not contain at least two analysis frames with this measure for pitch. Also, one data point for this outcome variable was removed from the English subset because the value was extreme (150 Hz). The mixed models for these outcome variables did not meet the normality assumption for regression, but this was resolved for both subsets with a nonlinear transformation of the data using a logarithm (Winter, 2019).

The remaining outcome variables did not contain missing data, but most residuals were not normally distributed and some models violated the assumption of homoscedasticity. The residuals from the model for the outcome variable duration in the Dutch subset were also not normally distributed, and this was resolved using a logarithm transformation of the data. Whilst, the residuals of the model for duration from the English subset were not homoscedastic, so 181 data points (11%) were removed to achieve a significant Levene's test result.

Additionally, the residuals of the models for the standard deviation of F1 frequencies for both English materials were not normally distributed and both models were resolved with a logarithm transformation. The residuals of the model for the standard deviations of F2 frequencies within the English materials breached the assumption of normality and homoscedasticity but the assumptions for the model were met when 14 data points were removed (<1%) and the data was transformed with a square root transformation. When modeling the standard deviation of the F3 frequencies in the English materials, the normality assumption was met with a logarithm transformation, but the homoscedasticity assumption could only be met when 346 data points (21%) were removed.

Chapter 3 Results

The duration and spectral stability of lengthened vowels and filled pauses was examined amongst female and male second language learners of Dutch and English. The mean, standard deviation (SD) and ranges of the outcome variables, i.e., the duration and standard deviations (SD) of F0-3 frequencies, for the Dutch and English speech samples can be found in Table 1 and Table 2 respectively. The results of the mixed effects models will be reported in this chapter per outcome variable and material sample.

Table 1 Descriptive statistics for the outcome variables of Dutch materials, tabulated per fixed factor

		Females (n = 19)		Males (n = 11)	
		Lengthened vowels	Filled Pauses	Lengthened vowels	Filled Pauses
Duration (s)					
	mean	0.34	0.39	0.34	0.37
	SD	0.19	0.20	0.23	0.23
	range	0.13 – 1.01	0.08 – 1.28	0.14 – 0.96	0.09 – 1.29
SD of F0 (Hz)					
	mean	9.83	9.65	1.72	7.08
	SD	13.49	11.94	1.2	10.72
	range	1.33 – 63.37	0.75 – 63.42	0.63 – 4.1	0.62 – 56.58
SD of F1 (Hz)					
	mean	96.37	102.47	85.67	106.59
	SD	30.91	44.71	29.96	56.04
	range	43.22 – 160.2	36.73 – 328.84	33.73 – 135.64	29.64 – 453.71
SD of F2 (Hz)					
	mean	368.55	318.26	290.79	294.62
	SD	104.68	110.06	79.31	105.98
	range	195.57 – 512.95	151.44 – 983.57	134.97 – 429.24	107.05 – 661.92
SD of F3 (Hz)					
	mean	592.41	543.01	551.1	509.8
	SD	160.65	168.97	140.94	167.24
	range	332.29 – 945.84	229.22 – 1404.81	274.09 – 824.94	190.67 – 1214.98

Table 2 Descriptive statistics for the outcome variables of English materials, tabulated per fixed factor

		Females (n = 6)		Males (n = 12)	
		Lengthened vowels	Filled Pauses	Lengthened vowels	Filled Pauses
Duration (s)					
	mean	0.38	0.34	0.38	0.35
	SD	0.11	0.12	0.12	0.13
	range	0.15 – 0.68	0.15 – 0.69	0.15 – 0.68	0.15 – 0.69
SD of F0 (Hz)					
	mean	7.08	12.03	4.87	7.41
	SD	8.49	16.71	7.69	11.7
	range	0.7 – 49.07	0.52 – 146.13	0.27 – 56.57	0.11 – 86.36
SD of F1 (Hz)					
	mean	103.77	129.94	86.2	106
	SD	38.37	47.65	31.07	41.93
	range	38.64 – 273.91	31.9 – 347.68	35.18 – 212.71	37.54 – 318.39
SD of F2 (Hz)					
	mean	289.52	286.51	261.17	287.35
	SD	80.44	82.79	76.83	101.34
	range	133.22 – 584.98	103.38 – 663.09	127.81 – 587	105.96 – 673.12
SD of F3 (Hz)					
	mean	428.61	401.39	399.58	400.39
	SD	68.8	80.13	72.08	81.15
	range	263.83 – 552.95	200.45 – 554.72	206.5 – 554.23	149.6 – 554.98

3.1 Duration

3.1.1 The Dutch Second Language Sample

Table 3 Summary of Mixed Effects Model A

Model A	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-1.139	0.112	261.055	-10.127	>.001
disfluencyfilled pause	0.057	0.112	386.622	0.506	0.613
groupmale	-0.11	0.2	289.231	-0.551	0.582
disfluencyfilled pause:groupmale	0.079	0.201	391.24	0.395	0.693

Model A fits a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. Table 3 shows a summary of the coefficients for the fixed factors, t tests and p values. There are no reasons to reject the null hypotheses that there are no effects of these variables.

3.1.2 The English Second Language Sample

Table 4 Summary of Mixed Effects Model B

Model B	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.389	0.025	23.362	15.461	>.001
disfluencyfilled pause	-0.029	0.014	1442.269	-2.087	0.037
groupmale	0.001	0.03	23.133	0.047	0.963
disfluencyfilled pause:groupmale	-0.013	0.017	1444.32	-0.746	0.456

Model B fits a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. The summary of the t tests and p values for fixed factor coefficients, in table 8, showed that there was no significant effect of the disfluency factor (t value = -2.087, p = 0.037) or the speaker sex factor (t value = 0.047, p = 0.963). The summary also shows no effect of the interaction term (t value = -0.746, p = 0.456). There are no reasons to reject the null hypotheses that there are no effects of these variables.

3.2 Standard Deviation of F0 Frequencies

3.2.1 The Dutch Second Language Sample

Table 5 Summary of Mixed Effects Models C, Ci, Cii

	Estimate	Std. Error	df	t value	Pr(>|t|)
Model C					
(Intercept)	1.925	0.224	202.179	8.605	>.001
disfluencyfilled pause	-0.209	0.217	357.198	-0.963	0.336
groupmale	-1.413	0.424	233.53	-3.33	0.001
disfluencyfilled pause:groupmale	1.088	0.42	355.08	2.588	0.01
Model Ci					
(Intercept)	1.773	0.221	146.483	8.039	>.001
disfluencyfilled pause	-0.017	0.219	253.295	-0.076	0.94
Model Cii					
(Intercept)	0.346	0.379	75.208	0.913	0.364
disfluencyfilled pause	0.954	0.364	109.771	2.619	0.01

Model C fits a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. The effect of the disfluency type factor was not significant (t value = -0.963, p = 0.336). Whereas, the sex factor had a significant effect (t value = - 3.33, p = 0.001). The interaction term also shows significant reason (t value = 2.588, p = 0.01) to reject the null hypothesis that there was no difference in the standard deviation of F0 frequencies for the disfluencies (lengthened vowels, filled pauses) uttered by the speakers who were grouped by sex (female, male).

Hence, Model Ci and Model Cii are used to investigate the effect of the interaction. Model Ci, in Table 3, is a mixed effect model fit to a subset of the outcome variable that contains measurements made amongst female speakers. In Model Ci, the random factor are the speakers and there is one fixed factor for the type of disfluency observed. The null hypothesis of Model Ci is that the disfluency types will have a similar standard deviation of F0 frequencies. There are no significant reasons to reject the null hypothesis for Model Ci (t value = -0.076, p = 0.94). Whereas, Model Cii, in Table 3, is a mixed effect model fit to a subset of the outcome variable measured amongst males. For Model Cii, the random factors are the speakers and the fixed factor is the type of disfluency observed. Similar to the hypothesis tested by Model Ci, the null hypothesis of Model Cii is that the disfluency types will have a similar standard deviation of F0 frequencies. The summary of Model Cii shows a significant effect of the fixed factor disfluency type filled pause (t value = 2.619, p = .01) and the coefficient is positive, indicating that filled pauses had larger standard deviations of F0 frequencies than lengthened vowels amongst male speakers.

3.2.2 The English Second Language Sample

Table 6 Summary of Mixed Effects Model D

	Estimate	Std. Error	df	t value	Pr(>|t|)
Model D					
(Intercept)	1.591	0.242	24.107	6.574	>.001
disfluencyfilled pause	0.347	0.109	1534.649	3.174	0.002
groupmale	-0.408	0.293	23.711	-1.389	0.178
disfluencyfilled pause:groupmale	-0.197	0.134	1535.934	-1.468	0.142

Model D fits a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. The t tests and p values for this model showed that there is a significant effect of the disfluency factor (t = 3.174, p = 0.002). The estimate for the disfluency factor category filled pause is positive (beta 1 = 0.347). This indicates that the filled pauses had a larger standard deviation of f0 frequencies than lengthened vowels. There was no significant reason to reject the null hypothesis that there was a similarity in the outcome variable for the speaker sex factor (t value = -1.389, p = 0.178). The interaction term also showed no reason to reject the null hypothesis, which is that the predictor factors combined probably had similar effects on the standard deviation of F0 frequencies (t value = -1.468, p = 0.142).

3.3 Standard Deviation of F1 Frequencies

3.3.1 The Dutch Second Language Sample

Table 7 Summary of Mixed Effects Model E

Model E	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.489	0.094	153.142	47.844	>.001
disfluencyfilled pause	0.106	0.084	394.461	1.255	0.21
groupmale	0.034	0.165	175.353	0.208	0.835
disfluencyfilled pause:groupmale	-0.114	0.15	394.394	-0.754	0.451

Model E fits a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. The t tests and p values for Model E showed no significant probabilities to reject the null hypothesis that the standard deviation of F1 frequencies were similar for disfluency types and the sexes.

3.3.2 The English Second Language Sample

Table 8 Summary of Mixed Effects Model F

Model F	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.557	0.079	25.437	57.911	>.001
disfluencyfilled pause	0.195	0.039	1622.27	4.992	>.001
groupmale	-0.164	0.095	25.012	-1.719	0.098
disfluencyfilled pause:groupmale	0.026	0.048	1623.5	0.551	0.582

Model F fits a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. This model showed that there is a significant effect of the disfluency factor (t = 4.992, p > 0.001). The estimate for the disfluency factor category filled pause is positive (beta 1 = 0.195), indicating that these had a larger standard deviation for both female and male speakers. There was no effect of the speaker sex factor (t = -1.719, p > 0.098). The mixed effects Model F fitting an interaction term showed no reason to reject the null hypothesis that the predictor factors had similar effects on the standard deviation of F1 frequencies (t value = 0.551, p = 0.582).

3.4 Standard Deviation of F2 Frequencies

3.4.1 The Dutch Second Language Sample

Table 9 Summary of Mixed Effects Model G

Model G	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	5.822	0.072	233.942	80.988	>.001
disfluencyfilled pause	-0.095	0.071	387.702	-1.343	0.18
groupmale	-0.136	0.128	262.306	-1.066	0.287
disfluencyfilled pause:groupmale	0.005	0.127	392.401	0.037	0.97

Model G fits a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. The tests of the Model G showed no significant reason to reject the null hypothesis that the standard deviation of F2 frequencies was similar for the disfluency types, sexes or the interaction thereof.

3.4.2 The English Second Language Sample

Table 10 Summary of Mixed Effects Model H

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	16.761	0.585	24.339	28.647	>.001
disfluencyfilled pause	-0.116	0.28	1607.514	-0.416	0.678
groupmale	-0.929	0.708	23.854	-1.311	0.202
disfluencyfilled pause:groupmale	1.012	0.34	1608.613	2.975	0.003

Model H fits a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. The summary of this model, in table 10, shows that there are no significant effects of the disfluency factor (t value = -0.416, p = 0.678) or the sex factor (t value = -1.311, p = 0.202). Whilst there is significant effect of interaction term (t value = -2.975, p = 0.003).

3.5 Standard Deviation of F3 Frequencies

3.5.1 The Dutch Second Language Sample

Table 11 Summary of Mixed Effects Model I

Model I	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	6.318	0.065	267.912	97.286	>.001
disfluencyfilled pause	-0.046	0.065	385.203	-0.716	0.474
groupmale	-0.031	0.116	296.887	-0.267	0.79
disfluencyfilled pause:groupmale	-0.068	0.117	390.191	-0.586	0.558

Model I fit a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. The summary of Model I in Table 11 shows no significant effects of the disfluency type factor, sex factor or the interaction of the two factors.

3.5.2 The English Second Language Sample

Table 12 Summary of Mixed Effects Model J, Ji, Jii

	Estimate	Std. Error	df	t value	Pr(>|t|)
Model J					
(Intercept)	6.043	0.051	23.055	118.817	>.001
disfluencyfilled pause	-0.063	0.023	1275.647	-2.72	0.007
groupmale	-0.085	0.062	22.784	-1.378	0.182
disfluencyfilled pause:groupmale	0.078	0.029	1277.295	2.742	0.006
Model Ji					
(Intercept)	6.044	0.038	9.881	159.953	>.001
disfluencyfilled pause	-0.064	0.023	493.239	-2.754	0.006
Model Jii					
(Intercept)	5.958	0.039	14.17	152.653	>.001
disfluencyfilled pause	0.016	0.017	786.053	0.935	0.35

Model J fit a disfluency type fixed factor (lengthened vowel, filled pause), a speaker sex fixed factor (female, male), an interaction term for these factors and a random factor for individual speakers. This model also showed that there is a significant effect of the disfluency factor (t value = -2.72, p = 0.007). Whilst, there was no significant effect of the speaker sex factor (t value = -1.378, p = 0.182). In the mixed effects Model L, the interaction term was significant (t value = 2.742, p = 0.006). Hence, the null hypothesis that the predictor factors had similar effects on the standard deviation of F3 frequencies can be rejected.

Model Ji and Model Jii are used to investigate the effect of the interaction amongst the sexes. Model Ji, in Table 12, is fit to a subset of the outcome variable that contains measurements made amongst female speakers. In Model Ji, the random factor are the speakers and there is one fixed factor for the type of disfluency observed. The null hypothesis of Model Ji is that the disfluency types will have a similar standard deviation of F3 frequencies. The summary of Model Ji shows a significant effect of the fixed factor disfluency type filled pause (t value = -2.754, p = .006) and the coefficient is negative, indicating that filled pauses had smaller standard deviations of F3 frequencies than lengthened vowels amongst female speakers. Whereas, Model Jii is fit to a subset of the outcome variable measured amongst males. For Model Jii, the random factors are the speakers and the fixed factor is the type of disfluency observed. Similarly, the null hypothesis of the test using Model Jii is that the disfluency types will have a similar standard deviation of F3 frequencies. There are no significant reasons to reject the null hypothesis for Model Jii (t value = 0.935, p = 0.35).

Chapter 4 Discussion

The objective of this study was to see whether lengthened vowels and filled pauses in Dutch and English exhibited similar phonetic traits within recordings of female and male L2 speakers. Samples of these disfluencies were collected from L2 speech corpora and were used to obtain measures of duration and the standard deviation of F0 – 3 frequencies. Firstly, in Section 4.1, the research questions from section 1.5 are reiterated. Per question, the results are compared to the findings in previous research. Following this, a discussion of the limitation of this study is presented in Section 4.2.

4.1 Main Findings

Question 1: *Are the vocalic parts of filled pauses and prolonged vowels similar in duration? Does this differ for female and male speakers?*

In this thesis, mixed effects models were used to test whether the duration of the vocalic parts of filled pauses and lengthened vowels would be similar for female and male L2 Dutch and English speakers. Stouten et al. (2006) found that vocalic parts of filled pauses were often longer than phoneme-like speech segments. The results of the mixed effects Model A fit to the Dutch materials and Model B fit to the English materials showed there were no significant main effects or interactions for disfluency type factor or sex factor. These results indicated that there were no differences in the duration of the perceived disfluency (lengthened vowel, filled pause) and the group of speaker sex (female, male). These results corroborate to what was presumed by Audhkhasi et al. (2009), when they suggested that lengthened vowels would also be detected alike filled pauses due to duration. Hence, these disfluencies shared duration when tested per language.

Question 2: *Do second language speakers have a stable fundamental frequency in filled pauses and in lengthened vowels? How does this compare for female and male speakers?*

Previously, Shriberg (2001) reported that filled pauses and disfluent speech prolongations would have a relatively flat pitch, or one with a slight fall. Hence, I hypothesized that these two disfluency types would have negligible differences in standard deviations of F0 frequencies. The results of the statistical tests (Section 3.1.2 and Section 3.2.2) did not show this. The mixed effects Model C for the standard deviation of F0 frequencies for the Dutch sample showed that there was a significant effect of the interaction term and there was an effect of the fixed factor sex. The difference in standard deviations of F0 between the sexes was not been reported in previous filled pause detection studies. The results of Model Ci and Model Cii showed that filled pauses uttered by males had a significantly larger standard deviation of F0 frequencies than lengthened vowels, whilst this effect was not present in for the female group.

Contrastively, the Model D fit to the outcome of the standard deviation of F0 frequencies for the English sample did not have a significant interaction effect, but there was a significant effect of the disfluency. The filled pauses of both males and females contained significantly larger standard deviations of F0 frequencies than that measured in lengthened vowels. The larger standard deviation of F0 frequencies found for the filled pauses in these test results do not confirm the finding of Stouten et al. (2006) that there are no differences in pitch variation for filled pauses and other parts of speech. However, these findings might corroborate with the findings of Audhkhasi et al. (2009) that filled pauses would not effectively be detected with the standard deviation of F0 frequencies, because speaker's intonations would not be more stable throughout a filled pause.

Question 3: *Is the stability of formant frequencies similar in filled pauses and prolonged vowels? Are there differences for formant stabilities in these disfluencies for female and male speakers?*

Researchers who looked at how to detect filled pauses claimed that these could be detected by obtaining the standard deviation of formant frequencies during vocalic articulation and the premises of the detection studies were that the formant frequencies would not change much over time because the articulators would not move for the duration of the filled pause (Audhkhasi, Kandhway, et al., 2009; Kaushik et al., 2010; Krikke & Truong, 2013). Audhkhasi et al. (2009) suggested that lengthened vowels too could be detected using their technique based on measuring the standard deviation of formant frequencies. I hypothesized that the standard deviation of formant frequencies measured within lengthened vowels and filled pauses of Dutch and English spoken as a second language would be similar for both sexes and that filled pauses would have more stable formants. The tests done on the Dutch materials did not show any effects of the fixed factors, indicating similarities for the standard deviations of formant frequencies for lengthened vowels and filled pauses uttered by female and male speakers. However, the results of the mixed effects models for the English materials were less uniform.

The mixed effects Model F, Model H, Model J, Model Ji and Model Jii showed that lengthened vowels and filled pauses uttered by the female and male English speakers were not the same in terms of the measures taken. Effectively, the filled pauses were measured having a larger standard deviation of F1 frequencies than that in lengthened vowels for both groups of speaker sex. The results of the testing done with the mixed effects Model H for the standard deviation of F2 frequencies showed that interaction term had a significant effect. However, there were no main effects of solely the fixed factors. Hence, the direction of the effects crossed over, even though there were no significant main effects of disfluency type or sex. The mixed effects Model J testing the standard deviation of F3 frequencies must be interpreted with caution, because much of the data was removed from the study in order to achieve assume homogeneity of variance. Model J had a significant main effect of the interaction term and there was a significant effect of the disfluency factor. Model Ji and Model Jii were used to further investigate the significant effects of Model J and these showed that filled pauses had smaller standard deviation of F3 frequencies for female speakers.

The results of the tests done on the Dutch data showed that lengthened vowels and filled pauses were similar in terms of the standard deviation of formant frequencies. Whereas, the results of the tests on the English material showed that filled pauses could also have larger standard deviations of formant frequencies. The finding that filled pauses had larger standard deviations of formant frequencies than in prolonged vowels might suggest that: (i) filled pauses uttered by L2 speakers of English were less steady, or (ii) that the study was limited due to a systematic error, involving the quality of the recordings (see section 4.2).

4.2 Limitations and Future Research

Each data point in the outcome variables for the thesis was based on the literature reviewed (in chapter 1), my perception of the beginning and end of a prolongation or a filled pause disfluency, the data and the collection procedures done in PRAAT. Therefore, this study was limited by both objective and subjective factors which could leave room for improving further research of disfluencies.

Firstly, there was little literature on the acoustic differences for varying pronunciations and productions of the filled pause in native or second language Dutch and English speech. The pronunciations and prosody of filled pauses could vary for a multitude of reasons, e.g. Dutch and English filled pauses containing neutral vowels that may be dialect bound, or speakers might say these with a vocal fry register (Wieling et al., 2016). Future studies might want to examine similarities of acoustic measurements in filled pauses and lengthened vowels by attaining additional information on what dialect the individual learners were aiming to acquire and consider testing whether vocal registers could factor the acoustic features being tested. Previously, vocal fry was found to have impacted F0 frequencies (Boersma, 2014; Kent & Vorperian, 2018). Hence, further research could also focus on distinguishing whether the stability of intonation is affected by vocal creak. Due to time restrictions, these premises were not examined for this thesis.

Although some of the papers reviewed were experimental filled pause detection studies, these gave insights to what acoustic features were present for the filled pauses and some lengthened vowels in their speech materials. Little was written about the vowel qualities of filled pause variants, but the reviewed papers suggested that these had a number of features, including stable formant frequency tracks (Audhkhasi, Kandhway, et al., 2009; Kaushik et al., 2010; Krikke & Truong, 2013). This suggested that the vowel quality of filled pauses was not variable for the two filled pause variants of English, described with more detail in Chapter 1. Moreover, these results suggested that filled pauses and some lengthened vowels had a steady state articulation. Yet, Hughes et al. (2016) found that the formant dynamics of filled pause variants (i.e. uh and um) would best be compared separately for forensic speaker identification. This thesis did not examine whether the filled pause variants could also factor the acoustic features due to time constraints. Future research could also examine if the different sounding filled pauses share standard deviations of formant frequencies.

Moreover, future studies could also test whether the acoustic measurements of filled pauses and lengthened vowels could be factored by a speaker’s first language and amount of exposure to the second language (De Boer & Heeren, 2019; Geeslin & Long, 2014). From what was known about the speech materials, the Dutch speech was taken from L2 speakers who had one of two possible L1s, whilst individuals who spoke in the L2 English materials had a wider range of L1s. Due to constraints in time and resources this thesis could not explore the effects of such factors.

Additionally, the speech materials were not always grammatical and the pronunciation of function words like “and”, in English, or “en”, in Dutch, could have made it difficult to tell whether a filled pause was said. To deal with this, I listened to the context of a presumed filled pause up to five times in order to determine whether it might be a function word. In the case of uncertainty, I annotated the sound as “LFP” which was not used in the present study. Future studies should aim at reliably finding filled pauses in recordings as was done in De Jong et al. (2020), where two listeners’ annotations could be compared.

Finally, a limitation for this thesis could have been the quality of the materials. The English materials were not free of background noise and the sampling frequency was low (11025 Hz) which meant that the automatic formant and pitch tracker used might have been more prone to erroneous measurements. The Dutch materials contained less background noise and the sampling frequency was higher, meaning that PRAAT would not be as prone to error. Future studies should maintain that acoustic analyses are done with uncompressed recordings made in a quiet room and where the speaker keeps the same distance from the microphone.

Chapter 5 Conclusion

The research of this thesis was done in light of previous findings that showed that the detectability of filled pauses and lengthened vowels could be based on spectral stability, and claims that pitch and duration could be features for distinguishing filled pauses and lengthened vowels (Audhkhasi, Kandhway, et al., 2009; Kaushik et al., 2010; Stouten et al., 2006). This thesis set out to examine whether there were similarities in the duration and other acoustic measures for lengthened vowels and filled pauses uttered by female and male speakers of L2 Dutch and English using mixed effects models. The effects of the disfluency type (lengthened vowel, filled pause) and the groups of speaker sex (female, male) were not present in the same way for the outcomes from the L2 Dutch and English speech materials. More research is required to uncover the acoustic features of filled pauses in L2 speech.

This thesis showed that in the analyzed material there were no differences in the duration of the vocalic parts of filled pauses and lengthened vowels for males and females speaking L2 Dutch. It also showed that there could be differences in the standard deviations of F0-3 frequencies. The standard deviations of F0 frequencies from the Dutch materials were different amongst the sexes, only male speakers had larger standard deviations of F0 frequencies in filled pauses. The standard deviations of F1-3 frequencies for disfluencies uttered by female and male speakers in the L2 Dutch materials were not different. This indicated that the vocalic parts of filled pauses and the lengthened vowels in the L2 Dutch materials shared a duration and the formant stability acoustic feature which was previously attributed as a cause for false positive detection of filled pauses (Audhkhasi, Kandhway, et al., 2009) and this could also be true for filled pause detection in L2 speech.

Moreover, the duration of filled pauses and lengthened vowels was also similar for females and males speaking L2 English. The standard deviations of F0 frequencies measured in the L2 English materials were larger for filled pauses of both females and males. Yet, the tests done for the standard deviation of F1-3 frequencies for the L2 English materials did not have uniform effects of the fixed factors, but the results of these tests must be interpreted with caution because the recordings were not equally as clean and there were many outliers for the standard deviation of F1 and F3 frequencies. The standard deviations of F1 frequencies were larger for filled pauses of both sexes. The standard deviations of F2 frequencies were different in filled pauses and lengthened vowels uttered by females and males but there were no significant effects of the disfluency type of speaker sex. Finally, the standard deviation of F3 frequencies were also different for the sexes, females had larger standard deviations of F3 frequencies when uttering filled pauses. With higher quality L2 English recordings, it is possible to expect more accurate results.

Bibliography

Audhkhasi, K., Deshmukh, O. D., Kandhway, K., & Verma, A. (2009). Automatic evaluation of spoken English fluency. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/ICASSP.2009.4960712

Audhkhasi, K., Kandhway, K., Deshmukh, O. D., & Verma, A. (2009). Formant-based technique for automatic filled-pause detection in spontaneous spoken English. In *ICASSP '09 Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/ICASSP.2009.4960719

Boersma, P. (2014). Acoustic analysis. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (pp. 375–396). Cambridge University Press. https://doi.org/10.1017/CBO9781139013734.020

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program].

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394

Bosker, H. R., Quené, H., Sanders, T., & De Jong, N. H. (2014). The Perception of Fluency in Native and Nonnative Speech. *Language Learning*, *64*(3), 579–614. https://doi.org/10.1111/lang.12067

Braun, A., & Rosin, A. (2015). On the speaker-specificity of hesitation markers. In *Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: the University of Glasgow*. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0731.pdf

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73–111.

Common European framework of reference for languages: Learning, teaching, assessment. (2001). Strasbourg. Retrieved from http://rm.coe.int/1680459f97

De Boer, M., & Heeren, W. F. L. (2019). The speaker-specificity of filled pauses: A cross-linguistic study (pp. 607–611). Australasian Speech Science and Technology Association Inc. Retrieved from http://intro2psycholing.net/ICPhS/%0Aurn:isbn:978-0-646-80069-1

De Jong, N. H. (2018). Fluency in Second Language Testing : Insights From Different Disciplines. *Language Assessment Quarterly*, *15*(3), 237–254. https://doi.org/10.1080/15434303.2018.1477780

De Jong, N. H., Pacilly, J., & Heeren, W. F. L. (2020). Praat scripts to measure fluency automatically. Retrieved July 6, 2020, from osf.io/w3r7t

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). FACETS OF SPEAKING PROFICIENCY. *Studies in Second Language Acquisition*, *34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

De Leeuw, E. (2007). Hesitation Markers in English, German, and Dutch. *Journal of Germanic Linguistics*, *19*(2007), 85–114. https://doi.org/10.1017/S1470542707000049

Ferriera, F. (1993). The creation of prosody during sentence production. *Psychological Review*, *100*(2), 233–253.

Foulkes, P., & French, P. (2012). Forensic Speaker Comparison : A Linguistic – Acoustic Perspective. In L. M. Solan & P. M. Tiersma (Eds.), *The Oxford Handbook of Language and Law* (pp. 1–17). https://doi.org/10.1093/oxfordhb/9780199572120.013.0041

Frota, S., Arvaniti, A., & D'Imperio, M. (2012). Prosodic representations: Prosodic structure, constituents, and their implementation; Segment-To-Tone association; Tonal alignment. In and M. K. H. Abigail C. Cohn, Cécile Fougeron (Ed.), *The Oxford Handbook of Laboratory Phonology* (pp. 1–33). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199575039.013.0011

Geeslin, K. L., & Long, A. Y. (2014). *Sociolinguistics and Second Language Acquisition*. Taylor & Francis Ltd.

Gósy, M., Gyarmathy, D., & Beke, A. (2017). Phonetic analysis of filled pauses based on a Hungarian-English learner corpus. *International Journal of Learner Corpus Research*, *3*(2), 149–174. https://doi.org/10.1075/ijlcr.3.2.03gos

Gussenhoven, C. (2006). Transcription of Dutch Intonation. In S.-A. Jun (Ed.), *Prosodic Typology : The Phonology of Intonation and Phrasing* (pp. 118–145). OUP Oxford.

Gussenhoven, C. (2016). Foundations of Intonational Meaning: Anatomical and Physiological Factors. *Topics* *in Cognitive Science*. https://doi.org/10.1111/tops.12197

Hughes, V., Wood, S., & Foulkes, P. (2016). Formant dynamics and durations of um improve the performance of automatic speaker recognition systems speaker recognition systems. In *Proceedings of the 16th Australasian Conference on Speech Science and Technology (ASSTA)* (pp. 25–28). University of Western Sydney, Australia.

Jun, S. (2006). Prosodic typology. In S.-A. Jun (Ed.), *Prosodic Typology : The Phonology of Intonation and Phrasing* (pp. 430–458).

Kaiser, E. (1997). Nasals. Retrieved November 22, 2019, from http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/speech.bme.ogi.edu/tutordemos/SpectrogramReading/cse551html/cse551/node35.html

Kaushik, M., Trinkle, M., & Hashemi-Sakhtsari, A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. In *Australasian International Conference on Speech Science and Technology* (pp. 98–101).

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths : A review. *Journal of Communication Disorders*, *74*(November 2017), 74–97. https://doi.org/10.1016/j.jcomdis.2018.05.004

Krikke, T. F., & Truong, K. P. (2013). Detection of nonverbal vocalizations using gaussian mixture models: Looking for fillers and laughter in conversational speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 163–167).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Ladefoged, P., & Johnson, K. (2015). *A Course in Phonetics* (7th editio). Cengage Learning.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech*. *Cognition*, *14*, 41–104.

Mcdougall, K., & Duckworth, M. (2018). Individual patterns of disfluency across speaking styles: a forensic phonetic investigation of Standard Southern British English, *25*, 205–231.

McDougall, K., & Duckworth, M. (2017). Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication*, *95*(May), 16–27. https://doi.org/10.1016/j.specom.2017.10.001

Meisel, J. M. (2009). Second Language Acquisition in Early Childhood *. *Zeitschrift Für Sprachwissenschaft*, *28*, 5–34. https://doi.org/10.1515/ZFSW.2009.002

Pacilly, J. J. A. (2018). Explore.

Peters, J., Hanssen, J., & Gussenhoven, C. (2014). The phonetic realization of focus in West Frisian, Low Saxon, High German, and three varieties of Dutch. *Journal of Phonetics*, *46*(1), 185–209. https://doi.org/10.1016/j.wocn.2014.07.004

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org/

RStudio Team. (2018). RStudio: Integrated Development Environment for R. Boston, MA. Retrieved from http://www.rstudio.com/

Segalowitz, N. (2010). *Cognitive bases of second language fluency*.

Shriberg, E. (2001). To ‘errrr’ is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(1), 153–169. https://doi.org/10.1017/S0025100301001128

Shriberg, E. E., & Lickley, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, *50*(3), 172–179. https://doi.org/10.1159/000261937

Sleebos, Y. M. A. (2018). *Filled pauses in first and second language users: How speaker-specific is u(h)m across languages?* Leiden University.

Stouten, F. (2008). *Feature extraction and event detection for Automatic Speech Recognition Kenmerkenextractie en eventdetectie voor Automatische Spraakherkenning*. Ghent University.

Stouten, F., Duchateau, J., Martens, J., & Wambacq, P. (2006). Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. *Speech Communication*, *48*, 1590–1606. https://doi.org/10.1016/j.specom.2006.04.004

Stouten, F., & Martens, J. (2004). *Benefits of Disfluency Detection in Spontaneous Speech Recognition*.

COST278 and ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction.

Tavakoli, P., Nakatsuhara, F., & Hunter, S.-M. (2017). Scoring Validity of the APTIS Speaking Test: Investigating Fluency Across Tasks and Levels of Proficiency. *ARAGs Research Reports Online*, 1–56. Retrieved from http://centaur.reading.ac.uk/73379/

Tottie, G. (2011). Uh and Um as sociolinguistic markers in British English*. *International Journal of Corpus Linguistics*, *16*(2), 173–197. https://doi.org/10.1075/ijcl.16.2.02tot

Trofimovich, P., & Baker, W. (2006). LEARNING SECOND LANGUAGE SUPRASEGMENTALS: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech. *SSLA*, (28), 1–30. https://doi.org/10.1017/S0272263106060013

Turk, A., & Shattuck-Hufnagel, S. (1996). A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of Psycholinguistic Research*, *25*(2). https://doi.org/10.1007/BF01708572

Weisberg, J. F. and S. (2019). An R companion to applied regression. Thousand Oaks. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Whiteside, S.P. (2001). Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *The Journal of the Acoustical Society of America*, *110*(1), 464–478. https://doi.org/10.1121/1.1379087

Whiteside, Sandra P. (1996). Temporal-based acoustic-phonetic patterns in read speech: some evidence for speaker sex differences. *Journal of the International Phonetic Association*, *26*(1), 23–40.

Wickham, H., François, R., Henry, L., & Müller, K. (2019). dplyr: A grammar of data manipulation. Retrieved from https://cran.r-project.org/package=dplyr

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and Change in the Use of Hesitation Markers in Germanic Languages. *Language Dynamics and Change*, *6*(2), 199–234. https://doi.org/10.1163/22105832-00602001

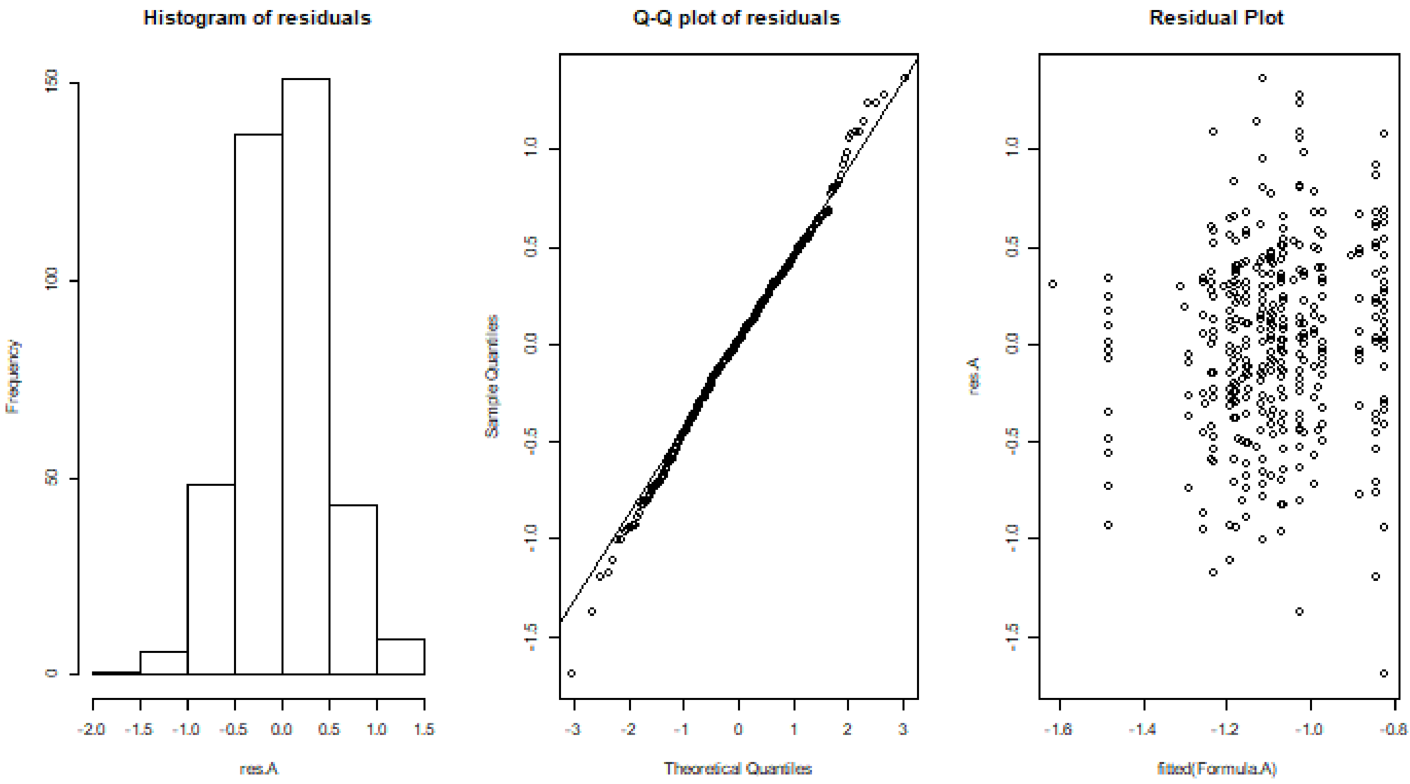
Winter, B. (2019). *Statistics for Linguists : An Introduction Using R*. Routledge.

Zsiga, E., & Podesva, R. J. (2014). Sound recordings: Acoustic and articulatory data. In R. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 169–194). Cambridge University Press.

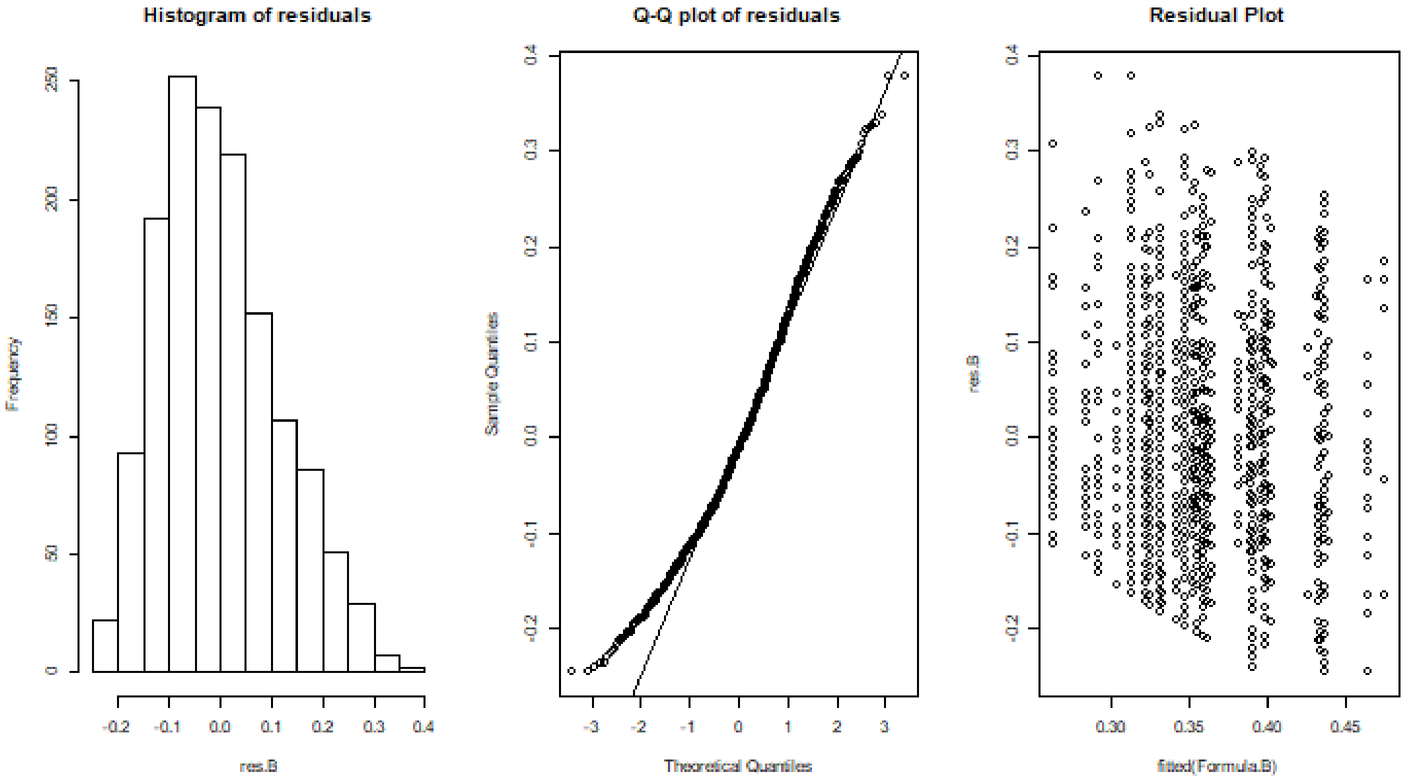
https://doi.org/10.1017/CBO9781139013734.010

Appendix A

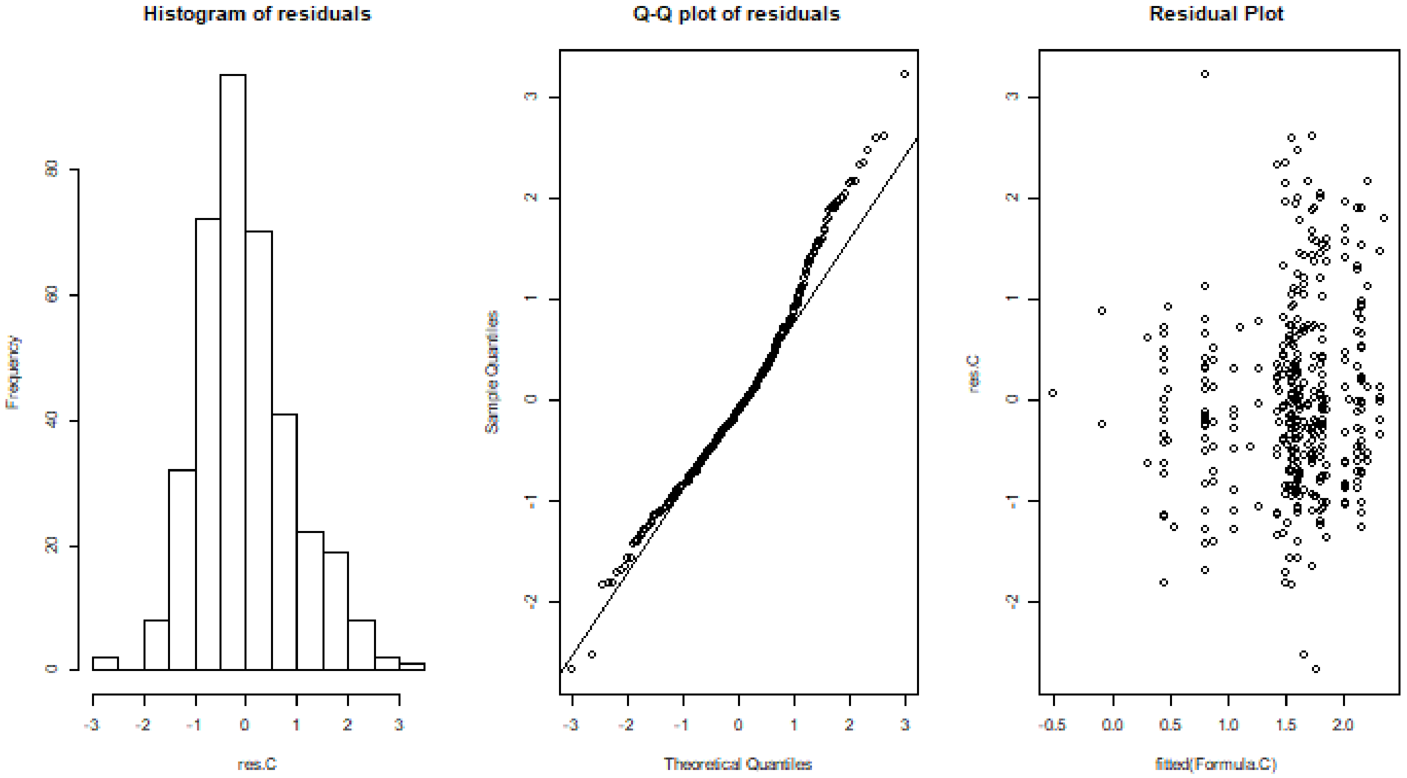
Residual plot for linear model from Section 3.1.1



Residual plot for linear model from Section 3.1.2



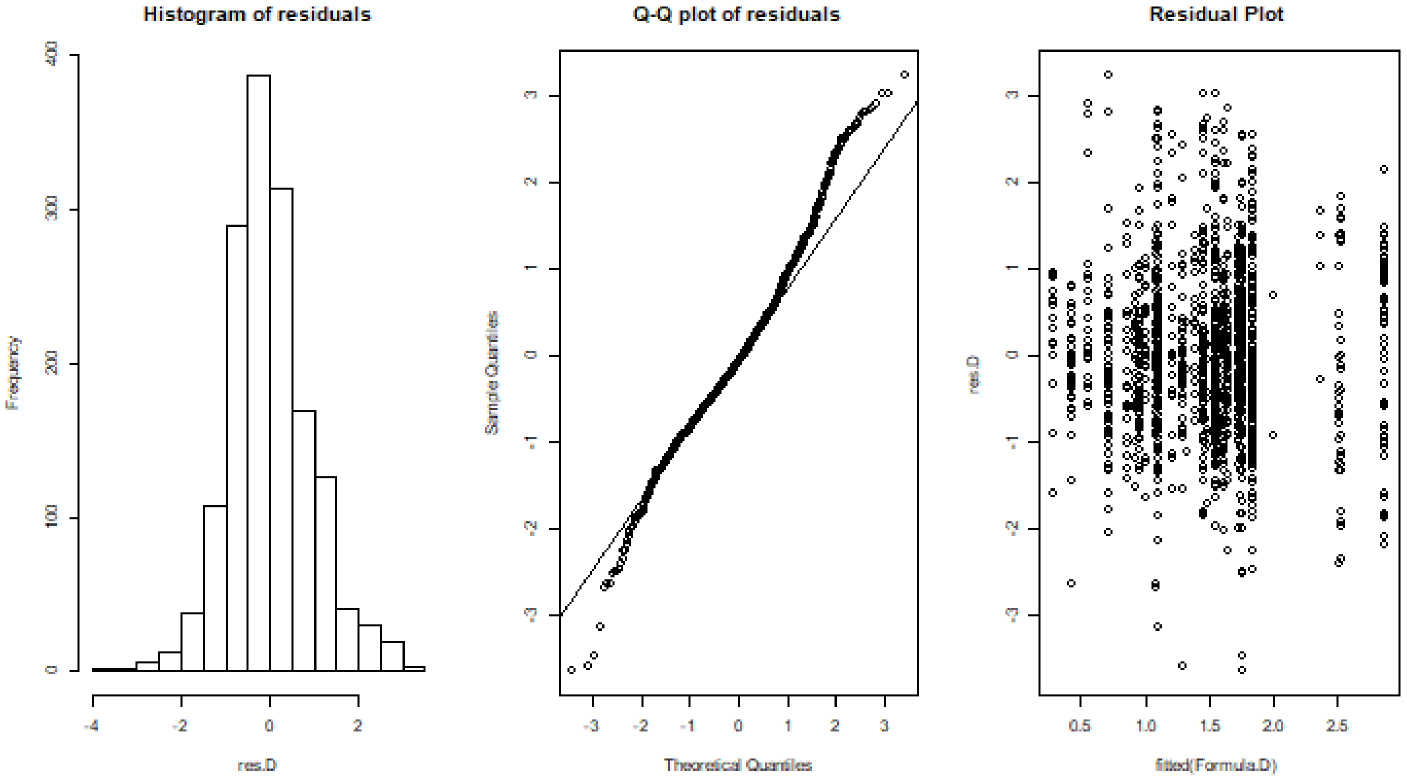
Residual plot for linear model from Section 3.2.1


Histogram of residuals

Q-Q plot of residuals

Residual Plot

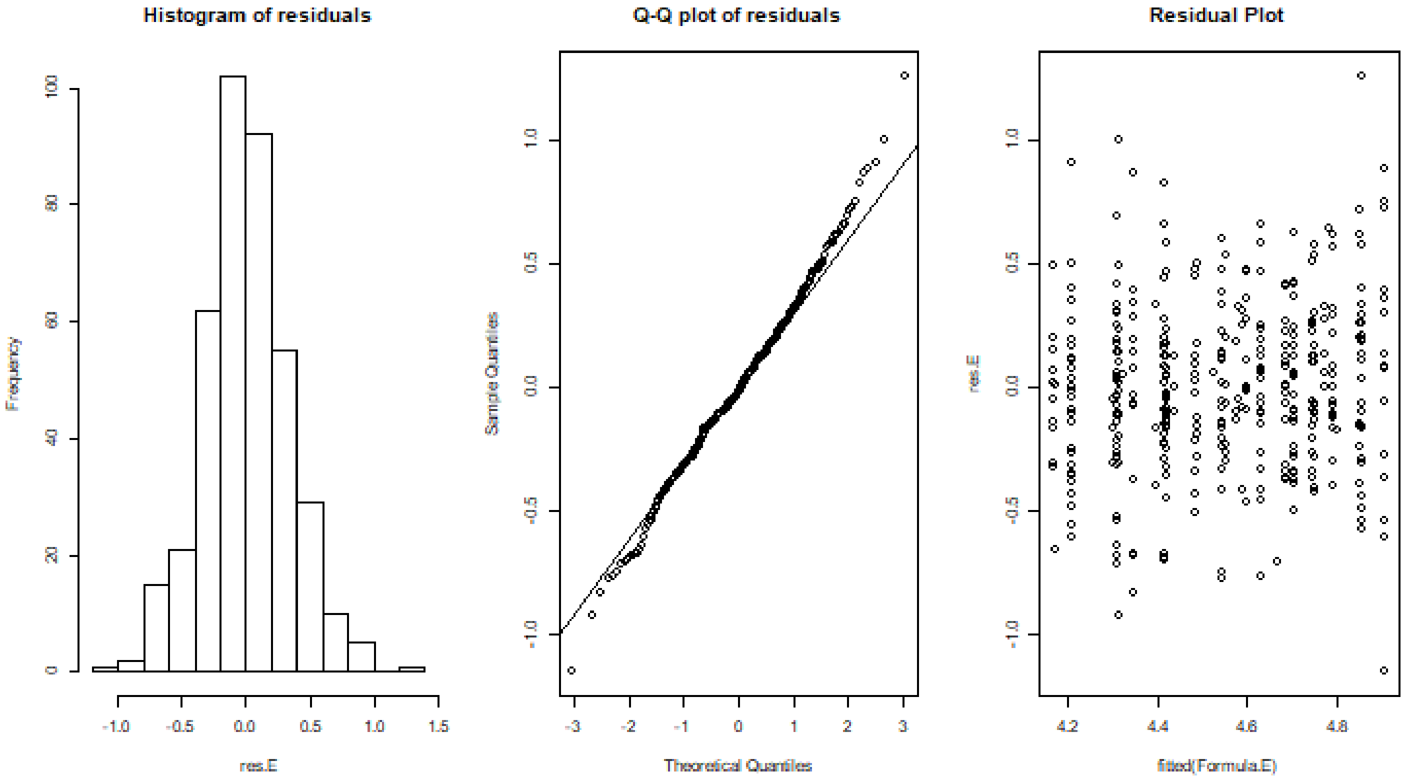
Residual plot for linear model from Section 3.2.2


Histogram of residuals

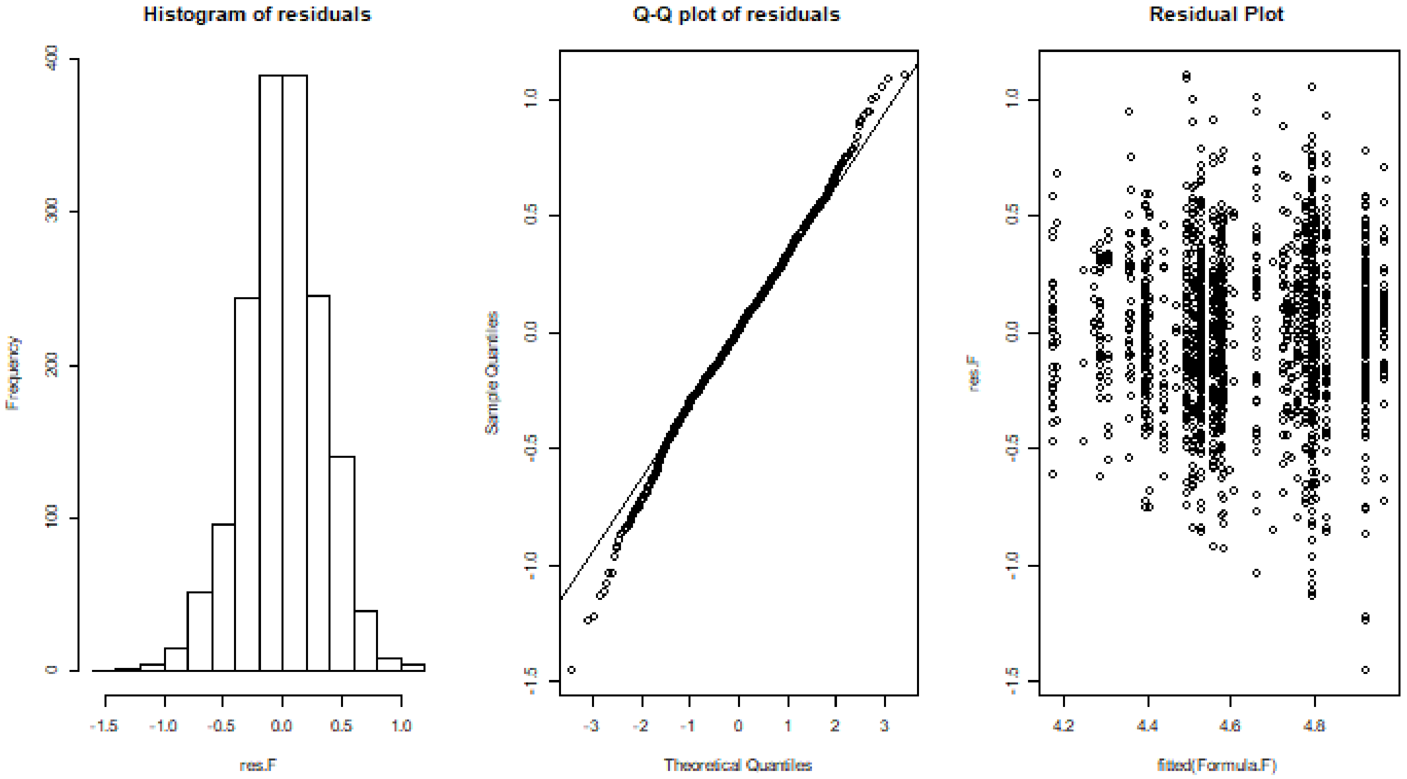
Q-Q plot of residuals

Residual Plot

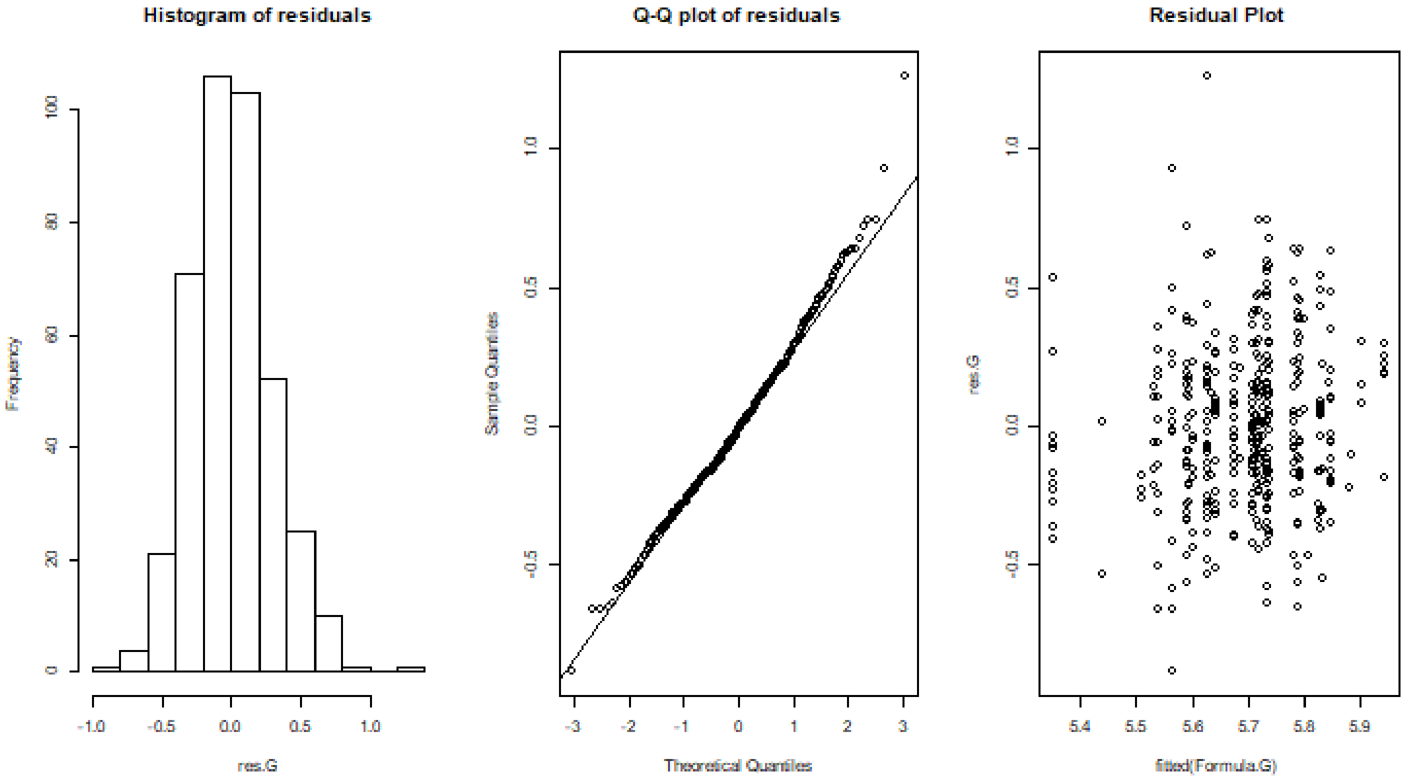
Residual plot for linear model from Section 3.3.1



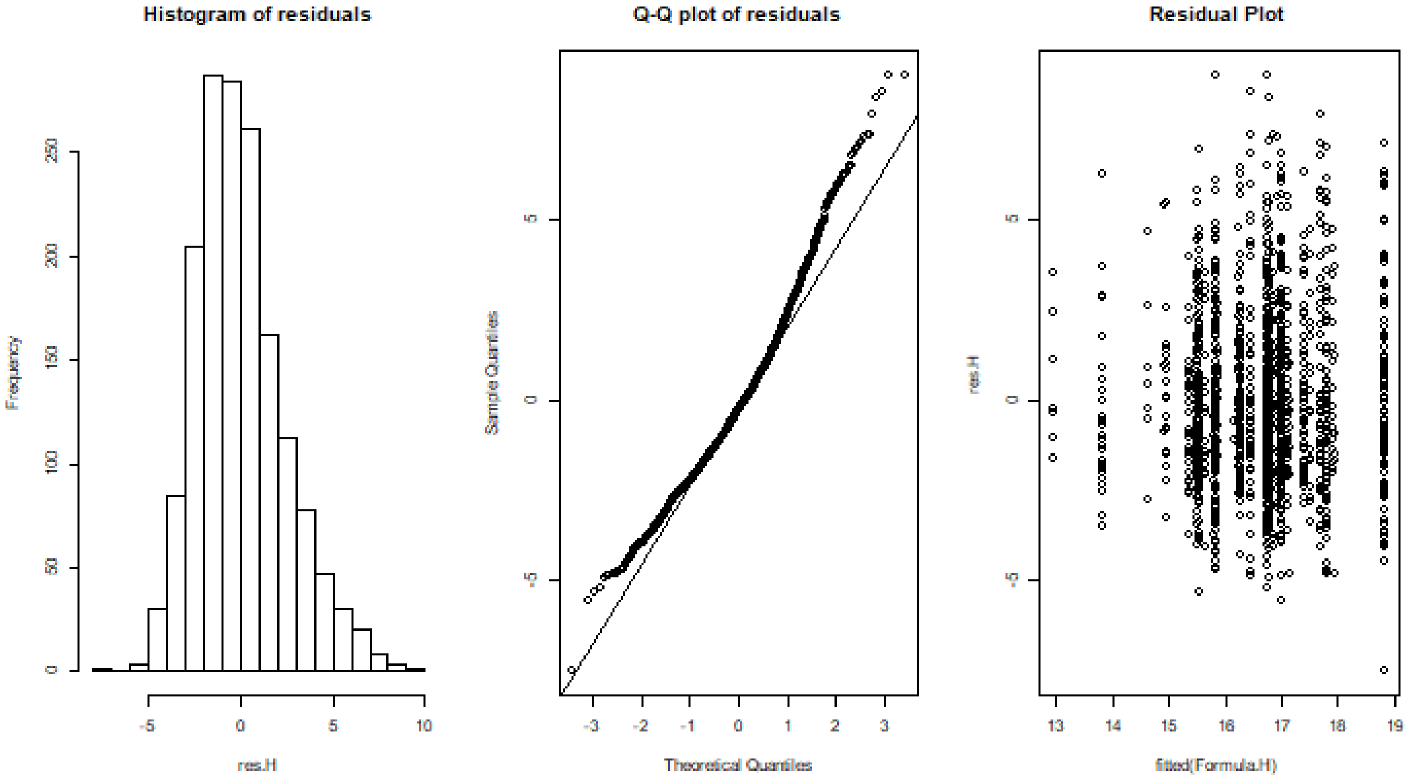
Residual plot for linear model from Section 3.3.2



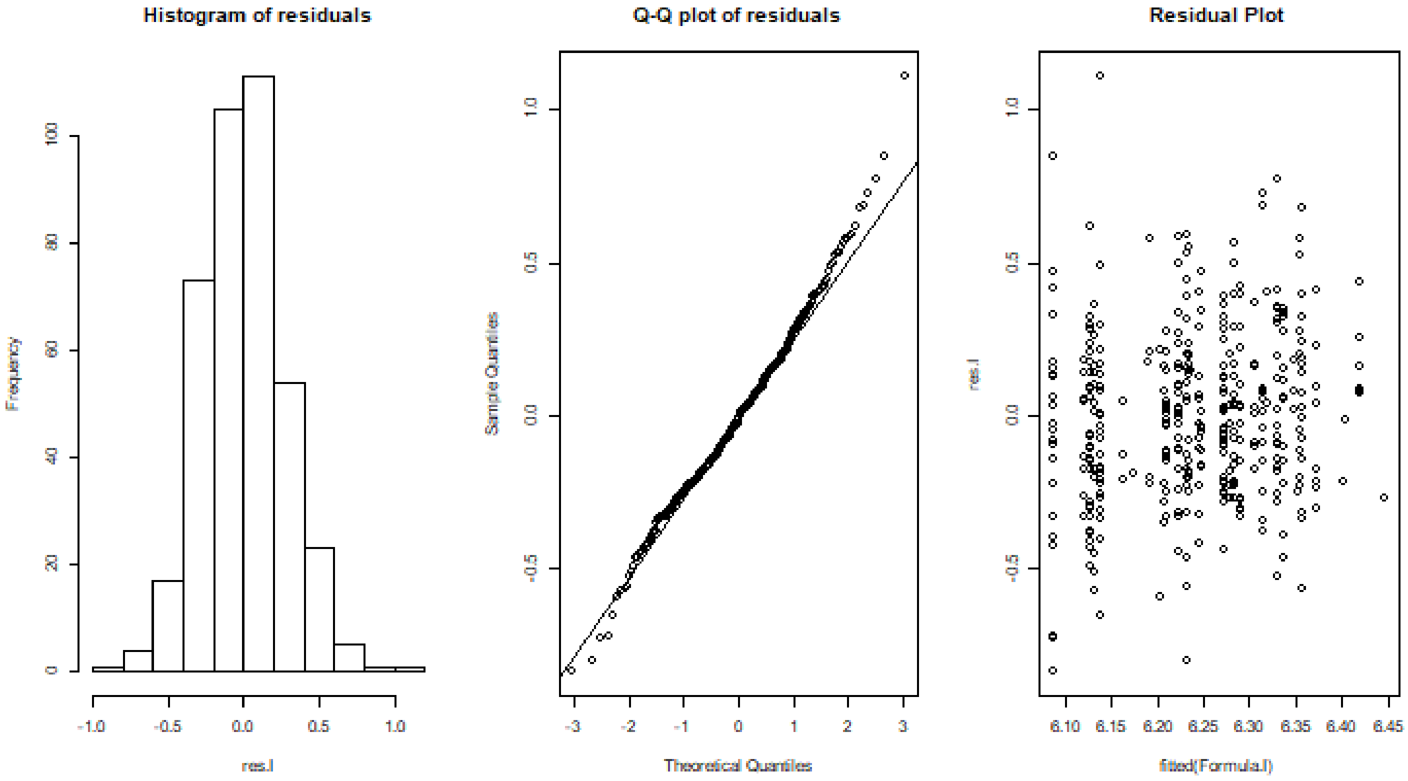
Residual plot for linear model from Section 3.4.1



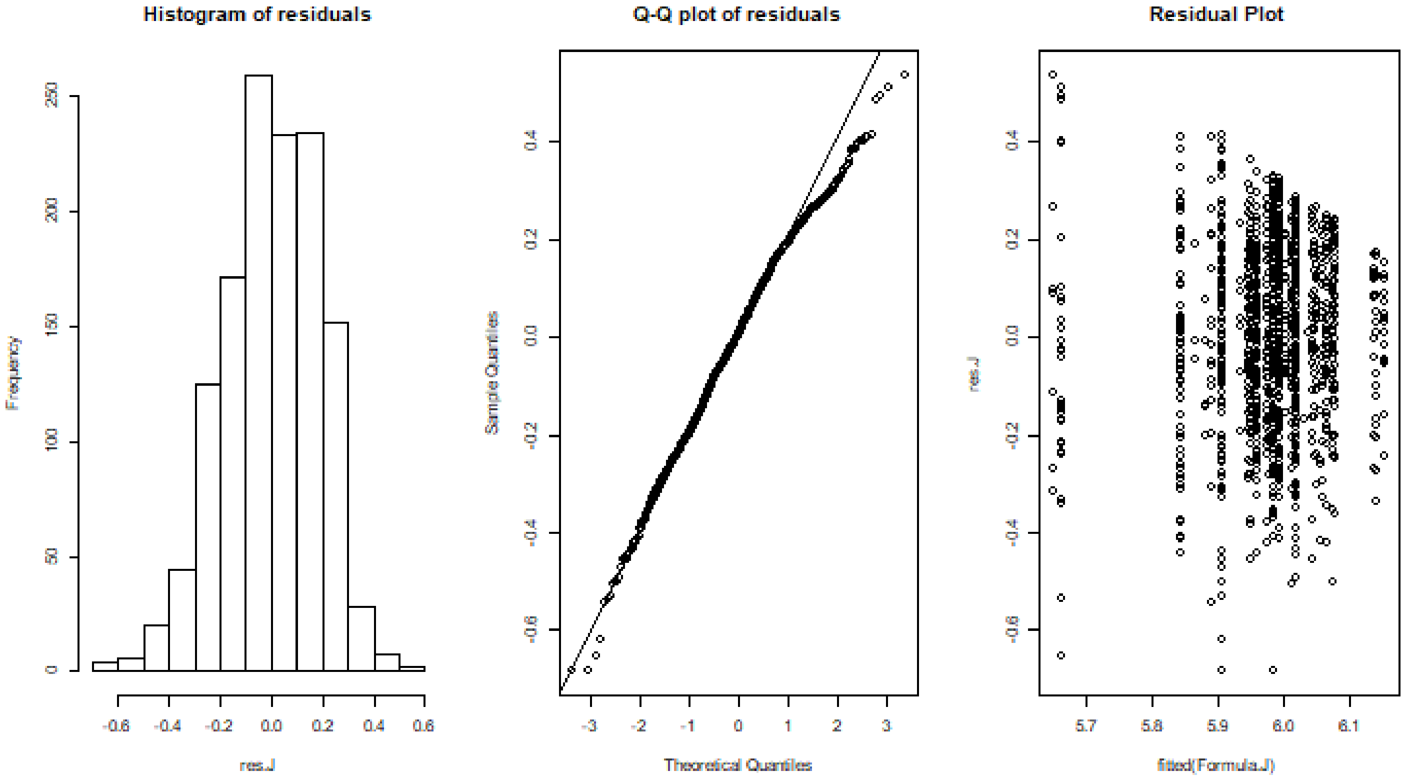
Residual plot for linear model from Section 3.4.2



Residual plot for linear model from Section 3.5.1



Residual plot for linear model from Section 3.5.2


Histogram of residuals
Frequency
res.J
Q-Q plot of residuals
Sample Quantiles
Theoretical Quantiles
Residual Plot
res.J
fitted(Formula.J)