THE DISTRIBUTION OF DISFLUENCIES IN SPONTANEOUS SPEECH: EMPIRICAL OBSERVATIONS AND THEORETICAL IMPLICATIONS

Hong Zhang

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

Mark Liberman, Christopher H. Browne Distinguished Professor of Linguistics

Graduate Group Chairperson

Eugene Buckley, Associate Professor of Linguistics

Dissertation Committee
Kathryn Schuler, Assistant Professor of Linguistics
Gareth Roberts, Assistant Professor of Linguistics

THE DISTRIBUTION OF DISFLUENCIES IN SPONTANEOUS SPEECH:

EMPIRICAL OBSERVATIONS AND THEORETICAL IMPLICATIONS

Dedicated to my beloved grandfather

ACKNOWLEDGMENT

I would like to use this space to express my gratitude to many people who helped me tremendously throughout the past five years. First and foremost, I am hugely indebted to my advisor, Mark Liberman, the person who has the most wisdom that I have met so far in my life. Mark not only helped me transform my dissertation from a much worse state to its current form, but also taught me how to solve problems both in doing research and beyond. This dissertation has also benefited greatly from discussions with my committee members, Kathryn Schuler and Gareth Roberts, as well as the audience of the DISS 2019 workshop, especially Melissa Redford and Ralph Ross. Of course, I need to acknowledge the help and mentorship from other faculty members at Penn Linguistics Department, particularly Jianjing Kuang, who reshaped my view on phonetics and phonology.

Over the last couple of years, I enjoyed every moment that I spent with the Phonetics Lab and the conversations that I had in that spacious and bright office. Every single person in that room deserves mentioning: Nari Rhee, Jia Tian, Wei Lai, Nattanun Chanchaochai, Yiran Chen, Hui Feng, Juhong Zhan, Ping Cui, Aletheia Cui and Sunghye Cho. I am also grateful to be able to spend five years in such a great community, and have found friendship with people like Hongzhi Xu, Andrea Ceolin, Milena Šereikaitė, Hassan Munshi, Andressa Toni, Amy Goodwin Davies, and my entire 2015 cohort of linguistics graduate students. Forgive me if I missed one or two names, but I would not survive without the emotional support from them.

Outside of the linguistics community, I am blessed to have received additional support from SASgov and GAPSA, the groups that I dedicated most of my non-academic life to. In particular, I would like to acknowledge Asminet Ling, Derek McCammond and Jacob Kaplan, with the help from whom I obtained my secret identity as a criminologist.

Finally, I must say thank you to my parents. Without their sacrifice over the years, nothing written on this paper would be possible.

ABSTRACT

THE DISTRIBUTION OF DISFLUENCIES IN SPONTANEOUS SPEECH: EMPIRICAL OBSERVATIONS AND THEORETICAL IMPLICATIONS

Hong Zhang

Mark Liberman

This dissertation provides an empirical description of the forms and their distribution of disfluencies in spontaneous speech. Although research in this area has received much attention in past four decades, large scale analyses of speech corpora from multiple communication settings, languages, and speaker's cognitive states are still lacking. Understandings of regularities of different kinds of disfluencies based on large speech samples across multiple domains are essential for both theoretical and applied purposes. As an attempt to fill this gap, this dissertation takes the approach of quantitative analysis of large corpora of spontaneous speech. The selected corpora reflect a diverse range of tasks and languages. The dissertation re-examines speech disfluency phenomena, including silent pauses, filled pauses ("um" and "uh") and repetitions, and provides the empirical basis for future work in both theoretical and applied settings. Results from the study of silent and filled pauses indicate that a potential sociolinguistic variation can in fact be explained from the perspective of the speech planning process. The descriptive analysis of repetitions has identified a new form of repetitive phenomenon: repetitive interpolation. Both the acoustic and textual properties of repetitive interpolation have been documented through rigorous quantitative analysis. The defining features of this phenomenon can be further used in designing speech based applications such as speaker state detection. Although the goal of this descriptive analysis is not to formulate and test specific hypothesis about speech production, potential directions for future research in speech production models are proposed and evaluated. The quantitative methods employed throughout this dissertation can also be further developed into interpretable features in machine learning systems that require automatic processing of spontaneous speech.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Speech disfluencies, which sometimes are also referred to as hesitation phenomena such as silent or filled pauses, false starts, repetitions and repairs, are prevalent in spontaneous speech. Even normally perceived fluent speakers can have a disfluency rate somewhere between 6 and 10 disfluent words per 100 words (Ferreira and Bailey, 2004; Shriberg and Stolcke, 1996; Rochester, 1973). The distribution of disfluencies has been shown to be dependent on factors related to both the message structuring and utterance planning in speech production. In addition, socioeconomic or discourse factors that are related to the speaker and the communication context are also proposed to be relevant. Numerous studies have demonstrated the non-random distribution of disfluency phenomena (Clark and Tree, 2002; Shriberg, 1994; Holmes, 1988). A common source of disfluencies is problems with speech planning. Speech planning problems can be surfaced as a change in disfluency in response to changes in sentence length, sentence complexity, lexical context, as well as prosodic phrasing (Goldman-Eisler, 1958; Tannenbaum et al., 1965; Beattie and Butterworth, 1979; Bell et al., 2003; Nakatani and Hirschberg, 1994; Lickley, 2015). In addition, hesitation markers are also frequently used as an expressive tool to signal a delay or mark the discourse structure of one's speech (Swerts, 1998; Clark and Tree, 2002). Evidence from research in the past decades has highlighted the need of a thorough understanding of disfluencies to benefit areas that require knowledge on spontaneous speech, such as in psycholinguistic and clinical investigations of language production, sociolinguistic research in language variation and change, as well as various aspects in speech technology.

Models of speech production often cite evidence from disfluencies. An explanation of why and how disfluencies occur based on summaries of linguistic variables can be used to infer the breakdowns during the hypothesized hierarchical process of speech production, hence supporting the specifics proposed in the model (Levelt, 1983; Holmes, 1988; Ferreira and Pashler, 2002). Evidence in support of the existence of a hierarchical process underlying speech production has been presented since the earliest studies on this issue (Maclay and Osgood, 1959; Goldman-Eisler, 1958; Levelt, 1983, 1989). Under this view, speech production is achieved through passing down an abstract idea through several stages involving syntactic planning, lexical selection and access, as well as phonological planning and motor control for articulation, resulting in the acoustic signal of speech. Disfluencies could therefore inevitably happen at all the stages involved in the production process. To understand the cognitive mechanism behind speech production is thus among the key motivations in disfluency research. On the other hand, models that exclusively focus on the motor control in the speech production process have also become a center for investigation in the neuroscience of human speech (Guenther, 2006; Bohland et al., 2010; Hickok, 2012).

In addition to breakdowns in the process of speech production, evidence has suggested that contextual variables, such as the topic of university lectures (Schachter et al., 1991; Moniz et al., 2014), familiarity with interlocutor (Bortfeld et al., 2001), speaking style (Moniz et al., 2014), the nature of the task (human-human vs human-machine communication) (Shriberg, 1994) as well as the processing load of the linguistic context (Bortfeld et al., 2001; Arnold et al., 2007), affect the rate and form of disfluent speech. This body of literature reports that higher disfluency rate is associated with situations with more demanding context, such as talking to a stranger, performing a harder task, or talking about more domain-specific jargon. However, some unfamiliar situations, such as during human-machine interaction, may have an effect in the opposite direction (Levelt, 1983; Blacfkmer and Mitton, 1991; Lickley, 1994). This difference between human-human conversation and human-machine interaction suggests that disfluencies can be affected by the perceived need of specific communication task (Broen and Siegel, 1972). Results from these studies indicate that disfluencies are more than a mere by-product of performance deficit in response to contextual variation. Speakers can actively control the production of disfluencies through more careful planning based on factors involved in the speaking task itself. For instance, certain sociolinguistic variables can potentially be among these factors. Recent studies looking at factors such as age, gender and English varieties, have shown that these variables do systematically correlate with the use and frequency distribution of filled pauses (Fruehwald, 2016; Tottie, 2011, 2014). The distributional difference across age and gender has also been interpreted as a change in progress led by females (Wieling et al., 2016).

Disfluencies as both a device for message restructuring and a symptom of breakdowns in the production process has broader implications beyond the interests of linguists and cognitive scientists interested in speech production and perception. Variations in the form and location of disfluent utterances can inform us about the potential cause of deficiencies in one's linguistic ability (Grossman and Ash, 2004). For example, variants of Frontotemporal Degeneration (FTD), a family of neural degenerative diseases with known effect of affecting human linguistic ability, can be characterized and distinguished in part by the surface language deficiencies. Effortful speech is a symptom for the non-fluent agrammatic variant of primary progressive aphasia (na-PPA) (Ash et al., 2010); patients with the semantic variant (svPPA) are often diagnosed with difficulties in lexical access (Ash et al., 2009; Mack et al., 2015). However, in addition to these apparent impairment with specific linguistic abilities, it has also been shown that temporal and prosodic features of such clinical speech are also crucial in distinguishing patients from healthy controls, and between different phenotypes (Nevler et al., 2017). Using linguistic information for the diagnosis of neural degeneration is an emerging field where the role of disfluent speech has already been highlighted. However, the understanding of the relation between the disruption in speech production and the underlying functional impairment is rather limited (Boschi et al., 2017).

The presence of disfluencies in natural speech poses great challenge in human language technology. One direct benefit from a robust understanding of the distribution properties of disfluencies is in fact to facilitate systems dealing with natural speech to accurately identify and eliminate the adversarial effects of the presence of disfluencies. Therefore substantial amount of effort has been made in automatic detection and removal of disfluencies from spontaneous speech to improve performance of both ASR and TTS systems (Liu et al., 2006; Shriberg and Stolcke, 1996; Qian and Liu, 2013; Siu and Ostendorf, 1996; Hough, 2014; Ostendorf and Hahn, 2013; Nakatani and Hirschberg, 1994). These studies both provided detailed pattern description of various disfluency phenomena with the goal of contributing to practical application (Stolcke and Shriberg, 1996; Siu and Ostendorf, 1996; Plauché and Shriberg, 1999), and developed statistical methods for the task of identification and removal of disfluencies (Liu et al., 2006; Hough, 2014; Qian and Liu, 2013). Goldwater et al. (2010) explicitly addressed the question of how disfluencies are related to the errors made by ASR systems. They suggested that repetition tokens, word fragments, as well as acoustically or prosodically indistinguishable disfluent segments are associated with higher error rate. Their results highlighted the need to fully explore the feature space of speaker variation to account for the observed error pattern. Although unlike early automatic speech recognition systems which were mainly trained on read speech or otherwise constrained speech format (Butzberger et al., 1992), the acoustic or language models trained on normally produced speech may still face the problem of generalizing across different domains, such as tagging twitter, blog post and spontaneous speech (Foster, 2010). It can be more challenging in the low resource domain where suitable data is not only sparse or expensive to acquire, but also presents large amount of deviation from the standard language, such as in the setting of clinical interviews. Thus detailed understanding of distributional properties of disfluencies across domains is still necessary for overcoming the constraints in modern speech technology.

Speech disfluencies are also an integral part in evaluating and improving dialogue systems that involve human-machine conversations. The difference among the three corpora used in Shriberg (1994, 2001) demonstrated that fewer disfluencies should be expected in human-machine communication in travel planning domain. In a human-robot communication scenario, Skantze et al. (2013) show that silence and filled pauses can inhibit user activity in a map task, realized as changes in user behavior in drawing. Modeling user disfluencies has also been shown to improve the engagement of human-robot conversations and the management of the flow of dialogue. Bohus and Horvitz (2014) proposed a forecasting and hesitation mechanism that leverages human disfluency information to predict user engagement, and generate proper response to facilitate a more fluid conversation. Skantze and Hjalmarsson (2010) showed that a dialogue system that incrementally incorporates filled pause and self-repairs can achieve shorter response time and generate more naturally perceive speech, even though the generated utterances tend to be longer.

Given the large volume of work in speech disfluencies from fields ranging from linguistics, psycholinguistics, sociolinguistics to natural language processing and language generation, there is still the need to further elaborate how the multivariate feature space jointly defines the distribution of disfluencies. On the one hand, pieces of information have been provided from researchers focusing on questions concerning primarily the interests within one's own field. However, due to differences in research methods, including not only experiment design, but also the classification and annotation of disfluent speech segments, cross-domain generalization of these results can be challenging. On the other hand, both the design issue and availability of suitable data and computing resource also limit the analyses performed on certain forms of disfluencies. In this dissertation, I will attempt to explore further into this joint feature space by addressing questions from the following perspectives: the unexplored covariates of disfluency variation, the under studied forms of disfluencies, and the overall lack of understanding beyond fixed communication settings and speaker's cognitive state.

A deeper understanding of this dynamic and complex feature space behind disfluencies is crucial for both practical and theoretical reasons. Practically, applications involving natural human speech have the need to accommodate the presence of disfluent speech or utilize the information contained in it. For instance, inserting filled pauses in synthesized speech has been shown to improve the perceived naturalness of the system (Adell et al., 2012). Information contained in disfluent interview response can be used for disease diagnosis. Theoretically speaking, such an understanding will not only help resolve, or dismiss, the dispute over whether disfluencies, or hesitation markers more specifically, should be considered part of human linguistic apparatus which, at least partially, convey lexical meanings (Clark and Tree, 2002) or more of a by-product when one is trying to maintain fluency (Lickley, 2015), but also inform us the dynamic role that disfluencies play in structuring the content of speech and buffer outside disruptions.

1.1. Research questions

The goal of this dissertation is to provide an empirical description of the distribution of different disfluency phenomena, as well as to explore the potential feature space that can be used to address both theoretical questions regarding the speech production process and applied issues dealing with spontaneous speech. A major deficit in previous research is the imbalanced attention received by different forms of disfluencies with regard to their natural distribution. More focus has been placed on silent and filled pauses than repetition and repair. This is especially true when it comes to large scale corpus studies. One direct consequence is that the knowledge on the distribution property with respect to the immediate linguistic context of more complex hesitation phenomena is rather limited. Among the studies dedicated to repetitions and other repair phenomena, less attention has been paid to explore the joint prosodic, lexical and phrasal factors that defines the space of variation. With the observations in, for example, Shriberg (1994, 2001), it should be argued that along with the surface form variations, individual variation and their underlying sociolinguistic and cognitive factors, should be more systematically examined to fully account for the variations in disfluency phenomena. Therefore, the core of the questions that I would like to raise in this study is how the variation in speech disfluencies can be jointly characterized and explained across multiple dimensions. This question is crucial in pushing the boundary of the understanding of the speech production process, which has great implications for both theoretical and applied research.

This fundamental question is explored through analyses from the following two perspectives in this dissertation. First, I elaborate on the features that help define disfluency phenomena, and how these features are related to measurements that are potentially associated with aspects of speech production. With regard to the feature space involved in speech disfluencies, it is less clear how elements in communication, such as the role of the interlocutor, conversation topics and speakers cognitive state, contribute to the variation of the observed disfluencies. These variables are nevertheless fundamental in understanding the speech production process. Efforts have been made in sociolinguistic literature to understand the variation with a limited feature space. Regarding certain hesitation phenomena as sociolinguistic marker, such as filled pauses, has received more attention in the past two decades. More recent studies such as Wieling et al. (2016) have demonstrated an interesting gender distinction in the choice of fillers which can be attributed a trend explained by a change in progress. Less understood in this domain is how the topic of conversations, as well as the interlocutors, affect variations in disfluencies, and how such meta-linguistic information interact with the immediate linguistic contexts in which disfluencies are realized. These variables are crucial in clarifying how much of the observed variation is indeed a change in the sense of sociolinguistics, and how much is a reflection on the cognitive process of speech production.

Second, I examine potential implications of these empirical observations on the advancement of the modeling of speech production. This is carried out through cross domain comparison of disfluency phenomena. The domains to be considered include the cross-linguistic, cross-communication context, and cross-speaker's cognitive state realizations of the same disfluency phenomenon. Looking back at the literature, our understanding of disfluent speech is predominantly based on studies in English within a single population group, with a handful papers concerning other languages such as German, French, Portuguese, Hebrew, Mandarin and Japanese (Fox et al., 2010, 1996). The lack of linguistic and speaker state diversity constraints researchers from discovering and exploring the disfluency phenomena that bear language-specific characteristics. For example, cross-linguistic studies on repetitions have generally acknowledged that function words, especially those immediately preceding a content word in a constituents, are more frequently repeated than other word classes (Fox et al., 1996, 2010; Clark and Wasow, 1998). However, it is not clear what would happen to a language that relies predominantly on morphological devices to realize agreements and indicate spatial or temporal relations. On the methodology side of the problem, attention to the cross-linguistic aspects of disfluencies, or self-repairs more specifically, is mostly from conversation analysis, such as in the form following Schegloff et al. (1977). In terms of communications settings, the brief review above has shown that few studies have explicitly compared the speech produced for different tasks, and even fewer addressed the question of how cognitive impairment would affect the production of disfluencies as compared to normative fluent speakers, even though the study of disfluencies as a clinical condition or symptom is abundant.

Underlying these questions is the need to enrich our current understanding of disfluencies comprehensively considering simultaneously both linguistic and meta-linguistic variables. Such descriptive work is essential in establishing the underpinnings for both the theoretical work on speech production models, and developing applications in various fields requiring knowledge about spontaneous speech.

1.2. Methods

Research in speech disfluency has been conducted through both quantitative observations of speech corpora and careful analyses of speech produced in controlled lab experiments. Both research methodologies have their unique advantages over the other, while facing particular challenges of its own.

Corpora of collections of spontaneous speech have been a primary source for disfluency research. The apparent reason in favor of using corpora of spontaneous speech is that speech disfluency primarily occurs in spontaneous and under-prepared speech. Given its relatively rare occurrence (as only 6% to 10% as the generally acknowledged disfluency rate) and great potential for individual variation, the amount of speech that is required to capture enough variance tends to be large. Thus corpus based research caters well with these demands. The widely used speech corpora include the Switchboard (Godfrey et al., 1992), ATIS (Dahl et al., 1994), and AMEX corpus (Kowtko and Price, 1989). These corpora represent a wide range of scenario where communication tasks tale place. For example, Switchboard consists of unguided spontaneous telephone conversations between two speakers under a provided topic, while ATIS represents human and machine oriented speech in the scenario of travel planning. Studies based on such wide representations of speech data (Shriberg, 1994, 2001; Shriberg and Stolcke, 1996) have been fruitful in identifying linguistic and contextual or discourse variables that correlate with surface variations of speech disfluency.

However, corpus-based studies of speech disfluency are often faced with two major challenges: the lack of properly annotated data and insufficient control for the environment in which the speech was produced. The first challenge is mainly due to the lack of awareness of creating properly annotated disfluencies during data collection. However, with disfluency information included in the transcription, significant performance gain has been reported in automatic part of speech tagging (Johnson and Charniak, 2004). Same applies when disfluent speech segments are removed (Kahn et al., 2005). The second challenge, though is of lesser concern, poses questions to the interpretation of the observed disfluency patterns. As it is often the case that corpora of spontaneous speech are comprised of conversations relatively freely conducted by task participants, causal interpretations between the linguistic variables and disfluency events are even harder to establish (Schnadt and Corley, 2006).

Experimental work has also been conducted to explore the nature of speech disfluency (Zvonik and Cummins, 2003; Tanenhaus et al., 1995; Arnold et al., 2007; Bortfeld et al., 2001; Schnadt and Corley, 2006). A typical production experiment is set up in a way that participants are expected to produce speech guided by certain speaking tasks. One research strategy in looking at silent pause in read speech is to look at synchronous speech (Zvonik and Cummins, 2003; Krivokapić, 2007), where speakers are asked to read some text either along or in synchrony with a partner. This method is meant to control for individual variation of text reading in an experiment setting. Some other research engages participants in tasks whose completion requires communication with a partner (Schnadt and Corley, 2006; Bortfeld et al., 2001; Arnold et al., 2007). For example, Schnadt and Corley (2006) adopted a network task developed from Levelt (1983) and Oomen and Postma (2001), where the task is based on describing a network structure. Bortfeld et al. (2001) asked pairs of speakers to describe and match sets of objects, where the processing load was controlled for by manipulating the familiarity of objects, familiarity between two speakers, and so on.

Language production tasks are widely used as tools for neural degeneration diagnosis. For example, the Boston Naming Test (Goodglass et al., 2000) and picture description tasks such as the Cookie Theft picture. However, the goal of these tasks is mainly as prompts to elicitate spontaneous speech within a more constraint environment without assuming factorial conditions. Therefore they are fundamentally of a different nature than the experimental methods listed above.

Although the speech produced in more controlled lab settings could provide many desirable properties for variable extraction and causal interpretation, it faces the problem of higher limitation on the amount of accessible data from experiments and lack of direct connection between lab conditions and real-world situations. This limitation can be illustrated through one heavily studied question: what is the effect of lexical frequency and sentence complexity on disfluency production. In a well-controlled experiment, several studies (Tannenbaum et al., 1965; Beattie and Butterworth, 1979; Jescheniak and Levelt, 1994) find significant effect of lexical frequency or context predictability on the fluency of speech or shorter response time in object naming. However, when lexical frequency is defined as the extent to which people agree on the name of particular nouns, strong effect of this measure of familiarity is also observed (Hartsuiker and Notebaert, 2009). Then the question of frequency effect on fluency becomes whether it is the speaker experience or global frequency of words that affect the fluency of speech. In another study (Tsiamtsiouris and Cairns, 2013), sentence length and structural complexity have been shown to interact with the disfluency of produced speech through an experiment of repeating sentences of 6 and 30 words long. Unfortunately, this study failed to discuss other covariates that might as well explain the observed group difference, or how much variation in disfluency can still be explained by their controlled variables when other unobserved effects are present.

In addition, the use of lab speech doesn't necessarily solve the problem of proper annotation. While it can be claimed that careful transcriptions of disfluency events are made possible as the speech data is under the full control of the researcher, the transcriptions are often compromised with the size of speech material and hurdles for cross-lab generalization. Since most speech does not happen in well controlled environment as responses to various production tasks, practical implications from lab studies might be limited.

1.3. Research methods for the current study

Comparing the two streams of research methodology, corpus-based analyses is still preferred over fine-controlled lab speech, for at least three reasons. First, the two major challenges faced with corpus data, namely the lack of known control and inadequate annotation, can be mitigated in feasible manners. The concern for the lack of causal interpretation can be circumvented if the right data is used in the study, such as through consistently annotating disfluent speech, proper sampling method, and using large quantity of balanced speech data. The issue of disfluency annotation can be solved at least in part through semi-automation. On the one hand, hesitation markers such as filled pauses and full word repetitions can be relatively accurately identified automatically, while applying certain optimization procedure can significantly reduce the time needed to annotate other disfluent phenomena. Second, using publicly available corpus data facilitates collaboration within the field across labs and institutions, and preserves the integrity of research results. Finally, research from the naturally produced speech in realistic communicative settings can more easily be translated to applications that benefit various research communities and beyond. The observed patterns from careful description of naturally produced speech can also be more informative to answers to questions about speech production in psycholinguistic research. With these apparent benefits in mind, corpus-based analyses is adopted in the current study.

1.4. Corpora selection

A total of six speech corpora are selected to explore the proposed questions in this study: Fisher (Cieri et al., 2004), SCOTUS 2001 (Yuan and Liberman, 2008), Alcohol Language Corpus (Schiel et al., 2008), UCLASS (Howell et al., 2009), Czech Spontaneous Speech Corpus (Kolár et al., 2005) and a compilation of political speech from American Rhetoric (Rhetoric, 2020). Among this selection, the primary source of information is the Fisher corpus, which I will provide a brief overview below. Other speech corpora will be introduced when they are due for analysis.

1.4.1. Fisher

Fisher (Cieri et al., 2004) is a corpus of telephone conversations created in response to the unique needs for automatic speech recognition (ASR) systems. The corpus is built with consideration of controlling a broad range of factors that are essential in representing daily conversational speech. The entire corpus contains 16,454 conversations, totalling 2742 hours of speech. Unlike most of other speech corpora, Fisher balances speakers' age, gender, and represents a wide range of dialectal variation. To encourage the inclusion of large quantities of vocabulary, conversations were guided by 40 topics that are pertinent to both day-to-day life and current pressing social and political issues. The list of topics can be found in the appendix. The collection process also included a platform-driven protocol, with which the data collector initiates calls and matches between potential participants who expressed interest in the selected topic. This procedure not only maximizes inter-speaker variability, but also reduces the sampling bias to a great extent. Each participant was expected to complete at most three 10-minute conversations. However, the actual number of conversations each participant contributed may vary. Unedited transcriptions are also made available in the corpus.

Given the nature of the Fisher corpus, it is particularly suitable for exploring inter-speaker and inter-contextual variations in speech production. In this study, the subset of Fisher that contains native speakers of American English who completed exactly three conversations are chosen to evaluate these variations. The selected sample contains 9471 one-sided speech from 3157 speakers. The total duration of speech is about 790 hours.

1.5. Some definitions

Before moving to the actual corpus analyses, some clarification in terminology is necessary. The kind of disfluencies of my primary concern is same-turn self-initiated disfluencies, in contrast to disfluencies that may involve other initiated repairs. As reviewed above, the kinds of analysis in this study are highly dependent upon the amount of information annotated in the selected corpora, as the corpora are made up of different speech styles, collected following different protocol, and intended to serve very different demands. As outlined at the beginning, the Fisher corpus can be suitable for exploring individual variation in disfluencies as a function of sociolinguistic variables as well as the effect of discourse factor such as conversation topic. The primary disfluency phenomena being addressed with this corpus are silent pauses, filled pauses, and repetitions, given the constraints proposed by the amount of data and available annotated information. On the other hand, SCOTUS is mainly used to answer questions about individual variation in repetition and repair, because it contains ample speech from each of the eight supreme court justices, while maintaining a reasonable amount for proper annotation based on verbatim transcription.

Two clarifications have to be made with regard to the analysis window considering the nature of speech being analyzed. The questions to be answered are first what is the basic unit of analysis and how to define it? And what is the criteria for labeling the disfluency categories? Shriberg (1994) finds her disfluent instances from individual “sentence”, where a sentence is defined as a unit that can otherwise be marked with a period or question mark. Her judgement is somewhat arbitrary and hard to implement with large quantities of data that have not been properly segmented. In the remainder of this section, I will define the window of analysis for the three kinds of speech: telephone conversations, court debate, and radio interviews. The definitions are derived not only from the characteristics of different speaking styles, but more with considerations on the data collection process of different speech corpora. Disfluency phenomena classification and annotation are defined in later type-specific discussions, with specific considerations of the analysis and speech domain variation.

1.5.1. Key definitions in spontaneous conversations

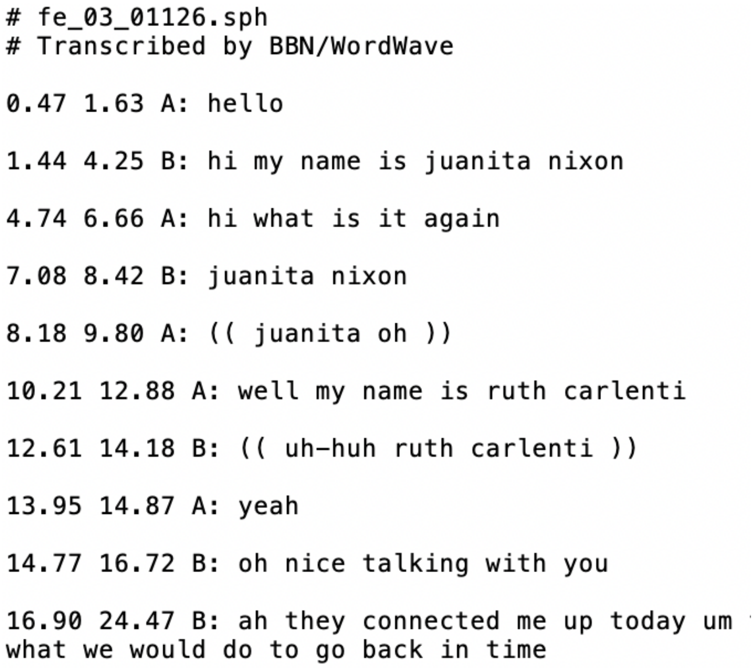
Key definitions concerning conversational data in the current set up are turns and utterances. The definition of an utterance can be fuzzy, especially in conversation settings. It has to be acknowledged that an utterance does not necessarily consists of a complete sentence: the sentence can be incomplete, or multiple sentences can form a large utterance group. One alternative to identifying utterances basing off rhythmic groups as well as considering the connection between pausing and syntactic constituency (Zellner, 1994). However, identifying utterance groups en mass based on these fuzzy definitions and correspondence also faces the problem of the uncertainty of the nature of the speech context.

Another critical definition in the current set up is what constitutes a silent pause? In addition to the debate over a binary threshold, in the context of telephone conversation, there could be pauses (or gaps) at the juncture of turn taking Beattie and Barnard (1979).

These pauses are not necessarily related to problems with speech production, but could be about courtesy to the interlocutor, or floor-holding. Filled pauses, similarly, can be used to hold the floor when two speakers were not facing each other to avoid dead air. More elaborated considerations on the structure of turn-taking are discussed in Heldner and Edlund (2010). Apparently these silences are not the silent pauses of the primary interest in this study.

With all the considerations above, I define the silent and filled pauses in the discussion on conversational below as within-turn pauses, which means that silences at the beginning of a turn, without preceding speech segments, as well as speech segments that solely consist of filled pauses and non-content words are excluded from the analyses. A turn is then defined as the contiguous speech segments from one speaker between two speech segments of their interlocutor.

Although Fisher transcripts only consist of arbitrarily segmented chunks of speech material, mainly for the ease of transcribers rather than offering any linguistic insight, they do offer relatively clear clue for turn identification. Figure 1 illustrate the format of raw transcripts in Fisher. With the information readily available, turn segmentation is decided based on sorting and merging the time stamps by conversation sides provided in the transcriptions. One crucial observation that facilitates this segmentation is that floor holding or back channel talking often consist of short segments (less than four words long) of fillers or non-content words (such as *that's right, yeah*), which can be discarded without disrupting the overall integrity of speech transcripts. Although relying on the time stamps provided in transcriptions is not immune to segmentation errors, by excluding floor holding and back-channel talking through a simple rule, this segmentation method should return at least almost correctly segmented turns. Silent pause identification is then based on forced alignment results using Penn Forced Aligner (Yuan and Liberman, 2008) on the turn segmented speech.



```
# fe_03_01126.sph
# Transcribed by BBN/WordWave

0.47 1.63 A: hello

1.44 4.25 B: hi my name is juanita nixon

4.74 6.66 A: hi what is it again

7.08 8.42 B: juanita nixon

8.18 9.80 A: (( juanita oh ))

10.21 12.88 A: well my name is ruth carlenti

12.61 14.18 B: (( uh-huh ruth carlenti ))

13.95 14.87 A: yeah

14.77 16.72 B: oh nice talking with you

16.90 24.47 B: ah they connected me up today um
what we would do to go back in time
```

Figure 1: An example of the raw transcription format of Fisher corpus.

Turn identification in other corpora such as SCOTUS is an easier task compared to that in telephone conversations, as these conversations or monologues are face-to-face, which eliminates back channel talking and floor holding fillers. The time-aligned verbatim transcriptions are also segmented into contiguous speech from different parties in a session. Thus a turn is simply a segmented transcription file. Compared to telephone conversations, the turns in these corpora are more likely to contain complete sentences. This observation may facilitate analysis of disfluencies at different syntactic or prosodic junctures. However, it does not warrant changing the minimal analysing unit from a turn to an utterance.

The Czech corpus is the most well annotated corpus among the corpora I propose to use in this study. Sentence boundaries as well as boundaries of syntactic phrases will be determined based on corpus annotation.

1.6. The structure of the dissertation

The dissertation will be structured as the following. In chapter 2, I will set up the ground by reviewing the literature on both the descriptive analysis of speech disfluencies and related theoretical and practical discussions on speech production related questions. More attention will be paid to the types that I will focus on in this study. Chapter 3 explores the sociolinguistic and discourse variables that potentially play a role in variations in the production of silent and filled pauses. They include the sociolinguistic contexts that have been more thoroughly examined in the literature, such as age, gender and English dialects, as well as the less discussed questions regarding the role of conversation topics and speaker accommodation. Chapter 4 through 6 primarily explore questions regarding repetitions in spontaneous speech. I will examine the linguistic features, including lexical, phrasal and prosodic features, that help define different repetition phenomena. Implications of the effect from these linguistic variables to language production models, as well as practical applications in identifying speaker's cognitive states, will also be discussed. Final discussion and remarks will be made in Chapter 7.

Chapter 2

Background

In this chapter, I will review the literature pertaining to speech disfluencies from three distinctive yet highly intertwined perspectives: Efforts that aim to distinguish different disfluency phenomena, research on speech production of which speech disfluencies are an integral component, and the practical implications of speech disfluencies. It will be argued that further descriptive work on speech disfluencies is still needed for both the interest of a refined speech production theory, and the development of future empirical applications.

2.1. Classification of disfluencies

Disfluency phenomena have been discussed from different perspectives, and have followed different descriptive paradigms. The two widely approached points of view are the forms of the disfluencies and their functions in the process of speech production (Lickley, 2015). Formal descriptions of the form of disfluencies normally do not assume particular functional underpinnings that explain the surface variation, although the two broad perspectives are never cleanly separate. In this section, I first review some of the existing classification schemes of both the form and function of disfluencies, then I elaborate on the disfluency phenomena that are the focus of this dissertation.

2.1.1. Classification systems of disfluency phenomena

Speech disfluency, as implied through the terminology itself, refers to disturbance or disruption during speech production. The source of this disturbance can be pathological, but is also attributable to occasional break-downs during the production process. Levelt (1989) proposed a model which points out the locations in this process which the break downs may happen, and how disfluencies can be informative for our understanding of speech production. In this model, speech production is accomplished incrementally through three basic stages: formulating the message to be delivered, organizing the linguistic materials that are essential for communication, including syntactic planning, lexical access and phonological selection, and finally controlling the motor system to produce the intended linguistic output. Disfluencies thus reflect the disruptions that occur at different stages in this process. The cause of such disruptions can be pathological.

Unsurprisingly, some early classification systems are motivated initially to serve the needs of particular clinical population. Mahl (1956) refers to the disruption in fluent speech as "disturbance" for the main interest in distinguishing normal and schizophrenic speech. This classification recognizes nine distinct categories of "disturbance", which are: *ah, sentence correction, sentence incompletion, repetition of words, repetition of partial words (stuttering), intruding incoherent sound, tongue slip, whole or partial word omission.* On the other hand, Johnson (1961), with the focus on comparing stuttering and non-stuttering speech, proposes another set of eight categories of "disfluency". In this system, Mahl's *ah* is categorized as *interjection*. Broader terms, such as *revision*, *incomplete phrases*, *broken words*, are also used to replace some of the similar but more specific classes in Malh's system, such as *sentence correction*, *sentence incompletion* and *repetition of partial words*. This change in terminology shows the need for better generalization in describing disfluency phenomena. Johnson's system additionally reflects the special need for stuttering research, which can be seen through the inclusion of domain specific category of *prolonged sounds*.

With the goal of describing more general "hesitation" phenomena in spontaneous speech, Maclay and Osgood (1959) adopted Mahl's first four categories, while replacing the last three with *filled pause* and *unfilled pause*, *non-retraced false start*. The remaining categories in Mahl's system: *repetition of words* and *stutter*, have been consolidated into one category *repeat*. They believed that this categorization represents the most of hesitation phenomena that happen in spontaneous speech. Similarly, in yet another system, Blankenship and Kay (1964) largely follow Malh's first four categories, but change the rest into *word change* and *non-phonemic lengthening of phonemes*.

More recent studies in speech disfluencies often tend to cater particular needs in domains such as human language technology and cognitive science. The methodology of such research relies on analyses of large scale speech corpora. This approach requires more systematic and consistent annotation mechanism. Shriberg (1994) consolidates an array of classification systems (Mahl, 1956; Maclay and Osgood, 1959; Blankenship and Kay, 1964; Levelt, 1983; Blacfkmer and Mitton, 1991; Bear et al., 1993) into a 5-category scheme consisting the following basic forms: *filled pause*, *repetitions*, *substitutions*, *insertions* and *deletions*. This categorization is followed in later studies (Heeman, 1997; Lickley, 1998; Eklund, 2004).

Shriberg (1994)'s system also acknowledges the fact that disfluency phenomena are structured. In general, a disfluent segment consists of the word/partial word/phrase/partial phrase that to be repaired (reparandum), the interruption point, the hiatus (interregnum), the repaired or repeated word or phrase, and the resumption of fluent utterance. Figure 2 illustrates the structure of a disfluency segment in an utterance.

Unlike the classification system of disfluencies based purely on the form variation, a functional view of a system tries to identify what is the cause of the failure in fluency. Examples of such a view include Dickerson (1972) and Hieke (1981). Although these studies differ in their particular practical or theoretical goals, the classifications they adopted recognize the distinction between a way to gain more time for planning (hesitation), and a strategy to re-establish fluency after a break (the need for repair). This functional dis-

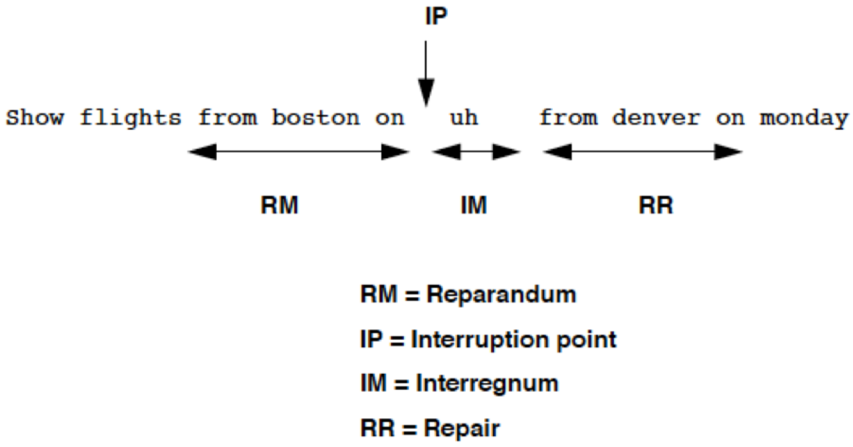


Figure 2: The structure of a typical disfluent region. Image taken from Shriberg (1994).

tinction has been claimed to be distinguishable from the surface patterns in relation with the relative sequencing of repetition and filled or unfilled pauses. A more elaborated and detailed functional classification for repairs is discussed in Levelt (1983), which will be reviewed in more detail when repair disfluency is considered. Since the goal of the current study is focused more on a substantial description of the patterns, a functional classification system is of secondary interest in this discussion. Thus it will only be brought up when an interpretation of the patterns makes it necessary.

In this study, I primarily consider three out of the five generally agreed categories of disfluency (Shriberg, 1994; Heeman, 1997; Lickley, 1998; Eklund, 2004): *filled pause, repeat*, and the un-annotated but closely related hesitation phenomenon *silent pause*. Repair will only be discussed in relation with filled pauses and repetitions when deemed necessary, for the reasons explained below. Prolongation is yet another major disfluency type that received a lot of attention. However, I will defer to another separate study to explore its categorization and distribution, due to its close correlation with other hesitation phenomena as well as the intrinsic properties of the syllable. The present study focuses on the three disfluency phenomena and starts from the existing categorization systems to provide detailed documentation of variations within each kind of disfluency, and discover overlooked properties that could define new disfluency categories which are meaningful for both theoretical and practical grounds.

2.1.2. Silent pause

Silent pause refers to the brief period of silence during speech production. As natural speech presumably contains silences of varying duration and for multiple reasons, such as marking the end of a prosodic group, closure during producing a stop consonant, or hesitating in responding questions, the first question about silent pause is what duration constitutes *brief*? And the second impending question is how could one distinguishes a pause that is truly hesitant from a pause that is resulted from the phonology or prosody of one's speech?

Neither of the questions has straight forward answers. To understand what is the duration that justifies a silence period as a silent pause, we should first acknowledge the large variability in terms of the pausing pattern. For example, Luce and Charles-Luce (1985) reported that the duration of closure in English word final stops can vary between 30 ms to 250 ms. Thus a pause as short as 50 ms could be interpreted as a silent pause in some context but not in others. From the perspective of processing time, Ferreira (2007) argues that pauses as short as 80 ms can be the result of planning difficulty or prosodic processing. The existence of a wide range of silent pause duration puts the significance of asserting a definitive threshold for what constitutes a silent pause into question. Nevertheless, one would still hope that a reasonably defined cut-off point could still provide some information about silent pauses and hesitation within particular domains.

In the early yet still influential work, Goldman-Eisler (1958) proposed that a threshold of 250 ms should be used in research addressing questions concerning the cognitive process of silent pauses, as pauses shorter than this threshold are generally accounted for by articulatory adjustments during speech production. Although adopted by many (Mack et al., 2015; Beattie and Butterworth, 1979; Greene and Cappella, 1986), some lower threshold values have also been used (Gee and Grosjean, 1983; Eklund, 2004; Martin, 1970). In some other work, a much lower threshold, as short as less than 100 ms, was used in annotation (Martin, 1970; Eklund, 2004). Butcher (1981) further demonstrates that what perceived as a silent pause does not only depend on the absolute duration, but is also conditioned on the prosodic context. Thus it is more informative to understand what is the distribution of silent pause duration, and selecting a threshold reflecting context specific properties.

To attend to the context sensitive nature of hesitation pauses, many studies ask what are the discourse, syntactic, prosodic and dialectal effects on pause duration (Levelt and Cutler, 1983; Krivokapić, 2007; Zvonik and Cummins, 2003; Kendall, 2009). Using a synchronous speech method, Zvonik and Cummins (2003) showed that silent pauses between syntactically more complex phrases and longer prosodic phrases (Krivokapić, 2007; Zvonik and Cummins, 2003) are longer. However, Krivokapić (2007) suggests that more complex prosodic structure doesn't have equivalent effect as an increase in syntactic complexity. Kendall (2009) argues that speakers from different dialect background also vary in their pause duration. On the question of silent pause distribution, through a large scale multilingual study, Campione and Véronis (2002) showed an existence of bi- or tri-modal distribution in silence distribution in spontaneous speech, and offered a sharp criticism on the handling of statistical analyses of duration contrast in most research practices.

Research into the relation between silent pause duration and its immediate syntactic or prosodic context inevitably touches on the second question: How to distinguish the pauses that are caused by hesitation from a fluent pause? Krivokapić (2007) and Zvonik and Cummins (2003)'s results may indicate that there is some relation between properties of fluent pause and the syntactic and prosodic structures in which the pause occurs. Attempts have been made to link the complexity of syntactic and prosodic structure to fluent pause du-

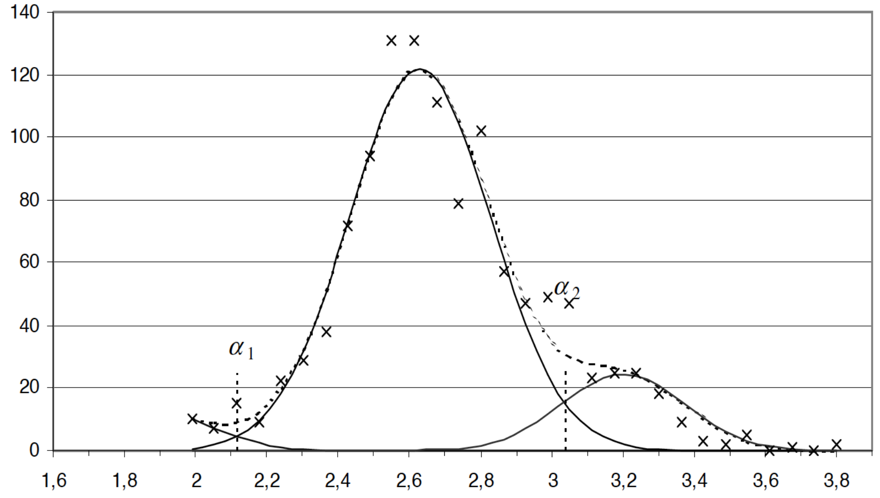


Figure 3: Bi-model distribution of pause duration in French spontaneous speech, as demonstrated in Campione and Véronis (2002). The time units (the horizontal axis) are in $\log_{10} ms$.

ration (Cooper and Paccia-Cooper, 1980; Gee and Grosjean, 1983; Watson and Gibson, 2004). However, Ferreira (1993, 2007) argue that there isn't a direct relation between the structure of syntactic and prosodic phrasing. According to Ferreira's proposal, prosodic based pauses have to be distinguished from performance based pauses. The distinction can be made through their relation to the phrase before and after the pause: performance based pauses are associated with the following phrase, while prosodic based ones are associated with the preceding phrase. Therefore only performance based pauses are related to difficulties in planning and can occur anywhere in an utterance. Thus performance pauses occurring at syntactic phrase junctures may reflect difficulties in syntactic planning, and within phrase pauses may be related to problems with lexical access, according to Levelt's model for language production.

Although Ferreira's distinction between performance and prosodic based pauses offers a clean explanation for understanding the pausing phenomenon, it doesn't really address the question of how to differentiate the two types of pauses in practice, especially in annotating speech corpora. Thus in practice, researchers still utilize perceptual based judgement (Nakatani and Hirschberg, 1994; Eklund, 2004; Clark and Tree, 2002) in identifying hesitation pauses, and it seems still the best practice before consistent and objective way to identify pauses has been developed (Lickley, 2015).

2.1.3. Filled pause

Filled pauses, which are also often referred to as fillers, is probably the most easily distinguishable and heavily studied disfluency phenomenon. A crucial distinction between filled pauses and filler words has to be made clear. Filled pauses, in most disfluency studies, refer specifically to *um* and *uh* and their counterparts in other languages. Filler words, on the other hand, may include discourse markers such as *well*, *you know* et cetera, which are not a concern to the current study.

Filled pauses, with more explicit forms compared to other disfluency phenomena, are more easily identified in larger scale speech corpora with verbatim transcriptions. A more accurate frequency count can subsequently be obtained. The average frequency of filled pauses is somewhere between 1.3 to 4.4 per 100 words, depending on the corpora being analyzed (Bortfeld et al., 2001; Shriberg, 1994; Eklund, 2004; Lickley, 2015). Cross-linguistically, as demonstrated in a range of Germanic languages as well as in French, Spanish, Hebrew, Japanese and Mandarin, a filled pause generally takes two forms: a pure (often times reduced) vowel (such as a schwa), or such a vowel followed by a nasal coda. The exact realization of the two alternatives varies in different languages. In American English, it takes the forms of *uh* and *um*, while it's often transcribed as *er* and *erm* in British English (Clark and Tree, 2002; Tottie, 2011).

The use of filled pauses is generally regarded as related to hesitation. Hesitations in spontaneous speech can be the result of the speaker being uncertain (Brennan and Williams, 1995; Smith and Clark, 1993), under high cognitive demand (Arnold et al., 2007), or facing a choice (Schachter et al., 1991). Many of the structural properties of filled pause are often understood with regard to the general role of disfluencies in speech production (Corley and Stewart, 2008). As suggested by several studies (Oviatt, 1995; Bortfeld et al., 2001; Smith and Clark, 1993; Brennan and Williams, 1995; Swerts, 1998; Swerts and Krahmer, 2005), the insertion of a hesitation marker such as a filled pause may not be purely automatic. For example, lower rate of disfluency is observed in human-machine communication than human-human conversation (Oviatt, 1995), and Bortfeld et al. reported that in their highly controlled production environment, disfluency rate, particularly the rate of filled pauses, is greatly influenced by the role played by the speaker in the dialogue. However, these results seem to conflate the hesitations that are primarily driven by the need for planning and message structuring, with the hesitations that are resulted from contextual constraints on performance. It is also not clear whether filled pauses have any idiosyncratic properties that are distinct from disfluencies in a broader sense.

Apart from their apparent identity as hesitation markers, elements of non-randomness in filled pause distribution can be used as arguments for it being lexicalized in the speaker's vocabulary. If filled pauses are to operate in parallel with other filler words that function as discourse markers, they should be treated equivalently as *you know* and *well* which marks transitions or turns in the discourse, and therefore argued to be encoded with explicit discourse meanings (Clark and Tree, 2002). This view of filled pause, however, is refuted by Lickley (2015). Citing acoustic evidence from Shriberg and Lickley (1993), where it has been shown that fundamental frequency of filled pauses is closely related to surrounding phrases, Lickley (2015) argues that it is unlikely that speaker intentionally insert filled pauses as signals for upcoming hesitation. If it was the case, disruptions of the prosodic structure of the otherwise fluent utterance should be expected.

The curious existence of two alternative forms of filled pauses across languages has also been put under scrutiny. As the variation in its forms may suggested, Clark and Tree (2002) argue for a differentiation between the meaning of the two filled pauses. They claim that the nasal filler corresponds to a major delay in production, while the oral version signals a minor delay. This claim is supported by the observation that *um* tends to occur at the beginning of an utterance, while *uh*'s location is more often utterance internal (Shriberg, 1994; Shriberg and Stolcke, 1996). Speakers also appear to have preference over one form than the other, and some even exclusively use only one form (Shriberg, 1994, 2001). On the other hand, abundant evidence points to different preference of one form over the other by people from different socioeconomic groups (Wieling et al., 2016; Fruehwald, 2016). The trend that younger, especially female, speakers prefer *um* over *uh* is thought to reflect a change in progress that has spread across several Germanic languages (Wieling et al., 2016).

From the review above, it can be concluded that filled pauses, like other disfluency phenomena that signal hesitation, function both as facilitators for production in case of a forthcoming hesitation and a sign of disturbance in performance conditioned on the context. There is a clear distinction between the two forms of realizations, which can be the result of lexical difference, speaker's intentional choice, sociolinguistic language variation, or a bit of something from this list.

2.1.4. Repetition

Repetition is another common form of disfluency which has received much attention in the literature. The speaker may repeat a single word or phrase, or a partial word or phrase due to hesitation or the need for repair. In this review, I mainly focus on repetitions in fluent spontaneous speech. Here "fluent" refers to the fact that this repetition phenomenon does not cause trouble for the listener in parsing the speech input given the presence of disfluency in fluent speech delivery. I distinguish *fluent repetition* from *pathological repetition*. A good example of pathological repetition is stuttering, although it can share certain similarities with repetitions produced by fluent speakers (Guitar, 2013). Notice that the terminology *fluent repetition* is also used to describe an emphatic strategy that is used to emphasize or making a contrast, such as in sentences like *I really really like your idea.* This kind of repetition can be distinguished from repetitions caused by hesitation or message structuring through prosody (Lickley, 2015). Emphatic pitch, for example, can be marked on the repeated intensifiers, and other disfluency phenomena, such as silent and filled pauses, or prolongations, are not expected in the neighborhood of these repetitions. Emphatic repetitions are also mostly used with a limited set of content words or phrases, and it is unlikely the case that pronouns, prepositions and conjunctions are repeated. On the contrary, repetitions caused by repetition or the need to repair often don't carry prominent prosody, and are mostly likely to be function words (Clark and Wasow, 1998; Fox et al., 1996, 2010). In fact, Clark and Wasow (1998) reported a 10 times higher frequency of functional words being repeated compared to content words (25.2 per 1000 words vs 2.4 per 1000 words), and Lickley (1994) reported a statistics of 96% of repeated words being functional words. In terms of the location of repetition, hesitation repetitions are often at the initial of an argument (Fox et al., 1996, 2010). Cross-linguistically, at least among English, German and Hebrew, the repeated words are function words that immediately preceding the main content word of a clause. The distribution of exact lexical category of the function words, however, varies across languages, which can be interpreted as conditioned on the syntax (Fox et al., 1996). In the discussions in this study, the term *repetition* is used exclusively to refer to the fluent repetition caused by hesitation or message structuring, as reviewed above.

The forms of repetition can vary among single syllable words, multi-syllable words, multi-word phrases and word fragments. Some examples of possible variations of repetitions are listed in the examples in (1). The repeated words are only marked with boldface for the moment, although detailed annotation strategy for variations in repetition will be given later. Although higher rate of single-syllable word repetitions in the languages reported in literature has been generally reported, repeating word fragments or multi-word phrases is not expected to be uncommon. Even though there isn't reported statistics on the frequency of these different types of repetitions, counts on the frequency of disfluencies involving word fragments are available in several languages. Levelt (1989) reported 22% of word fragments in a Dutch pattern description corpus; Lickley (1994) reported 36% from conversational speech in British English; and Bear et al. (1993) found 60% in ATIS corpus. The variation may be caused by difference in the nature of the corpus, but among the reported word fragments, it might be possible that a substantial amount would involve repetitions. A careful description of the distribution of repetition types beyond word classes seems to be necessary.

(1) a. ... and **i i** actually don't watch any sports on television.

b. ... and **it's it's it's** a big question and i don't know that you know bush had an answer for that.

c. ... actually i am. **i'm not i'm not** so afraid of it.

d. ... yeah yeah they were **pr- pretty** distance portions of the state

In an account offered by Clark and Wasow (1998), the presence of repetitions shows an effort by the speaker to preserve the continuity of the speech when faced with higher cognitive demand after the initial commitment to the initiated constituent. This argument explains two facts about repetition: it is more likely to occur at the beginning of long and complex clauses, and when people repeat, they tend to restart from the beginning of the constituent in which the interruption happened. The higher rate of function words in repetition is mainly because their location in the constituent. This theory can in principle cover some causes of repetitions, but is unlikely to be the only or a major explanation. Apart from serving as a sign to restore interruption, Hieke (1981) proposes that repetitions can be a strategy that speakers uses to coordinate the flow of speech. This distinction between passive and active control of speech is termed *retrospective* and *prospective* repetition respectively. Later acoustic analyses have suggested that the two categories can be distinguished when pause length and prolongation in the region of repetitions are considered (Plauché and Shriberg, 1999; Shriberg, 1995). A third category that parallels with Levelt and Cutler (1983)'s notion of *covert repairs* is also able to be distinguished according to Plauché and Shriberg's study. This type of repetition can be briefly described as a "cover up" of a pre-detected speech error before articulating the erroneous lexical item, therefore the lexical items before the error are repeated to preserve speech continuity.

However, results from these efforts in identifying the functions of repetitions are still incomplete. For example, one fairly common form of rapid repetition has not received much attention. In their corpus study of the time lapse between cut-off and repair, Blacfkmer and Mitton (1991) suggest that the time gap can be very short and frequently effectively zero to signal any noticeable delay in the speech signal. More notably, listeners are often unaware of the existence of such repetitions, which can be attested from an accuracy evaluation of careful transcriptions of spontaneous speech (Lickley and Bard, 1998). The question, then, is whether this type of repetition a sign of hesitation, just as most other repetitions are, or is it simply random additions to the output speech as a result of execution error?

One final deficiency in the current literature inventory on repetition is the lack of cross-linguistic and cross-speaker condition perspective, especially a lack of understanding from languages with distinctive syntactic structure compared to that of Germanic or Romance languages, as well as speakers with impaired cognitive ability in processing speech. Fox et al. (2010) have already suggested the tendency to repeat function words needs to be interpreted conditioning on the morphosyntactic restrictions and word order of the language. It remains to be seen what is the most likely unit for repetition in a language which has rather limited inventory of function words such as prepositions and personal pronouns, or ordering them after the main content word in an argument.

2.1.5. Repairs

At least some evidence has suggested that repetitions can sometimes be classified as a form of repair, which has been termed as *covert repair* by Levelt (1983). However, *repair* in fact covers a wider range of more complex disfluent phenomena, and is often the place of confusion. Shriberg (1994) developed a structural coding algorithm in relation to the basic structure of the disfluency region, as illustrated in Figure 2. Seven types of disfluencies were originally identified through this system. However, for more efficient representation, I combined the category *conjunction*, which essentially refers to repeating conjunctive adverbs, with *repetition*. As shown in Figure 4, components in a disfluent region are identified and annotated following a fixed order. Using this algorithm, the three types of disfluencies reviewed above can be unified within more complex repair phenomena with a single structural representation, differing only in terms of what slots in the region are occupied.

Such a structural representation implies that a faithful description that regards filled pause and repetition as special cases of repair can also be naturally well justified. Therefore to avoid confusions in terminology, I will only refer to the repairs involving replacing the phrase in reparandum (RM) with repair (RR) as *repairs* in the following discussion. In other words, the repair phenomena reviewed here refer to *hybrid, substitution, insertion* and *deletion*, as listed in Shriberg's annotation scheme summarized in Figure 5, which includes instructions on what components to look for in determining the type of disfluencies.

In addition to structural representation of repairs, functional accounts have also been proposed to offer a categorization through examining the different cause of repair phenomena. In the theory brought up by Levelt (1983), five categories can be identified based on the reason of repair: **D-repair**: *refers to when original ongoing speech is aborted in ex-*

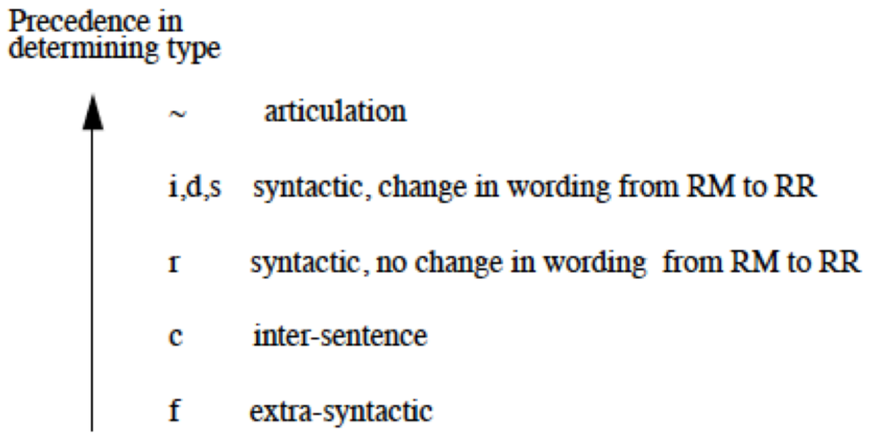
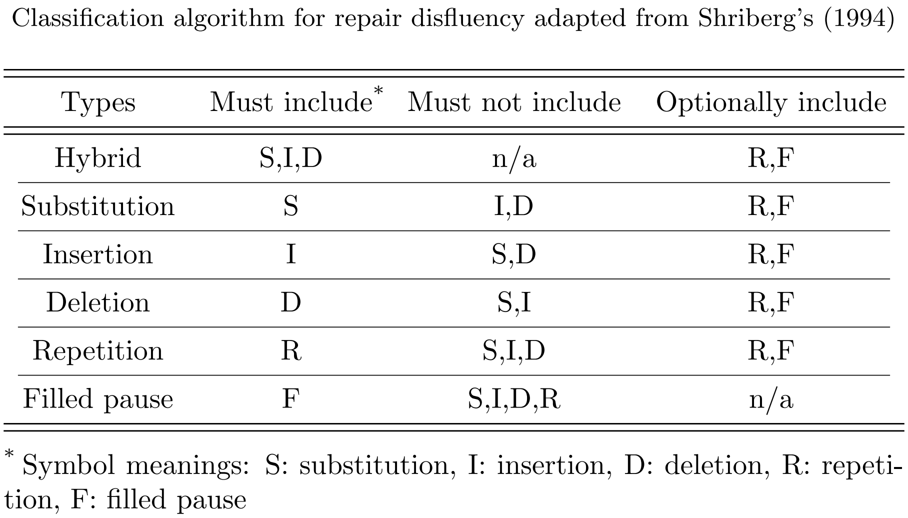

Precedence in determining type
~ articulation
i,d,s syntactic, change in wording from RM to RR
r syntactic, no change in wording from RM to RR
c inter-sentence
f extra-syntactic

Figure 4: The classification algorithm used in Shriberg (1994). f: filled pause, c: conjunction, r: repetition, i: insertion, s: substitution.

change for something different; **Appropriateness Repair**: *refers to when a speaker realizes that something in the speech is correct but needs to be modified for better communication*; **Error Repair**: *refers to the attempts to correct an error detected in the original speech*; **Covert Repair**: *refers to the repairs that are initiated before the error is produced, which result in repetitions of words right preceding the potential error site*. The last category is essentially all other repairs that don't fit the four established categories. According to Levelt, D-repair is equivalent to deletion. Appropriateness Repair can be initiated by the need to resolve ambiguity or offer further specification, while Error Repair can be further grouped into lexical errors, syntactic errors and phonological errors. Finally, Covert Repair may be the course to correct appropriateness or errors at the conceptual or planning stages.

However, this functional classification is highly subjective as it relies on judgements of speaker's model of listener's knowledge in the discourse (Shriberg, 1994). This judgement may inevitably lead to confusions in assigning classification labels. Blacfkmer and Mitton (1991) used a simpler functional based classification system, where only two major categories: *conceptual* and *production* based repairs are distinguished. However, this approach



Classification algorithm for repair disfluency adapted from Shriberg's (1994)

Types	Must include*	Must not include	Optionally include
Hybrid	S,I,D	n/a	R,F
Substitution	S	I,D	R,F
Insertion	I	S,D	R,F
Deletion	D	S,I	R,F
Repetition	R	S,I,D	R,F
Filled pause	F	S,I,D,R	n/a

* Symbol meanings: S: substitution, I: insertion, D: deletion, R: repetition, F: filled pause

Figure 5: The classification scheme used in Shriberg's algorithm for repair annotation.

by design is also not able to address the inherent problem of subjectivity in functional classification.

Through comparing the distribution of *deletion* and *repetition* in Switchboard, Shriberg (1994, 2001) found that speakers can be grouped into *repeaters* and *deleters* by the strategies they adopt in coping with the cognitive demands in talking while planning. She further argued that the possibility of relating this apparent strategic difference to cognitive processes in planning and production finds support from the prosody: repeaters have slower speech rate than deleters. This speed difference may be a reflection of the difference in the underlying processing speed between the two groups. However, since her primary intention was to offer a theory neutral description of disfluencies, the main take-away of this observation should be a stress on the significance of careful type descriptions of repair and repetition.

An extensive study of all the repair phenomena requires substantial effort in creating carefully annotated corpus. However, this requirement poses a major constraint on disfluency research in general. Automated annotation of disfluencies, especially repairs, is not impossible, but the performance of automatic systems is highly contingent upon the domain of annotated data that is used for training. By far, Switchboard is still the standard and primary source of annotated corpus for systems in disfluency annotation, such as Hough (2014), although advances in machine learning and natural language processing have been tremendous since 1992. A more extensive careful description of not only repairs, but disfluencies broadly, can help to push forward the efficiency of semi-automatic annotation, and benefit speech technology community by complementing the perspective that algorithmic advances may never catch. Such a description should be based on a sample with at least more than 20 speakers.

2.1.6. Summary

In this section, I reviewed the surface variations of the form of disfluent speech in normally fluent speakers. As discussed above, these categories are not independent from each other, and the nature of different disfluency categories determines the burden on researchers to identify and properly annotate the speech transcripts. This unfortunately limits the scope of analyses that can be efficiently performed on the categories that are harder to identify and correctly annotate. This limitation constraints researchers to conduct more extensive careful descriptions of the distribution and patterning of disfluency phenomena, especially repetition and repair disfluencies. Therefore in this study, I will combine automatic identification and extraction with semi-automatic manual annotation based on disfluency types. The primary focus of discussion will be on silent pause, filled pause and repetition. Due to its complexity in surface form variation, repairs will only be discussed in connection to the three disfluency types focused here. I will leave a detailed large scale description of repair disfluency for a future project.

2.2. Speech production and disfluencies

One important theoretical interest of the study of speech disfluencies is their close connection with speech production models. Speech production is a remarkably complex process. At some higher level, it requires the translation from abstract ideas to well-structured linguistic representation, which is then converted into articulation plans that are sent to articulators. The full process involves rapid yet precise coordination between message formulation, utterance planning and motor planning before articulation. The articulation process itself requires convoluted yet fast and accurate coordination between around 100 structurally and functionally distinct muscles for fluent delivery of speech (Kent, 2004). Literature on speech production can be roughly divided into two perspectives: From the psycholinguistic perspective, mainstream models assume a hierarchical structure for the production process and provide accounts from the abstract process of message formulation to the phonetic planning for articulation (Levelt, 1983, 1989). From the perspective of neural science, the interest focus on motor planning and execution, and how they are supported by the potential neural substrates behind the modeled processes (Guenther, 2006; Tourville and Guenther, 2011; Hickok, 2012). Recent proposals for the integration of a dual stream mechanism (Hickok and Poeppel, 2007), such as the Hierarchical State Feedback Control (HSFC) model (Hickok, 2012), offer a potential uniformed framework for the modeling of both higher level abstract processing and speech motor planning and control. Attempts to integrate the psycholinguistic, neurolinguistic and sensorimotor framework for speech production have hence been made (Hickok, 2014; Walker and Hickok, 2016).

Speech disfluencies are an integral part of the studies of speech production. Given the complex nature of the task of producing fluent speech, it is conceivable that break downs do happen. These glitches during speech production may result in disruptions or disfluencies during the delivery of speech. Speech production models from both perspectives offer theoretical and empirical groundings for the understanding of why and how speech disfluencies happen (Garrett, 1980; Levelt, 1983, 1989). On the reverse side, disfluencies have provided invaluable information for understanding how a speaker manages the flow of speech in varying conversational or cognitive contexts, especially when the context perturbs a smooth execution of a formulated utterance plan (Levelt, 1983; Clark and Wasow, 1998). Therefore speech disfluencies constitute an indispensable aspects in the study of speech productions, and a theory on speech disfluencies should also be grounded in the broader picture of understanding the full speech production process. In the following review, I outline the major design features for the psycholinguistic and neural models for speech production. From this review, it isn't hard to see how such models can be used to explain disruptions in spontaneous speech, and how speech disfluencies may offer insights for the further development of the theories.

2.2.1. Psycholinguistic models for speech production

Among psycholinguistic models of speech production, a predominant view is that the production process is hierarchically ordered. The hierarchical structure of the production process consists of stages such as message formulation, semantic planning, lexical selection, syntactic planning, phonological planning and phonetic planning and motor planning and execution. A widely accepted theory following this framework is Levelt (1989), as illustrated in Figure 6. In this model, speech planning proceeds in a stage-wise incremental fashion, where the plan of some planning unit from the previous level or stage of processing can either be formulated and passed down to the following stages or lower levels of processing, or formulated in parallel with other levels with incomplete formation. Although the architecture for such models are quite elaborated, it doesn't clearly address the timing or the window size within which planning is carried out. Rather, studies following this set up often treat the planning unit as given and static, ignoring the dynamic and sequential nature of human speech. Therefore later development of the theory placed more focus on lexical access and planning (Levelt et al., 1999).

Levelt's model regards disfluencies as production errors that are identified and corrected through a monitoring system. The monitoring system is enabled through an explicit speech monitor as shown in the graphical illustration of the model shown in Figure 6. This additional component serves as an error detection and correction machinery that responds to production errors and triggers the correction process or replanning. Although there isn't an agreement on the exact mechanism of how such a feedback loop might work, proposals have been made considering either an overt auditory feedback or some combination of auditory and internal feedback as a possibility (Postma, 2000). The proposed speech monitors can be grouped into three categories: intrinsic, response and external feedback (Borden, 1979). Among these proposals, as summarized in Postma (2000), a total of 11 logically possible monitors have been proposed, which are able to detect errors in preverbal message (monitor 1), responsible for detecting errors at different stages of production (monitor 2-4,6-9), or require additional assistance from the audition and speech comprehension (monitor 5, 10, 11). as depicted in Figure 6. Intrinsic feedback responds to internal feedback prior to movement, while response feedback operates on the motor output. External feedback relied on the auditory loop to perform monitoring.

The theoretical basis for the monitoring mechanism can be grouped into three main flavors: The perceptual loop theory (Levelt, 1983, 1989), production based monitoring (Laver, 1973, 1980; Schlenck et al., 1987) and the node structure theory (MacKay, 2012, 1992a,b). They differ in their assumptions about the accessibility of processing components in the production flow, and whether they try to find that the monitors function as a corrective role or are restricted to tuning and directive controlling of speech motor execution feedback. In perceptual loop theory, only certain end-products in the production process are monitored, in the same way as utterances even though some are considered in the inner loop. The

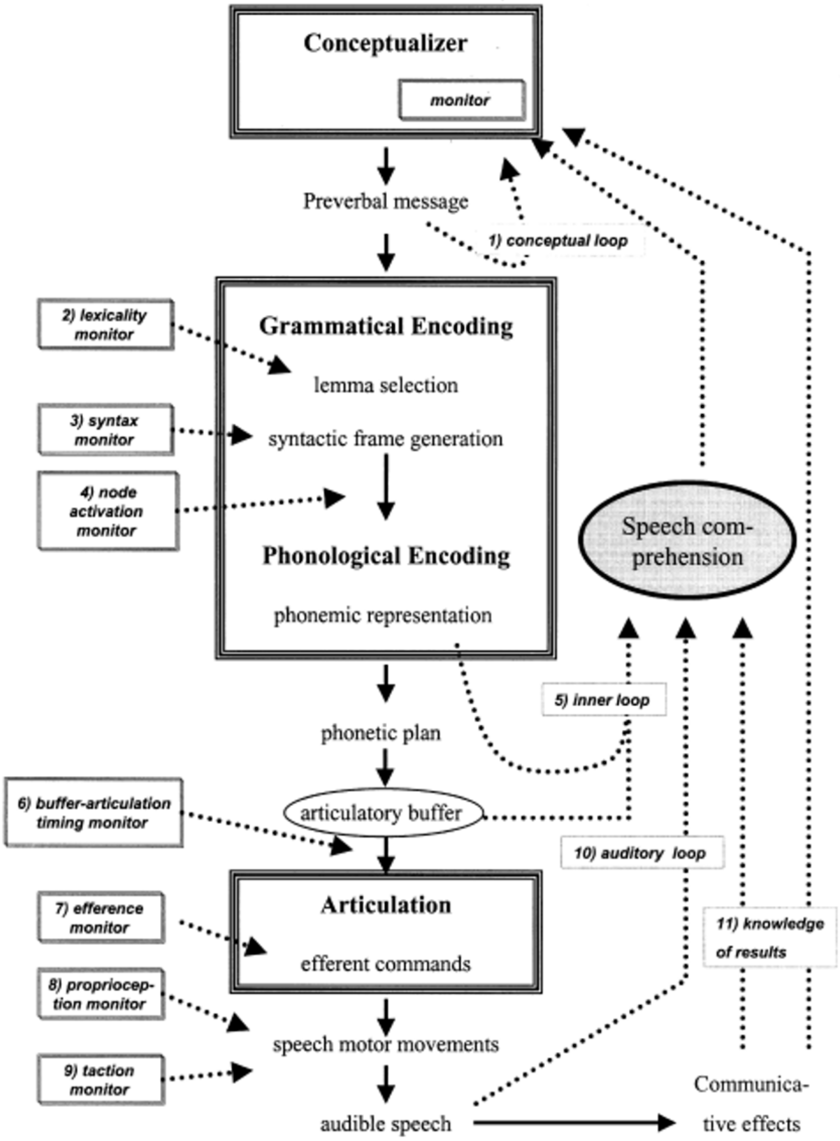


Figure 6: Schematic representation of Levelt's speech production model, with proposals for speech monitors outlined. The picture is taken from Postma (2000).

production based view holds that speakers have direct access to the processing components during production, so that components inside the formulator are accessible for monitoring. These monitors correspond to monitor 2 through 4, as well as 6 through 9 in Figure 6. Finally, the node structure theory accounts for monitoring by prolonged activation of uncommitted units, which are the novel components in the flow of production that are yet to be committed. This proposal doesn't require a separate structure of monitor. The prolonged activation automatically leads to error detection.

The diversity of speech monitors and monitoring theories proposed in the literature actually reflects the fact that speech errors are multitude: different surface self-repairs can be linked to disruptions at different levels in the production process. An overall consensus is that certain repair types, such as appropriateness repair, deletion and error repair (Levelt, 1983) are related to higher level processing including message reformulation. However, disagreement often arises with regard to the so-called covert repair (Levelt, 1983), where the delay or interruption in fluent speech delivery is minimal (Blacfkmer and Mitton, 1991). Thus theoretically these repairs do not necessarily need to trigger the monitoring mechanism. Prosodic cues such as the pitch contour and pause duration can offer further information in relating covert repair to particular monitoring mechanisms (Nakatani and Hirschberg, 1994; Levelt and Cutler, 1983).

2.2.2. Production models that focus on motor planning and control

The motor control system for speech production has independently received a lot of attention in neural science. The fundamental question in this stream of work is how the structure of the cerebrum is related to aspects of the physical movements of articulators that directly produce speech signal. One popular direction of research is to build physiologically plausible models through simulation and empirical experimentation. These physiologically plausible neural network models for speech production (Guenther, 2006; Bohland et al., 2010;

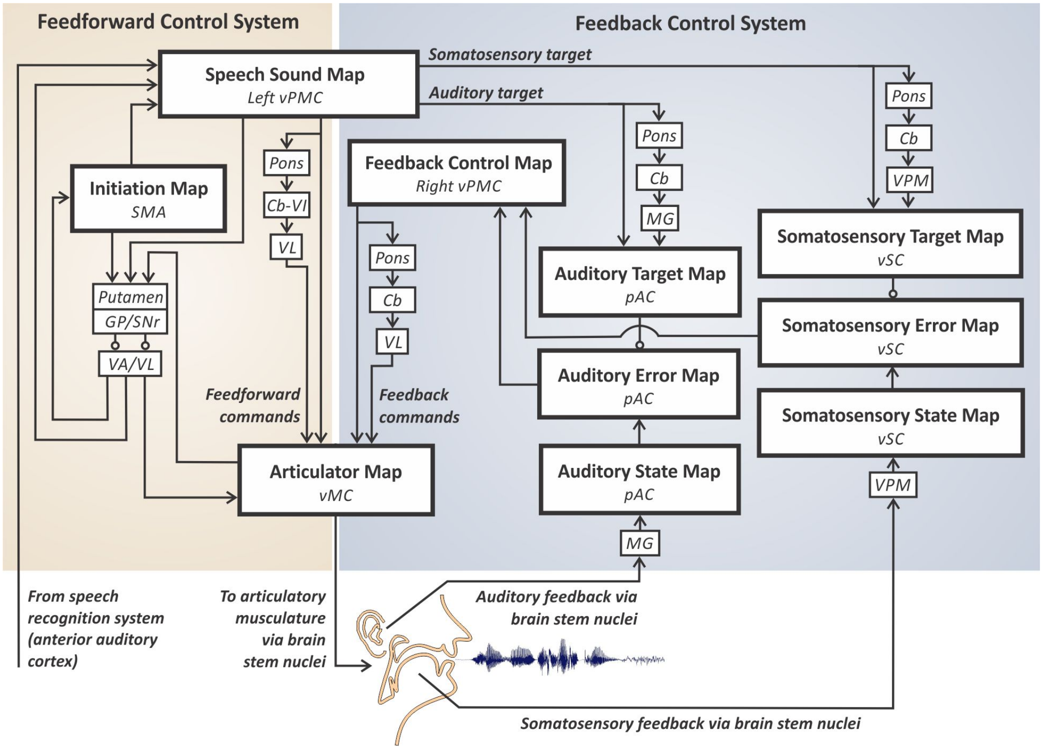


Figure 7: Schematic representation of the DIVA model. The image is taken from Guenther (2016).

Hickok, 2012) are able to provide detailed account on the motor planning and execution processes behind speech production. These models seek to find an abstract representation of the motor planning and execution process underneath speech production that can be empirically tested and refined. Supports for hypotheses made from such models have been found both in neural imaging and computer simulation studies.

One recent proposal in this tradition is the Direction Into Velocity of Articulators (DIVA) model proposed by Guenther (2006) and refined through later studies (Tourville and Guenther, 2011; Guenther and Vladusich, 2012). The schematic representation of the DIVA model is shown in Figure 7. In the graph, each block corresponds to a hypothesized neuron map resides in a particular region of the brain. The goal of the model is to learn a set of maps that eventually establish the connection between the phonological representation of speech sounds and articulator movements. The feedback control system examines the

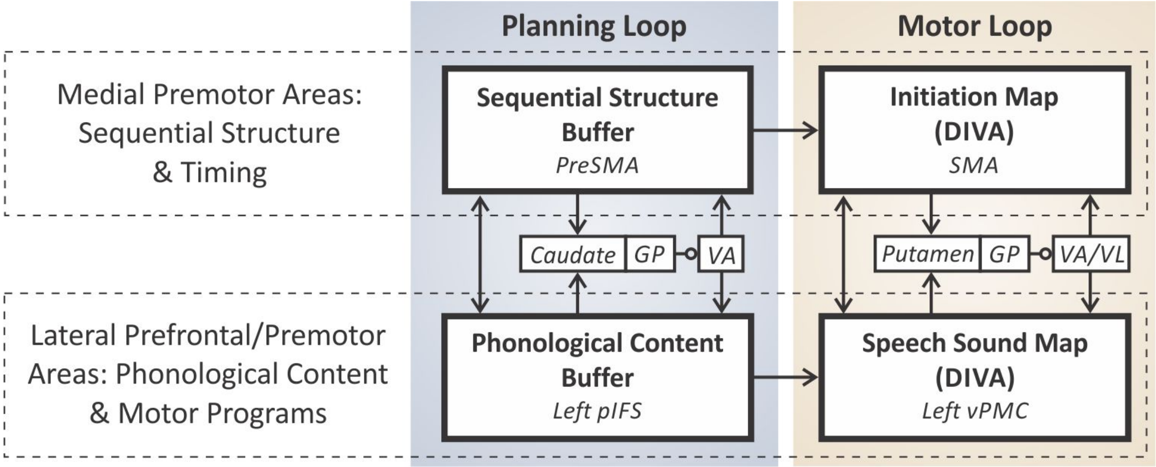

Planning Loop
Motor Loop
Medial Premotor Areas: Sequential Structure & Timing
Lateral Prefrontal/Premotor Areas: Phonological Content & Motor Programs
Sequential Structure Buffer
PreSMA
Initiation Map (DIVA)
SMA
Caudate
GP
VA
Putamen
GP
VA/VL
Phonological Content Buffer
Left pIFS
Speech Sound Map (DIVA)
Left vPMC

Figure 8: Schematic representation of the GODIVA model. The image is taken from Guenther (2016).

outputs at various stage in the forward control system and sends correction signals when necessary. Through simulation studies, this model is able to learn the movement control in the simulated environment to produce speech sounds. Theoretically, this model provides a basis for exploring localized functional correspondence in the brain to explain questions such as contextual variability and coarticulation.

Following the framework proposed in DIVA, the Gradient Ordered DVIA (GODIVA) model (Bohland et al., 2010) specifically details a mechanism for the sequencing of output phonetic sequence. The major contribution of GODIVA is its ability to connect an utterance plan to the motor program that is responsible for the production of phonetic realizations. This connection is established through the parallel planning and motor loop in the motor control system. An utterance plan is passed down to a motor plan through a competitive queuing mechanism, in which the activation of a phoneme is inhibited after firing. The schematic representation of the GODIVA model can be found in Figure 8.

An extension to the framework established by the DIVA model is the HSFC model as described in Hickok and Poeppel (2007) and Hickok (2012), with two major additions: it utilizes the ventral stream to give an explicit account on the somatosensory feedback, and it introduces the potential to extend the dual stream mechanism to connecting to higher level speech planning. In theory, the somatosensory feedback could potentially offer an elegant solution to certain covert repairs due to the relaxed processing time constraint. A yet more attractive property of the dual-stream model, as mentioned in Hickok and Poeppel (2007), is its potential in uniting the psycholinguistic and neurolinguistic traditions in modeling the speech production process.

Although not explicitly drawing connections to disfluencies in speech, neural network models could serve as a nice starting point to build a theory for the disfluencies that psycholinguistic models fell short in explaining. More specifically, these models can be applied or modified to account for disfluencies that do not overtly involve hesitation or error correction from higher level processing units, or do not allow sufficient time lapse for a full execution of the feedback loop as modeled in models involving speech monitors. In addition, details in these neural network models shed light on potential theories for the transmission of information or commands between levels of planning, and the generic role of inhibition in controlling the fluent delivery of speech. Because one motivation for these neural network models is to identify neural correlates of the proposed network structure, direct applications of these models in explaining neurological deficit of stuttering has been proposed, such as in Civier et al. (2010).

To sum up, neural network models for speech motor control provide theoretical groundings for both theoretical and experimental studies on the motor plan and execution in speech production. A potential advantage of such neural network models for the motor control process in speech production is the possibility to integrate higher level processing with lower level articulation plan. This ability can be helpful in dealing with speech repairs or disfluencies that cannot neatly fit into the predefined disfluency or speech repair categories in psycholinguistic literature. Success has also been made in applying these models to explain behaviors observed in clinical conditions such as stuttering.

2.3. Disfluencies under impairment

A practical implication of speech disfluencies is their potential in distinguishing speakers under various neurological or cognitive impairments from healthy population. Some use cases may include people under the influence of alcohol intoxication and fatigue (Dawson and Reid, 1997), as well as identifying patients with neurodegenerative diseases (Ash et al., 2009). The analysis of speech features such as disfluencies and their associated acoustic or prosodic properties hence becomes an area that has received increased attention (Cummins et al., 2018; Schuller et al., 2014). In this section, I would like to provide a brief overview of speech disfluencies in patients with neurodegenerative diseases, highlighting the prospects of the current research in improving health-related applications.

Many neurodegenerative diseases are directly related to degeneration in the brain regions that control human cognitive ability including language. The associated language deficits provide invaluable information for the identification and diagnosis for such diseases. For example, syndromes related to frontotemporal degeneration (FTD) include variants of progressive, selective language disorder (Boschi et al., 2017). Four (Gorno-Tempini et al., 2011; Rascovsky et al., 2007) phenotypes of this progressive primary aphasia: Non-fluent variant (naPPA), semantic variant (svPPA), logopenic varient (lvPPA) and behavioral variant (bvFTD), are currently diagnosed in consultation with their associated linguistic deficit (Ash et al., 2013). However, a consensus on the diagnosis criteria is not completely reached (Rascovsky et al., 2007). Although only being one of the cognitive domains that are impaired due to neurodegeneration, early stages of the Alzheimer's disease (AD) may also show symptoms of language impairment (Szatloczki et al., 2015).

Improving the efficiency of the identification and diagnosis of neurodegenerative diseases, especially at their early or predromal stage, has become a focus in the past decades. Research in this area generally takes two approaches: scientific investigations with a focus on understanding the correlates between measurable linguistic features and corresponding phenotype (Ash et al., 2009, 2010, 2012, 2013; Mack et al., 2015), and engineering effort to build systems for the classification problem (Pakhomov et al., 2010; Budhkar and Rudzicz, 2018). As will be reviewed later, the linguistic characteristics of speech produced under these clinical conditions are not mutually exclusive across the syndromes. Accurate diagnosis often requires careful manual analysis of the linguistic features based on clinician's own experience and judgements. These constraints nevertheless greatly impede the efficiency of clinical diagnosis at the cost of increased financial burden. Thus the goal for both research efforts is to identify and utilize the linguistic information obtained from patients' speech to guide the diagnosis or reduce the burden on human clinicians in processing the linguistic information they collect during a clinic visit. A commonly used method for elicitation of the speech is a picture description task using the Cookie Theft of the Boston Diagnostic Aphasia Examination (Goodglass et al., 2000). This task comes as a component of the standard cognitive assessment procedure, which generates about one minute of monologue from the patients.

The linguistic impairments can surface as deficits in the phonetic, phonological, semantic and morphosyntactic problems in the speech produced by FTD patients. These deficits may surface as disfluent speech compared to cognitively healthy speakers. Features that capture the variation in such surface disfluency would be informative in distinguishing different speaker groups, especially for cases that hard to tell by un-trained humans. Variation in the observed disfluencies presumably corresponds to the brain regions that are responsible for different processing tasks in speech production. Measurements such as speech rate, fundamental frequency, word distribution, well-formedness of sentences and discourse structure have been shown to be effective in drawing distinctions between the target and control group (Pakhomov et al., 2010; Mack et al., 2015). For example, among the four phenotypes of progressive primary aphasia, lower speech rate, as measured by the number of words per minute, errors with closed-class nouns and impoverished syntactic complexity (Grossman and Ash, 2004) are found to be associated with patients with naPPA. Patients with svPPA have been reported to have problems retrieving nouns (Ash et al., 2013), resulting in replacement with pronouns and simpler or more generic nouns. Reduced syntactic complexity among these patients and higher speech repair rate are also documented in the literature (Grossman and Ash, 2004; Ash et al., 2009). Reduced speech rate and higher rate of phonetic errors, together with increased speech repairs are reported to be the predominant linguistic deficit for patients with lvPPA (Ash et al., 2013). On the other hand, linguistic deficit in patients with bvFTD tend to show disorders at higher discourse level, rather than phonology and morphosyntax, compared to other FTD variants (Pakhomov et al., 2010; Grossman and Ash, 2004). For AD patients, more lexical errors, word finding difficulty, over use of indefinite terms, high frequency of repair and repetitions, as well as neologism are among the reported linguistic deficits (Boschi et al., 2017). As for Amnestic Mild Cognitive Impairment (aMCI, the predromal for AD) patients, their speech appears to be less pragmatically coherent compared to healthy controls, but less severe than AD patients (Ahmed et al., 2013).

Linguistic features reviewed above and their correspondence to types of neurological disorders are engineered to capture the phonetic, phonological, semantic and morphosyntactic properties of disordered speech. More focus has been placed in the task of MCI and AD detection (Pakhomov et al., 2010; Fraser and Hirst, 2016). Depending on the exact problem formulation, the classifiers achieved somewhere around 80% accuracy or F1 score in identifying target patients. More recently, integration of modern NLP technologies in the domain of speech based diagnosis assistance system is gaining more attention in the research community (Budhkar and Rudzicz, 2018; Nevler et al., 2017). However, with several limitations shared across these automation studies, the significance of the reported model performance should be interpreted with a grain of salt.

The primary limitation comes in the form of high selection bias in the construction of training samples. This selection bias comes into play both in terms of the limited number of selected patients as classification target and the lack of both quantity and diversity of patients' speech. The problem of small sample size and lack of diversity in the speech sample prohibit generalizations to broader use case of the developed systems. Given the limited access to a more representative sample for the population distribution of FTD patients, it is even harder to control for the potential confounders that are related to the progression of the disease: especially patients in later stages of the disease are potentially over represented due to the difficulty in the identification and diagnosis at earlier stages of the disease. Therefore it is unclear how much contribution is made with these automated systems to the identification and diagnosis at different stages of the disease. Unfortunately, studies generally do not disclose the distribution of the severity of symptoms in the patient group. It therefore has to be acknowledged that research in intelligent system for the detection of neurological disorder is still in its infancy.

2.4. Chapter summary

In this chapter, I reviewed the literature on speech disfluencies from three distinctive yet interrelated perspectives: Studies that are focused on describing and classifying different disfluency phenomena, that connect speech disfluencies with theories of speech production, and that investigate neurodegenerative diseases which draws close connection to disfluencies in spontaneous speech. Efforts in providing robust and reliable descriptions of different disfluency phenomena, as well as exploring correlates that potentially explains the observed variation, are crucial in both theoretical and practical sense. It is hoped that the current descriptive work could serve as a solid foundation for future theoretical and applied research in areas involving an understanding of speech disfluencies.

Chapter 3

The Variation of Silent and Filled Pauses

This chapter describes the variation of silent and filled pauses. I examine both sociolinguistic variables such as age, gender and dialects, and potential covarying contextual factors including speakers' cognitive state and the broader discourse of the utterance. I take as response variables silent and filled pauses at their face value, that is, measured by certain quantitatively defined value or count statistics based on the transcript. At the center of the discussion is how other measurable features, such as sociolinguistic and contextual features, can jointly explain the variation observed in the response variable. The discussion unfolds as the following. I first review the literature on the effect of sociolinguistic and discourse variables on silent and filled pauses. The actual discussion proceeds with two paths in parallel: reproducing the observations in the literature, and discovering new relations with speech from unexplored domains or feature representations. I will show that claims made about filled pauses in the literature are only partially correct, and provide detailed documentation of how the proposed feature space interact with the distribution of silent and filled pauses.

3.1. Background

The effect of sociolinguistic variables, such as age, gender and dialectal variation has received relatively little attention in the early literature on speech disfluencies. Although discussions on various issues related to disfluency phenomena can be traced back to Maclay and Osgood (1959), it is not until Levelt (1983) and Shriberg (1994, 2001) that reported gender difference in disfluency rate distribution in both a sample of six Dutch speakers and Switchboard. However, the question of how sociolinguistic variables affect disfluent patterns was not systematically investigated until Bortfeld et al. (2001). More research efforts have been directed to this topic in the past decade, with more comprehensive comparisons of the use of filled pauses across gender, age and socioeconomic groups (Tottie, 2011; Acton, 2011), as well as in different English varieties (Tottie, 2014; Kendall, 2009). In addition to age and gender, Laserna et al. (2014) also considered personality as a potential informative variable. The use of filled pause has also been treated as a sociolinguistic variable (Fruehwald, 2016) which in itself is proposed to be a language change in progress. A trade-off between the frequency of *um* and *uh* has been observed. This view is also upheld by a later study (Wieling et al., 2016), where the same trend appears to persist across several Germanic languages. Sociolinguistic variables are also examined in Yuan et al. (2016) in Mandarin, where gender effect was also been reported. However, even fewer studies have looked at other disfluency phenomena such as silent pause and include individual variation as an articulated research question. Among them, Kendall (2009) studied the distribution of silent pauses across dialect region in North America, with the goal of attaching social meanings to silent pause variation. Roberts et al. (2009) focused on repetitions in fluent male adults for the better treatment of stuttering adults. McDougall and Duckworth (2017) investigated individual variation of prolongation and repetition, in addition to silent and filled pauses, for speaker identification in forensic settings.

Bortfeld et al. (2001) approached the problem through a language production experiment, in which pairs of speakers were asked to pair sets of pictures through a mixed factorial design. The factors that were controlled for include familiarity of the picture, age, gender, education, and marriage situation. They analyzed both the overall disfluency rate and several individual disfluency categories. In terms of overall disfluency rate, more disfluencies have been found in more demanding planning tasks, such as unfamiliar domains and the role taken by the participant in the picture matching task. With regard to each disfluency category, there was a change in the predominant disfluency pattern conditioned on familiarity and roles taken in the task. They also reported that older speakers produced more fillers than younger speakers, and men had higher disfluency rate overall.

Bortfeld et al.'s study did suggest that sociolinguistic variables constitute a crucial component in determining how surface disfluencies in individual's speech may vary. However, some factors in their experiment, such as familiarity of the pictures and the role played by individual speakers, reflect less on the effects they called cognitive demand, especially if one wants to generate their results from lab speech to more general settings. One apparent example is that they used geometric shapes in their "unfamiliar" category, while actual kids in their "familiar" category. From their description of the task, it is unclear whether experiment participants' perception aligned with their specified conditions, and how this set up is paired with cognitive demand.

Among corpus studies of sociolinguistic variables' effect on disfluencies, the most heavily discussed topic is the variation in response to gender and age in the use of filled pause. Filled pauses in these studies are distinguished between the one with a nasal coda, transcribed as "um" or "urm", and the one that without ("er" or "uh"). In some earlier work, it has been acknowledged that *er* is the second most characteristic word for male speakers and fourth most characteristic word for speakers who are 35 years of age or older (Rayson et al., 1997), through looking at British National Corpus. *Erm*, on the other hand, is among the most characteristic words for people from higher socioeconomic classes. Using a collection of corpora of telephone conversations (including Switchboard and Fisher), Liberman (2005) observed that *uh* was used more frequently among male and older speakers, whereas *um* was more frequent among female and younger speakers.

Several more recent corpus studies have dedicated to discovering the effect of sociolinguistic variables in disfluency production. Tottie (2011) compared between two British corpora: the British National Corpus (BNC) and London-Lund Corpus (LLC), and across multiple speech styles and speaker groups therein. Although the statistics reported in this study is exploratory in nature, it points out several directions for future exploration. In addition to the observations that men use more fillers than female and a tendency for higher filler frequency among older speakers, she also raises the question of what is the socioeconomic status' effect on the use of fillers. Through comparing the socioeconomic stratification in BNC and LLC, she proposes that people with higher socioeconomic status tend to use more fillers as well. However, she doesn't further address how these factors interact, and stops at relating these sociolinguistic variables to the role of fillers in planning.

In a later study, Tottie (2014) compared the use of fillers between American and British English. She showed that in her sample of American English, the Santa Barbra of Spoken English Corpus, male speakers do not maintain a higher rate of fillers compared to female, inconsistent with the trend she reported for British English. She also argues for the existence of similarity between the distribution of *um* and *uh* and discourse markers, such as *well*, *you know* in her sample of American English, thus an evidence for the different discourse functions played by fillers in two varieties of English. Conversation topic and formality are also suggested to be influencing factors in filler production through this comparison. Although the variables she proposed can in principle be useful in identifying the distributional variation in the use of fillers across population groups, the corpora of her choice may not be optimal to offer unbiased claims about these factors. One fundamental problem is that the Santa Barbra corpus is much smaller in scale and sampled a different demographic group compared to BNC and LLC, even though all three corpora she used consist of conversational speech.

Acton (2011), on the other hand, documents the gender difference in *um* and *uh* distribution in American English from two spontaneous speech corpora: A speed-dating corpus collected from graduate students at an American university, and the Switchboard corpus.

In both corpora, he found a higher rate of *um* among female speakers than male. Both Tottie and Acton's work suggest that *um* is gaining currency in their respective speaker population, and Acton further proposes that this change is persistent across gender and age group.

Fruehwald (2016) and Wieling et al. (2016) further explored the initial observation of a potential change in progress in the use of *um* and *uh*, as documented in Tottie (2011) and Acton (2011). Fruehwald (2016) examined the frequency of *um* and *uh* by gender and age group using the Philadelphia Neighborhood Corpus (PNC). His apparent-time analysis shows that there is an apparent increase in popularity of *um* among younger generations in the past century, and female speakers appear to lead the change. This increase in popularity, however, is accompanied by a decrease in the relative frequency of *uh*, thus showing a trade-off between the two variants of fillers. This trend is clearly seen in Figure 9. Expanding upon Fruehwald (2016), Wieling et al. (2016) further looked into a range of Germanic languages, including English (British and American varieties), Dutch, German, Norwegian, Danish, and Faroese using a variety of spoken and written corpora. The selected languages in this study all have a similar binary contrast of filled pause (one with a nasal coda and one without). The trade-off of frequency was shown through a mixed-effect logistic regression model (for a detailed list of corpora, see Wieling et al. (2016)). In all the spoken corpora they examined, significant effect of age and gender on the likelihood ratio of *um* vs. *uh* has been found. However, great variation in terms of overall proportion of *um* and *uh* use is also observed across the selected corpora, even within a same language. This variation may be due to factors such as topic and domain variation in conversation, individual variation, dialectal variation, and some other unobservable endogenous variables. Two tentative explanations from the perspective of language contact and a potential extra-linguistic force that enables an independent yet parallel change have been proposed to account for this interesting change in progress.

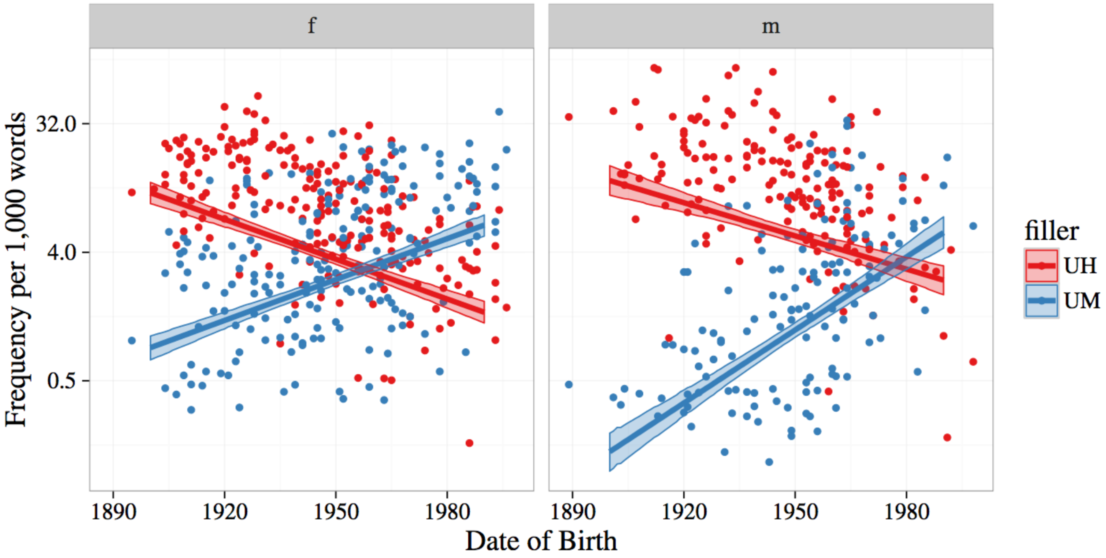


Figure 9: UM usage and UH usage trading in frequency in PNC, from Fruehwald (2016).

Compared to the extensive research on the variation and production of filled pauses, fewer studies have focused on the effect of individual variation and topic on silent pause distribution and other disfluency phenomena. Variation in silent pause has received some attention in earlier studies. The range of individual variation in silent pause distribution was reported as early as in (Goldman-Eisler, 1968), both in spontaneous and read speech. Duez (1982) compared both silent pauses and other disfluent non-silent pauses of French across three speaking styles: political interview, casual interview and political speech. Each speaking style contains 5 to 7 speakers, with average speech time around 30 minutes. It is found that silent pauses are longer and more frequent in both political and casual interviews compared to political speech. A wide range of individual difference is also noticed.

Studies reviewed thus far have all demonstrated that age, gender and other socioeconomic factors such as socioeconomic status and education potentially have an effect on the distribution of filled pauses, especially the relative frequency of the use of *um* and *uh*. This trend is persistent both across varieties of English and across the Germanic family.

However, as pointed out by several authors (Tottie, 2011; Acton, 2011), there exists a great potential for individual variability. The extent of this variability is less understood. In addition, our understanding of other discourse or extra-linguistic factors such as conversation topic is still rather limited. These lesser explored factors may be the underlying variables that explain the variation contributed by age and gender, through which we can connect sociolinguistic and cognitive factors that shape human language production. Such a comprehensive approach would build our knowledge towards a causal inference on why the surface variations are the way they are.

In the following sections, I address the two aforementioned questions: What is the range/effect of individual variation in the distribution of silent and filled pauses, and what is the role played by conversation topic? I will examine the variation in silent and filled pauses using Fisher corpus, focusing on the effect of conversation topic and the consistency within individual speakers across conversations. I will also present a case study on the variation of silent pause distribution induced by alcohol intoxication, using data from the Alcohol Language Corpus (Schiel et al., 2008).

3.2. Silent pause

In this section, I address the question of what is the individual variation in silent pause distribution, and how silent pauses are affected by conversation topic. The Individual variation addressed here mainly refers to how likely it is for an individual speaker to vary in their pausing patterns across conversations. I first raise the fundamental problem of defining silent pauses in spontaneous speech, then I proceed with the analysis using a redefined objective quantification of silence in speech.

3.2.1. An objective and robust representation of silence segments in speech

A challenge in silent pause research that concerns the very fundamental definition of silent pause is what is the appropriate threshold for separating true silent pauses that relate to hesitation or processing problems encountered during production from the silence that is the result of phonological or prosodic processes. As reviewed in previous sections, various thresholds have been used in the literature, ranging from 80 milliseconds to 2 seconds (Ferreira, 2007). Several studies have provided descriptive analyses of silent pause distribution across speaking styles (Zellner, 1994) and languages (Campione and Véronis, 2002). The findings on the question about proper selection of silence threshold are somewhat inconclusive, due to large amount of context-dependent variation and the predominant subjective judgement on what is a silent pause. However, it is generally acknowledged that the duration distribution of silence in spontaneous speech is bimodal or multimodal, and a threshold of 200 ms can serve as a sufficient cut-off point for most purposes. In a large-scale multilingual study of pause distribution in both spontaneous and read speech, Campione and Véronis (2002) showed that the distribution of pause duration bears language-specific traits, and the choice of threshold can subjectively change the result of statistic analyses. Faced with such a wide range of variability, one might still prefer subjective judgement as the better practice in determining whether a silent segment constitutes a silent pause (Lickley, 2015; Nakatani and Hirschberg, 1994; Eklund, 2004). Nevertheless, there is the need for a more objective quantification of silent pause duration that is robust to contextual and individual variation, so that cross domain comparison can be made possible.

In this study, I rephrase the question of what is the absolute cut-off duration for categorically separating silence from speech as what is the relation between silence duration and the duration of speech segment preceding and/or following the silence. This rephrase can objectively and consistently quantify the dynamics of speech production that an absolute separation of silence from speech through a hard boundary cannot accommodate. This approach acknowledges that silence is an amalgam of linguistic, cognitive and extra-linguistic factors that reflecting both the syntactic and prosodic, and the cognitive perspectives in language production. The relative duration between silence and speech segments implicitly incorporates these multivariate space and simultaneously preserves the identity of pauses, while releasing the burden of selecting an appropriate threshold for particular purposes. On the contrary, an absolute hard cut-off point would unavoidably over or under estimate of the rate of silent pauses for individuals with varying speaking rate, and potentially misrepresenting discourse or structural pauses as hesitation pauses. Thus analyses of the duration or distributional relations of silent pauses following this threshold would be biased based on the exact context and threshold chosen.

The relation between silence duration and preceding or following speech distribution is explored through estimating the joint probability density of 2D (bi-gram silence duration plus the speech segment duration before or following silence) or 3D (considering speech-silence-speech sequence) duration space. This method is non-parametric and assumption-free, meaning that biases imposed by researchers or particular research questions can be largely eliminated. Under this set up, multiple assumptions can be tested by directly working with a probabilistic distribution, controlling for the parameters of interests. In this manner, group differences in silent pauses can be easily observed from the joint distribution of pause duration and speech duration before and after the pauses, and parameterized using dimensionality reduction methods on the joint distribution space. Thus a compact representation of the pausing pattern can be achieved for each individual. Statistics on group differences can then easily be applied.

Figure 10 and 11 plot the joint density estimation of silence duration and following speech duration for a year's worth of president's weekly radio address for Obama in 2010 and Bush in 2008 (Liberman, 2016). These plots clearly show some structure that is the

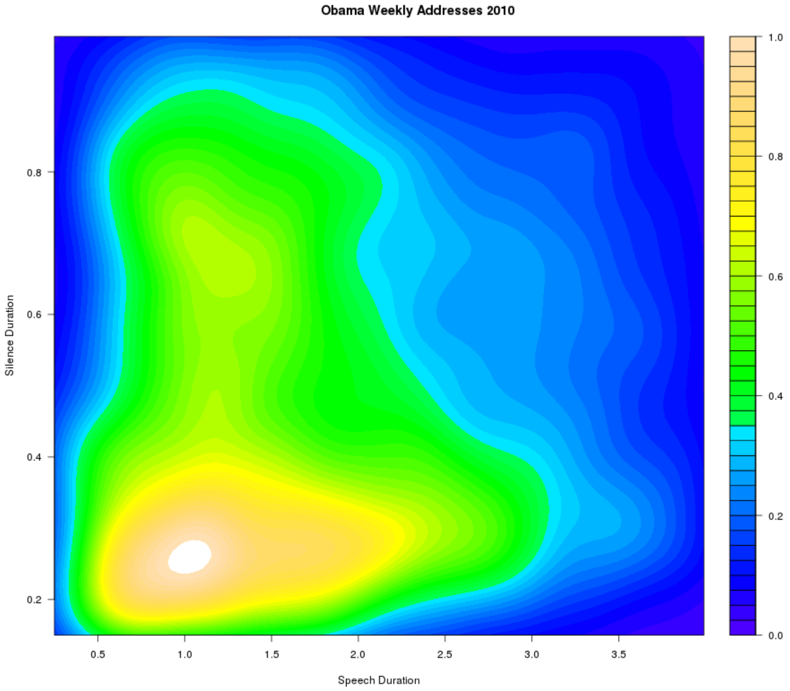


Figure 10: Joint density plots of silence duration (y-axis) and following speech duration (x-axis) in seconds of president weekly address for Obama.

result of individual variation in speaking style between the two former US presidents. For example, Obama's speech appears to have a peak at the coordinate around (1.0, 0.25), which suggests that his speech may be characterized by shorter speech segments between relatively short pauses. The secondary peak, at around (1.2, 0.7) may signal some longer pauses between paragraphs. Similar description can be made for Bush's speech, and a clear distinction between the speaking style of two presidents can be made basing off their distinctive patterns in pausing. Analysis of the relative duration between adjacent silence and speech segments in the other direction can be similarly carried out, so as the joint tri-gram speech-silence-speech duration.

With this simple demonstration, I have shown that a speaker's pausing characteristics can be captured by looking at the joint distribution of silence duration and following speech duration. This characterization can then be treated as the feature representation of individ-

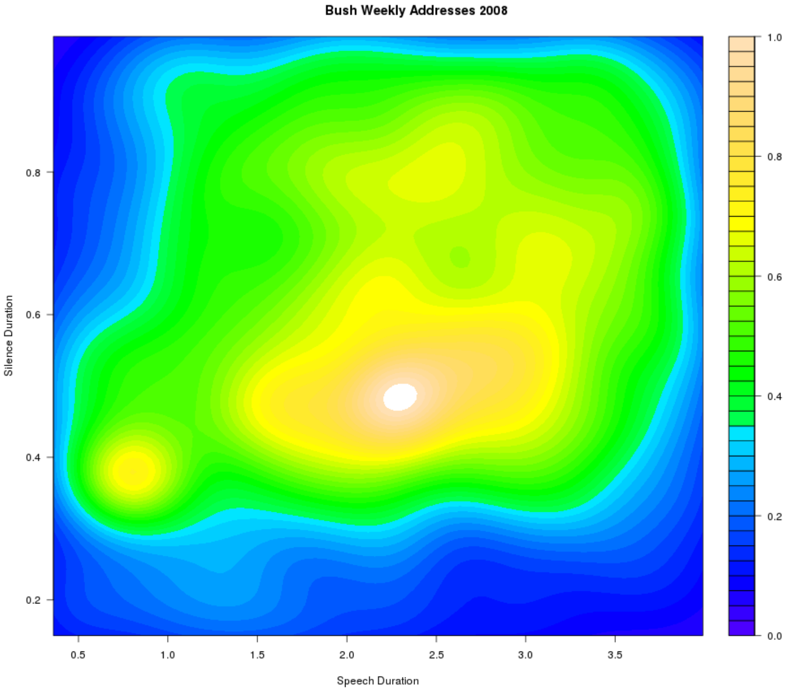


Figure 11: Joint density plots of silence duration (y-axis) and following speech duration (x-axis) in seconds of president weekly address for Bush.

ual speech. Therefore, we can perform some dimensionality reduction technique, such as Singular Value Decomposition (SVD), to achieve a compact representation of the information contained in these 2D density plots for each individual. Figure 12 is a joint density plot of the first two left singular vectors derived from over 5,000 read paragraphs in LibriSpeech (Panayotov et al., 2015). The input matrix to SVD is the flattened joint 100×100 2D density matrices of silence duration and following speech duration, obtained from a Speech Activity Detector (SAD) (Walker et al., 2015), for each read paragraph in the corpus. In this derived space, each combination of the values in the two latent dimensions represents a potential speaker in the population from which the initial sample is taken. The distribution in this derived space is clearly bi-modal, with a primary mode closer to the center of the graph, and a secondary mode towards the lower right corner. This bi-modal distribution could reflect two distinctive reading strategies that readers use when contributing to the

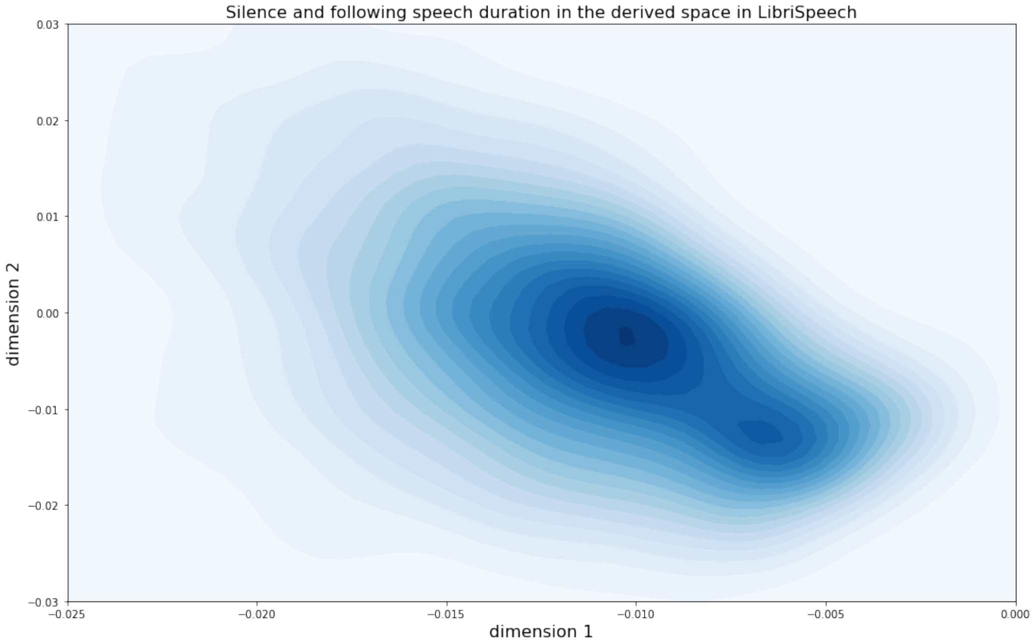


Figure 12: Plot of silence duration and following speech duration distribution of LibriSpeech in the derived 2D space.

corpus. This strategic difference may be the result of genre difference, gender difference, or whether the reader has received professional training. Thus, further explorations of the underlying explanatory factors can then be carried out.

To sum up, in this section, I proposed a more objective and non-parametric quantification of silent pause distribution in speech production. This quantification method first makes reference to the speech segment duration adjacent to the silent segment duration. The resulting joint density estimations are then passed to some dimensionality reduction method, here SVD, to achieve a compact representation for each individual in lower dimensional space. Through a simple demonstration with LibriSpeech and weekly presidential address, I have shown that this quantification method is effective in capturing individual variation, as well as the underlying structure in population distribution.

3.2.2. Individual variation in silent pause and the effect from conversation topic

Here, I use the quantification method illustrated in the previous section to address the question of individual variation in silent pause distribution. As discussed in the literature, both socioeconomic variables (Tottie, 2011; Acton, 2011) and conversation topic (Lickley, 2015; Bortfeld et al., 2001) may have an effect on the disfluencies in natural speech. The question then is how these variables affect silent pause distribution?

The data I use to answer this question is the sample from Fisher corpus. I first conduct an exploratory data analysis, with the goal to explore the in-sample group differences with regard to the socioeconomic variables reported in the literature, as well as conversation topic. Then I perform a regression analysis to test the observed group differences, and examine the potential interactions among explanatory variables. Individual variations are represented in the derived 2D space generated from the joint density estimation of silence duration and following speech segment duration for each individual speaker. The first two left singular vectors are used to construct the 2D space. The lower bound of silence duration is set at 150 ms to minimize interference from potential word-internal or within-phrase fluent pauses, but remain generous regarding all other pausing scenarios. In the following discussion, I will first explore the effect of each socioeconomic variable separately.

Gender The group difference in gender in the derived space is plotted in Figure 13. The entire sample contains speech from 1499 male speakers and 1658 female speakers. The density plots suggest that both male and female speakers have an overall similar shape of distribution in this derived space, while there is a larger spread among female speakers, but also a more clearly defined center of the distribution, compared to male speakers. However, event if the difference between the two distributions are significant, it is expected to be very small. Thus no difference is expected to exist between the two gender groups.

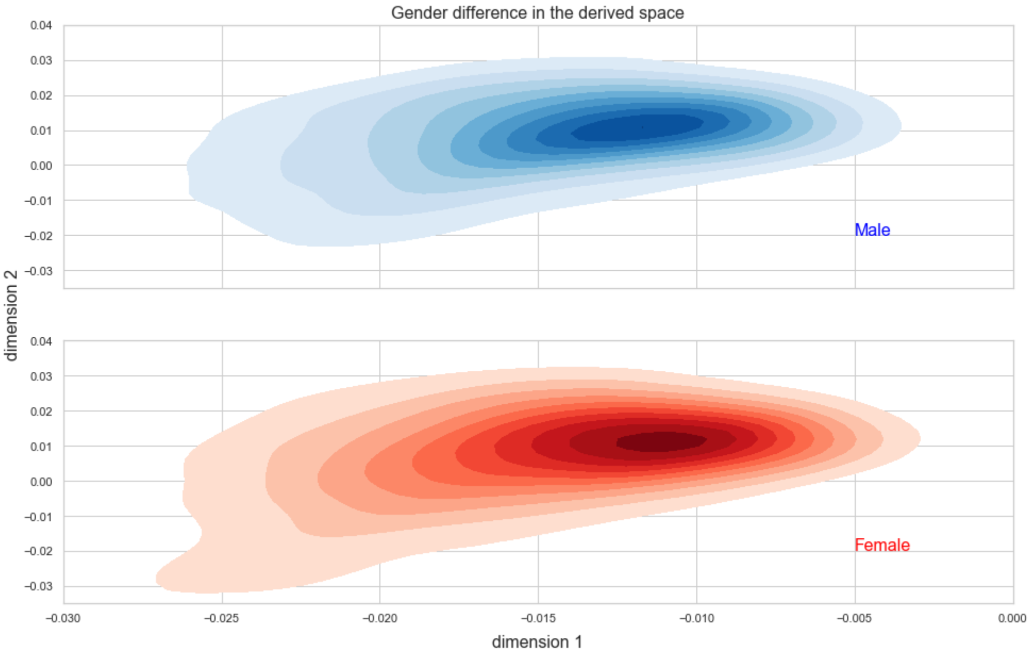


Figure 13: Gender and age difference in the relation between silence duration and following speech segment duration in the derived 2D space.

Age Group difference by age should also be expected, as the literature has suggested. Six age groups are arbitrarily defined: younger than or equal to 20, younger than or equal to 30, younger than or equal to 40, younger than or equal to 50, younger than or equal to 60, and older than 60 years of age.

Figure 14 plots the median values in the two latent dimensions in the derived space for each age group, controlled by gender. It can be observed that age does not seem to have a clear pattern among male speakers. However, there is a trend, although small, for speakers to move from the bottom to top along the second dimension among female speakers. Male and female speakers also appear to be roughly in two groups along the first dimension. Older groups, i.e., those older than 60 years of age, are likely to be outliers within the group of male or female speakers. Therefore with this observation, an interaction effect between age and gender on silent pause distribution can be expected. However, the effect size is also likely to be small, both across or within gender groups.

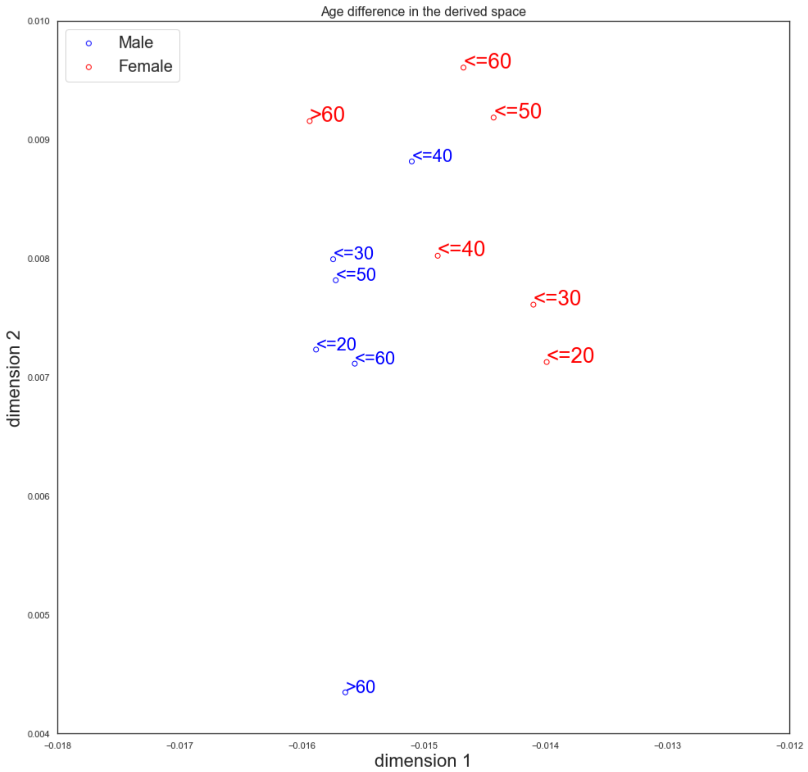


Figure 14: Gender and age difference in the relation between silence duration and following speech segment duration in the derived 2D space.

Years of education The years of education is binned into three categories in this analysis: those who received less than or equal to 12 years of education, who received less than or equal to 16 years of education, and who received more than 16 years of education. This categorization is intended to correspond to the general treatment of education variable in socioeconomic studies: people who received at most high school education, who attended some college level education, and who have attended graduate or professional education. On the other hand, due to the structure of education system, the distribution of years of education in years approximates a step function. One caveat here is that some interaction effect of years of education and age should be expected, on top of the interaction between gender and education, as some participants in the corpus were still attending high school or college at the time of their contribution.

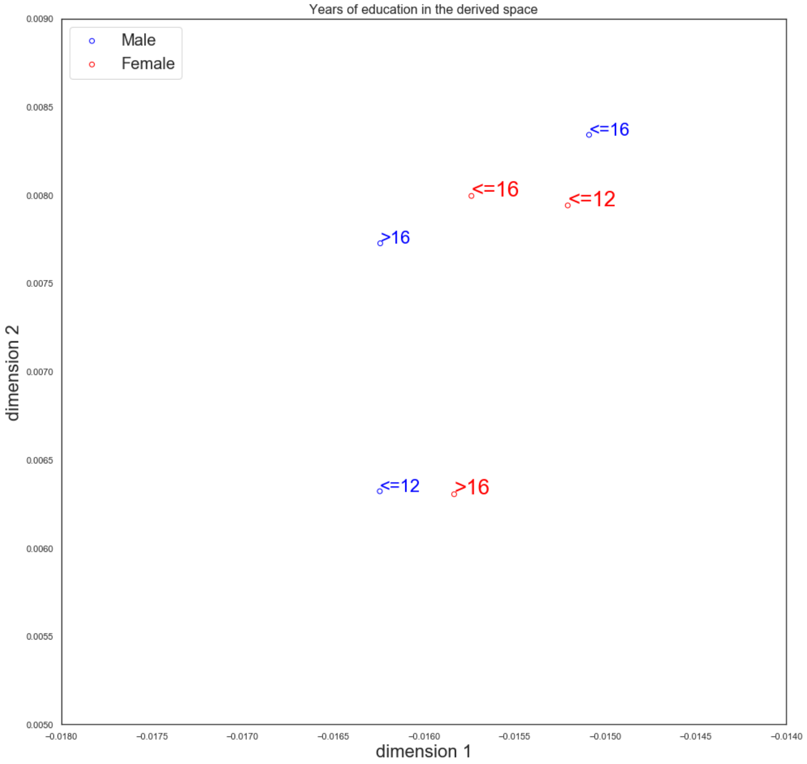


Figure 15: Density plots of the joint distribution of silence duration and following speech duration comparing medians of groups with different education background in the derived space. Education group is plotted conditioned on gender.

Figure 15 shows a clear interaction effect of years of education and gender, as among female speakers, there is a clear distinction between people who received at most college education and those who went to some graduate school. However, among male speakers, the group medians are more spread without clear pattern. Therefore, some interaction effect between education and gender can be expected, as well as the categorical effect of education level. The effect size, however, should also be expected to be small.

Dialect The last sociolinguistic variable to explore is dialect. Since it has been reported that dialectal variation does affect disfluencies in speech production both within North America (Kendall, 2009) and between American and British English (Tottie, 2014), and dialect dependent speech rate variation has also been recorded (Jacewicz et al., 2010), it is

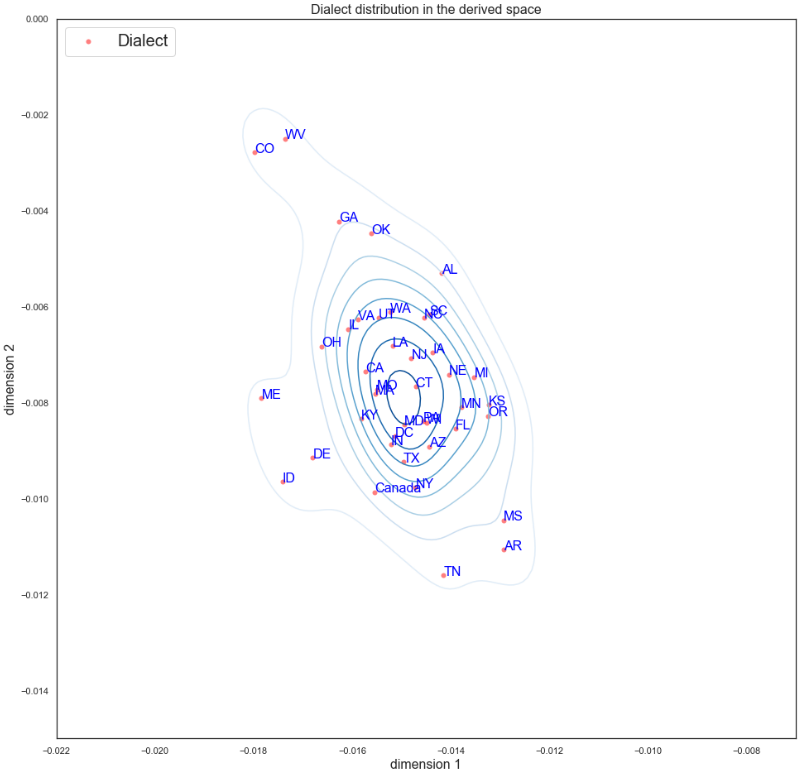


Figure 16: Density plots of the joint distribution of silence duration and following speech duration comparing medians of groups of self reported state where participants have been raised in the derived space.

worth asking if one's dialect has an effect on the silent pause distribution of their speech. In this study, I use the self reported place where participants of Fisher have been raised as the proxy for their dialectal background. In Figure 16, the median values of the first two dimensions in the derived space for each state are plotted. States with too few observations (less than 10) are excluded from this plot due to the potential high variance. The state variable can be considered as more close to a randomly selected sample, thus reflecting the overall population distribution in North America.

In Figure 16, a distribution of the medians across the states seem to be randomly spread in the derived space, approximating a Gaussian distribution with uniform yet differing variances in both dimensions. Further examination of the distribution of the medians doesn't reveal any correspondence to the actual geographic relations among known dialect regions in North America. Therefore, there is evidence that pausing, or the temporal structure of telephone conversations, does not vary across English dialect regions in North America.

Topic The Fisher corpus contains conversations conducted under 40 different topics, which were provided by the data collector, but voluntarily selected by the participants. As mentioned in chapter 1, this data creation process may introduce biases from two perspectives: The topic selection process by data collectors was not intended to construct mutually exclusive topics; rather the goal was to facilitate the unrolling of conversations. Thus, topic categories were not intended to be orthogonal to each other, and overlaps between topics are unavoidable. On the other hand, the topic selection process by participants introduces the second layer of bias, such that some topics are selected more often than the others, and the kind of selected topic is apparently a function of individual speaker's personal preference. Hence, an interpretation of topic effect has to take these biases into consideration. Nevertheless, the relative large sample size in Fisher can somewhat mitigate the effects from these biases, and the results are still informative given these biases are properly considered.

In Figure 17, each dot represents the median values of the joint distribution of conversations under the given topic. Although the overall shape of the distribution of the medians is approximately Gaussian, the variances appear to be non-uniform in the plotted dimensions. For example, mild evidence for two weakly separated clusters can be argued. Therefore considering some random effects from individual speakers, a main but very small effect of conversation topic on silent pause distribution can still be expected. One additional caveat is that this effect can be washed out by the existence of collinearity among some of the topics. Thus an effect of conversation topic on silent pause duration and distribution is expected to be minimal.

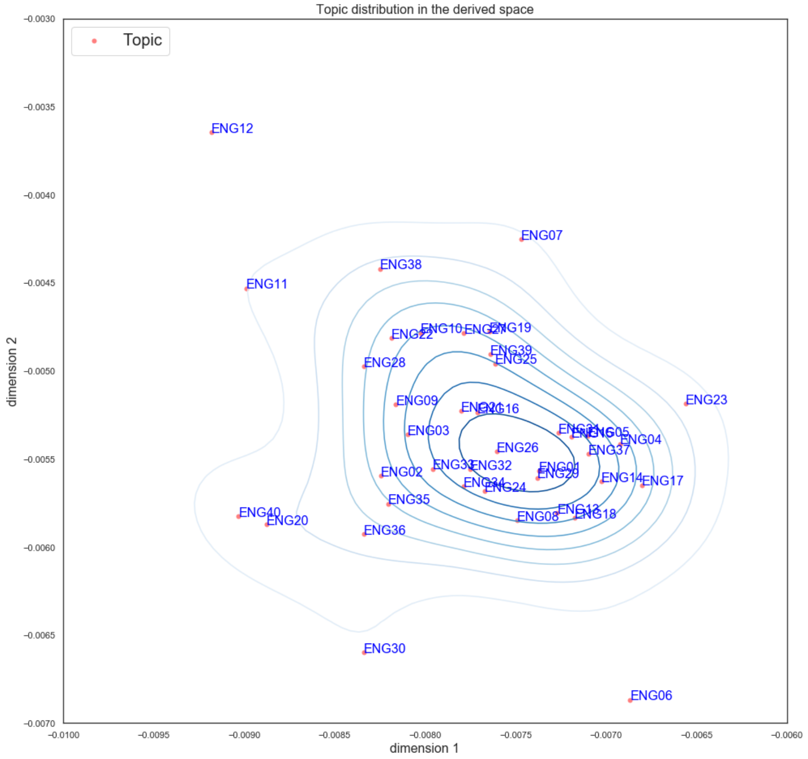


Figure 17: Density plots of the joint distribution of silence duration and following speech duration comparing medians of groups with different conversation topics in the derived space.

Variation across three conversations The last aspect of individual variation to address in this study is to what extend individual speakers would vary in terms of their silent pause distribution across conversations? The discussion of this question will be deferred to the regression analysis, where repetition is modelled as a random slope to the full mix-effect model.

Summary So far, I have demonstrated possible effects from the sociolinguistic variables and conversation topic on silent pause variation. In the derived space, gender appears to have limited effect on silent pause variation, but this effect is still expected to be significant. Age and years of education have been shown to interact with gender, where systematic change related to age has been found among female speakers, and variation in response to years of education among male speakers has also been observed, although these effects are also expected to be small. However, the structure of topic and dialect distributions are less clear.

Regression analysis A linear mixed effects model has been fitted to test the hypotheses formulated through the exploratory data analysis presented above. In this regression, the response variable is the ratio of total silence duration over speech segments duration per speaker per conversation. The same threshold for determining silence, 150 ms, has been used. This measurement is an aggregate of the 2D density estimation for each speaker in each conversation projected to a single dimension, which can be conceptualized as a representation of the average amount of silence contained in one's speech in a given condition. Values in the derived space are not used mainly due to issues with interpretation, and the potential non-unique singular values. The explanatory variables include the factors explored in the derived space, plus the ID of the call, which represents the call repetitions (the n-th call for the given speaker), the interaction between Age and Gender, Education and Gender, the three-way interaction among Age, Education and Gender. The random effect is specified as a random intercept for speakers and a random slope for repetitions. Therefore individual difference in conversation repetitions is effectively considered in the model. The continuous variable *Education* is transformed to categorical representation, following the same strategy shown in the exploratory analysis above due to little variation within this variable. The model is fitted using the *lmer* function in the popular R package *lme4*.

Results of the analysis are reported in Table 1. Under the view of a traditional F-statistics, all of the variables and interactions explored above are significant at $p < 0.05$. However, by this standard, call repetitions and the interaction between age and education are not significant in predicting the variation in the amount of silence within a segment of continuous conversational speech. The random slope of call repetition has very small variance ($\sigma = 0.0013$), suggesting that there is little variation across three conversations.

Table 1: Results from the mixed-effect analysis. [a][b]

	Df	Slope	Sum Sq	F value
Call ID	2	0.003	0.03	2.39
Topic	40	NA	0.56	2.33***
Sex(Male)	1	0.05	0.64	106.71***
Age	1	0.0002	0.12	20.73***
Educ	2	0.008	0.009	7.28***
State	51	NA	0.46	1.52*
Sex:Age	1	0.006	0.06	9.71**
Age:Educ	2	0.00	0.00	0.12
Sex:Educ	2	0.02	0.04	3.59*

[a] The *'s represents the significance level in a classical sense. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

[b] Education compares High School to Graduate School.

Looking at the effect sizes, male speakers on average have about 5 percent higher silence rate in their speech, while an increase by 1 year of age leads to a decrease of 0.02 percent of silent rate, when everything else is held equal. Thus ignoring all other factors, speakers across age groups do not vary much in the proportion of silence with regard to speech in their speech. This is also true when the interaction between age and gender is considered. The aggregated effect, when gender is controlled, is still only about 0.1 percent. The interaction effect of gender×education has a relatively larger magnitude, where for male speakers, one who has completed some graduate or professional education on average has 2.2 percent lower silence rate compared to someone with only high school diploma or less. The rate is 1.4 percent lower compared to college graduates.

3.2.3. Discussion

In sum, results from the descriptive analysis suggest that the distribution and duration of silent pauses are potentially weakly related to speaker features such as age and gender, as well as contextual features measuring the conversation itself. The regression analysis essentially confirms the observations made in the derived space. As predicted previously, the observed effects of age, gender, education, and their interactions are relatively small, maybe with the only exception of gender and education among male speakers. The significance shown in the topic variable suggests that larger discourse unit does affect the distribution and duration of silent pauses in one's speech, when all other variables are held constant. The observed effects are likely to be caused by the underlying cognitive factors that partially affect speech production. Although a direct and objective measurement of silence in speech still seems untenable, the proposed measurement in the derived space based on the relation between silence duration and the following speech duration could offer deeper insight into the nature of pauses in spontaneous speech. In the next section, I will further demonstrate that the proposed measurement of the distribution and duration of silent pauses can be telling about the cognitive impairment induced by alcohol intoxication. This discussion would further substantiate the hypothesis that changes in the distribution of silent pauses are in response to the underlying cognitive factors that affect speech production.

3.3. Alcohol intoxication and silent pause distribution

In this section, I present a case study of the effect of alcohol intoxication on the distribution of silent pauses. The results reported in this section are based on the speech produced in spontaneous monologue and dialogue tasks in the Alcohol Language Corpus (ALC). Through the discussion, it will become clear that a change in the motor control ability induced by alcohol intoxication has direct effect on the structure of silent pauses in spontaneous speech. It will further suggest that even with the presence of great individual variation, the proposed quantification of silent pause distribution is able to capture the change on an individual basis, as demonstrated through a simple classification task.

3.3.1. The problem and the data

Alcohol intoxication can cause deterioration in various aspects of cognitive processing, which may not only lead to problems in the motor control of speech production, but also result in deficits in speech planning (Peterson et al., 1990). Previous research has shown that speakers under the influence of alcohol intoxication tend to produce higher overall fundamental frequency (F0) (Baumeister et al., 2012), increased rate of disfluencies (Schiel and Heinrich, 2015) and changed short-term energy function and F0 contour (Baumeister et al., 2012; Heinrich and Schiel, 2014). Practically speaking, successful detection based on altered speech signal caused by alcohol intoxication can be helpful in the prevention of alcohol related health issues, such as drunk and drive. To facilitate the development of systems that improve the efficiency of alcohol intoxication detection, ALC (Schiel et al., 2008) has been developed and used for a speaker state detection challenge (Schuller et al., 2011). In the challenge, a common set of acoustic features were used to train systems on utterance level classification with a baseline test accuracy (Unweighted Average Recall, UAR) of 65.9%. The best system (Bone et al., 2014) following the paradigm of this challenge achieved a UAR score of 71.4%. Here we ask the question of how the distribution of silent pause duration changes when the speaker is alcohol intoxicated.

In this section, I take the same ALC and ask if the distribution of pause duration changes for individual speakers under alcohol intoxication. As suggested in Baumeister et al. (2012) and Heinrich and Schiel (2014), the effect of alcohol intoxication on speech is highly speaker dependent, meaning that the same effect may surface in the opposite direction on the same acoustic measures for different individuals. This property of intoxicated speech may partly explain the relatively poor performance of utterance level classifiers, even if trained using state-of-the-art neural network architecture with rich acoustic representation

(Berninger et al., 2016). Therefore I take a global perspective, with the goal of exploring the feature space that can efficiently represent the change induced by alcohol intoxication.

ALC is a collection of speech from a total of 162 German speakers (85 males, 77 females) produced in two conditions: sober and alcohol intoxication at a self-chosen intoxication level. The actual blood alcohol concentration (BAC) level was measured immediately before recording. Speech tasks used in the corpus include read speech, monologue (such as picture description, commands and instructions) and short conversations. Speech from the picture description task and short dialogues with the interviewer is chosen for the current study. The speech is recorded with a sample rate of 44.1 kHz with 16 bit rate. Verbatim transcriptions at phoneme level are available and the recordings are aligned.

3.3.2. Feature generation

The feature generation process follows the same method for the quantification of silent pauses detailed in this chapter. For this specific scenario, the joint density functions are estimated for each speaker separately for each of the two conditions: sober and intoxicated. All silent pauses longer than 50ms are included in the calculation. A 100×100 grid is used to sample from the 2-dimensional density function. Therefore the continuous density function is approximated by a 100×100 matrix per speaker condition.

To reduce the sparsity of this representation and achieve a compact representation of the distributions, each 100×100 matrix is flattened as a $1 \times 10,000$ vector. SVD is then performed on the full $162 \times 10,000$ matrix stacked from all the individual feature vectors in each intoxication condition. The left singular matrix (dimension 162x162) is used as the final feature representation of all the speakers in each state, where each row corresponds to an individual in the given condition.

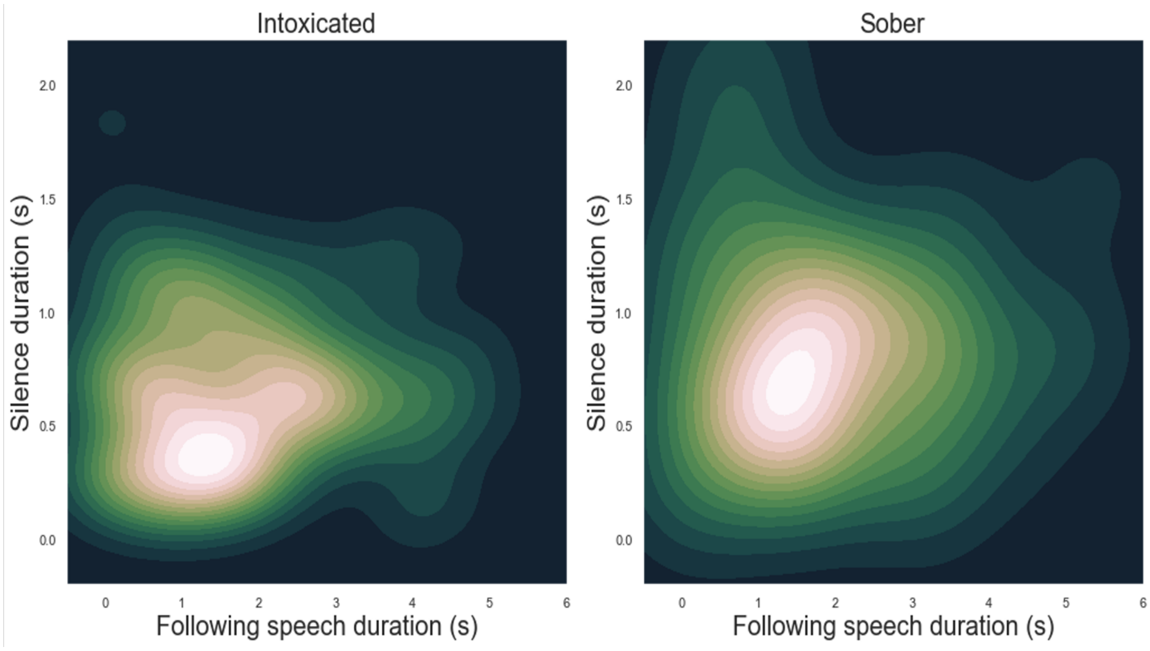


Figure 18: 2D density plot of the joint distribution of silent pause duration (y-axis) against following speech segment duration (x-axis) for a single speaker in intoxicated (left) and sober (right) conditions.

3.3.3. Results

Figure 18 illustrates the difference in the joint distribution of silent pause duration and the following speech segment duration for a single speaker in intoxicated (left) and sober (right) conditions. A clear distinction between the two joint distributions can be observed. Silent pauses produced in intoxicated condition appear to be shorter, and the overall distribution is multi-modal compared to the sober condition.

The scatter plots for all speakers in sober and intoxicated conditions in the first three dimensions in the derived space are plotted in Figure 19. In the coordinate defined by the first and second dimension, intoxicated speakers are distributed mainly in the lower right corner, while in the coordinate defined by the second and third dimension, intoxicated speakers are mainly distributed to the left of the vertical line as shown in the figure. Thus

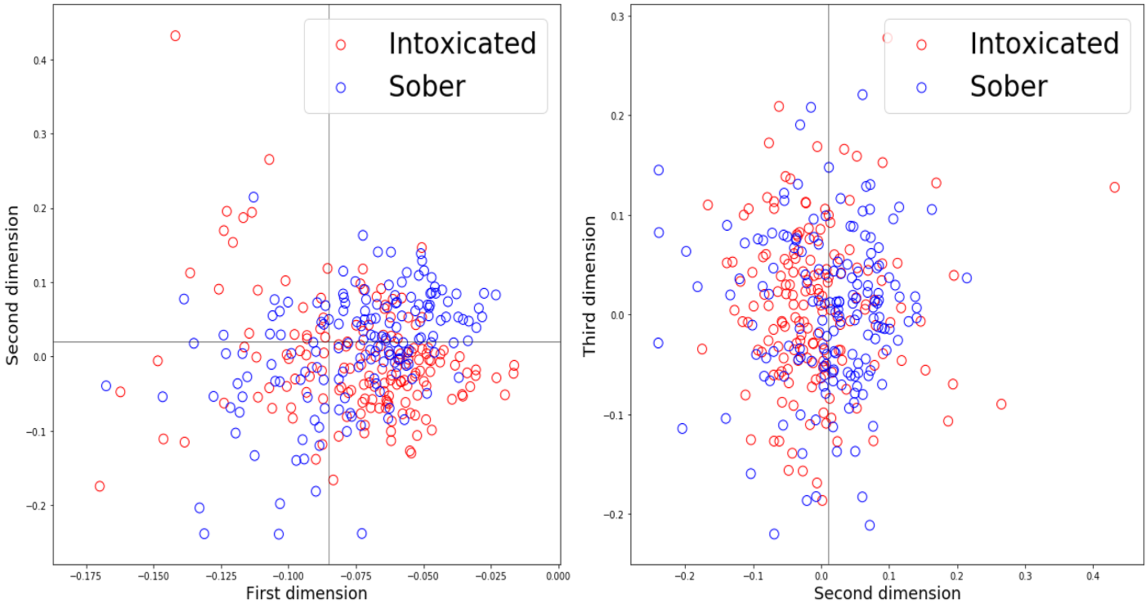


Figure 19: Scatter plots of individual speakers in the derived space in intoxicated and sober conditions. The left plot shows the distribution along the first (x-axis) and second (y-axis) dimensions, and the right plot shows the distribution along the second (x-axis) and third (y-axis) dimensions.

the derived feature space is able to represent the group difference in the distribution of silent pause duration as measured by its relation with the following speech duration.

To test the performance of this derived feature space in distinguishing speakers in intoxicated from sober condition, speakers are randomly divided into training and testing set with a 3-to-1 ratio. The training set contains 122 speakers in both intoxicated and sober states, where-as the testing set includes the paired intoxicated and sober states for the rest of the speakers. Therefore the task can be understood as distinguishing between sober and intoxicated states when the speaker is given.

To evaluate the feature representation derived here, a classification problem can be formulated: If the derived feature space is good at representing the change caused by alcohol intoxication, some discriminative classifier can achieve high accuracy in classifying a given individual as sober or intoxicated with this feature space on the testing set. Here, an out-

Table 2: Testing accuracy for SVMs trained on different percentages of the training data.

Training data	20%	40%	60%	80%	100%
Accuracy	65%	71.25%	88.75%	73.75%	93.75%

of-the-box SVM classifier with Gaussian kernel is used for this classification task. Tuning parameters are selected through a grid search with 10-fold cross-validation on the training set. Table 2 reports the simple testing accuracy for models trained on different percentages of the training data.

Results from this simple classification experiment suggest that the derived features are efficient in classifying speaker intoxication states between sober and intoxicated, as 20% of the training data can already achieve above-chance performance. Training on the full training set is able to yield a pretty high testing accuracy (93.75%) on the unseen test examples.

3.3.4. Discussion

In this section, I presented a case study in which the question of how alcohol intoxication affects the distribution of pause duration in spontaneous monologue and conversations at individual speaker level has been addressed. Using speech produced from the picture description and short conversation tasks in ALC, I demonstrated that the effect of alcohol intoxication on silent pause duration can be effectively represented through the relation between silent pause duration and the following speech segment duration. Dimensionality reduction techniques, such as SVD, are able to offer compact parameterization of the differences observed from the joint distribution. The derived features appear to be highly efficient in intoxication identification.

The good performance of this feature space as demonstrated through the simple SVM classifier also shows that, although individual variation in particular acoustic dimensions can be problematic in deriving good representations of alcohol intoxicated speech, features derived from rich characterization of joint distributions of related variables can generate robust parameterizations for speaker state detection.

The effectiveness of representing the pausing behavior in some derived space in the alcohol intoxication detection task further strengthened the proposal made earlier: latent structures related to both the distribution of location and duration of silent pauses can be revealed through feature engineering in some derived space, and the variation in the location and duration distribution of silent pauses is related to underlying processing mechanisms that are responsible for speech production. Methodologically, results from the discussion in this and the previous section highlight the potential for the idea of exploring the temporal structure of speech phenomena through modeling the probabilistic distribution in some higher dimensional space. This way, the nuances and interactions between linearly correlating units, which are otherwise overlooked in the original one dimensional space, can be robustly captured and represented. Dimensionality reduction techniques could serve as powerful tools in both visualizing and understanding the hidden structures in the higher dimensional probabilistic space. The major drawback of this approach is the lack of interpretability, which could potentially be remedied through follow-up experimental work on the phenomena of interest.

3.4. Filled pause

Similar to the discussion on silence distribution, in this section I first compare between groups within each of the proposed variable, but separately for the two variants of filled pauses: *um* and *uh*. Then I present two regression models independently built for the two filler words. The measurement for filled pause in this section is the word frequency per 100 words for each individual in each conversation for the exploratory analysis.

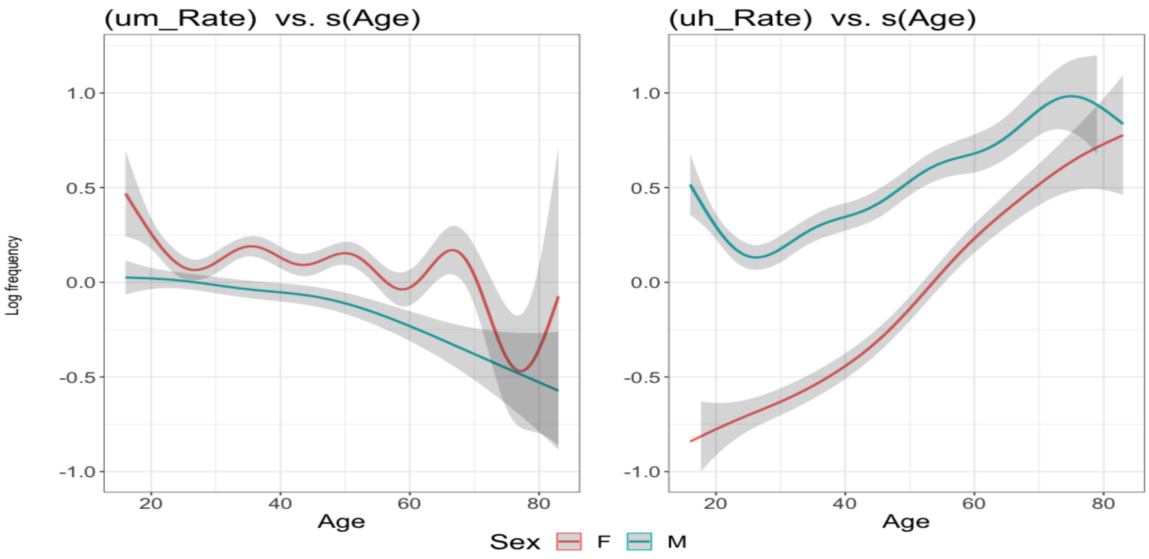


Figure 20: Log filled pause rates plotted as functions of age, grouped by speaker gender.

3.4.1. *Feature independent analysis*

Age and Gender Figure 20 plots the frequency of *um* and *uh* as a function of speaker age, controlled for gender. The regression lines were the mean estimators of a Generative Addative Model (GAM) fitted for each gender group under each condition using Poisson regression. The grey bands represents the 95 percent confidence band for the estimated mean in log space. The y-axis in each graph represents the log frequency of using *um* or *uh*, and x-axis represents age treated as a continuous variable.

The two graphs in Figure 20 clearly indicate an opposite trend of change of the frequency of two fillers: for *um*, the relation between frequency and age is slightly negative linear, while for *uh* the relation is almost perfectly positive linear, except for the oldest and youngest males group. The observations for the oldest age group are relatively sparse, while it's not the case for the youngest. Female speakers also almost consistently have higher estimated frequencies for *um* across all ages, while lower frequency for *uh*. This trend essentially replicates what have been reported in Fruehwald (2016) on a different data set and plotted along different dimensions, and in Wieling et al. (2016) with a different treatment of the age and frequency variables. Interestingly, the decrease in *uh* frequency, as age changes from older to younger, is higher among females, while the increasing rate is almost parallel between the two gender groups in the age range of about 22 to 60.

However, two details are worth mentioning. First, the trend for the change of filled pauses as a function of age is clearer and more stable for *uh* than *um*. The increase in popularity of *um* is actually not apparent among younger speakers (younger than 50 years of age). Therefore the seemingly increase in the popularity of *um* can be driven primarily by the low rate among older speakers. This can be problematic since the higher variance among older age groups may indicate the existence of unobserved heterogeneity. Second, the between gender frequency gap of *um* is also much narrower than the gap in *uh*. This difference in the trend of change with relation to gender may be aligned with the trend of a decreasing in the use of both *um* and *uh* over time as reported in Wieling et al. (2016) on Switchboard. The surface change or trade-off between the popularity of two variants of filled pauses can be argued to be driven by the higher variability in the use of *uh*, which is further attributable to other contextual or idiosyncratic factors that are not accounted for in the current and previous studies. A simple view of language change in progress is not able to offer adequate explanation to these nuances in the graph. Therefore the alternative explanation that the trade-off in the frequency of two fillers is associated age-related change in language processing and speech planning cannot be ruled out.

Education The relations between the frequency of *um* and *uh* and education, controlled by gender, are plotted in Figure 21. The box plot for *uh* does not show clear variation across education level in both gender groups. However, speakers with only at most a high school diploma have slightly lower frequency of *um*, and this seems to be true regardless of gender. The small difference in frequency distribution also seems to be stable. On the other hand, speakers with post secondary education background tend to have higher *um*

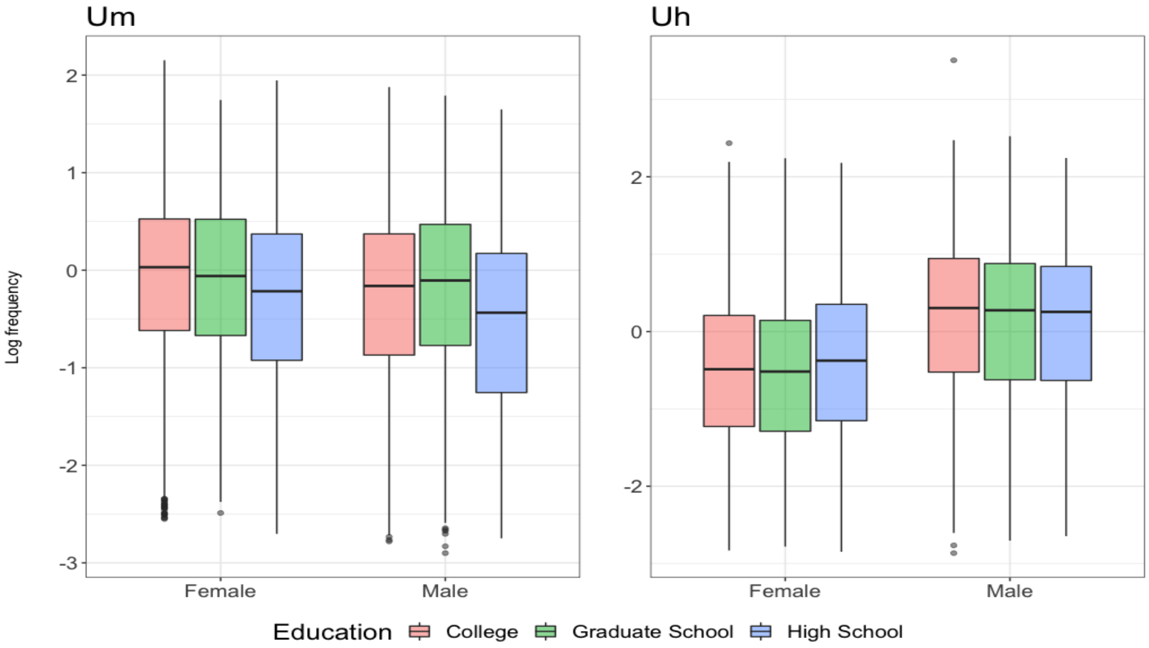


Figure 21: Box plot of *uh* (on the left) and *um* (on the right) frequency by education and gender.

frequency, and the between gender difference appears to be much smaller compared to *uh*. One possible explanation to this difference is a potential interaction between age groups and education level: people that are older may be less educated compared to younger age groups. Thus the apparent effect of education on *um* frequency may just be corroborated with the age effect that I have just shown.

Figure 22 and 23 plot age distribution by education groups: High School, Some college, and attended some Graduate or Professional education. If age is indeed the main factor underneath the observed difference between education levels, then one would expect an overall older age among High School graduates. In particular, since people older than 60 years of age have the highest *uh* frequency, one might expect more high school graduates in this age group as well.

A noticeable difference is not found in Figure 22, when people who are 60 or older are plotted separately or when the entire sample were plotted. Nevertheless, there does seem

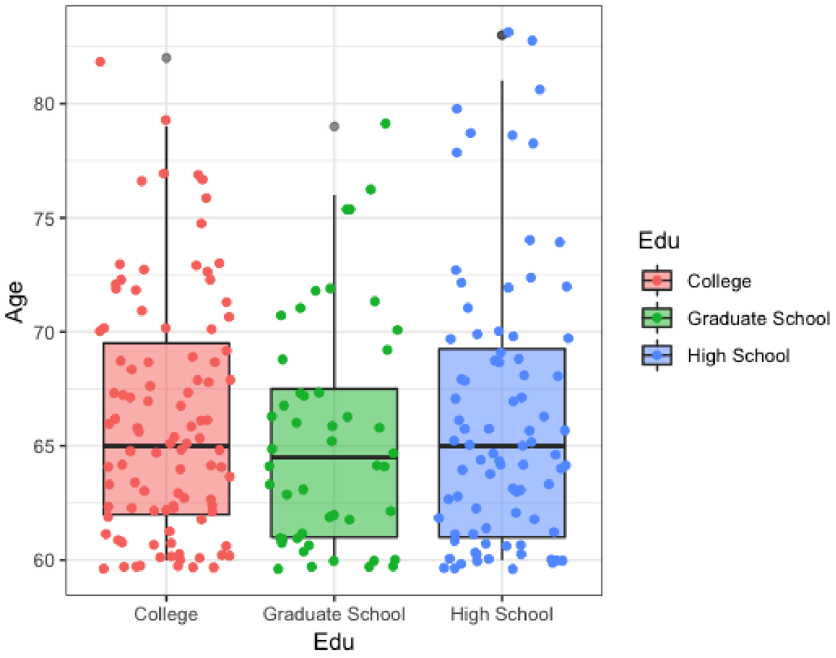


Figure 22: Age distribution by education level: speakers older than 60 years of age.

to be a slightly larger average age among people with only high school education, even though the variance among that group is also bigger. This difference is in fact confirmed through a one-way test of variance ($F = 22.878, p < 0.001$). However, the mean difference is only 3 years: 39 years of age for high school graduates, compared to 36 years of age for both of the other groups.

Dialect Variation across English dialects is plotted as the median *um-uh* frequency pair for each state against the joint contour of *um-uh* frequency pooled across the entire sample. As Figure 24 suggests, on average, there is a trade-off relationship between *um* and *uh* frequency. This relation indicates that for each individual speaker, there is likely to be a preference when choosing the filler word in conversation. The peak density on this graph essentially follows the direction of x-axis, which indicates that there are more predominantly *um* users in this speaker sample. Larger variance in *um* frequency across speakers is also apparent from the graph.

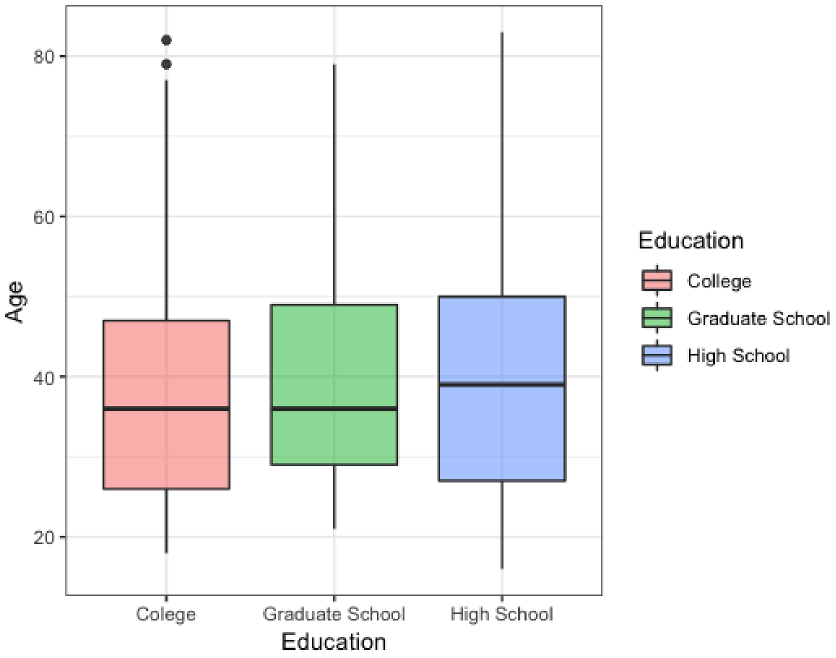


Figure 23: Age distribution by education level: all speakers.

By examining the distribution of states on this overall contour plot, it is found that the medians roughly follows the sample distribution and fail to show clear cluster structures. The distribution of states also appears to be at random, as no alignment between adjacent states and the acknowledged dialect regions can be identified. Thus, it is not likely that dialects would have a systematic effect on the choice of filled pause.

Topic Filled pause frequency of the two fillers is plotted across conversation topics similarly to the plot of dialects. In Figure 25, it can be observed that there is a larger variation along the *uh* frequency dimension of the median values. This suggests that there is greater variation in the frequency of *uh* across topics than the variation along the *um* dimension. Thus it is expected that topic mainly has an effect on the use of *uh*. But the existence of certain amount of variation across some topics along the *um* dimension, especially towards the bottom of the plot, suggests that some effect on *um* is also possible. The variance along

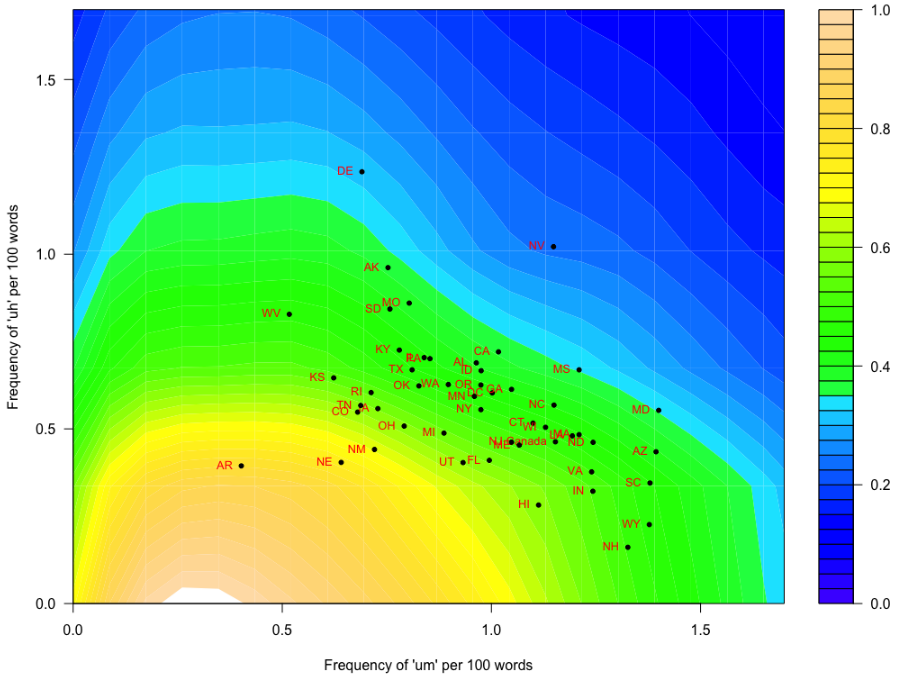


Figure 24: Contour plot of *um* frequency (x-axis) and *uh* frequency (y-axis) overlaid with speaker states.

the *um* dimension across topics is also non-uniform. These observed variations along the two axes can be potentially related to the content associated with each covariation.

Correlation matrices between pairs of conversation topics are derived to explore the (dis)similarities in terms of filled pause distribution across topics. The correlations are calculated based on the estimated density function from the frequency of filled pauses in each conversation in the given topic. The two types of filled pause are treated separately.

Figure 26 plots the two correlation matrices. In these plots, the lighter the color in the plot, the higher the correlation between pairs of topics. Figure 27 summarizes the cumulative distribution of correlation scores separately for the two filled pauses. One apparent difference between the two forms of filled pause is that the overall pairwise correlation of filled pause frequency between across topics in the frequency distribution is lower for "uh"

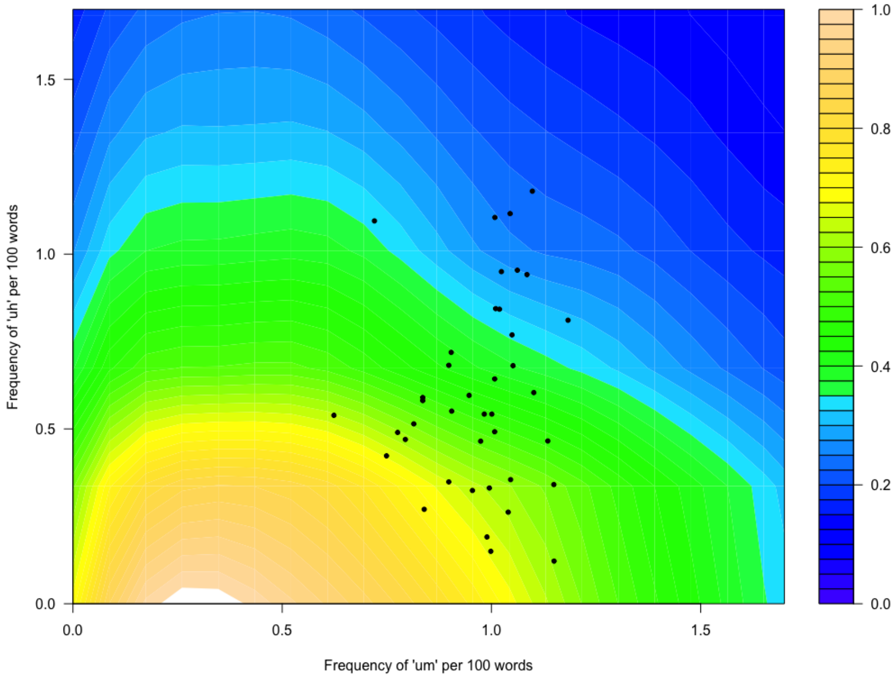


Figure 25: Contour plot of *um* frequency (x-axis) and *uh* frequency (y-axis) overlaid with conversation topics.

in comparison with “um”. In fact, the frequency distribution of “um” doesn’t seem to vary much across topics. The second observation is that the pairs of more (dis)similar topics also differ between the two forms of filled pauses. For example, in Figure 26, topic 26, 27, 28, 29 and 30 have very low correlation score (less than 0.4) with topic 7 and 8 in terms of the frequency distribution of “uh”. This difference may be attributable to the content of the actual conversation, as topics 26 to 31 are about more serious political or social issues such as Airport Security, Middle East and Foreign Relations, as well as Education and Family. On the other hand, topics 7 and 8 are on some hypothetical situations. However, the frequency distribution of “um” among these topics are highly correlated with correlation being over 0.9. One possible explanation to this difference is that the two forms of filled pause have different functions in the coordination of spontaneous conversations: “um” may

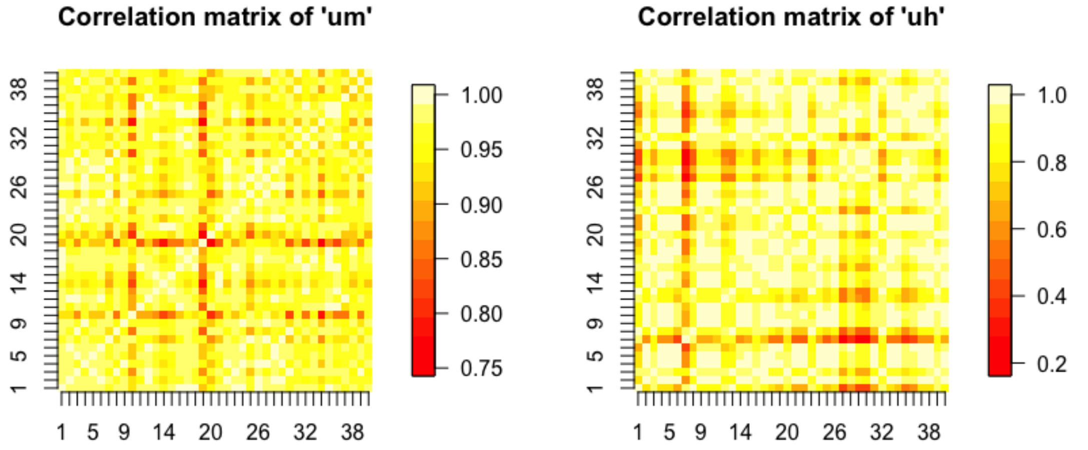


Figure 26: Correlation matrices of filled pause frequency between topics. Axes indicate topic numbers.

more likely be strategically used as a device for message structuring purpose, while "uh" tends to signal the variation in speech production due to changes in the discourse.

3.4.2. *Summary of the exploratory data analysis*

From the exploratory data analysis above, it can be expected that apart from dialect, other proposed explanatory variables, including age, gender, education and topic, will affect the choice of filled pauses in spontaneous conversations. The two variants of filled pause appear to have different sensitivity in response to the changes in the dimensions discussed above. In the rest of this section, I present two regression models, for *um* and *uh* independently, to address the question concerning what's the effect size of each of these variables, as well as the potential interactions. The two filled pauses are modelled separately, rather than jointly such as in Wieling et al. (2016), is mainly for the existence of primarily *um*-ers and *uh*-ers in the sampled corpus, and the analysis in this study concerns each speaker as

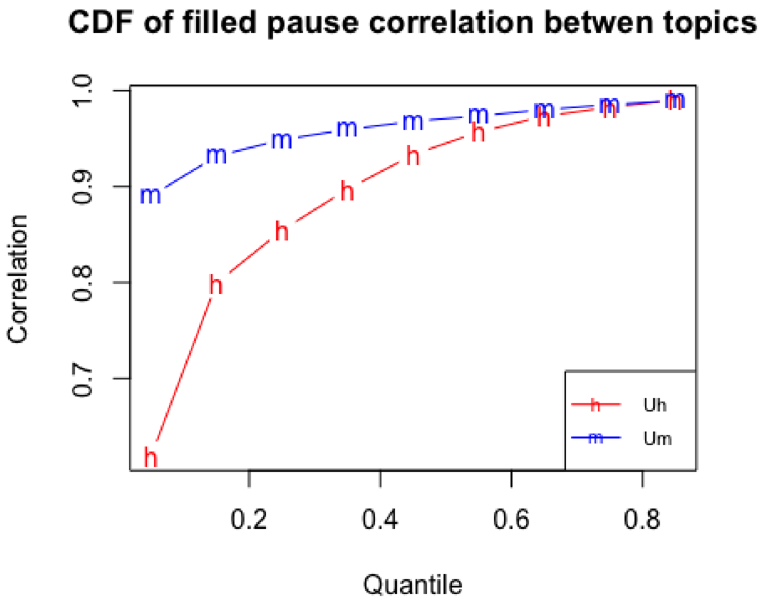


Figure 27: Quantile plot of the correlation scores.

one independent observation, rather than pooling across speakers to estimate group means. Therefore if the relative frequency of *um* and *uh* were taken, no valid ratio would be found for many observations. The potential high co-linearity between *um* and *uh* frequency also deems considering one filler as the explanatory variable of the other inappropriate.

3.4.3. Regression models

In this section I report the results from two Poisson mixed-effect regression models fitted for *um* and *uh* independently. The Poisson model regresses the log frequency of each fillers onto the space defined by the exploratory variables explored above. Speaker's idiosyncratic behavior in response to the conversation task and the variation across repetitions of the task

Table 3: Mixed-effect Poisson regression on *um* frequency.

	Df	Slope	Sum Sq	F value
Repetition	2	-0.18	3.69	1.84
Topic	40	NA	411.41	10.29***
Sex (Male)	1	-0.09	58.34	58.34***
Age	1	-0.005	49.52	49.52***
Education	2	-0.28	50.18	25.09***
State	51	NA	72.49	1.42
Sex:Age	1	-0.05	4.45	4.45*

*Slope for *Repetition* compares the third call to first call. Slope for education compares Graduate to High School education.

is modeled as the per-speaker random intersect and per speaker per repetition random slope. Thus within and cross speaker variation is accounted for in the model.

Table 3 summarizes the results for the filled pause *um*. As expected, except for the variable State, which is used to represent dialect variation, all other explored variables are estimated to have significant effect on the log frequency of *um*, if a threshold of $p = 0.05$ is chosen. As for the effect size, male speakers on average have 0.15 fewer *um* counts per 100 words of speech compared to female speakers, when everything else is held constant. For male speakers, an increase of 1 year of age corresponds to on average a decrease of about 0.06 count of *um* per 100 words of speech, while this decrease is about 0.005 count per 100 words for female. The sharper change among male speakers aligns with the steeper slope for male while more curvature for female observed in 20(a). In terms of education, comparing between those with graduate degree or higher, high school graduates on average uses 0.28 less *um* per 100 words, when age, gender, topic, repetition and dialect are controlled. Thus the estimated effects of these variables reliably reflect the differences observed in the exploratory analysis step.

In terms of the random effect, the estimated variance of slope is 0.082. Therefore, it appears that there is little variation within individual speakers across the three conversation

Table 4: Mixed-effect Poisson regression on *uh* frequency.

	Df	Slope	Sum Sq	F value
Repetition	2	-0.01	450.96	225.48***
Topic	40	NA	1955.92	48.90***
Sex(Male)	1	1.35	462.42	462.42***
Age	1	0.02	236.07	236.07***
Education	2	-0.06	2.48	1.24
State	51	NA	69.23	1.36
Sex:Age	1	-0.01	31.97	31.97***

*Slope for *Repetition* compares the third call to first call. Slope for education compares Graduate to High School education.

repetitions. This shows that *um* can be a filler whose per speaker frequency subjects more to speaker factors than to conversational or contextual factors.

The second model performs the same mixed-effect Poisson regression on the frequency of *uh*, with same model specification as the previous model. As summarized in Table 4, the model confirms the initial observations on the relations between each explanatory variable and *uh* frequency. In addition, the variable *Repetition* appears to be significant, which suggests that on average there is more cross-repetition variation in the use of *uh* for a given speaker.

An examination of the effect size is the following. Compared between male and female, a given male speaker on average uses 0.4 more *uh* per 100 words of speech, when everything else is held constant. In terms of age effect, for male speaker, an increase by 1 year of age corresponds to 0.01 more *uh* per 100 words, while this difference is about 0.02 per 100 words for females. Thus the trend observed in Figure 20(b) is also truthfully reflected in this model. As for the random effect, the estimated variance is about 0.48, which is substantially larger than the estimate for *um*. Therefore it can be hypothesized that the use of *uh* is more sensible to contextual variables, such as the familiarity of task, the identity of the interlocutor, or the nature of the conversation topic, to name just a few.

3.4.4. Accommodation between interlocutors

A follow up question in response to the observed interactions between sociolinguistic and discourse variables to ask is what is the role played by the interlocutor in the distribution of the two filler forms? In other words, what is the accommodation effect on filled pause distribution in telephone conversations? The effect of accommodation, or entrainment, has been studied in the past from the perspective of communication mode such as differences among human-human, human-machine, and dialogue versus monologue (Oviatt, 1995), the role played by the speaker in a communication task (Bortfeld et al., 2001), and in supreme court oral arguments (Beňuš et al., 2012) and conversations in a game setting (Beňuš et al., 2011). To investigate the temporal aspect of turn taking in spontaneous conversations, Ten Bosch et al. (2005) showed that the duration of between-turn pauses made by speakers in a dyad is statistically related, and gender appears to have an effect on the temporal aspect of turn-taking: male-male conversations tend to have more inter-turn overlaps than female-female conversations. Oviatt et al. (2004) suggested that, in a study of accommodation in human-machine communication among children, the largest adaptation comes from the pausing structure and acoustics of utterances. In human-human communications, converging patterns of pausing structure and the use of filler words or other signalling words contribute to the coordination of conversation and establishment of common ground (Beňuš et al., 2011, 2012).

To evaluate the effect of accommodation on the frequency distribution of the two filled pauses, I compare within each pair of speakers (speaker a and speaker b). The most obvious dimension for this comparison is gender: among all the variables considered so far, gender is the easiest identifiable covariate. Therefore I compare the frequencies in three different groups: male-male, female-female and female-male conversations. The comparisons are carried out using a sub-sample in which both sides of a conversation are present in the se-

Table 5: Correlation of filled pause frequency in conversation between speakers controlled for gender.

Filled pause type	Male-male	Female-female	Female-male
Um	0.113	0.103	0.056
Uh	0.368	0.205	0.192

lected full sample. This yields a sample consisting of 685 male-male conversations, 885 female-female conversations, and 675 female-male conversations. Correlations between speaker a and speaker b in each group are reported in Table 5. It is obvious from the correlation table that there is stronger correlation in the frequency of “uh” between interlocutors in a conversation than “um”, and this trend holds across all three conditions.

3.4.5. Discussion

In this section, I have explored the effects of socioeconomic variables, such as age, gender, years of education, dialect, and conversation topic on the use of filled pauses. The exploratory analysis of the frequency of *um* and *uh* first confirms the observation using other corpus data or statistic methods (Fruehwald, 2016; Wieling et al., 2016) that there is a trend for more *um* and less *uh* usage among younger speakers, which has been postulated as a change in progress led by female speakers not only in American English, but also across several Germanic languages (Wieling et al., 2016). It is also found that the rate of increase in *um* frequency is actually very slow especially among younger speakers, while the drop in *uh* frequency is steeper among female speakers compared to males. Amount of education and topic have also been suggested to affect filler frequency. However, the effect of education is mainly on the use of *um*, while the major effect by topic is on the use of *uh*. Little dialect effect has been suggested as well. Two Poisson mixed-effect regressions are then reported in support of the initial observation.

Although the results presented above have seemingly provided further evidence for a potential change in progress of the use of filled pauses, this analysis in fact shows that the loss of popularity in *uh* in return for more frequency of *um* is not a parallel process, in terms of both the pattern of trade-off within each gender group, and across genders. This asymmetric trade-off may be a result of a different sensitivity to contextual variation for the two fillers: as suggested by Figure 27 and 26, *uh* exhibits higher variability across conversation topics than *um*. Higher degree of accommodation effect has also been found in the use of *uh*, seen from the higher correlation of *uh* frequency between interlocutors. Furthermore, the regression models also offer evidence for larger expected variation for an individual speaker for *uh* than for *um*. Thus, different forces may exert different effects on the direction and magnitude of change for the two fillers. The observed change associated with speaker's age is therefore in fact likely to be fundamentally driven by the different function or meaning of the two forms of filled pause. This difference is also potentially related to their different response to cognitive factors involved in speech planning. Therefore a detailed examination of the feature space and how they influence speaker's decision in choosing between the filler words is warranted. The first next step would be to understand how the nature of different topics, such as the content of speaker's speech under the provided topic, affect the frequency distribution of the two filled pauses.

3.5. Chapter summary

In this chapter, I reported results from analyses of the effect of sociolinguistic variables and topic on silent and filled pause variation, considering the potential within and cross speaker variation in tandem. I demonstrated that a method based on robust representation of the probabilistic distribution of the relative temporal relationship in speech in linear time could reveal rich information about the dynamics of the phenomenon of interest. In the present case, this methodology has been shown to be effective in the classification task of alcohol intoxication identification. The result from the analysis and comparison between alcohol intoxicated and sober speech also suggests that underlying cognitive impairment can be a cause for the change in pausing behavior, which may indirectly explain the effect of other measurable sociolinguistic or contextual variables on the location and duration distribution of silent pauses. With regard to filled pauses, results in this chapter first replicated previous observations on the correlation between filled pause distribution and sociolinguistic variables, and supplemented new information considering the effect of topic and accommodation between interlocutors. Evidence from both the regression analysis and speaker and topic accommodation shows that the surface change in the relative frequency distribution of *um* and *uh* may in fact related to the different underlying cause of their production: one is potentially highly sensitive to the change in cognitive ability behind speech production, and the other is likely signalling certain pragmatic meaning or function in the communication context.

The existence of potential alternative explanation for the observed correlation between the frequency of filled pauses and sociolinguistic variables should not be considered as an argument against the proposed explanation in literature. Rather it points out a direction towards a causal explanation for the observed variation. In other words, the change in progress account on the correlation between filled pause frequency and speaker age is rather an elaborated description of the phenomenon conditioned on the available data, than a theory that explains why the observation is the way it is. As pointed out in Wieling et al. (2016), the seemingly unanimous pattern of change across at least Germanic languages poses a puzzle that lacks a clear path for solution. The discussion presented in this chapter raises the possibility that the observed change is potentially related in part to factors that covary with changes associated with the cognitive process of speech planning. Thus a joint view from both psycho-neurolinguistics and sociolinguistics should be regarded as a promising direction to move forward with.

Chapter 4

Repetitive Interpolation

As reviewed in Chapter 2, repetitions are generally regarded as a form of disfluency. However, there exists empirical evidence that some repetitions do not align with the assumption that replanning or hesitation is involved in speech disfluencies. In this chapter, I will present evidence for a potential category of a fluent version of repetition, which I will term as repetitive interpolation. The empirical observations to be discussed here suggest that: 1. Repetitive interpolation does not interrupt the flow of speech by introducing additional pauses or alter the prosody of the speech segment; 2. The frequency of repetitive interpolation is not correlated with the lexical frequency of the repeated words; 3. Repetitive interpolation occurs mainly in predictable lexical context; 4. Repetitive interpolation does not involve blockage of speech delivery as typically seen in stuttering. With these evidence, it is reasonable to hypothesize that repetitive interpolation is a distinctive phenomenon that should be categorized and understood separately. Unlike other disfluencies such as hesitation phenomena and speech repair, repetitive interpolation lacks features commonly found in disfluency phenomena such as no hesitation or replanning being involved. With this proposal, further hypotheses about the underlying production mechanism can be motivated.

The structure of the chapter is the following. I first review the literature on repetitions in spontaneous speech in more detail. In particular, I review previous studies on repetitions under the initial assumption that repetitions are a kind of disfluency. Then I argue that many observations of repetitions from both the literature and new empirical examples do not follow nicely from the assumption that they are disfluent. With such observations, I propose am informal definition of “repetitive interpolation” as “rapid repetitions of mostly single syllable function words at the beginning of a phrase”. I then present evidence from the acoustic, word frequency and lexical contextual perspectives to argue that this repetitive phenomenon should not be regarded as a kind of disfluency. Acoustically, this kind of repetition does not introduce apparent disruption in the flow of speech, as can be measured from the duration and the fundamental frequency of the relevant speech segments. From the lexical perspective, the likelihood of observing such repetitions is independent from the absolute word frequency of the repeated word, and the repetitions often occur in the context in which the following lexical item is more predictable compared to the same words in non-repeating contexts. These observations lead to the hypothesis that repetitive interpolation is a phenomenon in fluent speech that are distinctive from what has been traditionally understood as speech disfluency.

4.1. Background

As reviewed in Chapter 2, repetitions are often analyzed in speech disfluency literature as a symptom for problems encountered during higher level processing in speech production such as during message formulation. When the speaker stumbles in formulating an utterance plan, or needs to retract the yet-to-be produced speech for anticipatory error correction, they would repeat part of the phrase that has already been produced (Lickley, 2015; Clark and Wasow, 1998). In this section, I will first structure a detailed review from the conventional view of repetitions that they are a kind of disfluency phenomenon. Although this predominant view is able to explain some variations of repetitions, I will show, through both re-examining the analysis in literature and new empirical examples, that on the surface many repetition instances are in fact without signs of disruptions that typical of speech disfluencies. Thus an alternative analysis of such repetitions can be motivated.

4.1.1. Repetitions viewed as a typical disfluency phenomenon

In chapter 2, I have reviewed that repetitions can be further characterized with a two- or three-way distinction in terms of the potential functions they serve: it can function as filled pauses or filler words (Zellner, 1994; Maclay and Osgood, 1959), as a repair strategy for prospective and retrospective repair (Hieke, 1981), and additionally it may signal a covert repair (Levelt and Cutler, 1983). Such a two or three-way distinction among repairs and repetitions has found empirical support from acoustic studies (Shriberg, 1995; Plauché and Shriberg, 1999). In her analysis of single-word repetitions from speech of six speakers from Switchboard, Shriberg (1995) was able to classify the repetition instances into categories corresponding to the proposed *retrospective* and *prospective* repairs (Hieke, 1981) by considering the relation between word duration of the first and second repeat in a repetition. Interestingly, the distribution of the two types of repetitions is highly imbalanced, with retrospective repetitions taking the majority in their sample. Moreover, the duration of repeated words in retrospective repetitions is similar to the duration of fluent words. In a later work (Plauché and Shriberg, 1999), the authors particularly considered the repetitions of “i” and “the”, two of the most frequently repeated lexical items. Using an unsupervised hierarchical clustering algorithm, they were able to identify three clusters of repetitions with features measuring the duration of repeated words and silence within a repetition, as well as the F0 contour. This study is able to provide evidence for the existence of natural clusters among repetitions of “i” and “the” from the proposed acoustic features. However, the feature set in this study was constructed through discretizing continuous variables such as duration and F0, and it is likely that the clustering result is biased from the imbalance of feature distribution. Observations from these studies have been argued to support the existence of further subclassifications of repetition phenomena. The proposed subclasses can potentially fit into the hypothesized potential problems encountered during speech production. However, no direct evidence for the existence of the proposed production problems has been provided.

In addition to the acoustic investigations of repetitions, empirical work based on the textual features of speech has also been pursued. In their detailed analysis of repetition disfluency using Switchboard and London Lund corpora, Clark and Wasow (1998) attributed such a repetitive phenomenon as a strategy to maintain a balance between the demand for fluency and continuity of the flow of speech, as described in their Commit and Restore model, when the need for hesitation or repair is encountered. Their proposal is backed by the observation that repetitions are predominantly single syllable function words, and the rate of repetition is highly correlated with the syntactic complexity of the phrases in which repetitions occur. One intriguing observation of repetition phenomena is that the tendency to repeat predominantly single syllable function words exists cross-linguistically. This observation has been tested in a range of languages that have relatively diverse morphological and syntactic structures, such as English, German, Hebrew, and Japanese (Fox et al., 1996, 2010). Under the view of Clark and Wasow's model, this cross-linguistic trend is driven by the demand to preserve the completeness of the restored utterance, hence to repeat the left-most components in a restored phrase comes as a natural consequence. The analyses based on textual features have opened additional possibility for a theoretical explanation of repetitions. Although it has been shown that the left edge of a phrase is prone to higher rate of repetition, it is not yet clear whether this higher rate is proportional to the frequency of the word being repeated, or is also driven by the lexical contexts of repetition. As the cross-linguistic work has shown, the morpho-syntactic structure of the language does seem to affect repetition frequency, it is still possible that most repetition instances do not in fact reflect an issue with higher level speech planning, at least in the same way speech repairs do.

The acoustic and textual studies have suggested the need for subclassifications of repetitions. Such subclassification partly hinges on the relation between different repetitions and particular planning processes. Ultimately the theoretical interest in these studies lies in understanding how speech production processes explain speech disfluencies. However, the proposals made in the literature do not exclude alternative explanations from a non-disfluency perspective. Specifically, the following two problems that have been repeatedly mentioned are left not fully explained: the lack of replanning time between repeated words and the seemingly universal cross-linguistic trend of repetitions favoring single syllable function words. Although the functional variation of repetitions under the umbrella view of being singly a disfluent phenomenon has received supports from various acoustic studies (Shriberg, 1995; Plauché and Shriberg, 1999), the acoustic distinctions across functions reported in the literature contradict the timing requirement for a repetition to truly require a full repair and replanning cycle, even if only the internal feedback loop is what all it is needed (Civier et al., 2010). On the other hand, although the Commit and Restore model is able to explain why single syllable function words are the preferred target for repetition, it doesn't explicitly address the fact that these repetitions are not associated with other apparent disfluencies which provide stronger signal for the need for repair, and does not relate the model to the problem posed by pure acoustic analysis. Thus a single disfluent perspective on repetition is unable to explain all the observed variations for repetition.

4.1.2. An alternative view

The literature review above shows that regarding repetitions in spontaneous speech as a disfluency phenomenon is not able to account for all the potential variations of repetitive phenomena. In this section, I further illustrate how certain repetitions are arguably distinctive from a typical disfluency phenomenon. It can be argued that these repetitions may only be considered as disfluent because the same lexical item is repeated two or more times, which is not expected in fluent speech. However, other aspects of the speech delivery are in fact quite fluent in utterances with such repetitions. Although this kind of repetition has received some attention in previous research (Shriberg, 1995), it is generally regarded as a typical disfluency phenomenon. Intuitively, this kind of repetition can be defined as can be characterised informally as rapid fluent repetitions mostly single syllable function words at the beginning of phrases, which I propose to be named "repetitive interpolation". Repetitive interpolations are typically very short, consisting of mostly single syllable function words, and are very rapid, which one may be tempted to relate it to some forms of stuttering. Unlike other disfluency phenomena that introduce apparent disruptions to the delivery of speech, these repetitions are associated with none to minimal disruption of the flow of speech, and are not accompanied by other hesitation or repair phenomena. Some good examples of repetitive interpolations can be found in the following listings. The examples are drawn from SCOTUS 2001 (Yuan and Liberman, 2008) [1].

(2) but **it's it's** really the disability that we're focusing on and in the circumstances someone like that would be able

(3) **i i** didn't find it perhaps you could point me to where that was addressed

(4) though in this case **i'm i'm** sure you would contest it you would say there was no negligence

(5) poor people in the inner city **at at at** relatively low rates the state of ohio adopts a program

In addition to the impression that these repetitions are fluently integrated into the production of speech, the above examples also suggest that the context in which repetitions happen does not show signs of increased processing demand, such as words with lower frequency or more complex phrase structure. With these observations, a simple intuition

[1] Links to supplementary audio files for selected examples can be found in Appendix B

behind these examples is that such repetitions are in fact perfectly integrated into fluent articulations. Anecdotally, these repetitions are often filtered out by the perceptual system, although systematic perceptual studies are still needed to verify this anecdotal impression. Thus intuitively one plausible alternative explanation for repetitive interpolation is that that they are not speech disfluencies at least in the traditional sense.

In the following discussion, I will first describe the method I used to extract repetition examples from Fisher. Then I show that repetitive interpolations can be reliably distinguished from disfluent repetitions through an analysis of the prosodic features of repetition. A survey of the repetition type distribution will also show that repetitive interpolation comprises most cases of repetitions in spontaneous speech. I will provide two further evidence to support the proposal of establishing this new category of repetition phenomenon based on the observation of predominance of single syllable function word repetitions: First, the likelihood of single syllable function word repetition is randomly distributed across word frequencies, while repetitions of content words are correlated with word frequency distribution. The conditional entropy of the lexical context for repetitive interpolation is also lower when compared to the lexical context of same words not used with repetition, suggesting contexts that are more predictable given the existing words.

4.2. Identifying repetitive interpolations

Unlike the relatively clear surface distinction between two forms of filled pauses, repetitions are subject to more variability. On the surface, repetitions can be realized as full word repetition, partial word repetition and repetitions of short phrases. Another relevant dimension is that the number of repetitions a repeating unit has in a repetition. In this study, the identification task focuses on accurately identify the first three potential variations in the realization of repetitions. I primarily divide repetitions into three groups: *full word repe-*

tition, *partial word repetition*, and *other repetition*. As the names suggest, the distinction between the first two categories is mainly about whether the repeated segments are full words or partial words. The third category *other repetitions* is intentionally vaguely defined, as a better characterization and sub-classification within this category requires joint consideration across multiple disfluency types, while the goal of the current research is to offer in-depth type dependent analysis of different repetitions, although the practice of creating a garbage category in the building of disfluency classification systems is criticized by Shriberg (1994). Examples of other repetition include partial phrase repetitions and repetitions that include intervening filled pauses or other hesitation markers. Examples of each type of repetitions are given in Table 6. To further simplify the identification process such that a clear first picture of the phenomenon can be drawn, I restrict the discussion of full word repetitions to the instances which contains repeating the same word twice. Other dimensions of repetition, such as the number of repeats in a repetition, will be saved for future research.

4.2.1. The repetition identification algorithm

Identifying repetition instances can be formulated as a string matching problem given that detailed hand annotations for disfluencies are not available. In this study, repeated segments are identified using a modified version of the suffix tree algorithm (Weiner, 1973). A suffix tree is a trie data structure where all the suffixes of the string are represented as keys and the positions as values. A graph showing an example representation of the word “banana” using a suffix tree is shown in Figure 28. This algorithm can efficiently find duplicated characters in the suffix of a substring in linear time, with worst case time complexity being $O(n \log n)$ where n is the length of the string.

The main modification to the naive suffix tree implementation is that instead of representing each character in the string as keys for tree nodes, the minimum unit in a transcribed

Table 6: Classification of repetitions in this study.

Type	Description	Examples[*]
Full word repetition	Repeating complete word or phrase 2 or more times	it's a different context but **if if** it's something... and **it's it's it's** a big question... oh yeah **she loves she loves** the cats...
Partial-word repetition	Part of a word is repeated before the full word is delivered	if it's like in a **s- sexual** thing i think it that's where i draw the line... i'm not sure if **the qu- the question** i think says...
Other repetition	Partial phrase repetition, where the last word is replaced, or repetitions involving intervening filled pauses, or other situations involving repeating part of a phrase or word that not covered by the other two classes	instead of doing that **they'll play sil- they'll play politics** and say... **that's ah that's** true although it can be hard in our family sometimes... it is and **it is fascinat- it's no less fascinating** to watch...

[*] Examples are from Fisher corpus.

turn is set to be a word separated by space. The string matching problem is then defined as a word matching problem seeking to find the common sub-turn plus T words following the matched sub-turn. T additional words are considered in the matching algorithm so that partial word or partial phrase repetitions can be captured, as well as repetitions with filled pauses in the interregnum.

The algorithm runs as the following: Turns in the corpus are initially represented as lists of words. A search window of size N is firstly defined so that repetitions involving a phrase of up to N words can be captured. Before applying the suffix tree algorithm, a

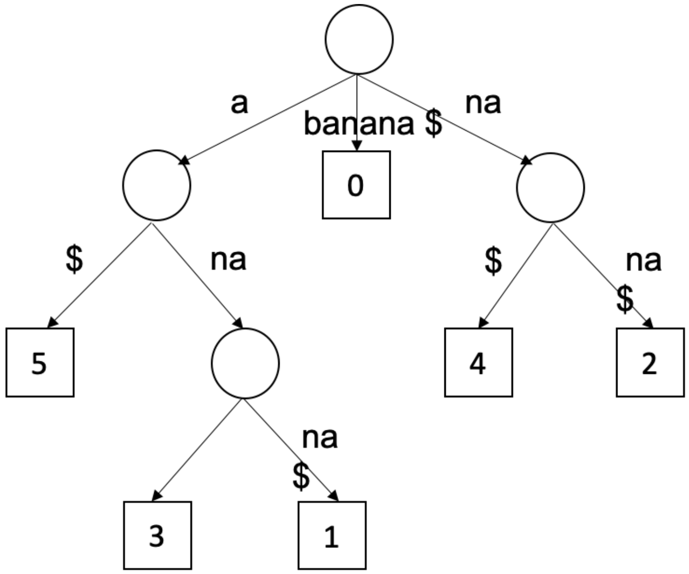


Figure 28: The suffix tree representation of the word “banana”. In this data structure, each edge represents a suffix. $ represents the end of string, and the numbers in square boxes indicate the starting index of the suffix.

turn is represented as a string of letters. A look-up table for each turn is made, which maps each word in the turn to the index of word-initial letter in the string. A suffix tree is then constructed for each turn based on the representation with individual letters as the minimal unit. Exhaustive searches are then performed starting from the left edge of the string using the suffix tree. A matching string is constructed using the words incrementally from the current window in the current turn. The search stops either when one or more matches are found or all the words in the current window have been added to the matching string. Then the left edge of the window is moved to the next word. Only the matched string that immediately follows the right edge of the matching string would be returned as the repetition of the matching string. A tolerance value T is introduced to account for non-exact matches, as described above. Therefore the returned repetitions include the repeated words plus T letters. The matching and matched strings are combined and translated back to an ordered word list through the look up table. If T is smaller than the length last word to be included in the returned word list, the last word is preserved.

Table 7: Performance check of the repetition identification algorithm.

Repetition type	Full word repetition	Partial word repetition	Other repetition
Instances	104,685	24,457	53,313
Precision	0.894	0.914	0.634

Classification of repetitions was achieved by examining the returned repeated word lists using two straightforward rules: full word repetitions contain only complete words and the smallest repeated units. Partial word repetitions contain word fragments (indicated by "-" in the transcription) at the end of the smallest repeated units. All other repetitions are classified as *other repetitions*. This procedure identified 104,658 full word repetition instances, 24,457 partial word repetition instances, and 53,313 other repetition instances. This simple search algorithm, however, should not be expected to identify all the instances of *other repetition* from the corpus, due to the high variability of such repetitions. Nevertheless, it should be able to cover the majority of full word repetitions and partial word repetitions. The performance of this identification mechanism is evaluated by measuring the precision of a random sample of 500 instances from each repetition category. The evaluation results are reported in Table 7.

An error analysis suggests that false positives in identified full word repetitions are mostly of two kinds: emphatic repetitions and floor holding repetitions, where the repeated phrases are filler words such as *right right* and *yeah yeah*. For partial word repetitions, errors occur when the identified partial word repetition is in fact part of some more complex repair structure, such as those involving insertion and deletion. Nevertheless, false positives in these two types of repetitions can be relatively easily identified. The precision indicates that this automated method can generate a relatively large sample of accurate full and partial word repetitions. Since the matching algorithm used in this study is greedy, i.e., all repetitions within the window for comparison are captured, with the identification check accuracy from a random subsample, using repetition instances identified to construct

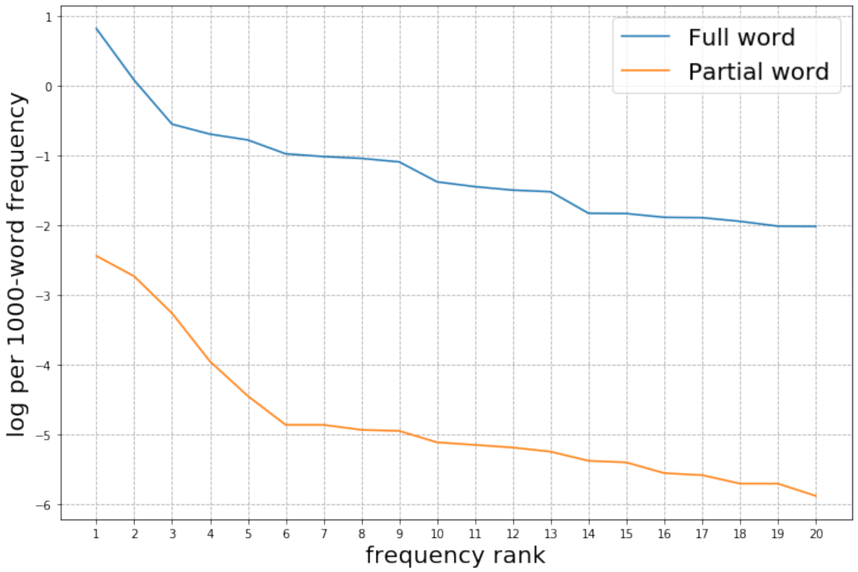


Figure 29: Lists of the most frequent full and partial word repetitions.

an analysis sample through this naive approach should not be expected to cause the problem of high bias.

Two potential realizations of repetitive interpolation based on this surface form classification system are full word repetition and partial word repetition. Partial word repetition is not yet ruled out from the analysis because repetitions of the word initial syllable could potentially be a realization of repetitive interpolation. In the following section, I will address the question of whether partial word repetition is a common form of repetition as the repetition of single syllable phrase initial words.

4.2.2. Frequency distribution of repetition forms

This section addresses the question of what is the distribution of partial word initial syllable repetition compared to single syllable word repetition, as they are the two potential candidates for repetitive interpolation identification. An understanding of this distributional contrast is crucial in formulating hypotheses about the underlying production mechanism responsible for repetitive interpolation: Is the minimal repetition unit a single syllable, re-

Table 8: 20 most frequently repeated full and partial words.

Full word	’i i’, ’and and’, ’the the’, ”it’s it’s”, ’that that’, ’a a’, ’it it’, ’they they’, ’you you’, ”i’m i’m”, ”that’s that’s”, ’to to’, ’in in’, ’is is’, ’if if’, ’what what’, ’my my’, ’we we’, ’but but’, ’or or’
Partial word	’y- you’, ’an- and’, ”i- it’s”, ”th- that’s”, ’ha- have’, ’e- even’, ’may- maybe’, ”y- you’re”, ”i- i’m”, ”i d- i don’t”, ”th- there’s”, ’j- just’, ’th- this’, ’s- some’, ’th- there’, ’o- other’, ’li- like’, ’be- because’, ’pe- people’, ’e- especially’

gardless of word boundary, or the repetition is indeed constrained by word boundary? The answer to this question is pertinent to the hypothesis about what is the minimal planning or activation unit in repetitive interpolation, which can further shed light on theories about planning itself and the transmission between stages of speech planning.

The frequency distribution of the 20 most frequently repeated single words and partial words is plotted in Figure 29. The list of most frequently repeated words is manually checked and repetitions for emphatic purposes (such as “very very”) and other non-fluency related cases (such as “the number is nine nine one”) are eliminated. Variants of transcription of same partial word, such as “i- it’s” vs. “it- it’s”, are also collapsed. The list of repeated words and partial words are summarized in Table 29.

Figure 29 suggests that the frequency of partial word repetition is much lower than that of single syllable word repetition. More interestingly, the word list, as in Table 8, shows that even among partial word repetitions, based on transcriptions alone, the majority come as repetitions of single syllable words but with somewhat reduced phonetic realization in the initial utterance. However, these repetitions can potentially just be the result of annotator variation or annotation errors. Nevertheless, the graph and table suggest that repetitions of the initial syllable of multisyllabic words are rare.

The comparison between full word and partial word repetition suggests a potential word boundary constraint in deciding the minimal unit for repetition. I will delay further discussions on this interesting observation until the next chapter, when I have established the acoustic and textual properties in support of treating repetitive interpolation as a distinctive category of repetitions.

4.3. The acoustic evidence of repetitive interpolations

The acoustic properties of repetitive interpolations are discussed through examining their prosody. Two prosodic features: the duration and F0 of relevant segments are examined, following the methods used in Plauché and Shriberg (1999) and Shriberg (1995). I will show that considering the proposed acoustic feature space, repetitions in general can be classified into three distinct groups: fluent repetitions, repetitions with a long silent pause between the repeated unit (delayed repetition), and disfluent repetitions, which occur adjacent to apparent hesitation or repair disfluencies. The main distinction between fluent and delayed repetitions is the silence duration between two repeated words, where delayed repetitions have sufficient duration for replanning. Examples of the three possibilities are listed in the examples below.

(6) ...doing that and even **as as** much as you try **to to** help them...

This is a typical example of fluent repetition under the classification criteria proposed in this section. The repeated words, **as** and **to** are perceptually well integrated into the flow of speech such that they do not cause delay and are not part of apparent hesitation or repair made by the speaker.

(7) ...i've decided to go back **and and** redo it...

The example above is what I term as delayed repetition, where the second repeats of the word **and** is the resumption of continued delivery of the speech after a brief silent pause, normally of 150 ms or longer.

(8) ...their thoughts on marriage **and and** ah ah i want to...

(9) ...all the money **that that**- the money is definitely a positive...

The two examples above illustrate the third possibility, where I term them disfluent repetitions. These repetitions are either hesitant (as in example 8) or part of a repair (example 9).

As reviewed above, one problem with models based on the disfluency assumption of repetitions is that they are unable to offer an adequate account in explaining the lack of delay in fluent repetitions for planning or replanning. In addition, I will also present evidence that repetitive interpolations do not modify the duration of repeated segments such that these repetitions cannot be alternatively viewed as elongation. F0 reset between the second and first repeat is also not observed in the repetitive interpolations. Therefore repetitive interpolation is likely to be driven by factors other than those underlying disfluent repair phenomena. This prosodic analysis thus supports the need for alternative accounts for repetitive interpolations.

4.3.1. Method

The annotation for different parts within a repetition segment is adopted from the model used in Levelt (1983) and Plauché and Shriberg (1999). As illustrated in Figure 30, the segments of interest include pauses before (P1), between (P2) and after (P3) the two repeats, as well as the first (R1) and second (R2) occurrence of the repeated phrase. The delayed repetition corresponds to long P2 under this system.

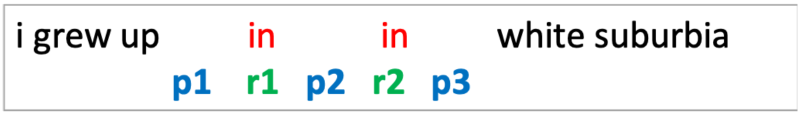


Figure 30: The structure of repetitions.

Table 9: Criteria used for repetition subclassification.

Repetition type	**Criteria**
Fluent	$P2 < 150ms$ or no perceptible pause between repeats; No other disfluencies in the same utterance.
Delayed	$P2 \geq 150ms$ or perceptible pause between repeats; No other disfluencies in the same utterance;
Disfluent	The repetition is part of other disfluencies in the same utterance; The repetition is immediately following, or followed by other disfluencies.

A subsample from the Fisher corpus (Cieri et al., 2004) was used for the current study. Speech from 200 native speakers of American English has been randomly selected to construct the analysis sample. Repetition instances were first automatically identified using the suffix tree algorithm described earlier. Then for each speaker, up to 5 random instances of repetitions were selected after manual examination and correction. The selected examples were further classified into three types of repetitions: *fluent*, *delayed*, and *disfluent*, based on the criteria summarized in Table 9. A total of 743 repetition examples were collected and classified for the analysis.

The threshold of 150ms silence between repeats was selected to acknowledge both the perceptual effect of a silent pause and the expected time that at least motor replanning would take Civier et al. (2010); Postma (2000). This threshold can be justified by the observation that P2 duration distribution is in fact bimodal, with a trough located around 200 ms, as shown in Figure 31. Therefore the *delayed* type can be regarded as representing repetitions that are potentially caused by minor breaks or brief hesitations. This is in contrast to repetitions that are co-occurring with other disfluency phenomena, which are supposedly linked to disruptions involving larger discourse structure.

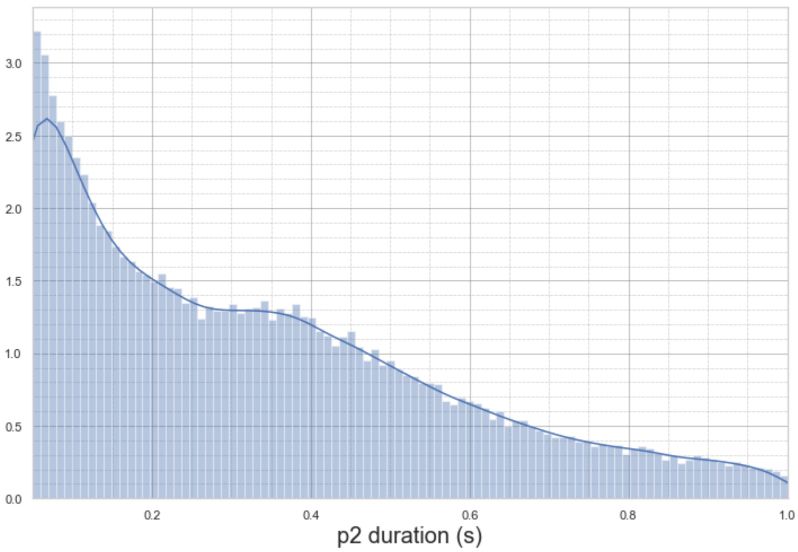


Figure 31: The bimodal distribution of P2 duration.

4.3.2. Acoustic measurements

Duration measurements were based on forced alignment result using the HMM-based Penn forced aligner. F0 measurements were obtained using a pitch tracker which implements the auto-correlation method described in Talkin (1995). Raw F0 measurements were smoothed via quadratic interpolation. To compare the F0 contours between R2 and R1, the smoothed F0 curves were projected onto an orthogonal functional basis defined by the first five Chebyshev Polynomials of the First Kind:

$$T_n(z) = \frac{1}{4\pi i}\oint \frac{\left(1-t^2\right)t^{-n-1}}{\left(1-2tz+t^2\right)}dt \tag{4.1}$$

A plot of the basis functions is shown in Figure 32. The coefficients attached to the basis functions after linear projection can get natural interpretations related to the overall F0 height, slope and higher order curvature of the contours under comparison.

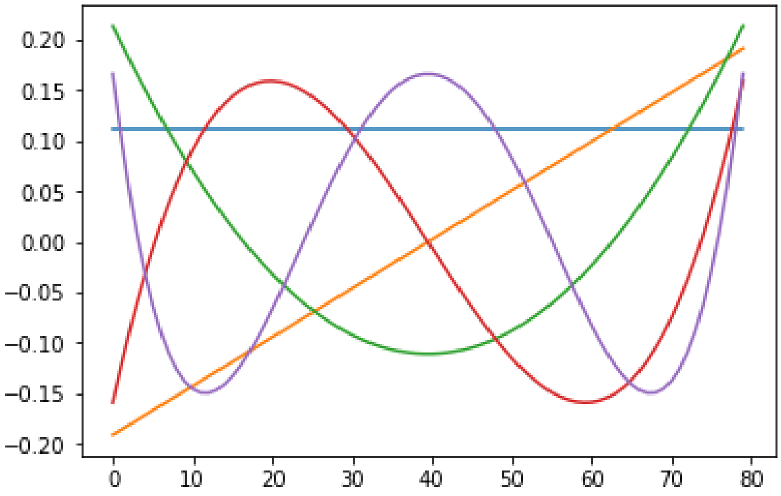


Figure 32: Plot of the shape of the basis functions.

4.3.3. Results

Results of this analysis are presented through answering the following three specific questions: 1. Can the artificially defined P2 duration predict the silence duration following a repetition (aka, P3 duration)? 2. Can the artificially defined P2 duration as well as fluency criteria predict the relation between R1 and R2, measured as their ratio? And 3. Can the artificially defined P2 duration as well as fluency criteria predict the F0 contour shape of the repeated words? It is expected that fluent repetitions correspond to short P3 duration, and no obvious modifications to the duration and F0 contour in a repetition. Thus the fluent repetitions defined here arguably belong to repetitive interpolations.

The distribution of repetition types in the speech sample is summarized in Table 10. The table shows that the majority of labeled repetitions are either fluent or containing some delay between repeats in otherwise fluent utterances. Since our classification criteria subjectively define the silent pause duration between two repeats (P2) and subsequently use this duration as defining characteristics between fluent and delayed repetitions, the variation of P2 duration is well predicted within each class. On the other hand, P1 is often

Table 10: The distribution of repetition type and repeated phrase type.

Repetition type	**Count/%**	**Single Syl. %**
Fluent	390/52.5	87.2
Delayed	164/22.1	82.3
Disfluent	189/25.4	86.2

represented as the silence of a long break by design, therefore its duration distribution is also known to have limited variability within each repetition type. Therefore the real question is whether the duration of P3 varies across the three possible types of repetitions. The other two variables worth considering are the relative duration and F0 trajectory change between R1 and R2. They are related to measures of prosodic boundary: repetitions that across a prosodic boundary are expected to have F0 and duration reset in R2. Therefore using these prosodic cues we can test the hypothesis that the three types of repetitions are subject to different disruptions during production. In addition, longer R1 compared to R2 can also indicate hesitation. We would predict that *fluent* type is at least not directly linked to hesitation or speech error, while *delayed* and *disfluent* types are related to minor and/or major (discourse structural) hesitation or repair.

Pause duration after repetition If the hypothesis holds, it can be expected that fluent repetitions will on average have very short delays in P3, while disfluent repetitions will have the longest. Delayed repetitions would have a P3 duration somewhere in between. Figure 33 plots P3 duration in three types of repetitions.

As expected, there is an increase in the overall P3 duration from *fluent* to *disfluent* repetitions, with the *delayed* somewhere in between. Most noticeably, the P3 duration is almost always shorter than 200ms in the *fluent* type, and the 4th quartile in the *delayed* type is also only 180ms. These short pauses are not sufficient for making an alternative speech plan or even motor plan. Therefore fluent repetitions are not only fluent in the sense that words within the repetition is free from major delay, but they are also an integrated part of

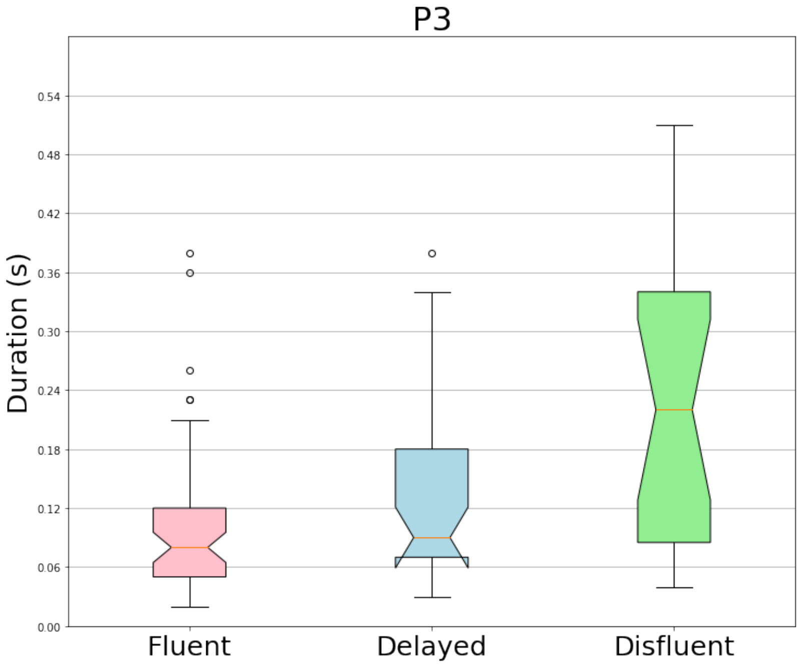


Figure 33: Silent pause duration following repetitions.

the overall fluent utterance free from other kinds of delay or disruption. On the other hand, the median duration of P3 in *disfluent* repetitions is over 200 ms, suggesting the possibility of some major planning between the end of repetition and the following delivery of speech. Thus observations from P3 duration distribution support our hypothesis.

Duration of repeated words Duration relationships between R1 and R2 can offer another piece of evidence to test whether repetitions involve major disruptions of speech delivery potentially caused by replanning. Under Hieke's dichotomy of retrospective and prospective repetition, longer R1 should be related to some latent repair process right before articulation, while longer R2 indicates hesitation. This distinction has been supported by Shriberg (1995) using Switchboard. Here we would like to ask if this dichotomy can be observed from our proposed repetition classification.

Figure 34 plots the density distribution of duration of R1 against R2 across the three types. It can be observed that in the *fluent repetitions*, the distribution of R1 duration is

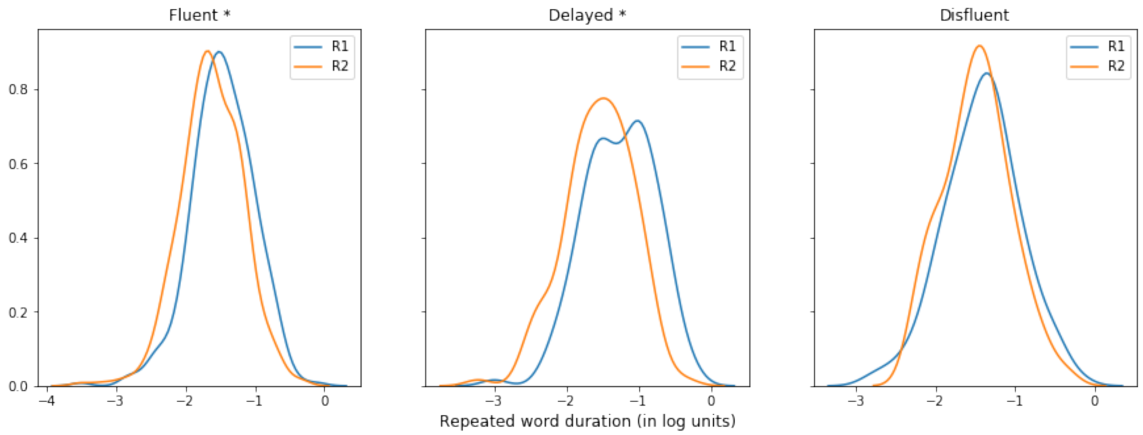


Figure 34: Duration difference between R1 and R2.

shifted towards right, showing a small but significant difference from the duration of R2 through a Wilcoxon signed rank test ($p < 0.001$, statistic=20255.5). This suggests that R1 is on average slightly longer than R2. Similar trend can be observed between R1 and R2 in *delayed repetitions*. This difference is also significant with Wilcoxon signed rank test ($p < 0.001$, statistic=1937.0). Notice that the distribution of R1 duration is bimodal, which may indicate different functions that such repetitions exert in different communication settings. Potentially they correspond to the cases of repetition with and without hesitation. However, in *disfluent repetitions*, the distributions essentially overlap with each other, and here the difference is not significant ($p = 0.2$, statistic=3830.5). This observation suggests that although for each R1-R2 pair, the duration is more likely to show larger difference compared to the fluent repetitions, there isn't a predominant trend: both hesitation or repair can be present in this type of repetition. The average non-significant difference between R2 and R2 also indicates a general reset of word duration in these repetitions, which may suggest a form of replanning.

The distribution observed from *fluent repetitions* is not likely to be associated with retrospective repetition as proposed in Hieke (1981). Comparing this graph with what have been reported in (Shriberg, 1995), where the duration difference is often found to be greater than 0.1s, the duration difference between R1 and R2 in our sample is quite small, such that the extra time is not sufficient for restructuring or planning for the coming repair. Thus the slight but consistent reduction in R2 duration can be better explained as shortening of R2 due to reduction caused by repeating the same word twice, rather than lengthening of R1. In addition, two ridges can be observed from the *delayed repetition*: one almost aligns with the diagonal, and the other is off diagonal in the lower right corner. This may suggest a further split within the delayed category: some repetitions are probably indeed caused by retrospective repair, while others are more similar to *fluent repetitions*.

4.3.4. F0 contour of repeated words

The next test of our hypothesis that the three repetition types are related to different fluency problem can be found through comparisons between F0 contours in R1 and R2. In the functional space defined by the first 5 Chebyshev Polynomials of the First Kind, as described above, coefficients attached to polynomials can get intuitive interpretations related to the shape of F0 contours. The first coefficient can be interpreted as representing the overall F0 height, and the second coefficient can be understood as representing the average slope. Coefficients of the third and higher order functions are then representing the more complex curvature properties. Pairwise Wilcoxon signed rank tests were run for each repetition type and in all five dimensions, but significant difference was only found in the first dimension. Therefore the change in F0 contour between R1 and R2, if there is any, is mainly present in terms of the overall F0 height, but not the slope or other higher order aspects of F0 curvature.

Figure 35 plots the coefficients of R1 in the first dimension against R2 across three types. The dotted line in each graph shows the estimated linear regression line. In both *fluent* and *delayed* repetitions, significant difference was found in the pairwise comparison between R1 and R2 ($p < 0.001$,statistic=19787.0 and 3747.0 respectively from Wilcoxon

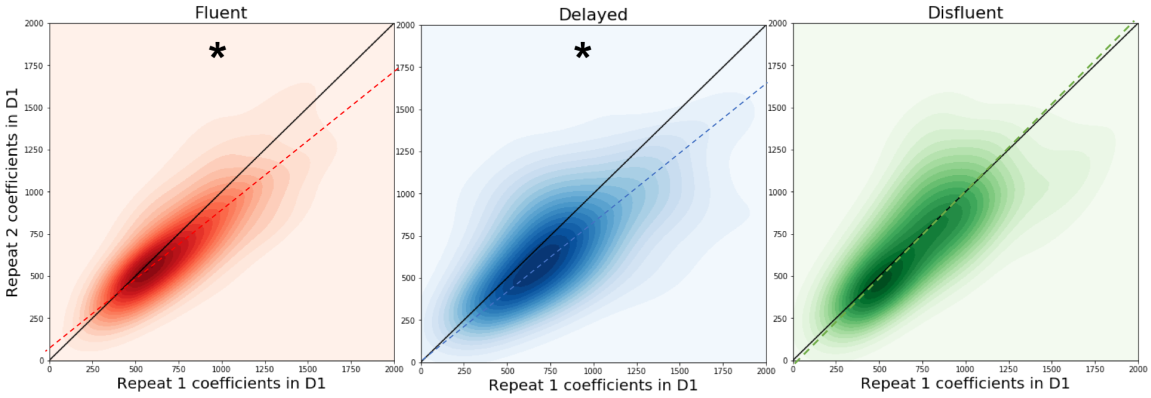


Figure 35: F0 difference in contour between R1 and R2.

signed rank test), suggesting that the overall F0 height is lower in R2 compared to R1. The estimated regression line also clearly shows an off-diagonal trend in both types. However, significant difference was not found in *disfluent repetitions* ($p = 0.203$,statistic=4072.0), and the regression line almost overlaps with the diagonal. Therefore the overall F0 height is about the same in R1 and R2 in this condition.

Lower F0 in R2 can be interpreted as relating to pitch lowering in speech production. The lack of pitch lowering in the *disfluent repetitions* indicates the potential existence of F0 reset in the second repeat within a repetition. Therefore it is probable that the second articulation of a same form is the result of replanning. On the contrary, the descending in F0 height in the other two types suggests the lack of reformulation of an utterance plan, at least at the level of some prosodic phrase. Thus the trend of F0 contour change reported here supports the proposal that fluent and disfluent repetitions involve different underlying production processes.

4.3.5. Discussion

In this section, I tested the hypothesis that acoustic features can be used to subclassify repetitions. Although similar approaches have been taken in previous studies, here I specifically examined whether the silence duration between repeated segments (P2 duration) correlates with other duration or F0 features. P2 duration is selected following the assumption that silent pauses within repetitions correlate with the hesitation or replanning in speech disfluencies. Other prosodic measurements, such as the F0 of repeated segments, silence duration following the repetition, and the relative duration of the repeated segment in a repetition, are also supposedly indicators of the presence of hesitation and repair. Thus the proposal for a separate fluent repetition category, the repetitive interpolation, would be supported if short P2 duration could predict other acoustic measurements that also indicate fluency.

Three arbitrarily defined types of repetitions were manually annotated based on P2 duration. The results reported so far all point to a clear prosodic distinction among the three types. In particular, all three measures: the silence duration following the repetition (P3 duration), the duration and F0 height difference between two repeats (R1 and R2), all suggest that the *fluent repetition* is distinctive from the other two types. However, the *delayed repetition* should be viewed as a mixture of more fluent repetitions and those that are caused by minor repair, in the sense of Hieke's retrospective repetition. The distribution of the three types in our sample shows that most of repetitions are either *fluent* or *delayed*, with fluent repetitions having the largest share. Thus it could first be argued that most of the repetitions found in spontaneous conversations are likely to be fluent. Characterizing all repetitions indiscriminately as disfluent is therefore not entirely appropriate.

With the acoustic descriptions presented so far, repetitive interpolation as a separate category of repetition phenomena should be motivated. In addition, the analyses based on a random sample of telephone conversations from the Fisher corpus found evidence that most repetitions are in fact not disfluent in the sense of being the result of hesitation or the need for repair. A closer look at other properties of such repetitions, such as the distribution of repeated words and the lexical contexts, would be necessary for formulating hypothesis on an explanation to this phenomenon. However, the current analysis is unable to draw a more clear picture of how such a theory would potentially look like.

4.4. The frequency distribution of repetitive interpolations

This and the next section address the textual aspect of fluent interpolations. As mentioned earlier, one interesting cross-linguistic observation for repetitions is that the repeated words are mostly single syllable function words. In addition, the relationship between phrase structure and frequency of repetition has been extensively discussed in the literature. As discussed earlier, the majority of repeated words in our current speech sample are indeed single syllable function words, and repetitions of single syllable partial words are also relatively rare. This observation motivates the question of whether repetition, or more likely repetitive interpolation, involves planning at phrase level or this trend is purely due to the high frequency of such words. Thus in this section, I'm asking if the high frequency of repetitions of single syllable function words is due to the mere fact that function words are much more frequently used in speech, or the different structural role played by different word categories.

One approach to understand this property of the distribution of repetitions in spontaneous speech is to examine whether the likelihood of a word being repeated is correlated with word frequency. If a higher frequency word repetition is mainly caused by the fact that this word has higher absolute frequency, then repetitions would likely be a random event that is irrelevant to the more complex planning process, such as those proposed to be related to hesitation and repair phenomena. In particular, since single syllable function words occur predominantly at the beginning of a phrase, and edges of a phrase are more prone to production error (Levelt, 1989; Clark and Wasow, 1998) that may or may not be resulted from planning issues, it is necessary to address if higher frequency of repetition of single syllable function words is indeed naturally expected. On the other hand, if the likelihood of repetition is not random when conditioned on the frequency distribution of words, other processes are then needed to account for this observation.

The distribution of the likelihood of observing repetition when the word is given can provide further support for the proposal that repetitive interpolation is a distinct fluent speech phenomenon that should receive different treatment compared to repetitions that are disfluent. As reviewed in Chapter 2, disfluencies are typically driven by the need for hesitation or repair, which is further led by the processing need in speech planning. Thus I would expect some frequency effect that affects the relative frequency of repetition compared to a word's absolute raw frequency. On the other hand, if the processing demand doesn't affect the likelihood of a word being repeated, the repetition itself is less likely an response to higher level planning process; rather it is an response that requires additional explanation from stages of production process that involve lower level processing.

The discussion in this section is carried out as the following. I first show that the overall distribution of the relative frequency of repetition compared to the absolute frequency of a given word, measured as the log-odds of repetition, has a non-uniform relation with regard to word frequency. In particular, different trends can be observed in high and low absolute frequency regions. Then I further inspect the hypothesis that this bipartite relationship can be partially explained by word class when the syllable structure is controlled: the relationship between the log-odds and word frequency appears to be random among function words, but linear among content words. I will argue that one potential interpretation for this observation lends support to the proposal for repetitive interpolation being a sign of fluency.

4.4.1. Overall distribution of log-odds of repetition against word frequency

A simple yet robust text-based test of hypotheses regarding whether repetitions are caused by similar factors as hesitation or repair disfluencies is the relationship between relative frequency of repetition and the absolute frequency of the repeated words. A lack of sys-

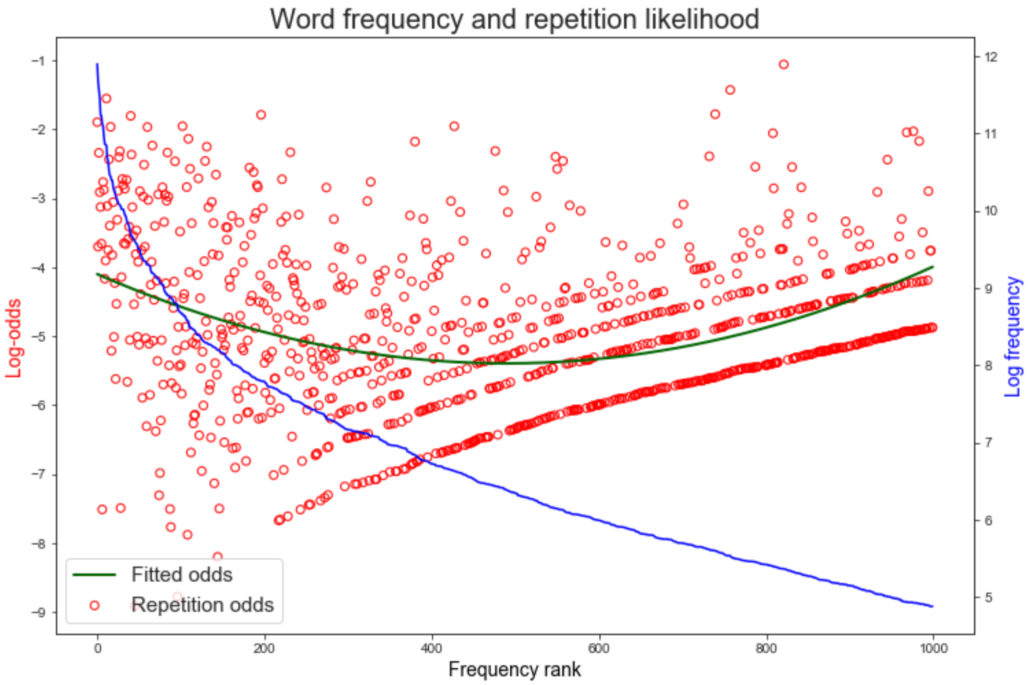


Figure 36: The relation between the log-odds of word repetition and word frequency for all words. The cut-off frequency rank is 1000.

tematic relation between the two variables could suggest that repetitions do not depend on properties of the lexical item, thus it's less likely the result of higher level planning problems typical of most disfluencies. This reasoning holds with the assumption that lexical frequency is indirectly related to the processing load required for the lexical item.

The relative frequency of repetition for a given word is calculated as the log-odds. The log-odds of repetitions of given words is calculated as the logarithm of the likelihood ratio for a word being repeated divided by the likelihood of the word not being repeated. The equation for this calculation is listed in equation 4.2:

$$log\ odds = \log\left(\frac{P(repetition_i|w_i)}{1-P(repetition_i|w_i)}\right) \tag{4.2}$$

where w_i refers to the ith word in the vocabulary, $repetition_i$ refers to w_i in repetition, and the log transformation is the natural log.

Figure 36 plots the distribution of the log-odds of word repetition as a function of word frequency for the first 1000 words ordered by their frequency. The phonological structure of the words are not yet considered. The overlaid blue curve represents the relationship between word frequency rank and the log frequency of a given word. The overlaid green curve is the fitted quadratic function of log-odds against frequency.

This plot shows that for words with higher frequency, especially for the first 300 most frequent words, the relationship between word frequency and the log odds is largely Gaussian: the distribution of the likelihood for a word being related does not seem to depend on its frequency, especially with the high variance for very high frequency words. The high variance drafts the significance of the negative slope in high frequency region of the regression line. For words with lower frequency, however, a slight inverse correlation, i.e., lower frequency words tend to have higher log odds of repetition, can be observed. This is suggested by the positive slope of the fitted regression line in the lower frequency region of the graph. One cautionary note for this observation is the stratification of the data points in this graph, which may indicate a lack of generalisation power due to limited sample size when the frequency count is low. However, the in-sample trend is still apparent.

One hypothesis from the trend shown in this graph is about repetition frequency and the word class of the repeated words. The distribution of different word classes in different word frequency regions partially explains the bifurcated relationship between word frequency and the log odds of a word being repeated, since function words, especially single syllable function words, are concentrated in the higher frequency region. The log odds for a given function word being repeated can be expected to be constant regardless of its word frequency. On the other hand, rarer content words have higher odds of being repeated, which may be related to word frequency effect and other syntactic or semantic correlates

Table 11: Word lists for monosyllabic function and content words.

word class	words
function words	i, and, the, it's, that, a, it, they, you, i'm, that's, to, in, is, if, what, my, we, but, or, so, they're, he, of, i've, are, how, for, there's, on, you're, this, with, just, where, she, we're, as, there, he's, your, their, do, at, then, who, some, one, our, from
content words	big, cause, go, long, most, true, real, see, six, make, way, got, say, great, same, let, ten, five, take, new, since, four, eight, huge, look, back, first, things, nine, hard, each, down, think, wait, kids, life, young, give, help, put, bad, work, want, try, come, part, went, high, made

of these words. Another statistical measure of this hypothesis is whether relative repetition frequency is correlated with absolute word frequency. Low correlation could indicate the existence of a word frequency effect.

4.4.2. Repetition frequency distribution controlled for the word class

To test the hypothesis that the relation between repetition frequency and absolute word frequency is dependent on word class, I compare repetitions of single syllable function words against single syllable content words among the most frequent words in each word class. Thus the words being compared is relatively phonologically comparable. The hypothesis is that if the repetition of the two broadly defined word classes is caused by the same underlying mechanism, similar frequency effect should be observed. That is, we would expect some relationship between word frequency and the frequency or likelihood of seeing the word in repetition, and the direction of this effect should be same with comparable effect size.

A sample comprising the 50 most frequently repeated single syllable function words and content words has been selected from all the repeated words. The selection of word lists excludes repetitions that are for rhetoric purposes (such as "good good") and floor holding (such as "yeah yeah" and "right right"). The selection also considers only the phonological

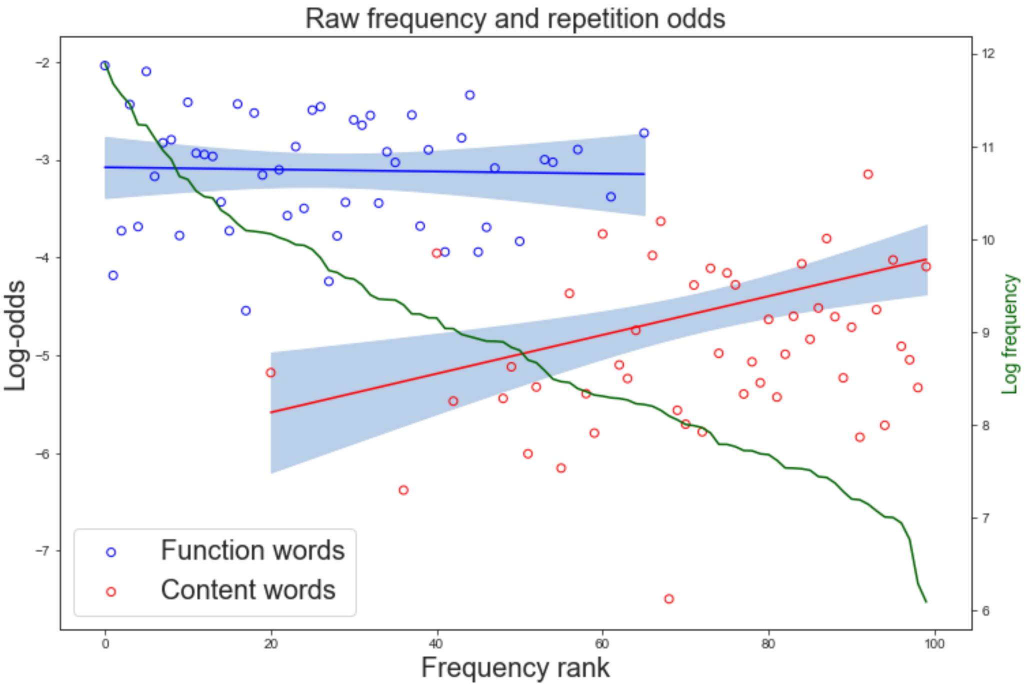


Figure 37: The relation between repetition odds and raw lexical frequency for single syllable function and content words.

forms of the content words; therefore inflections of a same word are treated as separate words. However, different grammatical functions for function words are distinguished, even though two distinct words may share the same pronunciation, such as “they’re” and “there”. The selected words are listed in Table 11. Similar to Figure 36, I plot the relation between word frequency and the log odds of a given word being repeated. Word frequency is ordered by combining the two word lists.

One observation of the selected content words is that they are mostly short verbs or adjectives that appear early in a constituent. Nouns are relatively rare in the list. In fact, only 2 out of the 50 words are uniquely used as nouns. This suggests possible preference for repetition for words toward the left boundary of a phrase, which is consistent with the high frequency of function word repetition in comparison to content words. The absolute frequency count of repeated content words, however, is much lower than function words in general. The lowest repetition count for content words is 8. The frequency counts are obtained from the Fisher sample that the analyses of this paper is based.

4.4.3. Results

Figure 37 plots the relation between repetition odds and repetition frequency. The frequency of selected words is ordered by the raw frequency of combined function and content words. Two regression lines, representing the relation between repetition log-odds as a function of word frequency, are fitted for the two word classes separately. The green line shows the raw log frequency for each word as a function of their frequency rank.

Two interesting trends can be observed from this graph. First, the log odds for single syllable word repetition is higher for almost all the function words. Since these words are most likely at the beginning of a phrase, the graph provides a further piece of evidence that words at the left edge of a phrase are more likely to be repeated. It can also be observed that the log odds for repetition is not dependent on word frequency for function words, but an inverse relation, i.e., lower frequency words have higher repetition odds, can be readily observed for content words, as suggested by the fitted regression lines. These observations are consistent with the first hypothesis based on the overall relation between repetition odds and word frequency as shown in Figure 36. Thus it can be argued that when the phonological structure of the word, here referring to the number of syllables, is controlled, the repetition of function words does not appear to be affected by absolute word frequency, while low frequency content words are more likely to be repeated. Based on these observations, it can be proposed that two factors are at play in determining the relative repetition frequency for a given word: the proximity to the left edge of a phrase and whether the word is a content word. The frequency effect observed among content words is potentially related to the frequency effect in general.

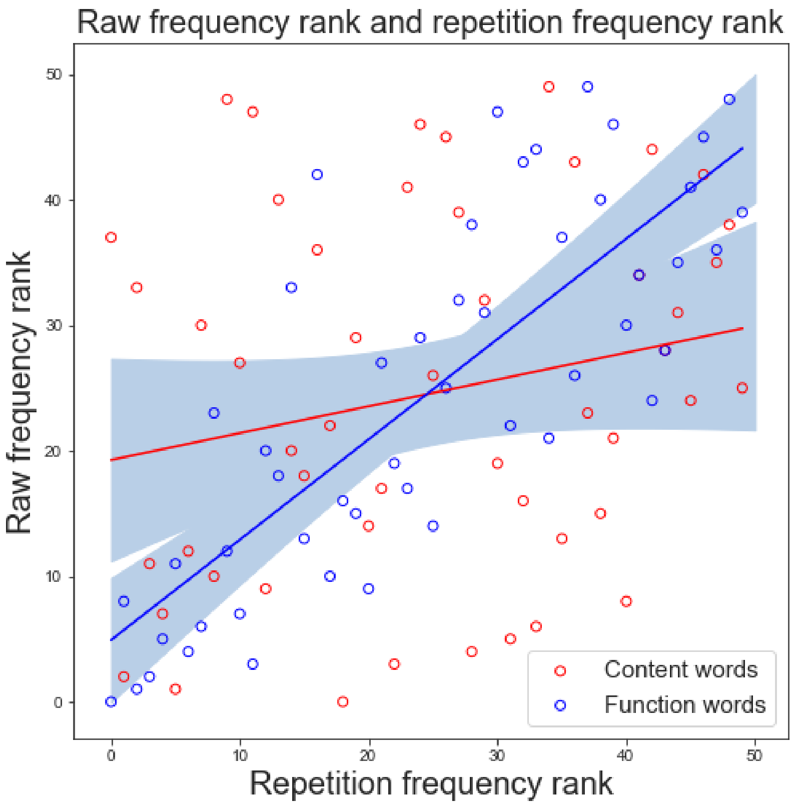


Figure 38: The relation between repetition frequency and raw lexical frequency for single syllable function and content words.

The next comparison to be made is the correlation between absolute frequency and repetition frequency of single syllable function and content words. This measure could reveal the randomness of repetitions in the following way: If the rate of repetition for a word is random, we would expect a strong positive correlation between raw frequency and repetition frequency, and the rank in two frequencies should follow the positive diagonal of the coordinate. If factors other than the frequency distribution of words play a role in determining the relative frequency, the correspondence between two frequency ranks would be expected to be deviated from the strong positive correlation. As Figure 37 shows, there is indeed a strong correlation between repetition frequency and raw frequency in function words, which roughly follows the diagonal, but not among content words. Combining the result from Figure 37, it can be argued that the raw word frequency of function words does not affect the likelihood of repetition for the given word, but it is not the case for content words. Thus this graph offers an alternative view of the distinction between repetitions of function and content words reported earlier.

This is an interesting result, which suggests that the repetition of function words may be only mildly affected by the identity of the word itself. However, considering the selected content words also tending to be at the left edge of the phrase, the repetition of content words can potentially be driven by factors that are associated with word frequency or other features pertaining to the word itself. Thus the repetition of content words is possibly driven by a different mechanism than the repetition of function words, although the syllable structure for the words compared are set to be comparable.

4.4.4. Summary

In this section, frequency related measurements have been examined to address the question of why single syllable function words dominate repetitions. The piecewise correlation function between the relative repetition frequency and absolute word frequency in low and high frequency rank region can be partially explained by the distribution of word classes in different frequency rank. Stronger word frequency effect on relative repetition frequency has been found among single syllable content words compared to function words. Given the observations that most repeated words are single syllable function words, and they are likely to be fluent, the lack of word frequency effect on relative repetition frequency motivates the hypothesis that the generation of repetitive interpolations is approximately stochastic.

The analysis above can be further summarized into the following two points, which have bearing on a possible theoretical explanation for the observations. First, the repetition of function words is more random compared to content words, when the syllable structure of these words are controlled as all the selected words are single syllable. The randomness suggests that features that are related to word processing do not appear to greatly affect the likelihood of a function words being repeated, while this is not the case for content words. Second, the result presented so far indirectly shows that the likelihood of a word being repeated increases when the word is closer to the beginning of a phrase. This is supported by the high log odds of function word repetition compared to content words, as well as the predominantly verbs and adjectives in the most frequently repeated single syllable content word. Therefore, although the repetition of function and content words can be attributed to different reasons, an explanation for both cases requires a component considering the relative position of the word inside the phrase.

4.5. The lexical context of repetitive interpolation

The lexical context of a target word or phrase indirectly predicts the likelihood of encountering planning problems during production (Arnold et al., 2007). One additional piece of evidence that repetitive interpolation is associated with fluency, rather than disfluency, is whether the lexical context for repetitive interpolation appears to be more challenging compared to the context with which the same words are produced without repetition. In the current discussion, a lexical context that requires higher processing demand refers to the context in which the word following the current word is more unexpected. Higher uncertainty in the lexical context can lead to the situation where the speaker is more prone to production errors or delay. In this study, the “unexpectedness” of the following word is measured as the conditional entropy of the next word given the current word. Entropy is a measurement of the amount of bit required to encode the information available from the outcome of a given random variable. The conditional entropy measures the entropy

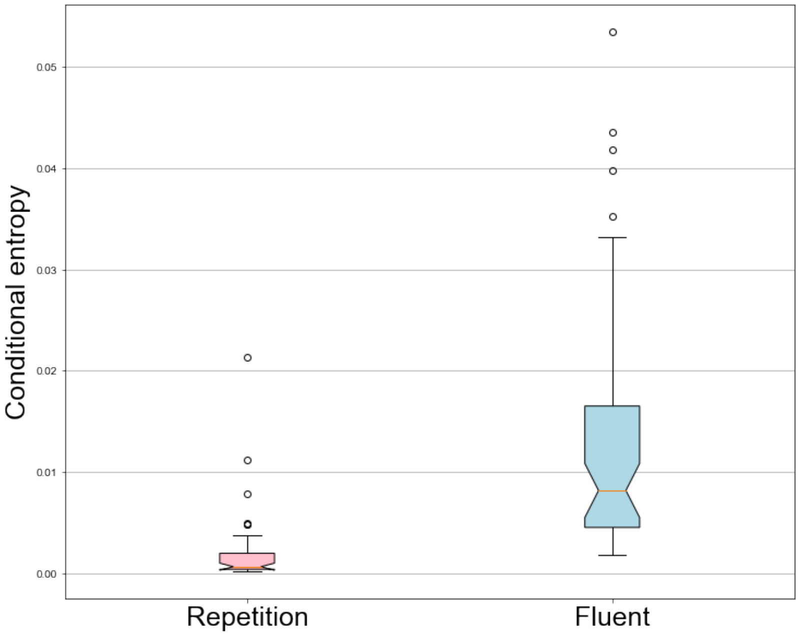


Figure 39: The distribution of conditional entropy of the 50 most frequently repeated words when used in repetition and non-repeating utterances.

conditioning on the results from a conditioning random variable, which can be calculated as the following:

$$\mathrm{H}(Y|X) = -\sum_{x\in\mathscr{X},y\in\mathscr{Y}} p(x,y)\log\frac{p(x,y)}{p(y)} \tag{4.3}$$

In the present case, Y refers to the collection of all possible next words given the present word X, with each individual next word denoted as $y \in Y$. Higher conditional entropy of the following words suggests that more bits are needed to encode the additional information provided by these lexical items, thus the lexical context is more unexpected. This higher level of unexpectedness is potentially related to higher processing demand in the given context. On the other hand, lower conditional entropy indicates that lower processing demand for planning is more likely.

In this section, the conditional entropy is calculated for each current word separately, so the collective notation X always has a cardinality of 1. The specific question to be answered is what is the distribution of the conditional entropy of the following words in repetition and fluent conditions? I use repetitions of single syllable function words as a proxy for repetitive interpolation, with the assumption that most of such repetitions are fluent as defined through their prosodic properties. The conditional entropy of words immediately following the 50 most frequently repeated single syllable function words when used in repetition is compared against the same set of words without repetition. It is hypothesized that if repetitive interpolation signals fluency rather than disfluency, its lexical context would require comparable or lower processing demand thus would show comparable or lower conditional entropy. However, if these repetitions reflect planning difficulty, it would be expected that they show at least comparable or even higher conditional entropy.

Figure 39 plots the distribution of the conditional entropy of the immediate following words in repeating (repetition) and non-repeating (fluent) conditions. This plot shows that, as predicted, the conditional entropy of the following words is on average much lower during repetition compared to their fluent counterparts. Since the majority of repetitions of single syllable function words are repetitive interpolations, as shown in previous sections, it can be argued that the low entropy for the repeating condition is attributed to repetitive interpolation. Thus repetitive interpolation happens in lexical contexts which are more predictable compared to fluent utterances in which repetition is absent. This result further shows that repetitive interpolation is likely not related to elevated processing demand, thus is less likely to be caused by the same underlying factors that induce hesitation or speech repair. Therefore the hypothesis that repetitive interpolation behaves more like fluency than disfluency is partially tested.

4.6. Chapter summary

In this chapter, I presented evidence from the prosody of repetition, repetition word frequency distribution and the lexical context measured by conditional entropy to show that the proposed repetitive interpolation is at least not a typical disfluency phenomenon in spontaneous speech. The observations made so far tend to support the alternative hypothesis that repetitive interpolation is not resulted from hesitation or the need for repair, thus likely to be a fluency phenomenon. Although the largely descriptive analysis presented so far is not yet able to lead to a theory on the underlying production mechanism, it is a starting point for related hypotheses that worth testing in future scientific studies of speech production. Here I attempt to lay out some possibilities, based on the available acoustic and textual observations.

The prosodic analysis of repetitions confirms the intuition that repetitive interpolations are a distinctive group of repetitions from those resulted from hesitation or repair. If the prosodic analysis based on the small but representative sample of spontaneous conversation hold at a larger scale, the current theory based on the disfluency assumption would not be able to account for the majority of repetitions found in spontaneous speech. This directly follows from the violation of the assumption that the existence of hesitation or the need for repair is involved in all repetitions. The existing speech monitoring models, such as those derived from the Nijmegen model of speech production (Levelt, 1989), apparently face the problem of not having sufficient time for the feedback loop to initiate the repair process. It has generally been assumed that the temporal commitment from processing the feedback from some monitor to reformulate an utterance plan is at least 200 to 250ms, which is much longer than our classification criterion for repetitive interpolation, and is also longer than the silence duration immediately following repetitions. Therefore it is not likely that fluent repetitions would trigger speech replanning all the way up to conceptualization.

Alternatively, an explanation from motor planning and execution perspective can potentially be tenable. Under this view, the production of fluent repetitions is fundamentally a motor control problem. Details of some popular speech motor control models are explained in (Guenther, 2006; Bohland et al., 2010; Hickok, 2012). A plausible scenario could be the following: the feedback loop in the motor control system mistakenly detects an error signal in the output from the forward model. This false positive then leads to the execution of an existing plan twice. This process is much faster than the one described previously, with an estimate somewhere between 60 and 120 ms (Civier et al., 2010), and is consistent with the duration of P2 and P3 that we observed from our sample. Unfortunately, relying solely on such a view may still only lead to an incomplete theory: if fluent repetitions are the result of motor control problem, it would not be unreasonable to expect similar repetitions occurring with phonologically more complex words, or smaller phonological unit such as the first or first several syllable of a word. However, this doesn't seem to be the case. As Table 10 shows, the majority of repetitions are single syllable function words. Thus the role of word boundary and word type should be jointly considered. The difference in relative repetition frequency distribution between single syllable function and content words also fails to support a full account from motor planning and execution alone, since the execution of an utterance plan is not supposed to be affected by the lexical category of the phonetic input (Levelt, 1989).

With the considerations sketched above, I propose that to achieve a plausible explanation to these fluent repetitions, the problem should first be formulated as a separate fluency problem concerning a different phase in the process of speech production compared to speech disfluencies. However, higher level processes are expected to affect how likely a duplication of motor command is for a given lexical input. As proposed in Hickok (2012), it is not unlikely that some motor control mechanism can be found in higher levels of the planning process, such as at the level related to morphosyntax or even semantics. Unfortunately, with our limited data, it is not possible to postulate further speculations on whether the duplication of command happens during lexical selection, morphosyntactic structuring or during the transmission of signals between levels of processing. The low conditional entropy of the lexical context of repetitive interpolation may suggest the existence of reduced inhibition and control in the transmission between an utterance plan and motor plan. However, the present analysis is not able to provide evidence to support this hypothesis.

Combining the evidence from these three perspectives, it can be assumed that repetitive interpolations do not involve higher level planning efforts that relate to message formulation. Quite on the contrary, this phenomenon is more likely to partially attributable to reduced motor planning and control, coupled with potential issues in the formulation and execution of an utterance plan. This argument can be further strengthened through a cross-linguistic investigation of repetition phenomena in general. Specifically, as I will show in the next chapter, the same predominance of repetitions of single syllable function words is also observed in Czech, a language that has rich morphology and flexible word order. The frequency distribution of repetitions is similar to that of English as well. Since the realization of repetitive interpolation is arguably similar to certain forms of stuttering, as both involve rapid repetitions of single syllables, and primarily relate to issues in motor planning and execution, one possibility is that both repetitive interpolation and stuttering can be explained with a single model regarding motor planning and execution. In the next section, I will give a brief exploratory overview of the repetitions observable from people who stutter. It will become clear that repetitive interpolation is a fundamentally different phenomenon that is distinctive from the clinical condition of stuttering. Thus separate theories should be developed to account for both repetitive interpolation and stuttering.

Chapter 5

The Role of Motor Control in Repetitive Interpolation

In the previous chapter, I presented evidence from the acoustics, the relative repetition frequency and the lexical context of repetitive interpolation to argue that this repetitive phenomenon should not be considered as a typical kind of speech disfluency. Both the acoustic and textual analyses presented so far hinted that a separate mechanism within the speech production system is responsible for repetitive interpolation. These evidences suggest that higher level processes in speech production such as message formulation is unlikely to be involved in these repetitions. To adequately explain these rather fluent repetitions, it is necessary to introduce a speech production theory that elaborates the mechanism involved in the formulation and execution of a plan for the production of the physical sound of speech. Models on speech motor control in speech production naturally become a candidate. In this chapter, I evaluate the potential role that motor planning and control plays in repetitive interpolation.

Although the descriptive analysis in this study is unable to provide evidence for a detailed production model for repetitive interpolation, I will show that a theory that considers motor control alone is not likely to be able to offer a full explanation of repetitive interpolation. This is demonstrated through a comparison between the realization of stuttering and repetitive interpolation. I will further show that when an utterance plan is readily available, such as in the case of read speech, repetitive interpolation also becomes less likely. Going back to the cross-linguistic observation that repetitions involve mostly single syllable function words, I finally show that in a language with complex morphology and flexible word order, such as Czech, the distribution of repetitions is still similar to what have been observed in English. Thus repetitive interpolation respects word boundaries, not syllable boundaries. This property further suggests that repetitive interpolation could potentially arise somewhere between the formulation and/or execution of an utterance plan and the formulation and/or execution of a motor plan for production.

The structure of this chapter is the following. I first compare the speech from the clinical condition of stuttering to repetitive interpolation. I will show that repetitive interpolation does not feature symptoms such as blockage of speech delivery and effortful speech, which are typical of stuttering. Then I compare the frequency of repetitive interpolation in read speech to spontaneous speech to show that repetitive interpolation is rare when an utterance plan is readily available. This is in contrast to what have been shown for spontaneous conversational speech. Finally, I report a case study of the repetitions in Czech to show that increased morphological complexity and word order are not associated with a different pattern and distribution of repetitive interpolation. I will conclude this chapter by a discussion of potential implications of repetitive interpolation on speech production theory.

5.1. An analysis of stuttering

Stuttering is a fairly common speech disorder seen in children, although only a minority of developmental stutterers have persistent stuttering in adulthood. Stuttering is primarily seen as a speech disorder caused by deficient motor planning and execution during speech production, although the exact cause of the disorder is still unknown. Typical stuttered speech include repetition of sounds, syllables, and words, prolongation of sounds and blocks in speech. These stuttering features typically occur predominantly at syllable initial position or word initial position. Clinical stuttering is said to show decreased disfluency with repeated reading of same passage and tendency for stuttering to recur in the same words/syllables in successive readings of the same text (Bloodstein and Ratner, 2008). However, comparisons between read and spontaneous speech among stutterers have not been explicitly addressed in the literature.

In this section, I would like to offer a description of repetitions in stuttered speech, so that repetitive interpolations found in normally fluent speakers can be properly distinguished from the clinical condition of stuttering. Qualitatively, I will describe the prominent symptoms of repetitions of syllables and phones in stuttering, and propose a crude classification for different symptoms that characterize repetitions in people who stutter. This approach, although subjective, suffices the purpose of drawing an intuitively motivated distinction between repetitions in stuttering and fluent repetition that I am focusing in this study. Quantitatively, I will provide statistics on the distribution of each qualitatively defined feature in the speech produced by people who stutter. It will become clear that fluent repetitions in normally fluent speakers are a distinctive kind of phenomenon in spontaneous speech. The data and illustrative examples are drawn from the UCLASS archive of stuttered speech (Howell et al., 2009).

5.1.1. A description of stuttered speech

As reviewed above, the literature usually describes stuttered speech as displaying repetition of phonemes, syllables or words, as well as prolongations and blocks. These features, however, do not occur independently. The repetition of phonemes and syllables is often accompanied by perceptually salient blocks, such as prolonged silence or noisy non-verbal vocalizations accompanied by uncontrolled facial or body movements, between repeated sequences. Prolongation of adjacent vowel or consonants of repetition is also fairly common. These symptoms are often interpreted as signs of hardships in articulating certain sounds and ways to avoid producing them (Mackay and Macdonald, 1984). These features of stuttering are observed from the selected sample of stuttered speech. Based on these observations, I propose a tentative working classification of three representative kinds of repetitions, based on perceptual judgement, that illustrate the distinction between stuttered repetitions and fluent repetitions in fluent speakers: repetition with block, initial phone repetition without block (Initial NB) and non-initial phone repetition without block (Non-initial NB). It should also be noticed that stutterers do occasionally produce repetitions that are similar to the fluent repetitions produced by fluent speakers. I term it quasi-fluent repetition for the current illustrative purpose and it is counted in later analysis. In the following examples, I give a brief description of the typical stuttering categories that I proposed above. These examples intuitively convey the distinctions that I am trying to make in this crude discrimination.

Repetition with bock can either be repetitions of syllables or repetitions of a single phone with interleaving blocks or non-linguistic vocalizations caused by the following block. An example of this kind of stutter is listed in the example below.

(10) ... and the work condition a s-[noise] a s-[noise] a s-[noise] [breath] a s-[noise]trong belief and motivate...

In this example, the speaker has trouble articulating the consonant cluster /tr/ in “strong”. An abrupt and noisy halting of articulation is immediately preceding this blocking consonant. In addition, as an effort to continue the articulation, this speaker resyllabified the previous indefinite determiner “a” with the first fricative “s” in “strong”. The speech is perceptually effortful and highly disfluent. The presence of blocks like this suggest troubles in continuation after articulating the current phone or syllable, and it is a typical symptom for stuttered speech.

Repetition of a single phone can be surfaced as prolongation, and does not necessarily occur at the beginning of a larger phonological unit. Therefore I distinguish between the remaining two types of stuttering repetition by whether the repeated phone initiates a syllable. I don't make explicit distinction between phone repetition and prolongation, as the underlining problem with the speech motor control is the difficulty of proper execution of the motor commands such that the desired phone or syllable is properly articulated. The cause of such repetition or prolongation is thought to be related to problems in the inhibition in the forward model of the execution of a movement plan, resulting in over reliance on the auditory feedback for excessive correction of the articulation command (Civier et al., 2010). Therefore on the surface, the symptoms could be elongating the phone or making rapid repetitions of it without apparent interleaving blocking sound; rather the block could come after the repetition or prolongation. An example of this stuttering symptom is illustrated below.

(11) what i enjoy the uh ssss- uh ss- uh ssatisfaction is the job satisfaction involved and...

In this example, the speaker had difficulty continuing after the initial consonant /s/ in the word "satisfaction". As an effort to carry on the articulation, he repeated or prolonged the initial fricative, and inserted filled pauses when this effort failed to lead to continuation in articulating the following speech segments. Unlike the first example, repetition of phones do not involve blocks between repeats.

On the other hand, the third kind of repetition, which also involves only repeating or prolonging a single phone, happens at the end of the syllable. Therefore a more proper description would be prolongation of the final vowel or consonant of a syllable or word. An example speech segment is the following.

(12) ...you either have to have a deg- greeee, orrr youuu can have been recommm-mended from all the g-g- officer.

A prominent feature of this speaker is that he prolonged almost all the final vowels, resulting in the speech sounding slow and slurred. In addition, he also prolongs the sonorant coda /m/ in “recommended” and creates a break in the continuous articulation of the word. Some extra non-linguistic vocalizations are also inserted in the articulation, causing unnatural breaks to the delivery of speech.

Although relatively infrequent, stutterers also produce repetitions that on the surface are similar to those produced by fluent speakers. An instance of fluent single syllable word repetition produced by a female adult stutterer can be found in the example below.

(13) owners hope that the customers would would come to to celebrate the the 20 years of of service.

In this example, the speaker repeats almost all the function words that starts a larger phrase. Her repetitions are rapid and do not create apparent delays or blocks in the overall delivery. It should also be noted that this short example is taken from read speech, rather than spontaneous monologue as the previous examples. For the same speaker, the spontaneous speech shows similar repetition pattern, as shown in the following example.

(14) um to- today I have been trying to to paint.

What is interesting is that for some adult stutterers, more frequent repetitions in reading and possibly in different forms are observed. However, repetition in read speech among fluent speakers appears to be rare, although more substantial quantitative research is needed to support this somewhat anecdotal claim. If these speculations are true, then it opens up a new avenue to address the question of why people produce fluent repetitions, and how this phenomenon can help to understand the complicated process of speech production.

Table 12: Distribution of stuttered repetition type by speaker.

Speaker	Block	Initial nb	Non-initial nb	Quasi-fluent	Total Rep.
spkr1	58 (4.8%)	1 (0.1%)	0	18 (1.5%)	77 (6.32%)
spkr2	59 (7.1%)	55 (6.6%)	30 (3.6%)	4 (0.5%)	148 (17.7%)
spkr3	75 (10.3%)	20 (2.8%)	6 (0.8%)	0	101 (13.9%)
spkr4	144 (8.8%)	9 (0.5%)	27 (1.6%)	27 (1.6%)	207 (12.6%)
spkr5	28 (2.9%)	2 (0.2%)	0	0	30 (3.1%)
spkr6	51 (3.7%)	10 (0.7%)	17 (1.2%)	6 (0.4%)	84 (6.1%)
spkr7	15 (6.1%)	2 (0.8%)	2 (0.8%)	2 (0.8%)	21 (8.5%)
spkr8	58 (7%)	7 (0.8%)	3 (0.4%)	10 (1.2%)	78 (9.4%)
mean	61 (6.3%)	13.3 (1.6%)	10.6 (1.1%)	8.4 (0.76%)	93.3 (9.71%)
σ	36 (2.4%)	16.8 (2.05%)	11.5 (1.1%)	9 (0.6%)	56.8 (4.48%)

5.1.2. Distribution of repetitions in stuttering

The goal of this section is to show that the qualitative definition of the repetition types seen in stuttering in fact captures the majority of repetitions in the speech of adult stutterers. Therefore the fluent repetitions observed in normally fluent speakers are an entirely different phenomenon that requires different explanations. Although the underlying cause of stuttering is still an unsolved problem and is an interesting one itself which may benefit from further understanding of the speech production in fluent speakers, it will not be addressed explicitly in this paper.

Monologues of 8 adult stutters in the UCLASS archive are used to show the distributions. The age of speakers ranges from 26 to 48. Repetition types are decided based on the transcriptions of the speech made available by professional speech and language pathologists. Table 12 shows the distribution of repetition types for each individual speaker. The measurement here is the percentage of total words with the given repetition type, along with the raw word counts.

This table illustrates that repetition with apparent interleaving blocks is a very distinctive feature among stutterers, regardless of the severity of the symptoms and their distributions. However, this is not observed in fluent repetitions. Other types of repetitions, especially those without apparent blocks during articulation, are relatively rare, regardless of the base rate of repetition and the presence of great inter-speaker variation. In addition, the overall rate of repetition is higher than the disfluency rate considering all kinds of disfluencies among fluent speakers (which is about 2.5% of all the words in spontaneous speech).

5.1.3. Summary

This cursory exploration of stuttering shows that a prominent feature of stuttered speech is some signs of difficulty in "getting things out". This has been largely assumed as a deficit in the motor controller, and could be surfaced as rapid repetitions of particular phonemes, prolongation, perceptually salient blocks in the delivery of speech, etc era. Although occasional rapid repetitions of single syllables or single syllable words do exist, these quasi-repetitive interpolations are relatively rare. With these examples of stuttering, it can be hypothesized that repetitions caused by underlyingly motor control problems are more likely to surface in a different form than repetitive interpolation, although in rare cases they can share more similarity. One proposal for the cause of stuttering is a deficit in the timing control in the motor controller (Mackay and Macdonald, 1984). Since more precise hypothesis about the relationship between motor control and repetitive interpolation is not yet able to be proposed, the possibility that a motor control deficit leads to fluent repetitions cannot yet be ruled out. Nevertheless the distinction between stuttering and repetitive interpolation can be reliably made at this point.

5.2. Repetitive interpolation in read speech

A more direct test of the role of motor planning and control in repetitive interpolation is to see its surface realization and distribution in a scenario which the utterance plan is supposedly given. An approximation of this scenario is the case of read speech. Unlike spontaneous speech where the utterance plan is generated on the fly, the speech produced in reading follows the scripts that is available to the speaker. Thus the process of reading aloud the available text primarily involves the process of translating an utterance plan to proper articulation gestures. As the examples in section 1 suggest, people who stutter still exhibit signs of repeating words or partial words in the task of reading. Such repetitions are likely to be resulted from problems in the motor control system (Civier et al., 2010; Mackay and Macdonald, 1984). Reading and repeating are both techniques used for eliciting speech from people who stutter for clinical evaluation (Van Riper and Emerick, 1984). Therefore it can be expected that read speech would contain repetitions that are similar to repetitive interpolation if repetitive interpolation were primarily related to motor control. Follow this reasoning, the first and fundamental question to be asked is what is the repetition frequency in read speech? If repetitive interpolation is primarily resulted from the motor planning and control stage of speech production, a comparable frequency and type distribution of repetitions should be expected from read speech.

To examine repetitions in read speech, I look at a collection of public political speech from the American Rhetoric database (Rhetoric, 2020). The collection, under the name “Rhetorical Literacy: 49 Important Speeches in 21st Century America”, contains 49 public speeches made by politicians, policy makers, professors and other public figures. These speeches are considered as a good approximation of read speech because the speakers are mostly reading off a rehearsed prepared script. Rehearsal can potentially reduce the occurrence of speech errors due to articulation error and misinterpretation of the text, but is not expected to be a confounding factor in the execution of a motor plan. 40 of the 49 speeches in the collection are randomly selected for this analysis, representing the speech from 33 speakers. The duration of the selected speeches ranges from 3 to 40 minutes. The total number of words in this constructed corpus is 88,201. Since transcriptions of the speeches were made post-speech delivery, the actual delivery of the speech is largely preserved. However, since disfluencies and speech errors are not consistently transcribed and labeled, Manual labeling of repairs and repetitions was done through editing the original transcriptions based on the audio files. The annotation follows an in-line annotation convention where only the reparandums are marked to indicate either a repetition or error repair.

5.2.1. Result

Repetitions in general are rare in this collection of public speech. Out of the 40 speeches, only 54 instances of repetitions, regardless of type, are observed. This translates to only 0.61 repetitions per 1,000 words, far less than the average count of 2.5 percent of the total words produced in fluent spontaneous speech. This percentage is even much lower than the overall repetition frequency of content words, which is around 2.4 per 1000 words. A caveat of these repetitions in read speech, however, is that a number of the repetitions may actually come from the impromptu speaking at the beginning of the public speech. Thus the frequency of repetition in read public speech reported here is an over estimate of the true frequency.

On the other hand, speech repairs due to mis-articulation are a more common phenomenon, with a raw count of 179 that corresponds to a frequency of 2.03 per 1000 words. These speech repairs are purely caused by articulation errors, such as mispronouncing a syllable, or omitting or inserting a word in the original script. More experienced and eloquent speakers, such as professional politicians, have fewer such speech errors in general.

5.2.2. Discussion

The lack of repetitions in general in read speech can be interpreted as an evidence that repetitive interpolation is at least not primarily a phenomenon related to the motor planning and control in speech production. If reading, especially rehearsed reading, is considered as the process of directly translating an utterance plan to a motor plan that directs speech articulation, then the lack of repetition in reading suggests that repetitive interpolation should at least involve the process of formulating an utterance plan, and potentially the transition between utterance planning to motor planning. On the other hand, this brief discussion is able to show that speech errors driven by motor planning and control do occur and at a similar frequency compared to the repetition of content words in spontaneous speech. This similarity potentially suggests that the repetition of content words is more likely to be resulted from issues in the motor controller, and should be interpreted as a strategy to cope with articulation errors. As for repetitive interpolation, an account purely from the stand point of speech motor control is not likely to provide an adequate explanation.

5.3. Repetitive interpolation in Czech

The next piece of evidence for the hypothesis that repetitive interpolation is not primarily a motor planning and execution problem can be found from a cross-linguistic perspective. As I discussed in the case of English, repetitions of partial words are rare, compared to single syllable words. Considering the fact that the majority of single syllable word repetitions are likely to be fluent, this could be interpreted as a sign that repetitive interpolations do not happen across word boundary. However, since single syllable function words predominantly occur at the beginning of phrases, the location where repetitive interpolation tends to occur, it cannot be ruled out that the lack of partial word repetitions is merely a frequency effect. Additional motivating evidence for our hypothesis can be found in languages that rely heavily on morphology to express the argument structure and have flexible word order.

This section provides evidence that repetitive interpolation is likely at least not only related to the motor planning and execution process from a cross-linguistic perspective. I ask whether word repetitions are dependent on the morphosyntax of the given language. The hypothesis is that if repetitive interpolation is more than an issue with the motor program in the production process, the morphosyntax of a language is not expected to show an effect unless language-specific motor planning and execution mechanism has been tested. I use a case study of repetition in Czech to explore this hypothesis. Czech has been chosen because it is a language whose morphology is complex with flexible word order, and the inventory of single syllable function words is limited.

5.3.1. Background

Discussions on the cross-linguistic variation in repetition are often found in the literature on discourse analysis, where the terminology for repetition and repair is *recycling* and *replacement*. Although the forms of repetition have been shown to be dependent on the morphology of the language, the repeating unit often respects the constituent boundary (Fox et al., 1996, 2010; Hayashi, 1994; Fincke, 1999). For example, in English, German and Hebrew, the repeating unit generally involves the function word immediately preceding the main content word of a constituent (Fox et al., 2010), while evidence of frequent partial verb repetition has been provided in Japanese, a language that is typically verb final and has tolerance of non-overt arguments in discourse (Fox et al., 1996). Japanese also contains morphological repetition and repair, where only the bound verbal suffix is repeated or replaced by another (Hayashi, 1994; Fox et al., 1996). The unit of repetition in Japanese is thus mostly within the constituents where repetition takes place, while repeating the full constituent is not impossible but rare. Similar observations have been made in Finnish

(Kärkkäinen et al., 2007). The examples below illustrate how repetitions and repairs may take place in Finnish and Japanese.

(15) tteyuuka koko denwa [kaket- kakete] kite sa
I:mean here telephone ca- call come FP
'I mean, (they) ca- called us here,'

In this example of Japanese, only the verb in the verb phrase is repeated in the repair, and proper inflection is added after the repetition. Therefore only the morphologically relevant segment in a phrase is repeated and repaired.

(16) mutta nyt [selvi-tä-än, -te-tä-än] nämä marka-t
but now manage-PASS-PERSCAUSE -PASS-PERS these mark-PL
But now let us manage, sort out these marks.

In this example of Finnish, the speaker initially produced a passive intransitive verb, while wished to replace the verb with a transitive form. The strategy employed here is just to insert the transitive suffix *-te* and repeat the rest of suffixes that stay unchanged.

In the literature of discourse analysis, the unit of repetition and repair has been claimed to be related to the projectability of constituents, as proposed in Wouk (2005) citing examples from Indonesian. The syntax of repetition and self repair has been formalized in Uhmann (2001), citing evidence from German that the start of a repetition or repair has a preference for the functional head in the phrase structure. Under this view, the degree to which early parts of the constituents project the syntactic structure and further indicate the completion of the clause determines the likelihood of the word or phrase being repeated or in the repair. For example, due to its right branching structure, the head of phrases in English project over the complements to the right, thus a wider scope of repetition is expected. On the other hand, in Japanese, this scope is rather limited due to its mostly left branching structure. Thus repairs need only be done with respect to the head. However, this claim does not justify the necessity of introducing the notion of scope of repetition into the picture: the null hypothesis should be that when it comes to repeating words in an utterance, a speaker would only repeat whatever is convenient: when there isn't small function words available at the left edge of a phrase, they would just repeat partial words, or possible filler words instead. Same logic also applies to repair, and is congruent with the assumption of preserving the continuity of delivery (Clark and Wasow, 1998).

Since in the discourse studies literature, statistics of the distribution of repetitions and repairs is rarely found, and the sample size being worked with tends to be very small, the null hypothesis cannot be quantitatively rejected. One other problem in the discourse literature is that it is not clear whether the reported examples are representative of spontaneous speech. Sampling bias may have potentially under-counted more common but less distinctive features of repetition of the language under discussion. Nevertheless, from examples cited in the qualitative literature, we have reason to expect that for a language with richer morphology, more flexible word order, and potentially different head directionality, the overall repetition frequency and type distribution would more likely to be different from what have been known in English and related languages. Knowledge of the cross-linguistic pattern distributions of repetition and repair will not only help to enrich the current theory on the relation between repetition and the syntax of a language, but also offer new insights to syntactic planning from a cross-linguistic perspective.

Czech is a good candidate for exploring the issues presented above. As a west Slavic language, Czech has relatively flexible word order and rich morphology. The case, gender, and number systems are almost exclusively expressed through a complex inflection system (Janda and Townsend, 2000). The trade-off of this complex inflection system is its limited inventory of function words such as pronouns and prepositions. Therefore what can be expected in terms of repetition in Czech is something like Finnish, a language with very complex morphology and relatively high freedom of word order. With the knowledge from

Finnish and Japanese, it can be expected that repetitions of partial word and partial repairs of inflectional suffixes can be fairly common. One other possibility is that the possible variation of repetitions is more limited, while other types of disfluencies such as filled pauses would take place where repetition would otherwise occur. However, because unlike colloquial Finnish, where speakers tend to insert isolated pronouns, spoken Czech is even more restricted in the use of function word categories, a stronger pattern may also be expected in Czech.

In this section, I hope to untangle some of the interesting questions regarding repetitions in spontaneous speech through analysis of larger collection of speech corpus. The first question to be addressed is whether the distribution of repetition in Czech is different from what I have reported for English. In particular, are there fewer single word function word repetitions in Czech? Then I will also explore if repetitions in Czech are systematically related to word frequency. The corpus for this analysis is the Czech Spontaneous Speech Corpus (Kolár et al., 2005).

5.3.2. Repetitive interpolation in Czech

The data we use to explore repetitions in Czech is the Czech Spontaneous Speech Corpus (Kolár et al., 2005). This corpus consists of 72 recordings of radio discussion program called Radioforum. The total speech duration is 24 hours. It broadly falls into the same speech style, spontaneous conversations, as Fisher and SCOTUS. However, noticeable difference between the mode of conversation, such as face-to-face conversation in a more formal setting, unlike that in Fisher, but less formal than SCOTUS. The total number of transcribed tokens is 225.3K with 25.3K unique words. Speech for both the host and guests of the show are transcribed at token level. The number of interviewees ranges from 1 to 3, while the number of host is up to 2. In total, speech from 94 speakers are recorded, with 77 males and 17 females. Turn-level alignments are also transcribed in detail, along with annotations of speech disfluencies and syntactic phrase boundaries. The annotation of the corpus follows the "Simple Metadata Annotation Specification" for English by LDC. This system annotates edit disfluencies (repetitions, revisions, restarts and complex disfluencies), fillers (including, e.g., filled pauses and discourse markers) and SUs, or syntactic/semantic units (Data Consortium, 2009). Thus the information provided in this corpus will ensure accurate and consistent analysis of repetitions in Czech, and the results can be comparable to that from the English corpora, when style difference is properly acknowledged.

Before moving forward with the quantitative analysis, some examples of repetitions in Czech extracted from the corpus are useful in establishing an intuitive understanding of the forms that repetitions in Czech can take. The utterances below illustrate some more common possibilities, where the repeated words are marked in bold. The transcription is made at token level, i.e., each space separated substring is an independent token, rather than a partial word or morpheme.

(17) že EE vzít **si si** advokáta **a a** nějaké náklady v tom případě pak uplatnit posléze

(18) jesli jsem byl zvyklý třináct let komunikovat s občany **a a** zúčastňovat se nekonečného množství

(19) EE **že že** naopak tento nárůst

(20) ale jaksi při nejmenším **na na** osobní svobodě

The intuitive observation from these examples is that just like in English, Czech also contains a fair amount of repetitions that can be described as repetitive interpolation. The repeated words are also mainly single syllable function words, rather than partial words or partial morphemes. The question for Czech is then, similar to that for English, is the distribution of repetition odds for a given word dependent on word frequency? Although Czech has a smaller inventory of single syllable function words, the same reasoning holds:

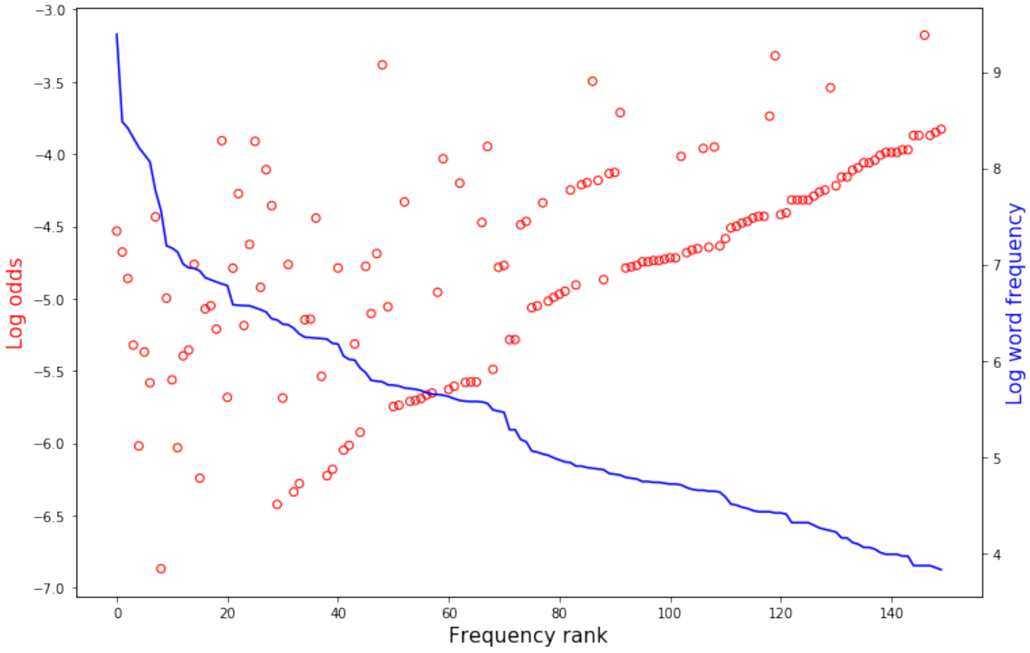


Figure 40: The relation between repetition odds and raw lexical frequency in Czech.

If the repetition of the kind as illustrated in the examples above is independent of word frequency, then an explanation from a motor planning and execution perspective is more appealing. On the other hand, repetitions that dependent upon word frequency might be better explained by higher level processing demand in the planning process.

Figure 40 plots the distribution of the log odds of repetition as a function of word frequency. The overlaid blue line shows the trend for the change of log word frequency as a function of frequency rank. As expected, the number of unique repeated tokens in Czech is much smaller compared to English, where only 255 unique repetition patterns are observed, and the majority of them have a repetition frequency of essentially 1. In Figure 40, only the first 150 repeated tokens are plotted so that a closer look at the higher ranked region of the graph can be made. In particular, a similar trend of independence between repetition odds and frequency can be found for the first 40 most frequently repeated words.

Table 13: Top 20 most frequently repeated words in Czech.

Words (English gloss)
EE (uh), a (and), že (that), na (on), to (it), ten (the), je (them), ta (the), v (in), ty (you)
jak (how), s (with), se (themselves), já (I), aby (that), co (what), těch (those), i (and)
pro (for), do (to)

This is similar to what have been discussed about repetitions in English. Thus it is likely that similar repetitive interpolation phenomenon can be expected in Czech as well.

Table 13 lists the top 20 most frequently repeated words ordered by rank and their corresponding English gloss. Interestingly, although the inventory of available single syllable function words is much smaller in Czech compared to English, the most common repeated words in two languages appear to be highly similar in terms of both form and meaning. Thus repetition again shows a tendency towards single syllable function words, even if the inventory in the language is rather limited.

Figure 41 compares the repetition log odds of the 20 most repeated words in English and Czech. Although the distributions roughly follow the same trend, repetition odds in Czech is consistently lower than that in English. Thus for a given single syllable function words, it is less likely to be repeated compared to English. One explanation to this observation can be that the structure of the language affects the absolute likelihood of a word being repeated. Since the word order of Czech is more flexible compared to English, the likelihood of a single syllable function word appearing at the left edge of a phrase is lower, compared to English. Thus the repetition odds is also lower. However, the alternative explanation does not rely on linguistic structure; rather the lower repetition odds is the result of different speech style, which can again have several confounding factors. For example, Czech conversations were conducted as three or more multi-person face to face conversation, while the English examples are from telephone two-person conversations.

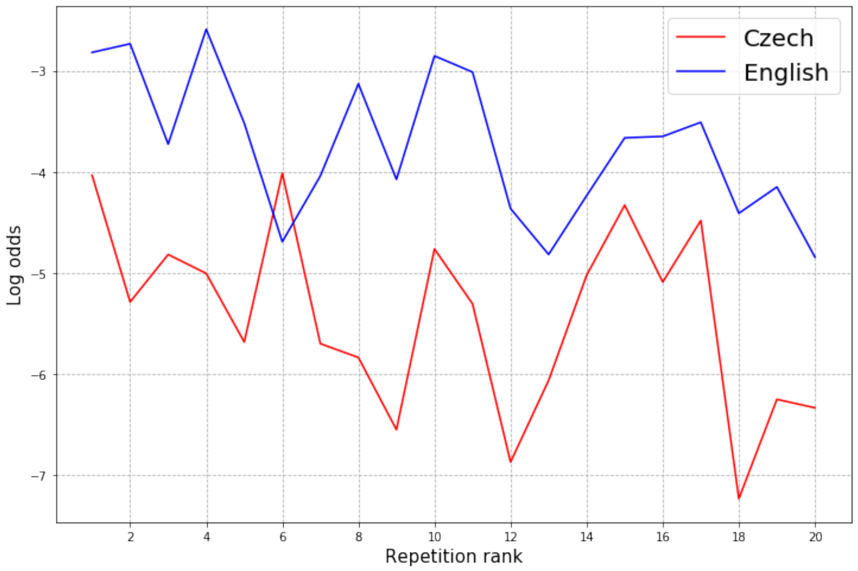


Figure 41: Comparing repetition odds of the most frequently repeated words in Czech and English.

The English examples are recorded in a much more informal setting as well. These potential confounding factors should be further controlled for a better picture of the difference between English and Czech.

Our observation of repetitions in Czech suggests that, in addition to the possibility that fluent repetitions are a distinct phenomenon from other forms of repetitions, the cause of this form of repetition may be related to higher level speech planning process above motor planning and control, for two reasons. First, we observed a much lower repetition rate in Czech than in English, which parallels the fact that Czech has a smaller inventory of single syllable function words. Second, Czech does not see an increase in the likelihood of observing partial word repetitions compared to English. This suggests that although the morphology of the language is mode complex, speakers of Czech, like those of English, process words holistically. Although more acoustic evidence is needed to further strengthen our claim, a cursory look at repetition examples in Czech shows that the repetition pattern of single syllable function words in Czech is also similar to English.

5.4. Chapter summary

In this chapter, I considered the theoretical explanation for repetitive interpolation from the perspective of speech motor control in speech production. One potential explanation is that repetitive interpolation arises at the motor planning and execution stage of speech production. Under this assumption, repetitive interpolation should be equally likely and be produced in similar forms when an utterance plan has been formulated. I tested this hypothesis by first comparing stuttering, a known speech disorder originated at problems with the motor control system, to repetitive interpolation. It is apparent that although both phenomena involve repeating certain speech segments, repetitive interpolation does not involve obvious blockage of the delivery of speech, and perceived difficulties in producing certain phonemes and/or syllables. In fact, as I have shown in Chapter 4, repetitions of partial words are rare compared to repetition phenomena in general. Therefore repetitive interpolation is at least not caused entirely by the same motor control problem that is responsible for stuttering. I further compared the distribution of repetition in read speech to spontaneous conversations. In the setting of read speech, it can be assumed that an utterance plan is already given: it is just the written text that to be read by the speaker. Thus speech errors and repairs can be assumed to relate to the motor controller in the production process. A lower rate of repetition in this setting could thus indicate the existence of other factors in the production system that induce repetitive interpolation. As discussed earlier, a much lower repetition rate is indeed observed in read speech. Thus the discussion so far has found two pieces of evidence for the hypothesis that repetitive interpolation is not merely related to speech motor control, but likely involving the formulation of an utterance plan.

Finally, I briefly discussed the cross-linguistic similarity of repetitive interpolation by comparing the distribution of repetitions between English and Czech. Unlike English which has a big inventory of single syllable function words and predictable word order that tends to place these single syllable function words at the beginning of phrases, Czech is morphologically rich with highly flexible word orders. If the assumption that repetitive interpolation is only a motor control problem holds, then one would expect higher rate of partial word repetition in a language like Czech. However, the results suggest that repetition in Czech is again predominantly single syllable function words, and the frequency distribution of repetition is related to the frequency of the repeated function words. A higher rate of partial word repetition is also not observed. Thus this cross-linguistic observation also motivates the hypothesis that repetitive interpolation likely involves the formulation of an utterance plan or the transmission of the utterance plan to motor controller.

Although the observations made so far do not fully support hypotheses based on a theory that only deals with the motor planning and control stage of speech production, neurologically plausible network models of the motor control of speech production may shed some light on the details of the specific hypotheses to be tested. For example, models of speech motor control, such as the DIVA model laid out in Guenther (2006), could lead to further hypotheses on segment duration properties relevant to repetition production, based on the proposed feedback loop which considers both auditory and somatosensory feedback. The question that might be addressed in this regard would be whether the repetition in repetitive interpolation involves coordination with auditory or somatosensory feedback. Furthermore, models dealing with sound sequencing, such as the GODIVA model proposed in Bohland et al. (2010), could offer ideas on how units in an utterance plan are excited and inhibited after excitation when being transmitted to the motor controller. Combining with the forward model proposed in DIVA, mechanisms on the coordination of syllable level and phoneme level planning could be proposed. Finally, the HSFC model as described in Hickok and Poeppel (2007); Hickok (2012, 2014) could be a source for more unified proposals for speech production across levels at least involving utterance planning and motor control, and how the two stages are coordinated. A final theory on the underlying mechanism behind repetitive interpolation might be a weighted sum of ideas presented in these production models centered around motor control. Such a theory would be crucial in linking hypotheses about the more abstract speech planning processes, such as those described in Levelt (1989) and its followers, to observable and measurable phenomena in spontaneous speech. Eventually the hypotheses along this line would be backed up by experimental results.

To sum up the discussion, I have laid out three pieces motivating evidence for the hypothesis that repetitive interpolation is not only a speech motor control problem. Although the discussion presented so far is not able to lead to more concrete and testable hypothesis directly relating to the neuro- and psychological mechanism of speech production, I briefly outlined a path moving forward. Future observational or experimental work could use the documentation of repetitive interpolation in this study as a starting point towards a rich theory on speech production.

Chapter 6

The Variation of Repetitive Interpolation

In this chapter, I explore the variation of repetitive interpolation among both normally fluent speakers and speakers who are impaired by alcohol intoxication. As I have established in chapter 4, repetitions in spontaneous speech should not be considered as a single phenomenon of disfluency; rather at least two broad types of repetitions: disfluent repetition and repetitive interpolation, should be acknowledged. Repetitive interpolation is also arguably the norm in fluent utterances rather than anomaly, in the sense that its production is less affected by higher level planning problems as one would assume for disfluency. In chapter 5, I started the discussion on how a theory that only focuses on the motor control of speech production is also inadequate in fully explain the empirical observations made so far. Another important component towards a theory of repetitive interpolation is an understanding of how this repetitive phenomenon varies across speaker groups defined by some measurable feature space. I will shift the focus to this variational aspect in this chapter.

Following up on the argument for repetitive interpolation being a distinctive feature of fluent articulation, I address how repetitive interpolations may vary as measured by their textual and prosodic features. Both measurable speaker features such as age and gender, and the linguistic context that potentially explain the observed variation are explored. By looking at the absolute variation measured through frequency distribution of different kinds of repetitions and their prosodic features, I would like to address the question of whether repetitive interpolation is indeed more frequent than disfluent repetitions. Thus repetitive interpolation can further be argued as a prominent feature which can potentially be used to quantify speech fluency. On the other hand, a probe into the potential explanatory power of certain measurable speaker and contextual features on these observed variation could point to directions for abnormality detection across different speaker groups. With these two questions in mind, I hope the discussion in this chapter would serve as an attempt for establishing a robust understanding of repetitive interpolation such that it can serve as a well-defined feature for fluency. Finally, I present a case study that compares repetitions between typical fluent speech and the speech produced under alcohol intoxication. It will be shown that impairment caused by alcohol intoxication has prominent effect on both the textual and prosody of repetitions.

The results of this chapter can be briefly summarized as the following. Fluent repetitive interpolation predominates the observed repetitions produced in multiple communicative settings. This observation partially justifies the use of the overall repetition rate of single syllable function words as a proxy for the true rate of repetitive interpolation for a given speaker. The bimodal distribution of the silence duration between repeated words reported in chapter 4 is found to be similar to the distribution of silence duration following filled pause “uh”. This observation supports the categorical distinction between repetitive interpolation and disfluent repetition. With these descriptive analyses, the proposal that repetitive interpolation is a prominent feature of fluency is further supported. However, most speaker and contextual features failed to show correlation with the frequency and duration distribution of repetitions, thus this property of fluency could potentially serve as a strong and stable indicator for the state of speaker’s speech production system. This hypothesis is further tested with the speech produced under alcohol intoxication.

6.1. Background and outline

The central question here is whether a variation or contrast similar to what have been discovered with filled pause can be found. If there is any, then arguments can be made about how repetitions reflect on the planning process and speaker's cognitive ability. If there is not within normally fluent population, but a contrast can be found across speaker groups or mental states, then it can be argued that this phenomenon can be potentially informative about speaker state, and the theoretical treatment should be carried out differently. Literature on speaker dependent variations of repetitions is scarce. The literature review in chapter 4 has shown that the discussion on repetition is mostly centered around building an understanding of how repetitions in speech is related to error repair in speech production and the phrase structure. How the lexical category of repeated words and their grammatical roles or positions in phrases are correlated with the distribution of repetitions is often involved as evidence for the claims about particular production process that explains repetitions. As reviewed above, function words are expected to be most heavily repeated, while the absolute frequency of occurrence, relative to all the instances of repetitions, does vary (Foster, 2010). This observation has also been substantiated by the comparison between repetition frequency of single syllable function words and content words in the previous chapter. With regard to the duration properties of repetitions, it has been suggested in Blacfkmer and Mitton (1991)'s study on the timing of repair gaps that the cut-off-to-repair time may be too fast to fit into a self-monitoring model, as the one proposed in Levelt (1983, 1989). Although later studies (Shriberg, 1995; Plauché and Shriberg, 1999) have attempted to pair natural clusters found among repetitions with Hieke's classification paradigm, it is still not clear what is the distribution of the duration properties of repetitions and their relation to other speaker or contextual features. In other words, the question remains as to how the duration distributions are related to other explanatory variables that can offer insights into the reported duration variation. As for cross-speaker variation, if we follow the Commit and Restore model for repetition (Clark and Wasow, 1998), it can be expected that the socioeconomic features of speakers would potentially have an effect on the distribution of repetition, in a way similar to their effect on the distribution of two forms of filled pauses due to the parallel bipartite distinction among repetitions I have argued so far.

The structure of the chapter is the following: I will first report a speaker dependent analysis of the variation of repetitions. I use SCOTUS 2001 to illustrate that similar distribution of repetition phenomena is observable from speech produced in a different conversation setting, and how individual variation may manifest across speakers. I then follow up with the analysis by looking at how measurable speaker and contextual features may affect the frequency and prosodic properties of repetitions. Then I do a speaker independent analysis to argue that speaker dependent effect on the distribution of repetitions may be explained by a unified account that the production of both repetitive interpolation and certain forms of filled pause are caused by the same underlying mechanism, and the cognitive demand during production is likely a factor in influencing repetition distribution. With these descriptive analyses, it will become clear that repetitive interpolation is indeed a sign of fluency in normally fluent speakers, in the sense that little variation can be observed along dimensions that measures speaker features. Thus a theory on fluent speech production should be able to account for this phenomenon, and in practice this property of fluent speech is helpful in understanding the underlying problems of people with impaired fluency. The chapter will conclude with a case study of repetitions in alcohol impaired speech. It will be shown in this case study that cognitive impairment cause by alcohol intoxication affects the distributional property of repetition phenomena, which is potentially driven by repetitive interpolation.

6.2. Speaker dependent analysis

In this section, I address the question of how measurable speaker features, such as age, gender and years of education, may be used to account for variations in repetition phenomena. I first demonstrate, through an analysis of SCOTUS 2001 corpus, the overall trend across speakers, as well as the potential surface variability that can be observed across individual speakers. Then I present a more rigorous description with a large sample of speech from Fisher. This two-step strategy is due to the challenge posed by annotation efficiency and making proper distinctions agnostic of subjective judgement. In chapter 4, I have shown that with certain amount of subjectivity in the annotation process, it is possible to capture additional dimensions that are informative of variations among repetitions independent of the subjectively defined annotation categories. In this section, by looking at the frequency and duration features of a sample of well-annotated corpus without subjectively defined repetition categories, an objective and cross-domain baseline for comparison is made available.

6.2.1. Repetitions in Supreme Court Oral Debates

As mentioned above, repetitions are a more challenging group of disfluency phenomena because their high variability in their realized forms. Similar to speech disfluencies in general, corpora that provide consistent and accurate repetition annotation are lacking for large scale analyses. As an attempt to address such challenges, I first restrict the discussion to a much smaller sample that can be hand-labeled. The goal of the analysis of a small sample is to devise an efficient and objective annotation system. This system should be sufficient in evaluating individual variations of repetition and other relevant questions. The analyses in this section is based on the speech from eight US Supreme Court justices serving their term in 2001, using data from the SCOTUS 2001 corpus. In the following discussion, I will first describe the annotation procedure that I used for repetition and repair labeling. Then I will demonstrate that repetitive interpolations is able to account for the most of labeled repetition instances. Individual variation of the distribution of other repetitions does exist and will be explored in the following section.

The full SCOTUS corpus (Johnson and Goldman, 2009) contains 38 years of recordings linked to transcripts of oral arguments at the Supreme Court of the United States. The subset that contains the speech from 8 US chief justices in 2001 will be used for the present study. This subset was originally compiled for a speaker identification task (Yuan and Liberman, 2008). Verbatim transcriptions of the speech material after diarization are available. This corpus contains about 3 hours of speech from each justice. Unlike Fisher, this corpus provides ample speech material from single speakers, thus it is a good source for more detailed analysis of individual variation.

Data annotation The annotation is done on the time aligned SCOTUS 2001 corpus, where the speech from each justice in court sessions has been diarized, grouped and segmented into turns. The annotation follows an adapted version of Shriberg (1994)'s pattern labeling system (PLS), but focusing mostly on the reparandum and repair annotation. In the present case, reparandum explicitly refers to the repeated segments, and repair refers to the last repetition or repair for preceding repetitions that is integrated with the following fluent utterance. In-line annotation is adopted, mainly for efficiency considerations. Annotation symbols are summarized in Table 14.

In this system, annotations are organized in two levels: The primary symbols are used to represent the type of the disfluent word, and the secondary symbols are designed to mark the region a disfluent word belongs to. Primary and secondary symbols are ordered linearly from left to right, separated by the symbol <. The primary symbol can be omitted in the case of a complete restart, where the interruption point is immediately following the previous fluent utterance. The secondary symbol can be optional when the disfluency does

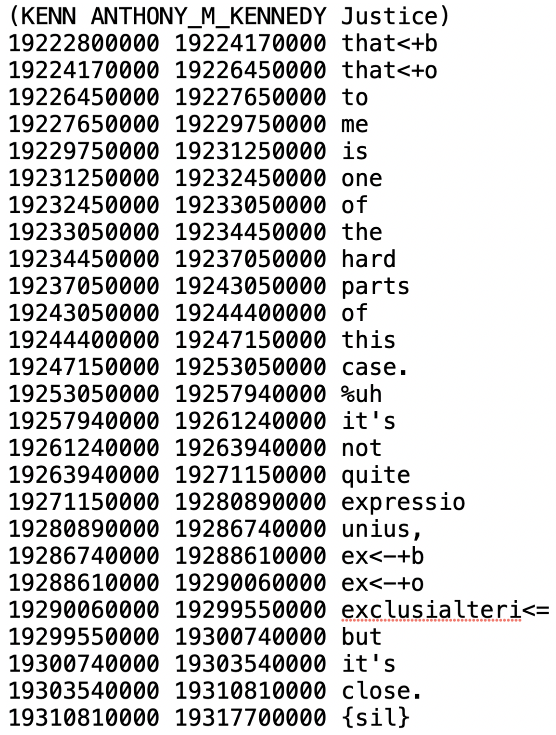
Table 14: The proposed detailed annotation system for repetitions and repairs.

Symbol	Explanation	Example
Primary symbols		
Unmarked	Fluent word	
“+”	Repeated word	that<+
“=”	The substituting word	exclusionary<=
“-”	Word fragments	ex<-
“∼”	Substituted or deleted word	expression<∼
“e”	Explicit edit or other words or vocalization between RR and RP	%um<e
Secondary symbols		
“.”	Interruption point	that<+.
“b”	The beginning of a repeating unit	that<+b
“o”	The middle of a repeating unit	in<+b your<+o
Other symbols		
“<”	Separator between word and annotation	that<+b
“%”	Filled pause	%um

not involve repetition and not immediately after the interruption point. Both primary and secondary symbols can be stacked, but primary symbols are always annotated to the left of the secondary symbol. A snapshot of the annotated transcript can be found in Figure 42. The first row of the transcript records the speaker information, and the first two columns contain the time stamps of the starting and end time of the word segment.

A simple script is written to parse the labeled transcripts. Duration measurements are based on the time stamps of speech segments provided in the time-aligned transcripts, where the first and second column correspond to the start and end point of the segment, as shown in the example Figure 42. Measurement unit is translated from sample index to second.

Analysis In chapter 4, I have proposed that by looking at the duration of repeated words (R1 and R2), silence duration between repeated words (P2) and after a repetition (P3), three



```
(KENN ANTHONY_M_KENNEDY Justice)
19222800000 19224170000 that<+b
19224170000 19226450000 that<+o
19226450000 19227650000 to
19227650000 19229750000 me
19229750000 19231250000 is
19231250000 19232450000 one
19232450000 19233050000 of
19233050000 19234450000 the
19234450000 19237050000 hard
19237050000 19243050000 parts
19243050000 19244400000 of
19244400000 19247150000 this
19247150000 19253050000 case.
19253050000 19257940000 %uh
19257940000 19261240000 it's
19261240000 19263940000 not
19263940000 19271150000 quite
19271150000 19280890000 expressio
19280890000 19286740000 unius,
19286740000 19288610000 ex<-+b
19288610000 19290060000 ex<-+o
19290060000 19299550000 exclusialteri<=
19299550000 19300740000 but
19300740000 19303540000 it's
19303540000 19310810000 close.
19310810000 19317700000 {sil}
```

Figure 42: A screen shot of the annotated transcript from SCOTUS.

types of repetitions can be reliably distinguished, with two of them potentially corresponding to repetitions that involve replanning of message structuring and the other to repetitive interpolation. Following this proposal, here I ask whether the proposed duration-based measurements can reveal natural clusters of different kinds of repetitions. I mainly look at two measurements of the duration features of repetitions in the supreme court oral debate speech: The duration of P2 and P3, and the ratio between R1 and R2. Following the assumption made in the previous chapter, short P2 and P3 duration, and lack of readjustment of the duration between two repeated words can be argued to indicate fluent interpolation, rather than disfluency. Furthermore, I show that in the scenario of supreme court debate, the distribution of repetitions suggests that the majority of repetitions should be considered fluent based on the proposed duration measures.

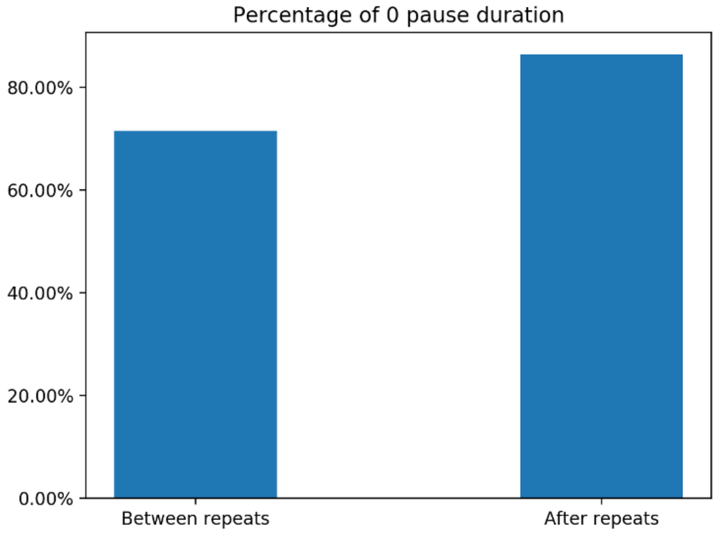


Figure 43: Delay duration adjacent to repetitions. Between repeats measures P2 duration, and after repeats measures P3 duration.

Figure 43 plots the percentage of zero silent interval between and after repeated words. Due to the small sample size in SCOTUS 2001, excluding these none-delayed repetitions would result in a sample that does not provide enough variability to show the variation among repetitions with non-zero delays. Therefore this plot illustrates how common repetitive interpolations are in the speech sample. As Figure 43 shows, high percentage of both P2 and P3 duration are essentially zero, which suggests that the majority of repetitions are fluent in the sense that no apparent delay is observable within the repetition and no apparent delay is observable after the repetition. This lack of delays at both P2 and P3 provides evidence that hesitation due to message restructuring in the replanning process is not common in repetitions, echoing the point made by Blacfkmer and Mitton (1991). The second measurement I use is the duration ratio between R1 and R2 in a repetition. A ratio greater than 1 indicates prolongation of the first repeat while a ratio less than 1 suggests the prolongation of the second repeat. A ratio distribution that centers around 1 would show that modification of word duration is less likely. Therefore the lack of change means that

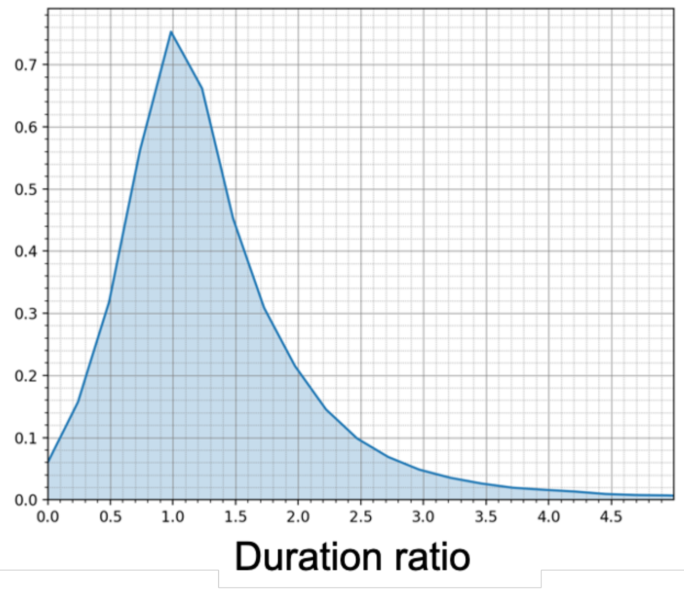


Figure 44: Distribution of the ratio between R1 and R2 across all the speech from the chief justices.

repetitions do not involve hesitation or replanning at higher level of processing. A R1 and R2 duration ratio distribution that centered around 1 is indeed observed from this sample of speech, as shown in Figure 44. This suggests that the majority of repetitions do not have major changes in the duration of repeated words. Thus repetitive interpolation is the majority class of the annotated repetitions in SCOTUS 2001 corpus.

The duration ratio measurement is further plotted for each Justice separately. The question is how the individual justice may vary in their production of repetitions. The answer to this question can help understand whether and how individual speakers may vary in their production of repetitions so that more specific questions can be motivated for large sample analysis. These questions would be crucial in formulating the theory on the production mechanism behind repetitive interpolation. As Figure 45 shows, although the overall density is concentrated around the ratio of 1 across all Justices, individual variation in the overall distribution is also apparent: a tendency for bimodal distribution can be observed

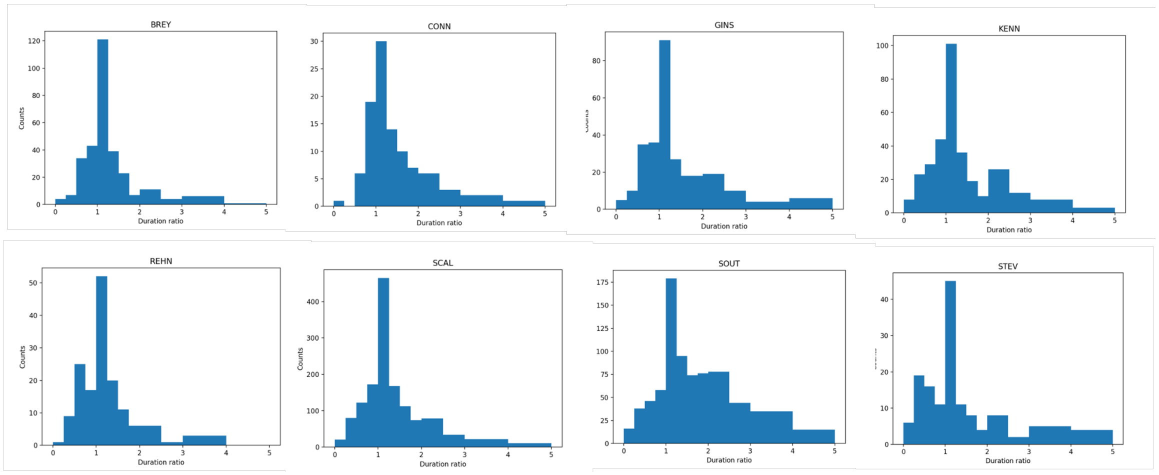


Figure 45: Distribution of the ratio between R1 and R2 for each Justice.

in the speech of Justice Kennedy, Justice Rehnquist, Justice Sout and Justice Stevens. The existence of bimodal distribution for repetitions further suggests that repetition phenomena are potentially correspond to different processing demands during speech production. In other words, some repetitions are more disfluent than others. In particular, as the figure suggests, this variation surfaces as both the shape and location of the secondary peak in the overall distribution. For some Justices, such as Justice Breyer and O'Connell, the tendency for a secondary peak is almost absent, implying that the repetitions in their speech are mostly fluent interpolations. This variation can potentially be linked to both properties of the speaker and the cognitive demand for message formulation.

To sum up, this brief analysis of the duration properties of SCOTUS 2001 corpus suggests that among repetitions, repetitive interpolation is likely to be the norm. This is tested with speech across all the eight supreme court justices. Repetitions caused by hesitation or other higher level planning problems likely follow a speaker dependent distribution, which may be explained by features of the speaker themselves, such as age and gender, as well as contextual factors unique to when the speech was produced.

6.2.2. Individual variation of repetitions in Fisher

In this section I further explore the question of how features of individual speakers would explain the speaker-dependent variation of repetition with a large sample of speech. As suggested from the duration distributions described in the previous section, repetitive interpolation appears to be the norm among repetition phenomena in general. However, the distribution of repetition phenomena does tend to vary across speakers. One hypothesis is that speaker features are able to explain the observed variations. Here I present an analysis of the variation in repetitions, assuming that the frequency of repetitions can proxy the frequency of repetitive interpolation given the prevalence of repetitive interpolation among all repetitions, as functions of speaker features. I show how speaker features, including age, gender, and education correlate with the frequency and P2 duration of repetitions.

Methods Unlike the small sample approach where more accurate annotations of repetitions are available, I resort back to frequently repeated single syllable function words as the set of repetitions for examination. It has been hypothesized that the majority of single syllable function word repetitions are repetitive interpolations, so the frequency and duration variation within this subgroup can be treated as variation among repetitive interpolation. The relative frequency of repetitive interpolations is estimated through the repetition rate of repetitions of the 50 most frequently repeated single syllable function words. Thus for frequency features I ask how speaker features are related to the relative frequency of repetition among the selected indicator examples compared to the absolute frequency of the same set of words. The duration properties are evaluated using the same set of words comparing within the repeated words only.

Speaker features are taken from the documented speaker fields in Fisher. These features include: age, gender and years of education. The frequency counts and duration measurements of repetitions of the 50 most frequently repeated single syllable function words in the

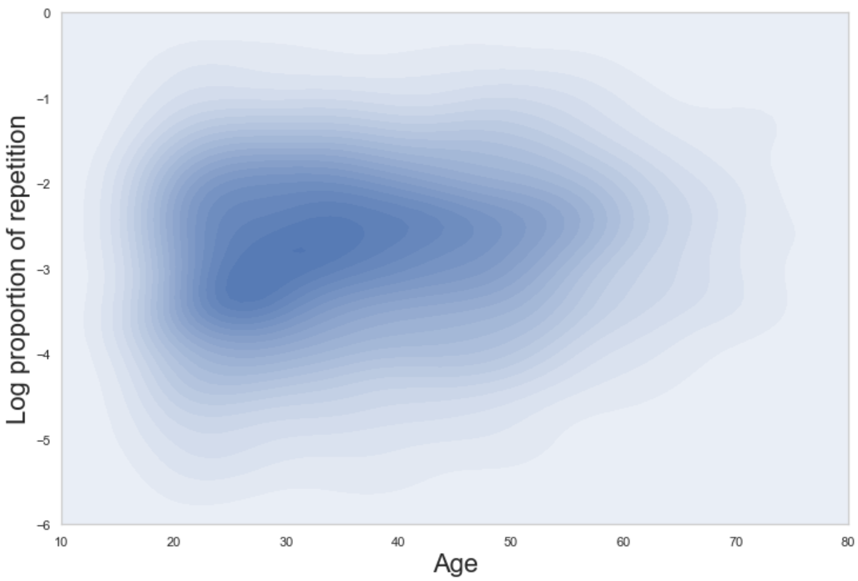


Figure 46: The relationship between proportion of repetitions for given single syllable function words and speaker age. Proportion is log transformed.

corpus from all the speech produced by a given speaker are used as the response variable in the correlation analysis. Speech from the full 3272 speakers is used for the analysis. In the remainder of this section, I first describe the relation between the proportion of repetition among the selected set of words and speaker properties including age, gender and years of education. Then I report the results from the same analysis with the proportion of P2 duration being zero as the response variable. Finally I explore how the relative duration between the repeated words in a repetition varies across speakers.

Speaker features and repetition frequency The relation between the log of relative repetition rate and age is plotted in Figure 46. This density plot suggests that the relation between repetition rate given words and age is mostly gaussian. However, a slight positive slope for the high density region can be argued for the age group between 20 and 40 years of age. The spill over of density towards the upper left region in the graph for the same age group may suggest the existence of other unobserved variables that correlate with age

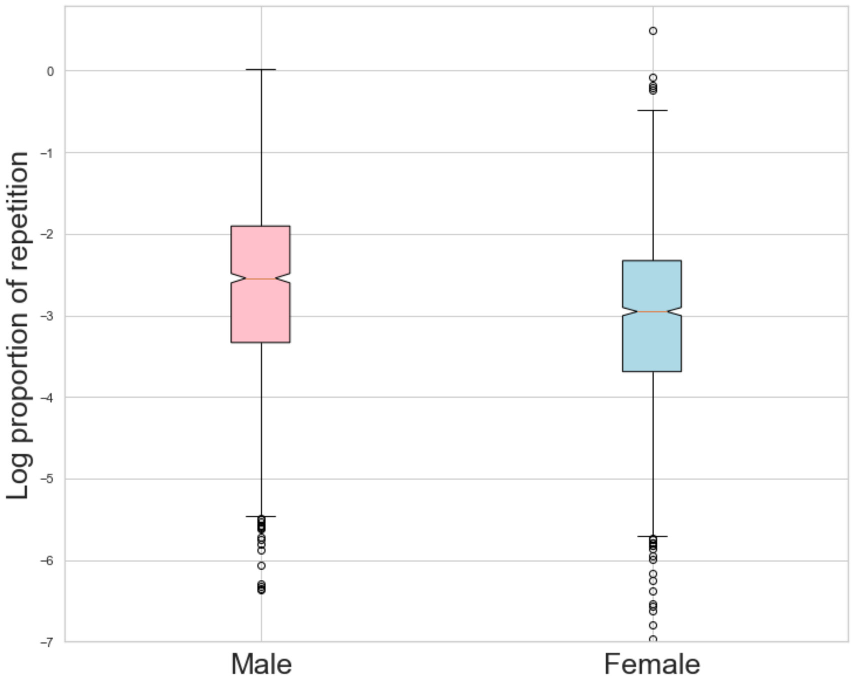


Figure 47: The relationship between gender and repetition rate for frequent single syllable function words.

which jointly affect the rate of repetition. However, the average effect is expected to be small.

A potential measurable confounder, i.e., the variable that is unaccounted for in the graph presented in 46, is speaker's gender. Figure 47 shows the box plot of group difference in the log proportion of repetition. It is clear that male speakers on average have higher repetition rate than female speakers. This difference is statistically significant with a two-tailed t-test ($t = 10.135, p < 0.001$). Therefore male speakers are potentially more likely to produce repetitive interpolations.

But how this gender-related difference in the production of fluent repetitive interpolation interacts with age? Figure 48 and Figure 49 plot the median and 75 percentile of repetition rate grouped by speaker gender. The median plot shows no clear trend of repetition rate as a function of age, even when gender is controlled. However, the 75 percentile

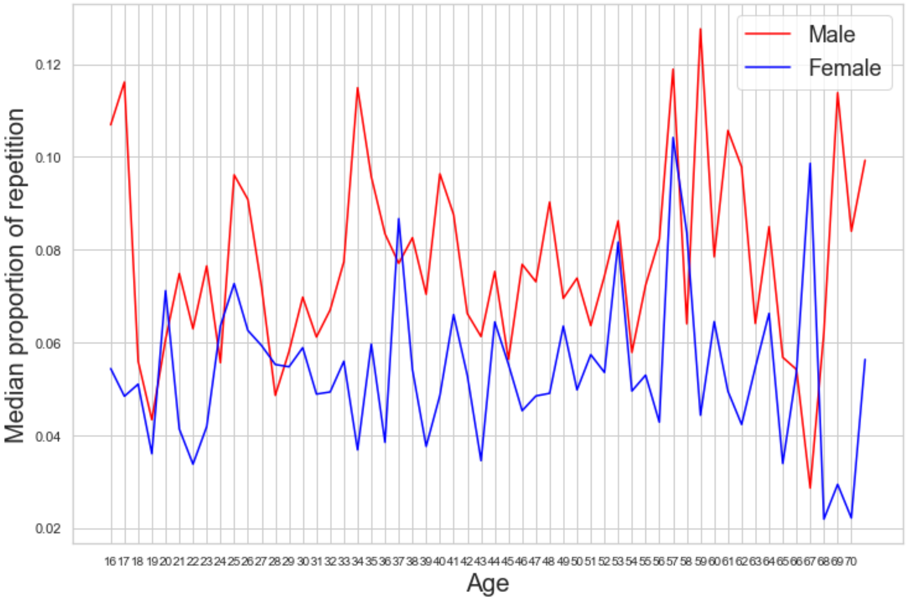


Figure 48: The relationship between age and median repetition rate for frequent single syllable function words grouped by gender.

trend has a slight increase, from about 7.5 percent to 12.5 percent for females, and from about 12.5 percent to 17.5 percent for males, between the age of 16 and early 30s. This age range overlaps with the slight increase observed in the density plot in Figure 46. However, other older age groups do not seem to have an effect even when gender is controlled.

The observations above indicate that gender does seem to interact with age in affecting the rate of repetitive interpolations. However, this interaction effect is only slightly observable at or around 75 percentile of the population and restricted to younger population. Thus it can be hypothesized that the effect of age is not uniform across the cross-section of age groups. However, if there is any truly measurable effect it would likely be small. So the overall conclusion is still that age probably doesn't affect the rate of fluent interpolations.

The effect of gender on the rate of repetitive interpolation is interesting. A similar effect of gender has been reported on the use rate of two forms of filled pauses (Wieling et al., 2016). In chapter 3, I have shown that there exists a strong gender effect on the use of filled pauses, especially of the form "uh" where male speakers have higher rate compared

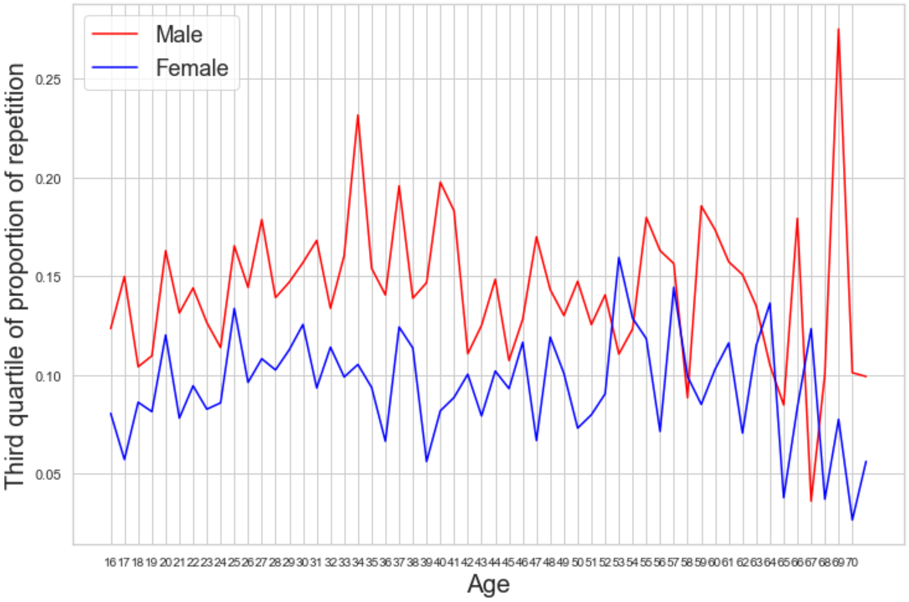


Figure 49: The relationship between age and third quartile of repetition rate for frequent single syllable function words grouped by gender.

to female speakers. The distinction between "um" and "uh" with respect to gender could potentially be related to the function or meaning distinction between the two filled pauses, where "um" signals a major and potentially intentional delay while "uh" signals a minor yet automatic one, as also suggested by Clark and Tree (2002). This meaning or function distinction can further be hypothesized as a correlation between different planning problem, where "um" relates to planning issues at the stage of higher level message formulation, and "uh" is related primarily to the planning and execution of an utterance plan. Under this context, a hypothesis is that there might be a parallelism between "uh" and repetition, since both these phenomena are proposed to partially correlate with utterance plan and motor control. Thus in some sense repetitive interpolations can be conceived as functioning in a similar manner with certain filled pauses in the phonetic form of "uh". Whether this is indeed supported by the speech in Fisher will be examined in later.

The last potential variable for explaining variation in repetition rate is years of education. Figure 50 plots the relationship between log proportion of repetition against the

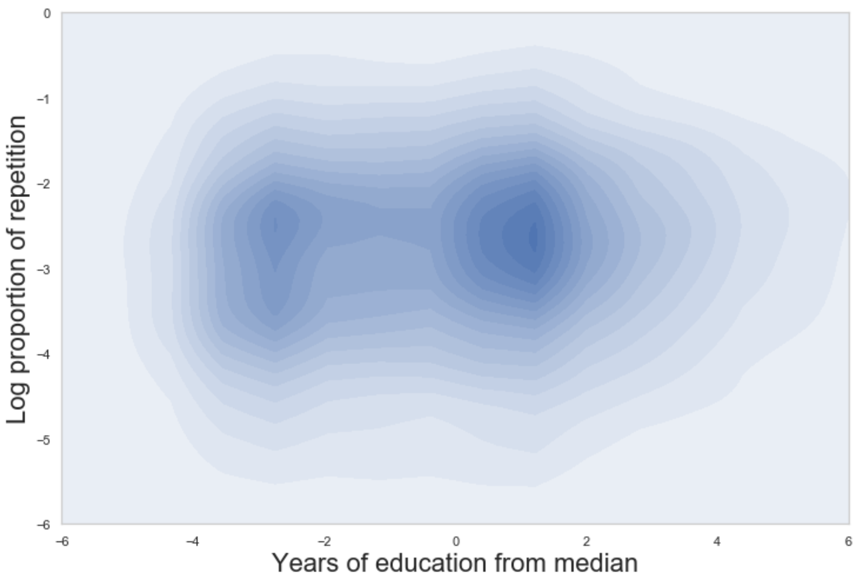


Figure 50: The relationship between years of education and repetition rate.

standardized years of education. In this plot, the horizontal axis represents the deviance, or the positive or negative difference a speaker's years of education from the mean. Since most people either received a college education or not, the sample is concentrated in two major groups. As this density plot suggests, more years of education does not seem to affect the rate of repetition.

Speaker features and duration distribution The second aspect of the problem is whether speaker features are correlated with the distribution of duration features. As P2 duration in a repetition has been used as a proxy for determining whether a repetition is fluent, independent of the identity of repeated words, the question can be formulated as whether there is any association between speaker features and the frequency of repetitions that are fluent, as measured by P2 duration in a repetition. The analysis of SCOTUS 2001 suggested that the proportion of repetitions in which P2 duration is zero is a good indicator of the distribution of repetition types. Therefore the dependent variable for the duration measurement is set to be the proportion of zero P2 duration of repetitions produced by a given speaker. In other words, here I ask if the proportion of zero P2 duration is affected by speaker features.

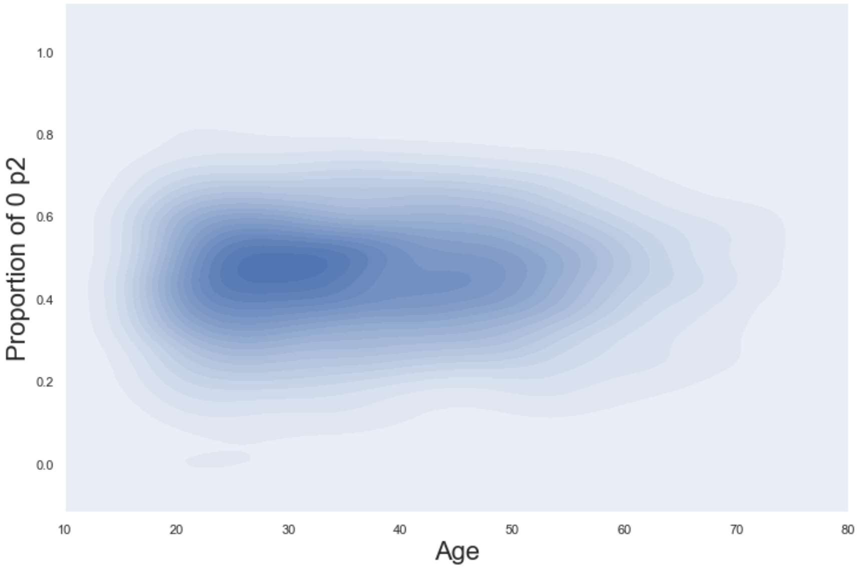


Figure 51: The relationship between age and the proportion of P2 whose duration is 0.

As shown in Figure 51, 52 and 53, these measured speaker features do not seem to have an effect on the proportion of zero P2 duration. The relationship between the relative frequency of fluent repetitions and age, as well as years of education, are both Gaussian, and a group difference between speaker gender is also not observed. Thus it is unlikely that the proposed speaker features are correlated with the distribution of repetitive interpolations.

6.2.3. Interim summary

To sum up the discussion of speaker features' effect on the distribution of repetitions, I looked at both the relative frequency of repetitions compared to the absolute frequency of the same set of single syllable function words and the distribution of P2 duration among the same set of words. The idea is that these two measurements can provide a rough proxy for the distribution of repetitive interpolations. Combining the results from this independent analysis of two measurements of the distribution of repetitive interpolation, the general take-away is that the proposed speaker features almost never have an effect on the rate of repetitive interpolation. The only exception is the effect of gender, where male speak-

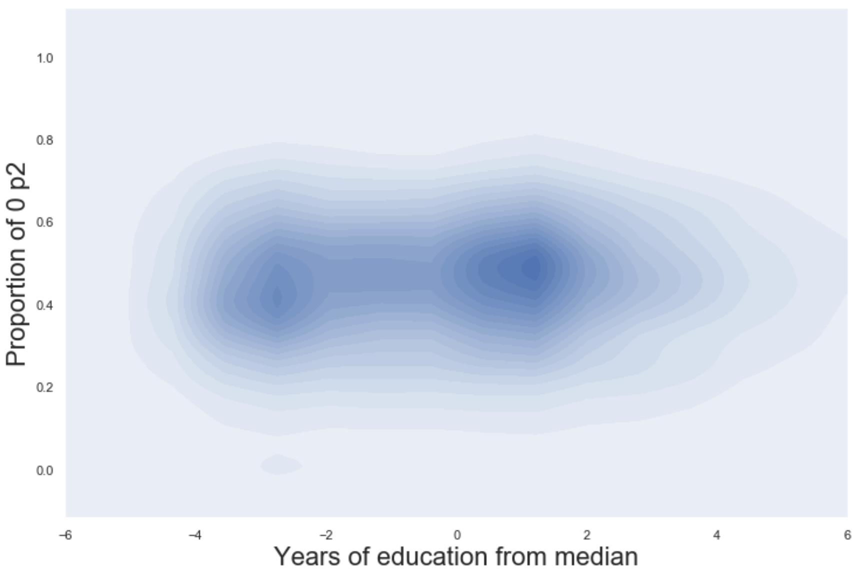


Figure 52: The relationship between years of education and the proportion of P2 whose duration is 0.

ers have higher rate of repetitions when words are controlled. However, they don't show higher proportion of zero pausing between repeated words in a repetition. The uniformity of fluent repetitive interpolation across different speaker groups could suggest that such interpolation phenomenon is indeed a sign of fluency: regardless of speaker's age and socioeconomic or intelligence status, cognitively healthy speakers are expected to produce similar repetitive interpolations with similar prosodic properties. The observation that this property of spontaneous speech is not correlated with age and years of education, the two parameters thought to reflect to certain extent the speakers' cognitive or intellectual ability, also suggests the possibility that repetitive interpolation can be purely mechanical. The observation that male speakers produce more repetitions could potentially be linked to the observation that male speakers produce larger amount of filled pause "uh" compared to female speakers. This potential parallel between "uh" and the distribution of repetitive interpolation prompts the hypothesis that in certain repetitions, the first repeated word repeated in fact serves as a hesitation marker. Since the distinction between "um" and "uh" is

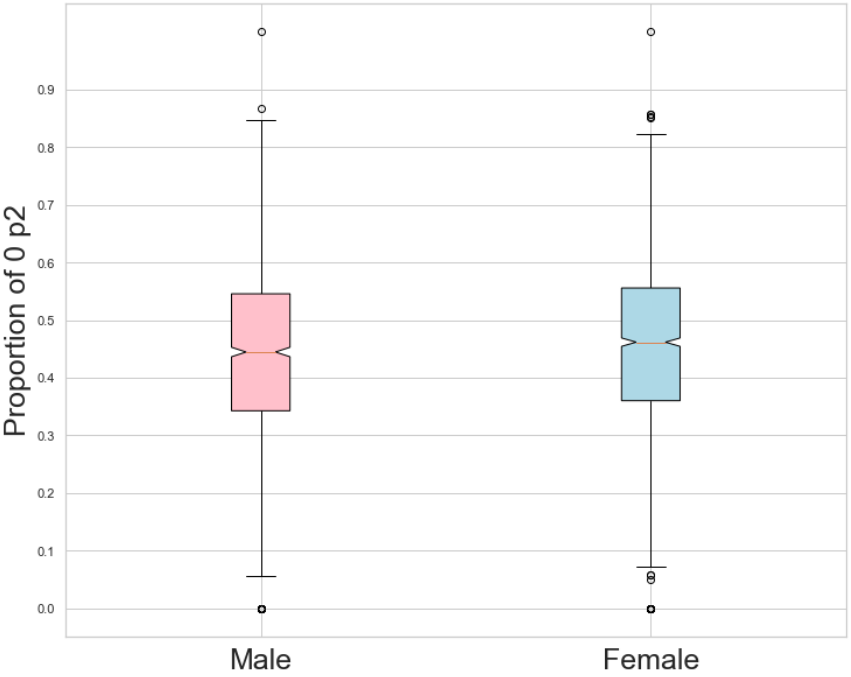


Figure 53: The relationship between gender and the proportion of P2 whose duration is 0.

better documented in the literature and people have reached better consensus on the nature of their distinction, a more convincing argument can be made with regard to the repetitive interpolation by looking at the two phenomenon together.

6.3. Speaker independent analysis

In this section, I shift the focus to speaker independent features that quantify the variation of repetitions. I first follow up on the observation made at the end of the last section and ask whether a parallel between repetitions and filled pause “uh” can be found. A parallel bimodal distribution of pause duration pertaining to repetitions and “uh” may suggest that certain repetitions serve similar function as certain filled pauses, likely indicating a delay in production. If this parallelism holds, it could further strengthens the proposal that certain repetitions, but not all, are a sign of hesitation. Such repetitions caused by hesitation should be distinguished from repetitive interpolation. Then I explore the contextual features that quantify the distribution of repetitive interpolation. First I ask the phrasal context in which these interpolations occur, then I address how conversation topics affect the distribution of repetition frequency and duration feature distributions, as a crude measure for the contextual effect on repetition.

6.3.1. Parallel between repetitive interpolation and "uh"

To test the hypothesis that repetitive interpolations are not signals of production delays, I first ask whether the distributions of repetitive interpolation and "uh" are independent from each other. This question is answered by testing if it is more likely to observe a filled pause either before or after a repetitive interpolation. If the distribution of the two is independent, it can be argued that repetitive interpolation does not happen in the context which some kind of delay is expected. On the other hand, if the distributions do not seem to be random, then repetitive interpolation would likely in fact be a hesitation induced phenomenon, under the assumption that repeating words and using filled pauses are mutually inducing factors in the face of hesitation.

The second question of whether disfluent repetitions and filled pause "uh" being potentially linked to the same production problem is addressed through examining whether similarities can be drawn between the acoustic properties of the two phenomena. More precisely, I would like to ask if "uh" shows a bimodal distribution of the silence duration following the filled pause. As I will discuss later, pause distribution between the repeated words in a single repetition clearly follows a bimodal distribution, suggesting the existence of two natural categories of repetitions: One in the form of fluent repetitive interpolations that I am proposing, and the other the repetitions that cause by delays in formulating a production plan. The existence of such parallelism would align well with the dichotomy between repetitive interpolation and disfluent repetition.

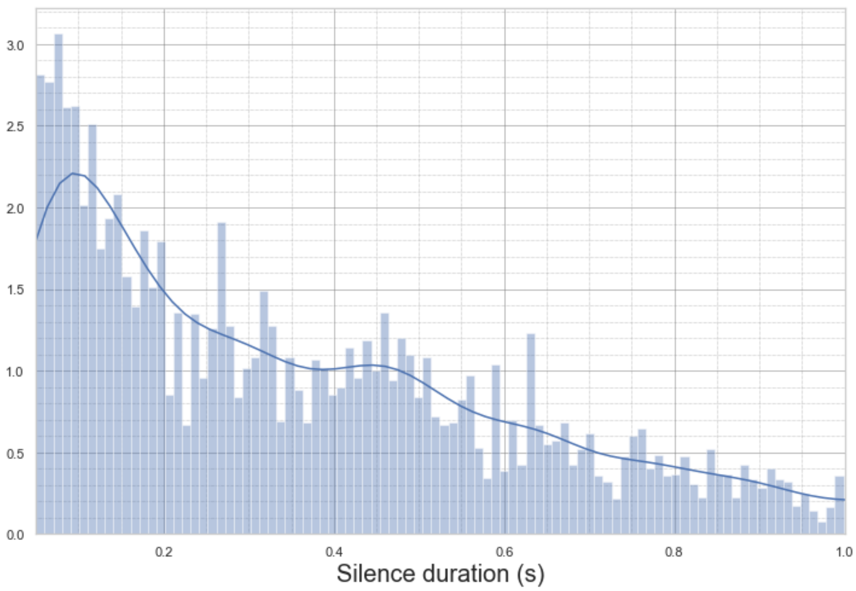


Figure 54: The distribution of silence duration following the filled pause “uh”.

Analysis The first hypothesis is examined by comparing the repetitions with a filled pause immediately before or after against the stand-alone repetitions. Repetitions of the 50 most commonly repeated single syllable function words are again used as the proxy for repetitive interpolation and treated as 50 independent observations. Wilcoxon signed ranked statistics is used to examine whether the frequency count are different. The test is not significant for both the comparison between filled pause before repetition and stand alone repetition ($p = 0.202$, statistic=15), and between filled pause following repetition and stand alone repetition ($p = 0.139$, statistic=13). Therefore there is evidence that repetitions occur independently of filled pauses.

The second question is answered by looking at the distribution of silence duration following the filled pause “uh”. This distribution is plotted in Figure 54. In this plot, only pauses with duration greater than 0 are plotted. The proportion of this duration being 0 is 62.4% out of a total of 26,465 examples. This distribution, however, should not be confused with the high proportion of 0 P2 duration for repetitive interpolation, as the filled pause it-

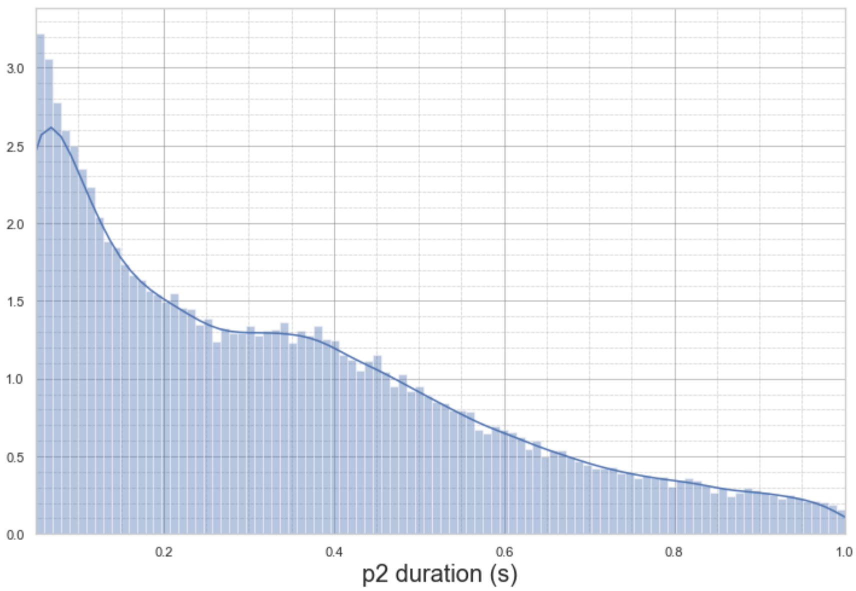


Figure 55: The distribution of silence duration between two repeated words in a repetition.

self signals delays in production. The histogram in Figure 54 shows that a weak secondary peak in the distribution around 0.4s can be observed in addition to the peak at around 0.1s. In comparison, the distribution of silence duration between two repeated words is plotted in Figure 55. The two distributions are indeed similar to each other, with a weak secondary peak around 400 ms.

Comparing the two plots of silence duration distributions, it is crucial to notice that they both have two peaks on each side of 200ms, with the peak greater than 200 ms less prominent. The processing time for a complete re-planning has been estimated to be around 200 ms (Civier et al., 2010). Therefore the bimodal distribution observed here supports the hypothesis that both repetitions and filled pause “uh” are related to delays potentially caused by replanning. This parallelism between repetition and “uh” supports the hypothesis that certain repetitions are disfluent and functionally similar to a hesitation marker, such as “uh”. However, the delay as represented by the secondary peak for “uh” is slightly longer than that of P2, and the distribution is more spread with a less well-defined categorical distinction between two peaks. This distinction might be related to the processing process which determines which hesitation marker to be used in the given context.

With the three pieces of evidence presented so far in terms of the parallel between "uh" and repetition: the same effect of gender, independence of occurrence and the similarity of adjacent silence duration, a uniform production mechanism can be proposed to account for "uh" and the repetitions that are disfluent. Under this hypothesis, unlike repetitive interpolation which indicates fluency, disfluent repetitions can be extended to a broader category of strategy that speakers can resort to when faced with problems in message formulation or replanning, in which "uh" and disfluent repetitions are both hesitation markers. I will delay the detailed discussion of the implications for production modeling to chapter 7. However, it is not clear how the selection of hesitation marker is determined in speech production, and how to understand the effect of gender on the relative frequency of repetition. In particular, it is yet to be tested that the difference between male and females in the relative frequency of repetition is driven by the small number of disfluent repetitions in the sample.

6.3.2. Context for repetitive interpolations

In this section, I examine contextual features for repetitive interpolation from two perspectives: the relative frequency of repetitive interpolations as the function of the repeated word location in an utterance, and the effect of conversation topic on the frequency distribution of repetitions. As has been discussed earlier, a primary drive for repeating single syllable function words is their proximity to the left edge of a phrase. The high frequency of repetition of single syllable function words has been shown to correlate with the high frequency of these words occurring at the beginning of the phrase. Here I further hypothesize that the beginning of an utterance would see higher rate of repetitive interpolation compared to elsewhere. On the other hand, repetition can also be a symptom for hesitation and speech repair. Unlike repetitive interpolations, disfluent repetitions are hypothesized not primarily

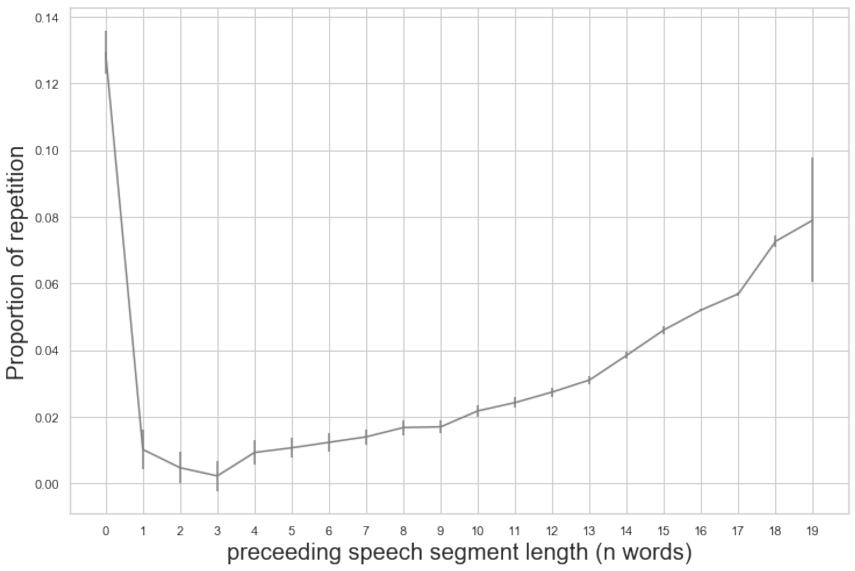


Figure 56: Proportion of repetitions of single syllable function words at different locations of a turn, measured by the number of words from the beginning of the turn.

driven by motor planning and execution issues. Therefore it is expected that disfluent repetitions may occur elsewhere in an utterance, although they could still be phrase initial. As for topic effect, it is hypothesized that certain topics may induce higher processing demand for the speaker, thus are correlated with higher rate of repetitions in general. However, since the design of conversation topic in the collection of the corpus was mainly as a way to encourage speakers talk, the effect size of conversation topic, at least as shown in the present dataset, is expected to be small and maybe inconsistent.

Results Figure 56 plots the proportion of repetitions of the 50 most frequently repeated single syllable function words among the same set of words at the same position in an utterance. An utterance is defined as a continuous speech segment between silent pauses of 250 ms or longer within a speaker turn. Utterances that are shorter than 5 words have been excluded from the analysis because they are most likely to be longer floor holding phrases such as “right right I see”. A turn is defined as a turn in a two-person conversation excluding back-channel talking and short floor holding, such as “yeah yeah”, “right right”.

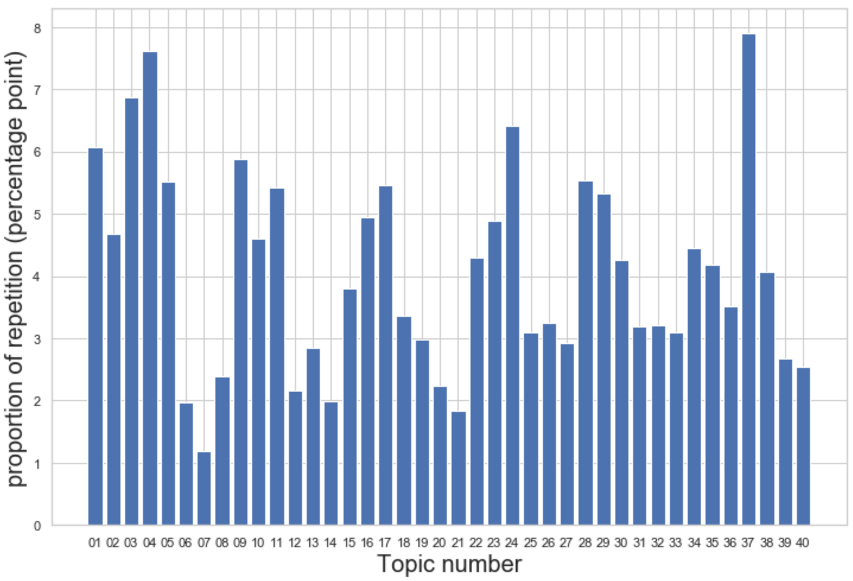


Figure 57: Bar plot of the variation of relative repetition rate across the 40 conversation topics given in Fisher. X-axis represents the arbitrarily assigned topic number.

The position of the repetition is measured as the location of the first word in a two-word repetition instance. The location of a word in a turn is measured as the number of words the current word is away from the first word of the turn. Thus position 0 refers to the first word in a turn.

As expected and consistent with the observation that higher frequency of repetition is observed at the beginning of a phrase, the beginning of an utterance also sees the highest relative frequency of repetitions. The relative frequency quickly drops towards the middle of the utterance, but steadily increases towards the end. Longer utterances are therefore expected to have high rate of repetition among single syllable function words towards the end. This trend in fact follows from the assumption that repetitions are related to two distinct underlying processes: repetitive interpolations are hypothesized as the result of issues related to lower-level utterance planning motor control, while disfluent repetitions are the result of higher level planning problems that directly lead to hesitation or repair. The steady increase of the relative repetition rate towards the end of a turn may thus suggest

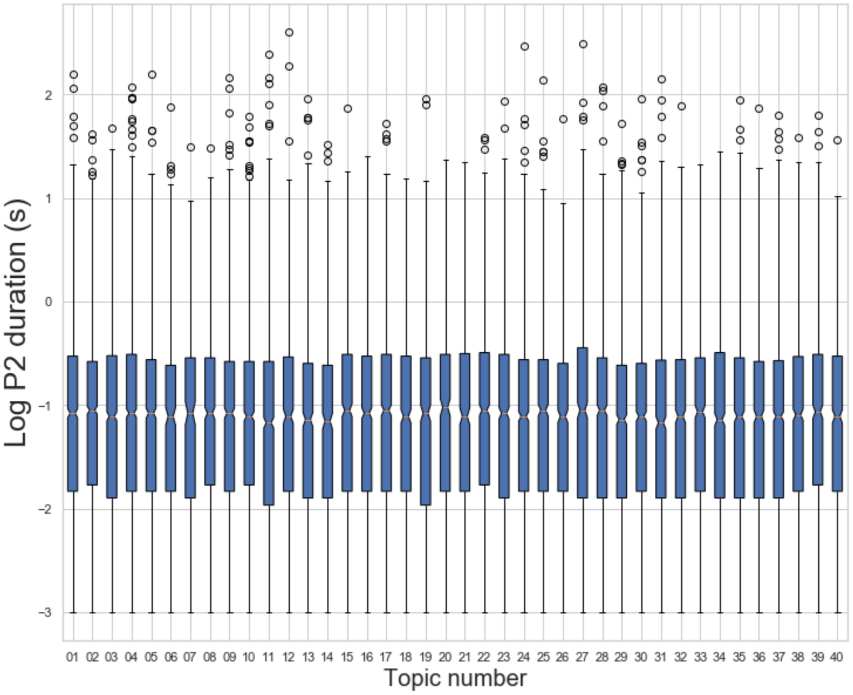


Figure 58: Bar plot of the distribution of non-zero P2 duration across topics given in Fisher. X-axis represents the arbitrarily assigned topic number.

the increased planning difficulty as a turn becomes longer. This increase is likely to be associated with more complex utterance structure with phrase conjunctions and embedding, which does not conflict with the general observation that repetitions tend to occur at phrase initial position.

In terms of the effect of conversation topic, I ask whether topic variation affects the relative repetition rate of the same set of single syllable function words used throughout the analysis, and whether it affects the duration properties of repetitions measured though the duration of P2 and the ratio of R1 and R2.

Figure 57 plots the variation of relative repetition rate across the 40 conversation topics in Fisher. The plot suggests a fair amount of variation across topics, with the lowest relative rate at around 1 percent and the highest reaching 8 percent. However, this variation is not significantly different from a random permutation of the same observed percentage values

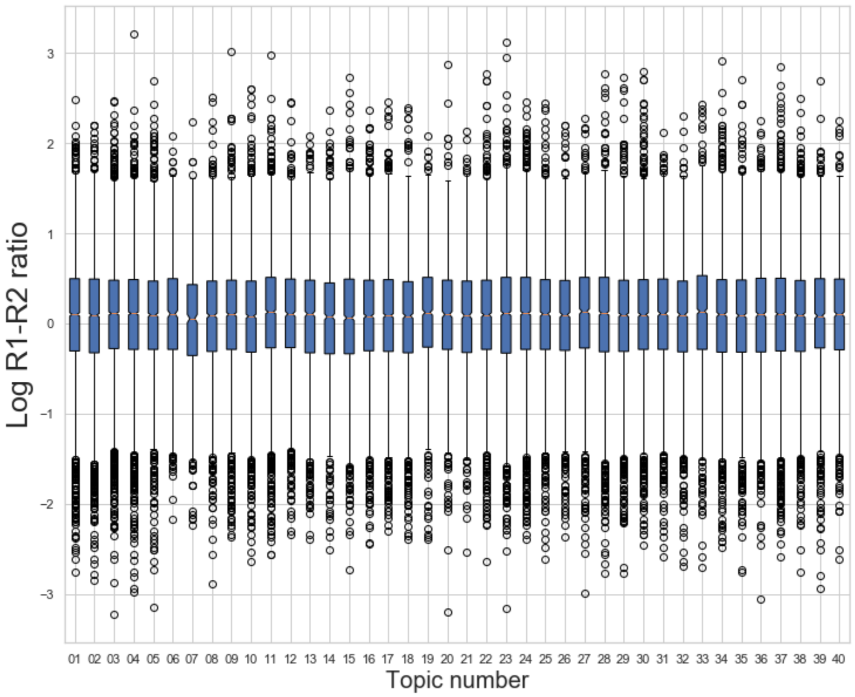


Figure 59: Bar plot of the R1/R2 ratio across the 40 conversation topics given in Fisher. X-axis represents the arbitrarily assigned topic number.

with a Wilcoxon signed rank test (statistic=367.5, p-value=0.753). Therefore it cannot be concluded that conversation topic as a measure of the discourse context of conversations has effect on the distribution of repetitions in general.

With regard to the variation in duration measurements, conversation topic also fails to show an observable effect, as shown in the box plots of Figure 58 and Figure 59. Figure 58 plots the distribution of the log of non-zero P2 duration across the pre-defined conversation topics, and Figure 59 plots the distribution of the log of R1-R2 ratio. The P2 measure is essentially uniform with the median at around 0.3s, and the ratio is uniform with a median slightly greater than 0, meaning roughly equal duration.

Summary To sum up, I have shown that the relative frequency of repetitions for the given set of single syllable function words is dependent on their location in an utterance. However, the results from the other proposed measurements of the contextual properties of the speech is inconclusive: although a fair amount of variation in the relative frequency of repetition has been observed across topics, this variation is not statistically different from a random permutation. The duration properties of repetitions also do not seem to vary as a function of conversation topic. Combing the observations from both the variation at utterance and larger discourse level, it can be proposed that changing cognitive demand during speech production is likely to have an effect on the rate of repetition phenomena and their typological distribution (repetitive interpolation or disfluent repetition). However, this effect is likely to directly stem from the changing demand during utterance formulation and delivery, while the effect from larger discourse context may be small or at least indirect. Since the current set up is not able to offer a more conclusive account on this latter issue, future work is needed to test the effect of broader communication context on the production of repetitions in a more controlled setting.

6.4. Summary of the analysis of normative speech

The discussion so far has addressed the questions regarding variations in repetition phenomena from two perspectives: A speaker dependent analysis in which I asked how features of the speaker correlate with the distribution and duration properties of repetitions, and a speaker independent analysis where I probed how the utterance context affect repetition. The speaker dependent analysis further strengthened the hypothesis that repetitive interpolation is the norm for repetition, with the majority of repetitions display traits of fluency rather than disfluency. However, both the distribution and duration features of repetitions do not seem to be strongly correlated with speaker's age, gender and years of education, except for an apparent effect of speaker's gender on the relative repetition rate. This observation is in parallel with the gender effect on the distribution of filled pause discussed earlier. A comparison between the distribution of pause duration between repeated words and the distribution of pause duration following “uh” suggests that disfluent repetitions are potentially driven by the same mechanism as what affects the production of “uh”. In this view, disfluent repetitions can serve as a hesitation marker similar to “uh”. This observation supports the proposal of a bipartite differentiation between fluent repetitive interpolations and disfluent repetitions. Further analysis of the utterance context also seems to support this hypothesis, although the question of how the cognitive demand for planning affect repetition is left with inconclusive results.

Combining the discussions from both perspectives, the discussion so far offers further evidence that repetition is indeed a sign of fluency, and repetitive interpolations should be distinguished from disfluent repetitions in terms of the underlying planning and production. This fluency feature is consistent across available measurements for speaker features and is potentially under the influence of the cognitive demand for speech planning. This consistency indicates that repetitive interpolation can be a robust and stable feature for the fluency of speech, such that cognitively healthy speakers, regardless of age and gender, should show similar patterns of repetitive interpolation production. Thus the existence of this repetitive phenomenon in one’s speech can potentially be used to detect cognitive decays that affect one’s linguistic ability. A potential parallel between repetitive interpolation and the use of filled pause “uh” may suggest a uniform mechanism that fluent speakers resort to when problem with motor planning and control arises. This potential mechanism for dealing with delays in production could serve as a bridge in understanding the transmission between higher level planning stage and the motor planning and control in speech production.

6.5. Repetitive interpolation in atypical speech: A case study of the effect of alcohol intoxication on repetitions in spontaneous speech

One practical implication of repetitive interpolation is its potential ability in distinguishing speakers with different disfluency-inducing conditions that related to impaired cognitive functions. Here impaired cognitive functions may refer to a neurological disorder, or other clinical or non-clinical conditions that affect one's ability to produce normative fluent speech. Some examples of such states include alcohol intoxication, sleep deprivation and certain types of neurological disorder. As I have shown in previous chapters, repetitive interpolation should be regarded as a phenomenon of fluency rather than disfluency in spontaneous speech, and the variation across speaker groups within normative population is expected to be small. This observation indicates that repetitive interpolation could be a strong cue for anomaly detection.

In the remainder of this chapter, I specifically look at one potential disfluency-inducing conditions: alcohol intoxication. This condition has been chosen for the following reasons. First, alcohol intoxication has been shown to affect brain functions that relate to speech production. For example, alcohol has been shown to lead to impaired movement control (Dawson and Reid, 1997). Second, alcohol intoxication could result in salient perception of reduced fluency or disfluency in the produced speech. Although empirical quantification for "drunk speech" is largely lacking, anecdotal experience suggests that speech produced by someone who is drunk is qualitatively distinctive from the speech produced by someone sober. Last but not least, since the underlying neurological correlates of alcohol intoxication are better understood, the deviations observed from this atypical condition can help with developing hypotheses and theories on repetitive interpolation, and speech production in a broad sense.

Results from the analysis in this chapter suggest that difference between speech produced under alcohol intoxication and normative fluent speech can be consistently detected through both acoustic and textual measurements. Reduced relative repetition frequency is observed, suggesting that impaired speech production system may in fact lead to reduced repetition, or repetitive interpolation. Measurable difference can also be found in terms of R1 and R2 duration, as well as the distribution of repeated lexical items. Therefore, quantitative measurement of repetitions can provide valuable information in discerning speakers with disfluency inducing cognitive conditions.

6.5.1. Background

Alcohol is known as a general depressant of the central nervous system. Alcohol consumption may lead to reduced inhibition and impaired movement control (Dawson and Reid, 1997). Although the direct effect of this impairment on speech-related movement control has not been directly tested in the literature, it is a valid hypothesis that the impaired movement control could potentially affect the motor control system in speech production (Grimme et al., 2011). Research on the health impact of alcohol mainly focuses on the short term, medium term and long term effect following excessive alcohol consumption (Cargiulo, 2007), and cognitive and physical impact of alcoholism (Dawson and Reid, 1997). A short term consequence of excessive alcohol consumption is drunk and drive, and its associated hazard to public safety. However, preventative measures to identify drunk drivers can be expensive and ineffective, as these measures mostly rely on physical inspections, such as through breath or blood alcohol concentration tests, of suspected drivers conducted by law enforcement who always does not have the resource to screen every possible offender. As a result, alcohol is often found to be a factor in car accidents only after the crash. Thus it has become an interest to many to understand the behavior changes after excessive alcohol consumption without the need to conduct breath or blood alcohol concentration tests. Alcohol-related change in speech is among the areas of interest in this line of research.

As reviewed in Chapter 3, ALC has been constructed to meet the demands for practical applications outlined above, and has been used as a standard database for shared tasks such as speaker state detection (Schuller et al., 2014). Using this corpus, previous research has examined alcohol induced changes of speech disfluencies, including silent and filled pauses, repetition, and repair. Early studies have focused on finding reliable acoustic cues for alcohol intoxication detection (Behne et al., 1991; Cooney, 1998). Although the effect of alcohol intoxication on speech production has been investigated earlier, such as in Tisljár-Szabó et al. (2014), the only systematic study of disfluencies with spontaneous speech isSchiel and Heinrich (2015). However, Contrary to the expectation that reduced inhibition and impaired movement control may lead to more disfluencies, only minor changes in the rate of silent and filled pauses, false starts, interruptions, and the duration of pauses have been found. However, changes in the rate of repetitions and phonemic lengthening appeared to be much greater (Schiel and Heinrich, 2015). Particularly, the rate of repetition has been found to decrease under the condition of alcohol intoxication. Considering repetitions as a kind of speech disfluency as discussed in the previous research, this change in repetition rate as reported in Schiel and Heinrich (2015) is in the opposite direction to what in theory one would expect.

Although the decreased repetition rate with alcohol intoxication may sound puzzling under the disfluency assumption of repetitions, this observation in fact follows nicely from the proposal made in this study that most of repetitions, i.e., repetitive interpolations, are in fact a sign of fluency. Here I take a step further to try understand why reduced repetitions are observed in alcohol speech by looking at variables that potentially covary with the change of frequency. Following the analysis on repetitions in typical fluent speech, in this section, I examine both the acoustic and textual covariates. I will show that comparing repetitions in typical fluent speech and alcohol speech would further demonstrate that repetitive interpolation can be a dimension in measuring the fluency of speech. The associated acoustic and textual features have the potential to be engineered for feature based classification systems for alcohol intoxication detection.

In the following analyses, I compare the speech produced with alcohol intoxication to typical fluent speech by asking the following two specific questions: What is the group difference caused by alcohol intoxication, and how alcohol intoxication affect the speech production for individual speakers. We ask both how the distribution of repeated forms may differ in two intoxication conditions, and how their acoustic manifestations, mainly in measurements of the duration of relevant segments, are different. In the following discussion, the alcohol condition will be abbreviated as A-condition, and non-alcohol condition as NA-condition.

6.5.2. Data and methods

The ALC (Schiel et al., 2008) corpus serves as a good resource for the current purpose. As reviewed above, ALC is a corpus of spoken German initially collected for the specific task of in car alcohol intoxication detection. The corpus contains speech produced by same subjects in both sober and intoxicated conditions from 162 speakers. For each speaker, recordings in two intoxication conditions were made with most potential confounding factors controlled, such as the recording microphone, room acoustics, and the kind of tasks they were asked to perform. In the alcohol intoxicated condition, the alcohol intoxication was controlled for by self-identified desired alcohol consumption level. Both blood and breath alcohol concentration were measured immediately after alcohol consumption, and the speech tasks were performed thereafter. Due to the high individual variation in alcohol tolerance, using subjectively defined alcohol consumption level explains away the potential additional variation introduced by a fixed physical measurement of alcohol consumption. These design considerations allow the data set to be used to make causal interpretations.

Speech tasks in the ALC corpus include reading out sentences, sequences of digits, describing the route to a place, short monologues prompted through both written questions and picture descriptions, and short conversations with the investigator. Thus the range of tasks covers both more realistic conversational settings and the ones that reflect the cognitive functions and motor control. For the current discussion, only the speech produced in spontaneous settings, including the monologue and short dialogues, are used for analysis. Detailed transcriptions were made available in the corpus, including word-by-word transcriptions of the speech and annotations of disfluencies. Repetitions are identified from the manual annotations provided in the corpus.

The textual measurements of repetitions include the absolute per-hundred word frequency of repetitions, group token similarity between conditions across speakers, and token similarity for a given speaker between two conditions. The acoustic measurement follows the previous discussion and concerns primarily on the duration of repeated segments (R1 and R2) and well as the duration between repeated segments in a repetition (P2).

The overall frequency of repetition is calculated as the number of repetitions in the speech produced by each speaker in each condition divided by the total number of words. Repetition tokens refer to the form of the repeated segments. Here the interest is how the token form distribution differs between the two condition. To compare token difference, I construct a binary vector for each speaker condition whose indices correspond to token forms that are repeated at least twice in the combined alcohol and non-alcohol conditions. The difference between repeated segment can be calculated as the spectral norm induced by the 2-norm of the difference matrix, $||\mathbf{B_A} - \mathbf{B_{NA}}||_2$, where $\mathbf{B}$ represents the $nspeaker \times nform$ $0-1$ matrix in each condition. Thus each row in the matrix corresponds to a row vector of token form distribution. Larger spectral norm of a matrix suggests the larger scale such a matrix can stretch a vector. Thus this matrix has higher variance, which with the current set up can be interpreted as larger distance between the two matrices. For duration properties, following the simplification made throughout this dissertation, only repetitions that repeated the same token form twice are considered. Segment duration information is extracted from the alignment timestamps provided in the corpus.

6.5.3. Results

In this section, I will first compare the overall frequency and duration differences between A and NA conditions. Then I will discuss in more detail how measurements of word distributions can reveal the effect of alcohol intoxication on the production of repetitions both within individual speakers and across speaker groups.

The frequency of repetitions replicates the observation reported in the literature. With 12 fewer speakers from the same corpus, Schiel and Heinrich (2015) reported a drop from about 0.7 percent to 0.45 percent absolute repetition frequency in spontaneous speech, measured as the number of repetitions divided by the total number of words. In this study, the frequency of repetition in NA condition is higher (7.27 per 1k words) than in A condition (4.58 per 1k words). In terms of duration features, I compare the difference between the repeated segment duration, as well as the pause duration between the repeated segment, between A and NA conditions. The absolute duration change reflects prolongation which may be caused by reduced motor control due to alcohol intoxication. On the other hand, changes in the ratio of R1/R2 could indicate the potential change in utterance planning. It has been reported that prolongation rate increased from about 0.3 percent to 0.5 percent in A condition (Schiel and Heinrich, 2015). Consistent with this increase in prolongation rate, both R1 and R2 are on average 56 ms longer under alcohol intoxication, and both differences are statistically significant ($p = 0.007$ *and* $p = 0.002$ *respectively*). However,

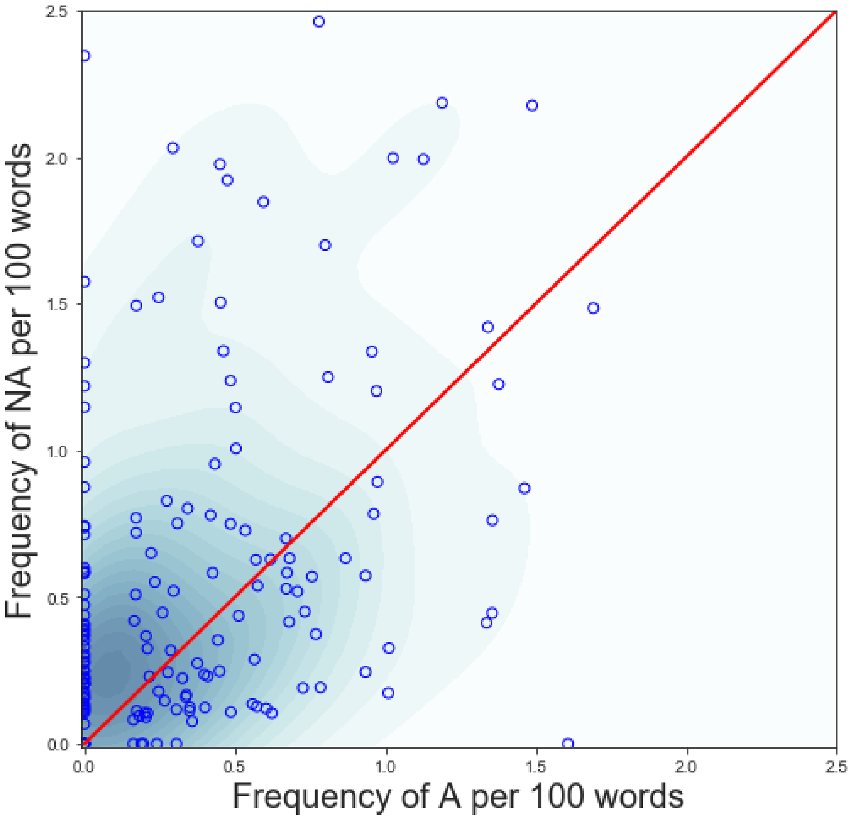


Figure 60: Repetition frequency distribution between alcohol and non-alcohol conditions.

the ratio of R1/R2 is not significantly different. Therefore alcohol intoxication is not likely to affect word-level or phrase-level replanning in repetition. As for the by-speaker cross-condition comparison of duration features, since most people do not have enough samples to make duration comparisons meaningful, duration features are not compared here.

One interesting question to ask is whether natural clusters can be identified by looking at the distribution of repetition frequency for each individual. Frequency difference between A and NA for each speaker is plotted in Figure 60. In this plot, each point represents an individual speaker plotted in the space defined by repetition frequency in A and NA conditions. In this scatter plot, three broad individual groups can be observed: those who repeat predominantly in the NA condition, represented by those closely follow the vertical axis (i.e., do not repeat in A), those who repeat more in the A condition, represented by

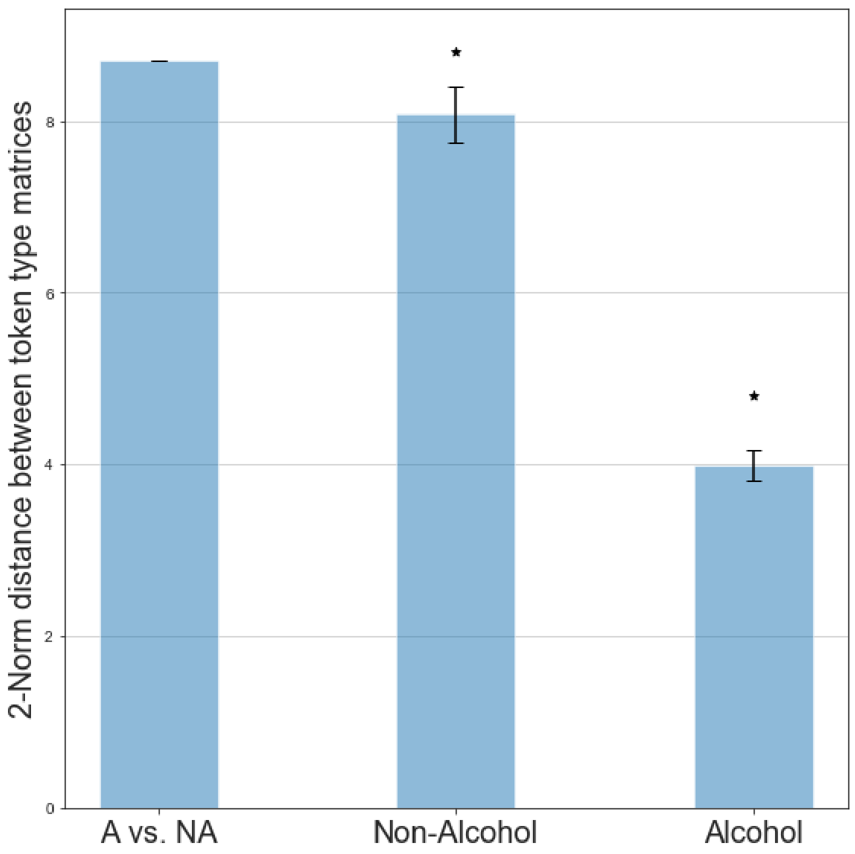


Figure 61: Word form difference between alcohol and non-alcohol conditions.

those dots roughly parallel to the horizontal axis, and those who repeat more or less with the same frequency, represented by those scattered along the equal distance line. The overlaid density plot is shifted towards the upper-left corner of the graph, suggesting that most speakers produce more repetitions in the NA condition. In particular, many of them only produce repetitions in the NA condition, while only a few speakers produce repetitions in the opposite direction. Thus for a given individual, it is also the case that more repetitions can be expected when they are not under the influence of alcohol.

A related question is whether the decreased repetition rate is associated with the change in the distribution of repeated segments. If the effect of alcohol is disproportionately heavy on the motor control system, it could be expected that this distribution would be different, so that otherwise non-repetitive segments are repeated due to the reduced movement control. To test the segment distribution difference between A and NA, three comparisons are made: two within condition comparisons where 50 random half-samples in each condition are compared to the other half of the same state (within-condition difference) in addition to the between-condition comparison. This within-condition comparison provides a baseline for how any random subsets of speakers are compared to the rest of the sample in terms of the repeated form distribution. Larger within-comparison 2-norm value indicates higher variability in repeated segment form distribution in the given condition. Then the between-condition variation can be compared to the within-condition baseline, and the two within-condition baseline can be compared to each other. Results show, as plotted in Figure 61, that between-condition difference is greater than both of the average within-condition baselines, while the difference between the between-condition 2-norm score and the NA condition within-condition is smaller than the difference between the within-condition score for the A condition. NA also has larger within-state difference compared to A, suggesting greater variability in repeated forms. Thus the observation here is that speakers repeat a different set of words and phrases when they are not under the influence of alcohol compared to when they are alcohol intoxicated. They also repeat a wider range of segments when they are sober.

The next question to be addressed is individual variation: for a given speaker, how much difference can be expected comparing their speech with and without the influence of alcohol intoxication? This within-speaker cross-condition difference for each speaker is measured as the cosine similarity between the two form vectors in two intoxication conditions. As Figure 62 shows, for majority speakers, their similarity scores are essentially 0, suggesting that the repeated words and phrases in the A condition do not overlap with the word and phrases produced in the NA condition. Although this comparison might be highly biased and discretized due to the limited number of repetition examples observed for each

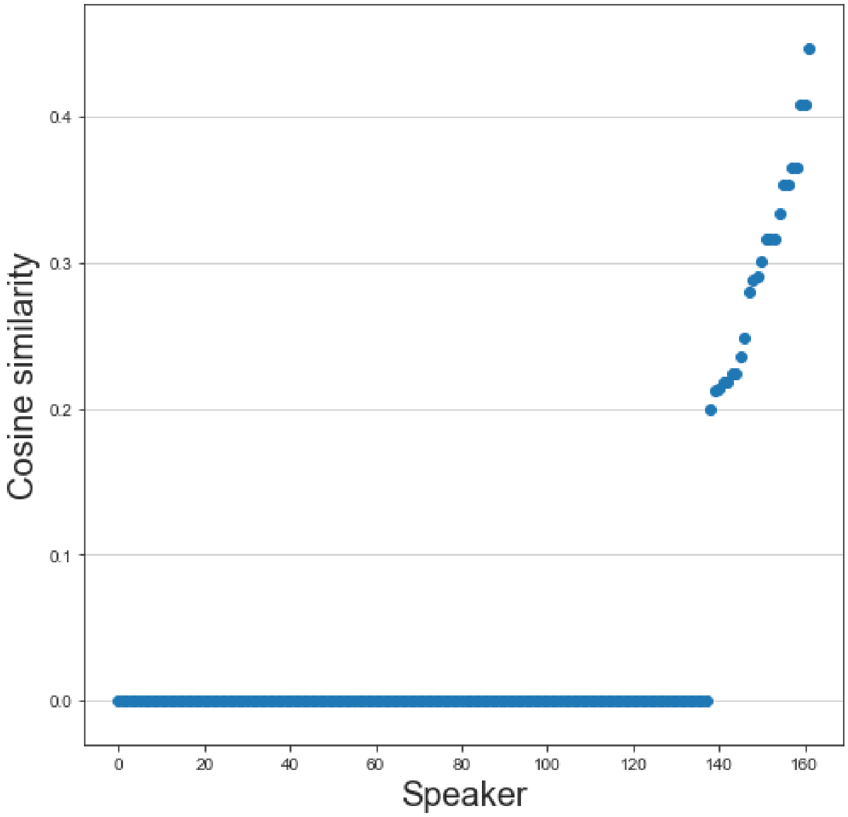


Figure 62: Similarity between repeated words in alcohol and non-alcohol conditions.

individual, especially in the A condition, the true by-speaker cross-condition similarity can still be expected to be low if not completely zero. Therefore for a given speaker, it can be expected that they are likely to repeat different words and phrases when they are under the influence of alcohol.

6.5.4. Discussion

In this section, I examined the effect of alcohol intoxication on the production of repetitions using ALC corpus. The design of the corpus enabled direct causal interpretation of the results reported above. I addressed this question by looking at the textual and duration features of spontaneous speech produced in sober and alcohol intoxicated conditions. Consistent with previous research, the overall repetition frequency is higher when the speaker is not intoxicated. Although elongations of repeated segment are observed, the duration ratios do not show a difference between the two conditions, suggesting the lack of impact on utterance planning. As for textual features, greater variation in the form of repeated segment is found in the non-intoxicated condition. For each individual speaker, the repeated segments are expected to be quite different in two conditions.

The comparisons reported above suggests that alcohol intoxication can directly cause the change in repetitive behavior in the production of spontaneous speech. The apparent elongation effect can potentially be explained by the reduced movement control caused by alcohol intoxication. The decreased repetition rate under alcohol intoxication may provide a further piece of evidence that most repetitions are not just a result of motor control problem. Combining the textual analysis, one possible interpretation could be that due to reduced motor control under alcohol intoxication, certain words or phrases with low repetition frequency in typical fluent speech are repeated with higher frequency. On the other hand, alcohol intoxication may have affected other the cognitive processes that are related to speech production, such as message formulation or utterance planning, such that the repetition rate appears to become lower. With this potential inhibitory effect on certain modules in speech production, reduced amount of repetitions are produced. This potential hypothetical explanation for the decreased repetition rate is consistent with the proposal that repetitive interpolations can be mostly regarded a fluent phenomenon, and the majority repetition instances are repetitive interpolations.

Although the observed differences in this section can be directly attributed to alcohol intoxication, it is still not clear how the intermediate steps are affected by alcohol. Nevertheless, the hypothesis that repetitive interpolation is a fluency phenomenon can lead to future experimental work that are able to test potential explanations more directly and efficiently.

6.6. *Chapter summary*

An understanding of the variation of repetitive interpolation has both theoretical and practical implications. Theoretically, inter-speaker and context-dependent variations could motivate hypotheses on the production processes that are responsible for the observed repetitions. Conversely, a theoretical account for the production of repetitive interpolation would also contribute to our understanding of the cognitive processes involved in speech production. Practically, group difference or similarities could provide invaluable information for applications such as speaker state detection and cognitive assessment. In this chapter, I provided an initial descriptive analysis of the potential individual and contextual variation of repetitive interpolation. I also showed that measurements of this repetition phenomenon can provide interesting insight in practical applications such as alcohol intoxication detection. The description of the variation presented so far has provided a good starting point for future work in speech production in more controlled settings.

Separate analyses have been performed to address questions regarding individual variation, the variation that is related to the immediate and discourse context, as well as speakers with and without the influence of substance. In terms of individual variation, the overall take away from the descriptive analysis is that variations, both in terms of the frequency and duration features of repetitions, are very limited across speaker groups traditionally defined in sociolinguistic literature. The only exception is the higher relative frequency of repetition among male speakers. This observation could potentially be related to the functional parallelism between disfluent repetitions and filled pause “uh”. Contextually, it appears that only the position within an utterance is correlated with the relative frequency of repetitive interpolation. Thus it is likely that the production of repetitive interpolation mainly involves utterance level planning and maybe the formulation and execution of a motor plan. The comparison between repetitions produced with and without alcohol intoxication could also be explained by a mechanism that involves utterance and motor planning but not higher level message formulation. A possible explanation could be found at the coordination between utterance planning and motor control.

Chapter 7

Conclusion

In this chapter, I will summarize the results and discuss their implications from discussions presented so far in this dissertation. Since the broad term “speech disfluencies” touches upon the interests across a wide variety of fields, a thorough understanding of both the potential variations and their empirical distributions is necessary for subsequent applied and theoretical work on problems regarding human speech. Although research on speech disfluencies has gained a strong momentum in the past decades, descriptive studies based on large collections of real spontaneous speech data across multiple languages and communication settings are still lacking. To meet this end, this dissertation has contributed to the research community by providing an empirical description of major disfluency phenomena across multiple communication settings, languages, as well as speaker’s cognitive states. Although the goal of this dissertation is fundamentally descriptive, patterns discovered throughout the discussion would be proven invaluable for both the theoretical understanding of speech production and empirical applications such as intoxication detection and cognitive assessment.

7.1. A general summary

The empirical descriptions in this dissertation started off by examining the variation of silent and filled pauses. I have shown that the structure of the temporal relationship between silence and speech segments can be effectively captured through exploring the latent dimensions of the joint probabilistic space defined by the relative relation between silence and speech segments. Interpretations of the differences found in the latent space can be obtained through establishing correlations between the one-dimensional projection of the joint space and sociolinguistic and contextual variables. I then provided a thorough description of the distribution of filled pauses by examining the distribution properties of the two forms of filled pause with regard to both speaker and contextual features. Through jointly considering the speaker and contextual features, I have argued that the proposed change-in-progress account on the relative frequency of “um” and “uh” in sociolinguistic literature is not able to rule out the possibility of age-related change in hesitation behavior, and the different discourse meanings encoded in the two variants of filled pause. This analysis further raises the question of how the variation of filled pauses can reflect variations and changes in the underlying speech planning process. Answers to this question would have deeper implications for both theoretical and applied interests in speech production models.

The discussion of repetitions has been centered around documenting empirical evidence for a separate class of repetition phenomenon that I proposed to be called repetitive interpolation. I have shown that this is a kind of repetition that a typical disfluent view of repetitions is incompatible with. This phenomenon can be described as rapid repetitions of single syllable function words in otherwise fluent speech delivery. Specifically, both acoustic and textual analyses have suggested that typical symptoms that are associated with speech disfluencies, such as disrupted speech delivery, increased semantic complexity in adjacent to the repetitions, and the semantic complexity of the repeated word, were not observed. I further examined the possibility of an explanation purely from the perspective of speech motor control. I have presented evidence, both through comparisons with repetitions in stuttering, read speech and a morphologically complex language, and descriptions of its distribution property across different speaker groups, that an account from speech motor control alone is not likely to offer an adequate theory of the production mechanism behind this phenomenon. However, current state of the art models on speech motor control can be a good starting point for establishing the connection between the processing at the level of utterance planning and the execution of a formulated utterance plan.

7.2. Implications for speech production modeling

The three major speech disfluency phenomena that I have explored all have close connections to different levels of processing involved in the planning of speech. The good performance of silence and speech duration based feature representation in distinguishing speakers under alcohol intoxication suggests that silent pauses are salient indicators for problems in the motor planning and control of speech production. The potential effects of speaker and contextual features on the distribution of two forms of filled pause can be interpreted as the different discourse meanings or functions of the filled pauses that associated with higher level message structuring and planning processes. This interpretation, if tested in future research, could cast doubt on the change-in-progress explanation of the frequency distribution difference exhibited between two filled pauses as a shallow account conditioned on the availability of data. Therefore a complete picture for the use of filled pauses in spontaneous speech should consider both the cognitive and pragmatic aspects of the problem.

A major contribution of this dissertation to the modeling of the speech production process is the introduction of a previously under-discussed repetition phenomenon: repetitive interpolation. The lack of signs for typical disfluency yet rapidly repeating single syllable function words in fluent speech delivery, as elaborated in the discussions above, suggest that this form of repetition can potentially become a test ground for theories about the coordination between higher level message formulation and utterance planning process and lower level motor planning and control. As I have argued in previous chapters, neither a model of speech motor control along nor a model that includes message replanning is able to explain the variation of both the form and distribution of repetitive interpolations observed in the available speech samples. Although the descriptive analysis presented so far is not able to lead to specific theoretical claims about the nature of repetitive interpolation, as well as detailed proposals for a model of speech production, the available information should serve as a foundation for future experimental and simulation works on the pertinent issues regarding speech production. In particular, repetitive interpolation can be informative for the modeling of the transmission of information between stages in the planning process. In this regard, recent research on neural-based models of speech motor control can offer some invaluable initial ideas on how to move forward.

The existing computational models for speech motor control, such as DIVA, GODIVA and HSFC models (Guenther, 2006; Bohland et al., 2010; Hickok, 2012), have drawn a rather detailed picture of the neurological basis of the planning and execution of speech motor commands. The inclusion of a forward and feedback model for both the physical and somatosensory output of motor commands, as illustrated in DIVA (Guenther, 2006), could be extended to model more abstract speech planning and monitoring processes. The problems of syllabic sequencing in the execution of a speech motor plan could potentially be reformulated to fit the need of the sequencing of more abstract planning unit. The solution to this problem, which has been proposed in the GODIVA model (Bohland et al., 2010) as a competitive queuing (CQ) mechanism, is also extendable to abstract sequencing problem. This mechanism utilizes the binary operation of excitation and inhibition of previously unused and excited unit to achieve the desired sequencing outcome. The selection of the next unit to be executed is accomplished through comparing the activation potential of the remaining available phonological unit. The dual stream mechanism (Hickok and Poeppel, 2007) essentially provides an assumption to the pathways that connects higher level and lower level processing stages, such that the forward mapping and monitoring can be achieved efficiently across multiple processing levels. Pieces from models of speech motor control mentioned above could serve as initial hypothesis for the underlying process for repetitive interpolation, which might be refined with further evidence from more controlled experimental or simulation settings.

One practical implication for such modeling effort is to develop robust yet interpretable feature space that could serve in applications such as speaker state detection, screening and diagnosis of neural degeneration, as well as evaluation of the proficiency of second language learners. The present study has already shown the effectiveness of multiple dimensions of speech disfluencies in distinguishing speakers under the influence of alcohol intoxication. It is equally promising in developing applications in other areas concerning spontaneous speech with an in-depth understanding of key disfluency phenomena.

7.3. Future directions

Future research on speech disfluencies should follow two fundamental streams of thoughts: A line of theoretical work can be developed to experimentally test empirical observations from speech corpora across different languages, conversation contexts and speaker's cognitive states. On the other hand, feature engineering aiming at developing representations of aspects of speech disfluencies for potential practical applications should also receive its fair share of attention. Although the two streams of research effort intrinsically cater to communities interested in rather different questions, they nevertheless complement each other in a mutually beneficial if not dependent manner. Efforts in automatically identifying and processing speech disfluencies with both textual and acoustic input are also naturally indispensable, especially in an era that massive amount of data can be made available easily yet proper annotation is still somewhat prohibitively expensive.

In the introduction session of her dissertation, Shriberg (1994) acknowledged that although an all-encompassing theory of disfluencies is the ultimate goal of disfluency research, the field was in an early stage of discovering the regularities in disfluency production. Although the past 25 years have seen tremendous development in the field, I will keep stressing the need of the continuous effort in this pattern recognition enterprise. What an ultimate theory on speech disfluencies may still be out of reach in the current date and time, but every piece of finer description of the phenomena will bring us a step closer to the ever expanding goal. I would like to finally argue that a theory on disfluencies is one that unites streams of thought in the pursuit of knowledge about human speech from a diverse set of perspectives, and brings immense impact on the advancement of human language technology.

APPENDIX

A. Topic and prompt list in Fisher Corpus

- **ENG01** Professional Sports on TV: Do either of you have a favorite TV sport? How many hours per week do you spend watching it and other sporting events on TV?

- **ENG02** Pets: Do either of you have a pet? If so, how much time each day do you spend with your pet? How important is your pet to you?

- **ENG03** Life Partners: What do each of you think is the most important thing to look for in a life partner?

- **ENG04** Minimum Wage: Do each of you feel the minimum wage increase - to $5.15 an hour - is sufficient?

- **ENG05** Comedy: How do you each draw the line between acceptable humor and humor that is in bad taste?

- **ENG06** Hypothetical Situations. Perjury: Do either of you think that you would commit perjury for a close friend or family member?

- **ENG07** Hypothetical Situations. One Million Dollars to leave the US.: Would either of you accept one million dollars to leave the US and never return? If you were willing to leave, where would you go, what would you do? What would you miss the most about the US? What would you not miss?

- **ENG08** Hypothetical Situations. Opening your own business: If each of you could open your own business, and money were not an issue, what type of business would you open? How would you go about doing this? Do you feel you would be a successful business owner?

- **ENG09** Hypothetical Situations. Time Travel.: If each of you had the opportunity to go back in time and change something that you had done, what would it be and why?

- **ENG10** Hypothetical Situations. An Anonymous Benefactor: If an unknown benefactor offered each of you a million dollars - with the only stipulation being that you could never speak to your best friend again - would you take the million dollars?

- **ENG11** US Public Schools.: In your opinions, is there currently something seriously wrong with the public school system in the US, and if so, what can be done to correct it?

- **ENG12** Affirmative Action.: Do either of you think affirmative action in hiring and promotion within the business community is a good policy?

- **ENG13** Movies.: Do each of you enjoy going to the movies in a theater, or would you rather rent a movie and stay home? What was the last movie that you saw? Was it good or bad and why?

- **ENG14** Computer games.: Do either of you play computer games? Do you play these games on the internet or on CD- ROM? What is your favorite game?

- **ENG15** Current Events.: How do both of you keep up with current events? Do you get most of your news from TV, radio, newspapers, or people you know?

- **ENG16** Hobbies.: What are your favorite hobbies? How much time do each of you spend pursuing your hobbies? Do you feel that every person needs at least one hobby?

- **ENG17** Smoking.: How do you both feel about the movement to ban smoking in all public places? Do either of you think Smoking Prevention Programs, Counter-smoking ads, Help Quit hotlines and so on, are a good idea?

- **ENG18** Terrorism.: Do you think most people would remain calm, or panic during a terrorist attack? How do you think each of you would react?

- **ENG19** Televised Criminal Trials.: Do either of you feel that criminal trials, especially those involving high-profile individuals, should be televised? Have you ever watched any high-profile trials on TV?

- **ENG20** Drug testing.: How do each of you feel about the practice of companies testing employees for drugs? Do you feel unannounced spot-checking for drugs to be an invasion of a person's privacy?

- **ENG21** Family Values.: Do either of you feel that the increase in the divorce rate in the US has altered your behavior? Has it changed your views on the institution of marriage?

- **ENG22** Censorship.: Do either of you think public or private schools have the right to forbid students to read certain books?

- **ENG23** Health and Fitness.: Do each of you exercise regularly to maintain your health or fitness level? If so, what do you do? If not, would you like to start?

- **ENG24** September 11.: What changes, if any, have either of you made in your life since the terrorist attacks of Sept 11, 2001?

- **ENG25** Strikes by Professional Athletes.: How do each of you feel about the recent strikes by professional athletes? Do you think that professional athletes deserve the high salaries they currently receive?

- **ENG26** Airport Security.: Do either of you think that heightened airport security lessens the chance of terrorist incidents in the air?

- **ENG27** Issues in the Middle East.: What does each of you think about the current unrest in the Middle East? Do you feel that peace will ever be attained in the area? Should the US remain involved in the peace process?

- **ENG28** Foreign Relations.: Do either of you consider any other countries to be a threat to US safety? If so, which countries and why?

- **ENG29** Education.: What do each of you think about computers in education? Do they improve or harm education?

- **ENG30** Family.: What does the word family mean to each of you?

- **ENG31** Corporate Conduct in the US.: What do each of you think the government can do to curb illegal business activity? Has the cascade of corporate scandals caused the mild recession and decline in the US stock market and economy? How have the scandals affected you?

- **ENG32** Outdoor Activities.: Do you like cold weather or warm weather activities the best? Do you like outside or inside activities better? Each of you should talk about your favorite activities.

- **ENG33** Friends.: Are either of you the type of person who has lots of friends and acquaintances or do you just have a few close friends? Each of you should talk about your best friend or friends.

- **ENG34** Food.: Which do each of you like better–eating at a restaurant or at home? Describe your perfect meal.

- **ENG35** Illness.: When the seasons change, many people get ill. Do either of you? What do you do to keep yourself well? There is a saying, ”A cold lasts seven days if you don’t go to the doctor and a week if you do.” Do you both agree?

- **ENG36** Personal Habits.": According to each of you, which is worse: gossiping, smoking, drinking alcohol or caffeine excessively, overeating, or not exercising?

- **ENG37** Reality TV.: Do either of you watch reality shows on TV. If so, which one or ones? Why do you think that reality based television programming, shows like "Survivor" or "Who Wants to Marry a Millionaire" are so popular?

- **ENG38** Arms Inspections in Iraq.: What, if anything, do you both think the US should do about Iraq? Do you think that disarming Iraq should be a major priority for the US?

- **ENG39** Holidays.: Do either of you have a favorit holiday? Why? If either of you you could create a holiday, what would it be and how would you have people celebrate it?

- **ENG40** Bioterrorism.: What do you both think the US can do to prevent a bioterrorist attack?

B. Links to supplemental audio files of examples used in the dissertation

Information of selected examples used in the dissertation is summarized in the table below:

id	Speaker	Source	url
2	Breyer	SCOTUS	hozh3497.github.io/audios/breyers.mp3
3	O'Connell	SCOTUS	hozh3497.github.io/audios/oconnell.mp3
4	Kennedy	SCOTUS	hozh3497.github.io/audios/kennedy.mp3
5	Scalia	SCOTUS	hozh3497.github.io/audios/scalia.mp3
6	Unknown	Fisher	hozh3497.github.io/audios/fluentrep.mp3
7	Unknown	Fisher	hozh3497.github.io/audios/delayedrep.mp3
8	Unknown	Fisher	hozh3497.github.io/audios/hesitiationrep.mp3
9	Unknown	Fisher	hozh3497.github.io/audios/repairrep.mp3
10	Unknown	UCLASS	hozh3497.github.io/audios/syllableRepBlock.mp3
11	Unknown	UCLASS	hozh3497.github.io/audios/consonantRep.mp3
12	Unknown	UCLASS	hozh3497.github.io/audios/prolongationNonInitial.mp3
13	Unknown	UCLASS	hozh3497.github.io/audios/quasiFluentRepMono.mp3
14	Unknown	UCLASS	hozh3497.github.io/audios/quasiFluentRep.mp3
17	Unknown	Czech	hozh3497.github.io/audios/czech1.mp3
18	Unknown	Czech	hozh3497.github.io/audios/czech2.mp3
19	Unknown	Czech	hozh3497.github.io/audios/czech3.mp3
20	Unknown	Czech	hozh3497.github.io/audios/czech4.mp3

BIBLIOGRAPHY

E. K. Acton. On gender differences in the distribution of um and uh. *University of Pennsylvania Working Papers in Linguistics*, 17(2):2, 2011.

J. Adell, D. Escudero, and A. Bonafonte. Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication*, 54 (3):459–476, 2012.

S. Ahmed, C. A. de Jager, A.-M. Haigh, and P. Garrard. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology*, 27(1):79, 2013.

J. E. Arnold, C. L. H. Kam, and M. K. Tanenhaus. If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5): 914, 2007.

S. Ash, P. Moore, L. Vesely, D. Gunawardena, C. McMillan, C. Anderson, B. Avants, and M. Grossman. Non-fluent speech in frontotemporal lobar degeneration. *Journal of Neurolinguistics*, 22(4):370–383, 2009.

S. Ash, C. McMillan, D. Gunawardena, B. Avants, B. Morgan, A. Khan, P. Moore, J. Gee, and M. Grossman. Speech errors in progressive non-fluent aphasia. *Brain and Language*, 113(1):13–20, 2010.

S. Ash, C. McMillan, R. G. Gross, P. Cook, D. Gunawardena, B. Morgan, A. Boller, A. Siderowf, and M. Grossman. Impairments of speech fluency in Lewy body spectrum disorder. *Brain and Language*, 120(3):290–302, 2012.

S. Ash, E. Evans, J. O'Shea, J. Powers, A. Boller, D. Weinberg, J. Haley, C. McMillan, D. J. Irwin, K. Rascovsky, et al. Differentiating primary progressive aphasias in a brief sample of connected speech. *Neurology*, 81(4):329–336, 2013.

B. Baumeister, C. Heinrich, and F. Schiel. The influence of alcoholic intoxication on the fundamental frequency of female and male speakers. *The Journal of the Acoustical Society of America*, 132(1):442–451, 2012.

J. Bear, J. Dowding, E. Shriberg, and P. Price. A system for labeling self-repairs in speech, 1993.

G. W. Beattie and P. Barnard. The temporal structure of natural telephone conversations (directory enquiry calls). *Linguistics*, 17(3-4):213–230, 1979.

G. W. Beattie and B. L. Butterworth. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, 22(3): 201–211, 1979.

D. M. Behne, S. M. Rivera, and D. B. Pisoni. Effects of alcohol on speech: Durations of isolated words, sentences, and passages in fluent speech. *The Journal of the Acoustical Society of America*, 90(4):2311–2311, 1991.

A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024, 2003.

Š. Beňuš, A. Gravano, and J. Hirschberg. Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, 43(12):3001–3027, 2011.

Š. Beňuš, R. Levitan, and J. Hirschberg. Entrainment in spontaneous speech: The case of filled pauses in Supreme Court hearings. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 793–797. IEEE, 2012.

K. Berninger, J. Hoppe, and B. Milde. Classification of speaker intoxication using a bidirectional recurrent neural network. In *International Conference on Text, Speech, and Dialogue*, pages 435–442. Springer, 2016.

E. R. Blacfkmer and J. L. Mitton. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39(3):173–194, 1991.

J. Blankenship and C. Kay. Hesitation phenomena in English speech: A study in distribution. *Word*, 20(3):360–372, 1964.

O. Bloodstein and N. B. Ratner. *A Handbook on Stuttering*. Delmar Learning, 2008.

J. W. Bohland, D. Bullock, and F. H. Guenther. Neural representations and mechanisms for the performance of simple speech sequences. *Journal of Cognitive Neuroscience*, 22(7): 1504–1529, 2010.

D. Bohus and E. Horvitz. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 2–9. ACM, 2014.

D. Bone, M. Li, M. P. Black, and S. S. Narayanan. Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors. *Computer Speech & Language*, 28(2):375–391, 2014.

G. J. Borden. An interpretation of research on feedback interruption in speech. *Brain and Language*, 7(3):307–319, 1979.

H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147, 2001.

V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa. Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*, 8: 269, 2017.

S. E. Brennan and M. Williams. The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34(3):383–398, 1995.

P. A. Broen and G. M. Siegel. Variations in normal speech disfluencies. *Language and Speech*, 15(3):219–231, 1972.

A. Budhkar and F. Rudzicz. Augmenting word2vec with latent Dirichlet allocation within a clinical application. *arXiv preprint arXiv:1808.03967*, 2018.

A. Butcher. Aspects of the speech pause: Phonetic correlates and communication functions. *Arbeitsberichte Kiel*, (15):1–233, 1981.

J. Butzberger, H. Murveit, E. Shriberg, and P. Price. Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceedings of the Workshop on Speech and Natural Language*, pages 339–343. Association for Computational Linguistics, 1992.

E. Campione and J. Véronis. A large-scale multilingual study of silent pause duration. In *Speech Prosody 2002, International Conference*, 2002.

T. Cargiulo. *Understanding the Health Impact of Alcohol Dependence*. Oxford University Press, 2007.

C. Cieri, D. Miller, and K. Walker. The Fisher Corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71, 2004.

O. Civier, S. M. Tasko, and F. H. Guenther. Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production. *Journal of Fluency Disorders*, 35(3):246–279, 2010.

H. H. Clark and J. E. F. Tree. Using uh and um in spontaneous speaking. *Cognition*, 84(1): 73–111, 2002.

H. H. Clark and T. Wasow. Repeating words in spontaneous speech. *Cognitive Psychology*, 37(3):201–242, 1998.

O. Cooney. *Acoustic Analysis of the Effects of Alcohol on the Human Voice*. PhD thesis, Dublin City University, 1998.

W. E. Cooper and J. Paccia-Cooper. *Syntax and Speech*. Number 3. Harvard University Press, 1980.

M. Corley and O. W. Stewart. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602, 2008.

N. Cummins, A. Baird, and B. W. Schuller. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54, 2018.

D. A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Proceedings of the Workshop on Human Language Technology*, pages 43–48. Association for Computational Linguistics, 1994.

L. Data Consortium. Czech Broadcast Conversation MDE Transcripts. `https://catalog.ldc.upenn.edu/LDC2009T201`, 7 2009. accessed April 18, 2019.

D. Dawson and K. Reid. Fatigue, alcohol and performance impairment. *Nature*, 388(6639): 235–235, 1997.

W. B. Dickerson. *Hesitation Phenomena in the Spontaneous Speech of Non-native Speakers of English.* PhD thesis, University of Illinois at Urbana-Champaign, 1972.

D. Duez. Silent and non-silent pauses in three speech styles. *Language and Speech*, 25(1): 11–28, 1982.

R. Eklund. *Disfluency in Swedish Human–human and Human–machine Travel Booking Dialogues*. PhD thesis, Linköping University Electronic Press, 2004.

F. Ferreira. Creation of prosody during sentence production. *Psychological Review*, 100 (2):233, 1993.

F. Ferreira. Prosody and performance in language production. *Language and Cognitive Processes*, 22(8):1151–1177, 2007.

F. Ferreira and K. G. Bailey. Disfluencies and human language comprehension. *Trends in Sognitive Sciences*, 8(5):231–237, 2004.

V. S. Ferreira and H. Pashler. Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28 (6):1187, 2002.

S. Fincke. The syntactic organization of repair in Bikol. *Cognition and Function in Language*, pages 252–267, 1999.

J. Foster. cba to check the spelling investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384. Association for Computational Linguistics, 2010.

B. A. Fox, M. Hayashi, and R. Jasperson. Resources and repair: A cross-linguistic study of syntax and repair. *Studies in Interactional Sociolinguistics*, 13:185–237, 1996.

B. A. Fox, Y. Maschler, and S. Uhmann. A cross-linguistic study of self-repair: Evidence from English, German, and Hebrew. *Journal of Pragmatics*, 42(9):2487–2505, 2010.

K. C. Fraser and G. Hirst. Detecting semantic changes in Alzheimer's disease with vector space models. In *Proceedings of LREC 2016 Workshop: Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)*, number 128. Linköping University Electronic Press, 2016.

J. Fruehwald. Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics*, 22(2):6, 2016.

M. Garrett. Levels of processing in sentence production. In *Language Production Vol. 1: Speech and Talk*, pages 177–220. Academic Press, 1980.

J. P. Gee and F. Grosjean. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15(4):411–458, 1983.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992*, volume 1, pages 517–520. IEEE, 1992.

F. Goldman-Eisler. Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, 10(2):96–106, 1958.

F. Goldman-Eisler. *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press, 1968.

S. Goldwater, D. Jurafsky, and C. D. Manning. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200, 2010.

H. Goodglass, E. Kaplan, and B. Barresi. *Boston Diagnostic Aphasia Examination Record Booklet*. Lippincott Williams & Wilkins, 2000.

M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. Rohrer, S. Black, B. F. Boeve, et al. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014, 2011.

J. O. Greene and J. N. Cappella. Cognition and talk: The relationship of semantic units to temporal patterns of fluency in spontaneous speech. *Language and Speech*, 29(2): 141–157, 1986.

B. Grimme, S. Fuchs, P. Perrier, and G. Schöner. Limb versus speech motor control: A conceptual review. *Motor Control*, 15(1):5–33, 2011.

M. Grossman and S. Ash. Primary progressive aphasia: A review. *Neurocase*, 10(1):3–18, 2004.

F. H. Guenther. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5):350–365, 2006.

F. H. Guenther. *Neural Control of Speech*. Mit Press, 2016.

F. H. Guenther and T. Vladusich. A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, 25(5):408–422, 2012.

B. Guitar. *Stuttering: An Integrated Approach to Its Nature and Treatment*. Lippincott Williams & Wilkins, 2013.

R. J. Hartsuiker and L. Notebaert. Lexical access problems lead to disfluencies in speech. *Experimental Psychology*, 2009.

M. Hayashi. A comparative study of self-repair in English and Japanese conversation. *Japanese/Korean Linguistics*, 4:77–93, 1994.

P. A. Heeman. Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialog. *arXiv preprint cmp-lg/9712009*, 1997.

C. Heinrich and F. Schiel. The influence of alcoholic intoxication on the short-time energy function of speech. *The Journal of the Acoustical Society of America*, 135(5):2942–2951, 2014.

M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

G. Hickok. Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, 13(2):135, 2012.

G. Hickok. Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. *Language, Cognition and Neuroscience*, 29(1):52–59, 2014.

G. Hickok and D. Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007.

A. E. Hieke. A content-processing view of hesitation phenomena. *Language and Speech*, 24(2):147–160, 1981.

V. M. Holmes. Hesitations and sentence planning. *Language and Cognitive Processes*, 3 (4):323–361, 1988.

J. Hough. *Modelling Incremental Self-repair Processing in Dialogue.* PhD thesis, Queen Mary University of London, 2014.

P. Howell, S. Davis, and J. Bartrip. The University College London archive of stuttered speech (UCLASS). *Journal of Speech, Language, and Hearing Research*, 2009.

E. Jacewicz, R. A. Fox, and L. Wei. Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, 128(2):839–850, 2010.

L. A. Janda and C. E. Townsend. *Czech*. Lincom Europa Munich, 2000.

J. D. Jescheniak and W. J. Levelt. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4):824, 1994.

M. Johnson and E. Charniak. A TAG-based noisy channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 33. Association for Computational Linguistics, 2004.

T. R. Johnson and J. Goldman. *A Good Quarrel: America's Top Legal Reporters Share Stories from inside the Supreme Court*. University of Michigan Press, 2009.

W. Johnson. Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. *Journal of Speech & Hearing Disorders. Monograph Supplement*, 1961.

J. G. Kahn, M. Lease, E. Charniak, M. Johnson, and M. Ostendorf. Effective use of prosody in parsing conversational speech. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 233–240. Association for Computational Linguistics, 2005.

E. Kärkkäinen, M.-L. Sorjonen, and M.-L. Helasvuo. Discourse structure. *Language Typology and Syntactic Description*, 2:301–371, 2007.

T. S. Kendall. *Speech Rate, Pause, and Linguistic Variation: An Examination through the Sociolinguistic Archive and Analysis Project*. PhD thesis, Duke University, 2009.

R. D. Kent. The uniqueness of speech among motor systems. *Clinical Linguistics & Phonetics*, 18(6-8):495–505, 2004.

J. Kolár, J. Svec, S. Strassel, C. Walker, D. Kozlíková, and J. Psutka. Czech Spontaneous Speech Corpus with structural metadata. In *Ninth European Conference on Speech Communication and Technology*, 2005.

J. C. Kowtko and P. J. Price. Data collection and analysis in the air travel planning domain. In *Proceedings of the Workshop on Speech and Natural Language*, pages 119–125. Association for Computational Linguistics, 1989.

J. Krivokapić. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, 35(2):162–179, 2007.

C. M. Laserna, Y.-T. Seih, and J. W. Pennebaker. Um... who like says you know: Filler word use as a function of age, gender, and personality. *Journal of Language and Social Psychology*, 33(3):328–338, 2014.

J. D. Laver. The detection and correction of slips of the tongue. *Speech Errors as Linguistic Evidence*, pages 132–143, 1973.

J. D. Laver. Monitoring systems in the neurolinguistic control of speech production. *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand*, pages 287–305, 1980.

W. J. Levelt. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104, 1983.

W. J. Levelt. *Speaking: From Intention to Articulation*. MIT Press, 1989.

W. J. Levelt and A. Cutler. Prosodic marking in speech repair. *Journal of Semantics*, 2(2): 205–218, 1983.

W. J. Levelt, A. Roelofs, and A. S. Meyer. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38, 1999.

M. Liberman. Young men talk like old women. `http://itre.cis.upenn.edu/~myl/languagelog/archives/002629.html`, 3 2005. accessed March 31, 2019.

M. Liberman. Political sound and silence. `http://languagelog.ldc.upenn.edu/nll/?p=23990`, 3 2016. accessed March 31, 2019.

R. J. Lickley. *Detecting Disfluency in Spontaneous Speech*. PhD thesis, University of Edinburgh, 1994.

R. J. Lickley. HCRC disfluency coding manual. *Human Communication Research Centre, University of Edinburgh*, 1998.

R. J. Lickley. Fluency and disfluency. *The Handbook of Speech Production*, page 445, 2015.

R. J. Lickley and E. G. Bard. When can listeners detect disfluency in spontaneous speech? *Language and Speech*, 41(2):203–226, 1998.

Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540, 2006.

P. A. Luce and J. Charles-Luce. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America*, 78(6):1949–1957, 1985.

J. E. Mack, S. D. Chandler, A. Meltzer-Asscher, E. Rogalski, S. Weintraub, M.-M. Mesulam, and C. K. Thompson. What do pauses in narrative production reveal about the nature of word retrieval deficits in PPA? *Neuropsychologia*, 77:211–222, 2015.

D. G. MacKay. Awareness and error detection: New theories and research paradigms. *Consciousness and Cognition*, 1(3):199–225, 1992a.

D. G. MacKay. Errors, ambiguity, and awareness in language perception and production. In *Experimental Slips and Human Error*, pages 39–69. Springer, 1992b.

D. G. MacKay. *The Organization of Perception and Action: A Theory for Language and Other Cognitive Skills*. Springer Science & Business Media, 2012.

D. G. Mackay and M. C. Macdonald. Stuttering as a sequencing and timing disorder. In *In*. Citeseer, 1984.

H. Maclay and C. E. Osgood. Hesitation phenomena in spontaneous English speech. *Word*, 15(1):19–44, 1959.

G. F. Mahl. Disturbances and silences in the patient's speech in psychotherapy. *The Journal of Abnormal and Social Psychology*, 53(1):1, 1956.

J. G. Martin. On judging pauses in spontaneous speech. *Journal of Verbal Learning & Verbal Behavior*, 1970.

K. McDougall and M. Duckworth. Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication*, 95:16–27, 2017.

H. Moniz, F. Batista, A. I. Mata, and I. Trancoso. Speaking style effects in the production of disfluencies. *Speech Communication*, 65:20–35, 2014.

C. H. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *The Journal of the Acoustical Society of America*, 95(3):1603–1616, 1994.

N. Nevler, S. Ash, C. Jester, D. J. Irwin, M. Liberman, and M. Grossman. Automatic measurement of prosody in behavioral variant FTD. *Neurology*, 89(7):650–656, 2017.

C. C. Oomen and A. Postma. Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, 30(2):163–184, 2001.

M. Ostendorf and S. Hahn. A sequential repetition model for improved disfluency detection. In *INTERSPEECH*, pages 2624–2628, 2013.

S. Oviatt. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1):19–36, 1995.

S. Oviatt, C. Darves, and R. Coulston. Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11(3):300–328, 2004.

S. V. Pakhomov, G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D. S. Knopman. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23(3):165, 2010.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

J. B. Peterson, J. Rothfleisch, P. D. Zelazo, and R. O. Pihl. Acute alcohol intoxication and cognitive functioning. *Journal of Studies on Alcohol*, 51(2):114–122, 1990.

M. Plauché and E. Shriberg. Data-driven subclassification of disfluent repetitions based on prosodic features. In *Proc. International Congress of Phonetic Sciences*, volume 2, pages 1513–1516. Citeseer, 1999.

A. Postma. Detection of errors during speech production: A review of speech monitoring models. *Cognition*, 77(2):97–132, 2000.

X. Qian and Y. Liu. Disfluency detection using multi-step stacked learning. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 820–825, 2013.

K. Rascovsky, J. R. Hodges, C. M. Kipps, J. K. Johnson, W. W. Seeley, M. F. Mendez, D. Knopman, A. Kertesz, M. Mesulam, D. P. Salmon, et al. Diagnostic criteria for the behavioral variant of frontotemporal dementia (bvFTD): Current limitations and future directions. *Alzheimer Disease & Associated Disorders*, 21(4):S14–S18, 2007.

P. Rayson, G. N. Leech, and M. Hodges. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1):133–152, 1997.

A. Rhetoric. Rhetorical Literacy: 49 Important Speeches in 21st Century America. `https://www.americanrhetoric.com/`, 3 2020. accessed August 11, 2020.

P. M. Roberts, A. Meltzer, and J. Wilding. Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of Communication Disorders*, 42(6):414–427, 2009.

S. R. Rochester. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2(1):51–81, 1973.

S. Schachter, N. Christenfeld, B. Ravina, and F. Bilous. Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, 60(3):362, 1991.

E. A. Schegloff, G. Jefferson, and H. Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.

F. Schiel and C. Heinrich. Disfluencies in the speech of intoxicated speakers. *International Journal of Speech, Language & the Law*, 22(1), 2015.

F. Schiel, C. Heinrich, S. Barfüsser, and T. Gilg. ALC: Alcohol Language Corpus. In *LREC*, 2008.

K.-J. Schlenck, W. Huber, and K. Willmes. "prepairs" and repairs: Different monitoring functions in aphasic language production. *Brain and Language*, 30(2):226–244, 1987.

M. J. Schnadt and M. Corley. The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28, 2006.

B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski. The INTERSPEECH 2011 speaker state challenge. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben. Medium-term speaker states—a review on intoxication, sleepiness and the first challenge. *Computer Speech & Language*, 28(2):346–374, 2014.

E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, UC Berkeley, 1994.

E. Shriberg. Acoustic properties of disfluent repetitions. In *Proceedings of the International Congress of Phonetic Sciences*, volume 4, pages 384–387, 1995.

E. Shriberg. To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169, 2001.

E. Shriberg and R. Lickley. Intonation of clause-internal filled pauses. *Phonetica*, 50(3): 172–179, 1993.

E. Shriberg and A. Stolcke. Word predictability after hesitations: a corpus-based study. In *ICSLP 96., Fourth International Conference on Spoken Language Processing.*, volume 3, pages 1868–1871. IEEE, 1996.

M.-h. Siu and M. Ostendorf. Modeling disfluencies in conversational speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 1, pages 386–389. IEEE, 1996.

G. Skantze and A. Hjalmarsson. Towards incremental speech generation in dialogue systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–8. Association for Computational Linguistics, 2010.

G. Skantze, A. Hjalmarsson, and C. Oertel. Exploring the effects of gaze and pauses in situated human-robot interaction. In *Proceedings of the SIGDIAL 2013 Conference*, pages 163–172, 2013.

V. L. Smith and H. H. Clark. On the course of answering questions. *Journal of Memory and Language*, 32(1):25–38, 1993.

A. Stolcke and E. Shriberg. Statistical language modeling for speech disfluencies. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 405–408. IEEE, 1996.

M. Swerts. Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4): 485–496, 1998.

M. Swerts and E. Krahmer. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1):81–94, 2005.

G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski. Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 7:195, 2015.

D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech Coding and Synthesis*, 495: 518, 1995.

M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268 (5217):1632–1634, 1995.

P. H. Tannenbaum, F. Williams, and C. S. Hillier. Word predictability in the environments of hesitations. *Journal of Verbal Learning and Verbal Behavior*, 4(2):134–140, 1965.

L. Ten Bosch, N. Oostdijk, and L. Boves. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2):80–86, 2005.

E. Tisljár-Szabó, R. Rossu, V. Varga, and C. Pléh. The effect of alcohol on speech production. *Journal of Psycholinguistic Research*, 43(6):737–748, 2014.

G. Tottie. Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, 16(2):173–197, 2011.

G. Tottie. On the use of uh and um in American English. *Functions of Language*, 21(1): 6–29, 2014.

J. A. Tourville and F. H. Guenther. The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, 26(7):952–981, 2011.

J. Tsiamtsiouris and H. S. Cairns. Effects of sentence-structure complexity on speech initiation time and disfluency. *Journal of Fluency Disorders*, 38(1):30–44, 2013.

S. Uhmann. Some arguments for the relevance of syntax to same-sentence self-repair in everyday German conversation. *Studies in Interactional Linguistics*, pages 373–404, 2001.

C. Van Riper and L. L. Emerick. *Speech Correction: An Introduction to Speech Pathology and Audiology*. Prentice Hall, 1984.

G. M. Walker and G. Hickok. Bridging computational approaches to speech production: The semantic–lexical–auditory–motor model (slam). *Psychonomic Bulletin & Review*, 23(2):339–352, 2016.

K. Walker, X. Ma, D. Graff, S. Strassel, S. Sessa, and K. Jones. RATS Speech Activity Detection LDC2015S02. Hard Drive. Philadelphia: Linguistic Data Consortium, 2 2015.

D. Watson and E. Gibson. The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, 19(6):713–755, 2004.

P. Weiner. Linear pattern matching algorithms. In *14th Annual Symposium on Switching and Automata Theory (swat 1973)*, pages 1–11. IEEE, 1973.

M. Wieling, J. Grieve, G. Bouma, J. Fruehwald, J. Coleman, and M. Liberman. Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, 6(2):199–234, 2016.

F. Wouk. The syntax of repair in Indonesian. *Discourse Studies*, 7(2):237–258, 2005.

J. Yuan and M. Liberman. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5):3878, 2008.

J. Yuan, X. Xu, W. Lai, and M. Liberman. Pauses and pause fillers in Mandarin monologue speech: The effects of sex and proficiency. *Proceedings of Speech Prosody 2016*, pages 1167–1170, 2016.

B. Zellner. Pauses and the temporal structure of speech. In *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.*, pages 41–62. John Wiley, 1994.

E. Zvonik and F. Cummins. The effect of surrounding phrase lengths on pause duration. In *Eighth European Conference on Speech Communication and Technology*, 2003.