Early Prosodic Manifestations of Disfluency

Jixing Li, Sam Tilsen

Department of Linguistics, Cornell University, Ithaca, USA

jl2939@cornell.edu, tilsen@cornell.edu

Abstract

Theoretical models of speech production have hypothesized a relation between different types of disfluencies and the mechanisms responsible for them. Some disfluencies, such as filled pauses (e.g. 'um', 'uh') and repetitions (i.e. 'the the'), are argued to arise from difficulty in planning, while cutoff disfluencies (e.g. 'horiz-[ontal]') are argued to arise from self-monitoring. This distinction predicts that prosodic manifestations of disfluency, i.e. durational slowing and pitch/intensity modulation, should occur earlier for planning disfluencies than for self-monitoring disfluencies. The present study examined segmental duration, pitch, and intensity in speech produced just before filled pause, repetition, and cutoff disfluencies in the Switchboard corpus. The results showed that durational slowing occurs earlier and is more extensive before filled pause disfluencies than before repetitions and cutoffs. In addition, decreases of f0 and intensity occurred earlier before filled pauses than before repetitions, and intensity decreased more gradually before cutoffs than before repetitions and filled pauses. These findings support theoretical models in which cutoffs are associated with a self-monitoring mechanism and filled pauses/repetitions are associated with planning difficulties. Furthermore, differences in effect magnitudes between filled pauses and repetitions indicate that filled pauses may be associated with more severe planning difficulties than repetitions.

Index Terms: disfluency, prosody, duration, pitch, intensity

1. Introduction

Disfluency, such as filled pause (e.g., 'um', 'uh'), repetition (e.g., 'the the') and cutoff word (e.g., 'hori[zontal]-'), is a common part of human speech that occurs at a rate of 6 to 10 per 100 words [1, 2]. One influential model of speech production [3] suggested that disfluency can occur at the self-monitoring stage (e.g., self-correction of speech errors), such as cutoff disfluencies, or at the planning stage (e.g., word retrieval difficulties), such as repetitions and filled pauses. Repetitions and filled pauses have been claimed as different styles of hesitations, both serving the function of "holding the floor" before the troublesome item is finally retrieved [3, 4]. But evidence from processing studies suggested that listeners were sensitive to different types of hesitation disfluencies, and repetitions affected comprehension less than filled pauses [5, 6]. It is therefore hypothesized that filled pauses reflect a more severe planning problem than repetitions, and speakers are aware of the upcoming problems earlier before using 'um' and 'uh' than before repeating.

The theoretical model of planning and self-monitoring disfluencies predicts early prosodic modulation (e.g., increased duration, decreased pitch and intensity) before repetitions and filled pauses but not before cutoffs. The different degrees of planning difficulties between filled pauses and repetitions predict a larger magnitude of prosodic modulation before filled pauses than before repetitions. We tested these hypotheses by examining the segment duration, f0 and intensity of 500 ms of speech before cutoffs, repetitions and filled pauses in Switchboard [7]. The results showed that durational slowing occurs earlier and is more extensive before filled pauses than before repetitions and cutoffs. In addition, decreases of f0 and intensity occurred earlier before filled pauses than before repetitions, and intensity decreased more gradually before cutoffs than before repetitions and filled pauses. These findings support the distinction between planning and self-monitoring disfluencies and the distinction between the magnitude of difficulty associated with filled pauses and repetitions.

1.1. Mechanisms of disfluency

According to the "double perceptual loop" theory of self-monitoring [3], disfluency arises due to planning problems (e.g., word retrieval difficulties) or self-monitoring (e.g., self-correction of speech errors). At the monitoring stage, speakers detect the problem after hearing their own speech, resulting in an "overt repair" like cutoffs (e.g., "here is a hori[zontal]– a vertical line"). A problem can also occur before articulation, for example, speakers might have difficulty retrieving the word "vertical" (either the lemma or the lexeme [8]), resulting in a hesitation period filled with repetitions (e.g., "here is a – a vertical line") or filled pauses ("here is a – um vertical line") [9].

Under this model of disfluency, [4] hypothesized that words preceding the interruption point are lengthened before planning disfluencies but not before self-monitoring disfluencies, such as cutoffs. The rationale is that cutoffs are the results of failure to detect an error before implementing the phonetic plan; the error is only detected through self-monitoring after articulation. Thus there should be no abnormal durational pattern before the onset of the cutoff word. For hesitation disfluencies, the problem occurs at the planning stage for the troublesome item. Therefore, lengthening of the words in the current articulatory buffer is a strategy to buy time to retrieve the troublesome item, and if all words in the articulatory buffer has been used, a hesitation period would follow until an item is finally retrieved. [10] tested this hypothesis by examining the duration of four words before cutoff disfluencies and planning disfluencies (repetitions, silences and filled pauses) in Switchboard. The results showed a significant increase in the normalized duration of words before repetitions, silences and filled pauses, but not before cutoffs, supporting the distinction between planning disfluencies and self-monitoring disfluencies.

Despite the evidence for a distinction between cutoff and repetition/filled pause disfluencies, the difference between repetitions and filled pauses is less clear. Both [3] and [4] considered repetitions and filled pauses as a unified phenomenon of hesitation, and [4] explicitly stated that speakers vary in their prefer-

ence for using repetitions or filled pauses to hold the floor before continuing with the message content. But repetitions have also been argued to differ from filled pauses. The action of repeating a word in speech production has been suggested to be the result of a failure to inhibit the most highly activated word at the moment of repetition (the first repeat) [11]. Following this "activation hypothesis", a filled pause is therefore a successful inhibitory control of repetition. This hypothesized difference between repetitions and filled pauses has been observed in perception studies using EEG [5, 6]: repetitions did not affect the N400 effect associated with target words or the likelihood of later recognition of those words, but 'uh's clearly affected the N400 effect and memory for subsequent words. A straightforward interpretation of these findings would suggest that listeners were sensitive to different functions attributed to different types of disfluency, and filled pauses signaled a more serious problems in speech production as it induced more processing effort. Indeed, [10] showed that word duration increased more before filled pauses than before repetitions.

1.2. Hypotheses of the present study

To further test the differences between cutoffs, repetitions and filled pauses, we examined the prosodic properties (duration, pitch and intensity) of speech before the three different types of disfluencies in the Switchboard corpus [7]. As shown in [12], abnormal prosodic patterns of speech (i.e., more and longer pauses) occurred in more difficult speaking tasks (e.g., interpreting a cartoon rather than simply describing a cartoon). It is hypothesized that prosody of fluent speech immediately proceeding a disfluency is indicative of the degree of difficulty associated with the upcoming disfluency. For example, an earlier and larger increase of segment duration and decrease of pitch and intensity before disfluency would indicate a more severe planning problem.

The speech preceding a disfluency has been considered as the reparandum (the region to be repaired, e.g., 'hori-[zontal]' in "the hori-[zontal] vertical line") under the classical model of disfluency structure [3]. The other two components of a disfluency include the optional editing phase (the disfluent period, e.g., 'uh', 'um'), and the repair (the correction region, e.g., 'vertical' in "the hori-[zontal] vertical line"). However, not all disfluencies fit well in this framework. Hesitations, for example, might not have an overt reparandum to be repaired (e.g., in "the–um vertical line", there is no clear error before 'um', and no clear repair proper). In addition, such analyses typically identify an "interruption point" (the red line in Figure 1), which is particularly problematic as speakers may detect an upcoming problem well before the overt lexical manifestations of disfluency. Early detection of lexical access difficulty, for example, may result in duration slowing and pitch/intensity modulation in a reparandum. In contrast, disfluencies which arise from self-monitoring, may show no such effect. The present study therefore avoided the terms of reparandum and repair, and considered the pre-interruption speech as a sequence of segments and marked them as '-1', '-2', '-3', etc. (see Figure 1)

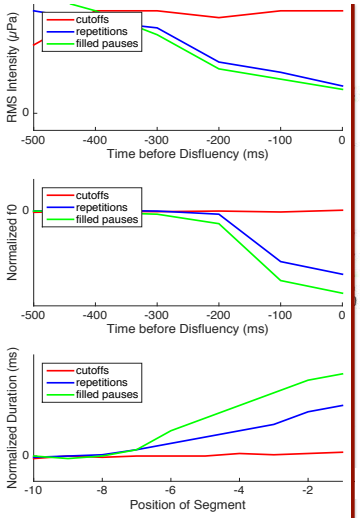
Under [3]'s distinction between planning disfluencies (e.g., repetitions, filled pauses) and self-monitoring disfluencies (e.g., cutoffs) and the empirical evidence that filled pauses indicate more severe planning problems [5, 6], we hypothesize (see Figure 1 for a schematic illustration):

Hypothesis 1: Pre-interruption manifestations of disfluency should be observed for planning disfluencies but not self-monitoring disfluencies.

Predictions: There is a significant increase of normalized segment duration and decrease of normalized f0 and RMS intensity before repetitions and filled pauses, but not before cutoffs.

Hypothesis 2: Pre-interruption manifestations of disfluency will be greater in magnitude for filled pause disfluencies than repetition disfluencies.

Predictions: There is an earlier and larger increase of normalized segment duration and decrease of normalized f0 and RMS intensity before filled pauses than before repetitions.



CUT:	to	the	hori-[zontal]			vertical... "
REP:	"Go	to	the			the ... "
FP:	"Go	to	the	um		vertical ... "

Figure 1: *Schematic illustration of the hypotheses*

2. Methods

2.1. Annotation of disfluency

The original annotation of disfluency in Switchboard [4] followed the classical reparandum-repair frame of disfluency [3] and left out a number of disfluent words that are not followed by a repair. There was also no information about the types of disfluency. We therefore re-annotated disfluency in Switchboard based on (1) the 'SIL' marker at the phone level; (2) the part-of-speech tag 'UH' (interjection, e.g., 'uh', 'um', 'well', 'like'); (3) the ']-' marker in the word orthography (e.g., 'e[ven]-'), and (4) repeated word strings ($\leq$ 4 words). Since many (about 95% in the corpus) disfluencies combine multiple interruptions and repairs and thus cannot be unique classified) e.g., "uh people who SIL uh ran SIL"), we restricted our analysis to simplex disfluencies, i.e cutoffs contain only one cutoff word; filled pauses contain only one 'um' or 'uh', and is not preceded by a cutoff word; repetitions contain only one repeated word, and is not preceded by a cutoff word.

2.2. Segment duration

Many factors influence the duration of segments in speech, including speaking rate, stress, accent, syllable position, discourse/phrase position, segmental interactions, syntactic/semantic factors, etc. [13]. To normalize the raw durations of segments at -1 to -10 positions before disfluency, we first calculated their expected durations based on a linear mixed-effects

model for durations of fluent segments at other positions. The fixed effects of the model include the natural logarithm of the token frequency of the segments in Switchboard, the stress level of the syllable that contains the segments (primary, secondary, none), and the position of the segment in the containing syllable (first, second or third segments in the onset (-3 to -1), the nuclei (0), and the first, second or third segments in the onset (1 to 3)). Speaker is included as random effect. The current formula used for the linear mixed-effects model is:

$$exp\_dur \sim 1 + log(freq) + stress + position + (1|spkr) \quad (1)$$

The normalized duration of segment is the difference between the raw durations and the expected durations, i.e. the residuals of the regression. Outliers with $|z| > 2.5$ were excluded from analysis (overall 3.2% of the data).

2.3. Pitch and intensity

The raw f0 contour of 1s of speech before each disfluency were extracted using the fxrapt function from the Voicebox toolbox for Matlab [14]. Each contour was further processed by removing any f0 values exceeding 4 s.d. from the mean, and filling the gaps by spline interpolation. The normalized f0 contours are the z-scores of the post-processed f0 contours. The root mean square (RMS) intensity for every 20 ms of the 1s speech before each disfluency was calculated to get the intensity contours.

3. Results

3.1. General description of the disfluency data

The current annotation of disfluency yielded 71667 disfluent stretches (including 25364 silences) and 124623 fluent stretches. The average length of a fluent stretch is 9.2 words (*SD* = 4), and the average raw duration of a disfluent stretch is 1115.2 ms (*SD* = 1.2). The total number of fluent words is 629285. The rate of disfluency under the current annotation is 7.6%, consistent with previous findings [1, 2]. The criteria for different types of simplex disfluency yielded 884 cutoffs, 1469 repetitions and 1140 filled pauses. The proportion of each type in the whole corpus and the examples are shown in Table 1.

Type	No.	%	Examples
CUT	884	1.2%	"gonna have to **un[load]-** ..." "was CNN the **on[ly]-** ..."
REP	1469	2%	"she was able **to - to**..." "I think that **they - they**..."
FP	1140	1.6%	"don't know I just - **um** ..." "read a lot about - **uh** ..."
OTHER	68174	95.2%	"**SIL uh SIL** about sending **SIL um** ..." "I mean **I I I SIL I hard[ly]**- I **SIL** ..."

Table 1: *Proportions of disfluency types in Switchboard*

3.2. Duration

As predicted by our hypothesis, the normalized segment durations were lengthening the most before filled pauses, an intermediate degree before repetitions, and lengthened the least before cutoffs. Before cutoffs, there were significant increases of duration from the -4 to -1 segments ($p < .001$), with the last two segments lengthened by 10.1 ms ($SD = 0.04$) and 27.5 ms ($SD = 0.05$), respectively; before repetitions, lengthening started from the -7 segment ($p < .001$), and the durations of last two segments increased by 19.7 ms ($SD = 0.04$) and 40.5 ms ($SD = 0.05$); before filled pauses, lengthening also started from the -7 segment ($p < .001$), but the last two segments were lengthened by 33.4 ms ($SD = 0.05$) and 64.3 ms ($SD = 0.04$) (see Table 2 for the one-sample *t*-tests results). The different degrees of lengthening before the three types of disfluencies are clearly shown in Figure 2a.

We further calculated the cumulative increase of duration at each segment position. The results showed that the slowdown began at about 160 ms before filled pauses, 90 ms before repetitions and 50 ms before cutoffs (see Figure 2b), suggesting that speakers were aware of their problems earlier before filled pauses than before repetitions than before cutoffs.

(a) **Cutoffs**

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
N	598	733	808	861	863	873	862	860	847	803
M	-2.1	0	2.1	0	1.5	1.8	5.1	7.3	10.1	27.5
SD	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05
t	-1.50	0.27	1.61	0.82	1.20	1.46	3.91	5.58	6.92	15.85
p	0.13	0.79	0.1	0.41	0.23	0.14	<.001	<.001	<.001	<.001

(b) **Repetitions**

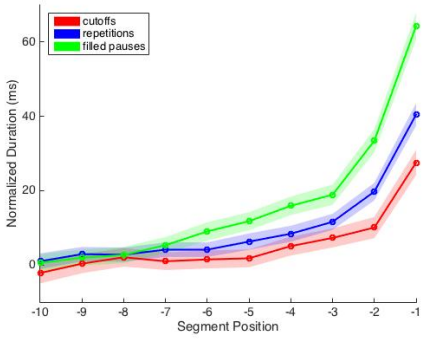
	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
N	1177	1341	1423	1448	1446	1431	1427	1429	1389	1248
M	1.0	2.9	2.8	4.1	4.1	6.3	8.4	11.6	19.7	40.5
SD	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05
t	0.93	2.95	2.86	4.11	4.11	5.60	7.65	10.48	16.41	26.26
p	0.35	0.003	0.004	<.001	<.001	<.001	<.001	<.001	<.001	<.001

(c) **Filled Pauses**

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1
N	988	1093	1126	1128	1122	1122	1113	1101	1087	969
M	0	2.0	2.7	5.3	9.0	11.8	15.9	18.9	33.4	64.3
SD	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05	0.05	0.06
t	0.48	1.78	2.41	4.77	7.19	9.57	12.57	13.63	20.88	34.43
p	0.63	0.08	0.02	<.001	<.001	<.001	<.001	<.001	<.001	<.001

Table 2: *One sample t-tests for normalized duration of segments -1 to -10 before different types of disfluency*

(a) *Normalized duration of segments before different types of disfluency*



(b) *Cumulative increase of duration before different types of disfluency*

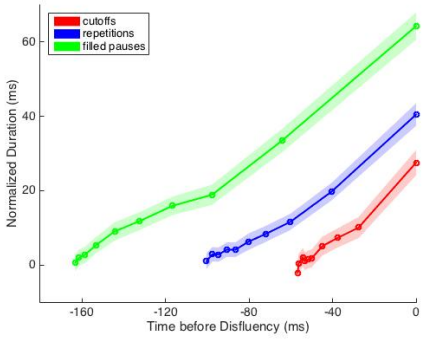


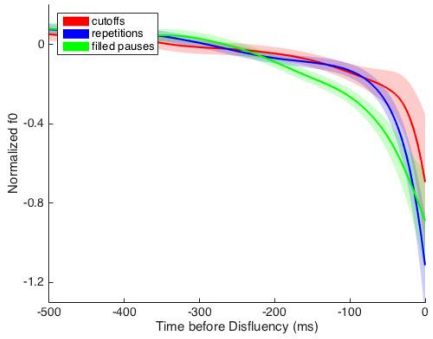
Figure 2: *Normalized duration of segments before different types of disfluency (bands indicate 95% confidence intervals).*

3.3. Pitch and intensity

The mean normalized f0 contours 500 ms before the three types of disfluency are shown in Figure 3a. There was a decrease in f0 before all three types of disfluencies, but the decrease started earlier (at approximately -150 ms) before filled pauses than before repetitions and cutoffs (at approximately -80 ms). A repeated measures ANOVA showed that, for the three disfluency types, the differences in f0 values at the five time point (-200, -150, -100, -50, 0 ms) were statistically significant, $F(4, 13960) = 79.34, p = 0$. Tukey's HSD tests showed that f0 of filled pauses differed significantly from that of cutoffs and repetitions at -150, -100, and -50 ms at $p < 0.01$ level, but f0 values of of repetitions and cutoffs were not significantly different at all of the five time points.

The RMS intensity contours 500 ms before the three types of disfluency are shown in Figure 3b. Similar to the f0 pattern, intensity started to decrease earlier before filled pauses (at approximately -200 ms) than before repetitions (at approximately -100 ms). However, before cutoffs, intensity decreased much earlier (at approximately -300 ms), although the slope was not as large. A repeated measures ANOVA showed a significant main effect of time (-200, -150, -100, -50, 0 ms) on RMS intensity ($F(4, 13960) = 394.51, p = 0$), and a significant interaction between time and disfluency type ($F(8, 13960) = 23.33, p = 0$). Tukey's HSD tests showed that RMS intensity differed significantly for the three types of disfluency at all the five time points at $p < 0.05$ level, except that at -50 ms, cutoffs did not differ from filled pauses ($p = 0.08$) and repetitions ($p = 0.45$), at -100 ms, cutoffs did not differ from filled pauses ($p = 0.32$), and at -150 ms, filled pauses did not differ from repetitions ($p = 0.07$).

(a) *Nomarlized f0 before different types of disfluency*



(b) *RMS intensity before different types of disfluency*

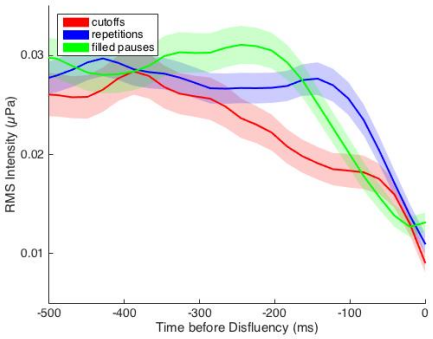


Figure 3: *Normalized f0 and RMS intensity before different types of disfluency (bands indicate 95% confidence intervals).*

4. Discussion and Conclusions

The hypothesis that lexical access disfluencies (filled pauses, repetitions) would have early prosodic manifestations than self-monitoring disfluencies (cutoffs) was supported: durational lengthening and f0/intensity decreases were more substantial in the lexical access disfluencies. Furthermore, early manifestations of disfluency were more extreme for filled pauses than for repetitions, supporting the hypothesis that filled pauses are associated with more severe lexical access/retrieval problems than repetitions are. In particular, the cumulative segment duration suggested that speakers began to slowdown about 160 ms before filled pauses, 90 ms before repetitions and 50 ms before cutoffs, supporting the hypothesis that filled pauses reflect a higher degree of planning difficulty than repetitions. Although cutoffs may not arise from planning problems, there is still a 50 ms lengthening effect, suggesting that a speech error may have already been detected before articulation at the lower-level of motor execution.

As predicted, f0 and intensity contours showed an earlier decrease before filled pauses than before repetitions. However, the onset of intensity decrease is much earlier before cutoffs, although the slope was more flat. One way of interpreting the early and gradual decrease of intensity before cutoffs is that they generally occur phrase-medially where there is a natural falling of intensity. Another possibility is that before filled pauses and repetitions, speakers are aware of an upcoming difficulty and are actively increasing their intensity level to hold the floor before the onset of the falling period.

The idea that different disfluencies reflect different degrees of difficulty has been proposed in a number of studies in the literature. For example, [15] argued that 'uh' and 'um' in filled pauses differ in meanings: 'um' signals a more serious problem than 'uh', and speakers select among 'uh' and 'um' to convey different messages. [12] showed that pauses occurred more often and were longer before words that were less predictable in the context of proceeding speech, whereas repetitions tended to follow unpredicted words. The notion of predictability includes word frequency and the proceeding syntactic and semantic context. There are well developed metrics for syntactic predictability, such as 'surprisal' and 'entropy reduction' [16, 17]; semantic predictability has also been modeled using the co-occurrence information under the Latent Semantic Analysis (LSA) approach [18]. It is interesting to further explore the planning difficulty associated with different types of disfluency under the syntactic and semantic complexity metrics.

In sum, we found different duration, f0 and intensity patterns for speech before filled pauses, repetitions and cutoff disfluencies. These prosodic patterns contribute to a more nuanced phonetic characterization of different types of disfluency, and sheds light on the nature of the underlying problems associated with filled pauses, repetitions and cutoff disfluencies. The study did not differentiate between the fillers 'um' and 'uh', which has been claimed to signal different degrees of difficulty [15]; it also excluded other types of disfluency, such as false starts (e.g., "he-she had to be put in nursing home") due to the lack of classification metrics, and silences due to their varying degrees of length. A more comprehensive metric for disfluency types should be developed in future studies to examine their prosodic properties and their underlying difficulties in speech production. Multi-disfluency (e.g., cutoff and filled pause) has also been excluded in current analysis. An examination of their prosodic properties might provide further insight into the disfluency mechanisms.

5. References

[1] H. Bortfeld, S. Leon, J. Bloom, M. Schober, and S. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and Speech*, vol. 44, pp. 123–147, 2001.

[2] J. Fox Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," *Journal of Memory and Language*, vol. 34, pp. 709–738, 1995.

[3] M. Levelt, *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press, 1989.

[4] E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, pp. 153–169, 2001.

[5] L. MacGregor, M. Corley, and D. Donaldson, "Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension," *Brain and Language*, vol. 11, pp. 36–45, 2009.

[6] M. Corley, L. MacGregor, and D. Donaldson, "It's the way that you, er, say it: Hesitations in speech affect language comprehension," *Cognition*, vol. 105, pp. 658–668, 2007.

[7] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," *Proceedings of ICASSP*, pp. 517–520, 1992.

[8] A. Caramazza, "How many levels of processing are there in lexical access?" *Cognitive Neurophisology*, vol. 14, pp. 177–208, 1997.

[9] H. Clark and E. Clark, *Psychology and language*. New York: Harcourt Brace Jovanovich, 1977.

[10] J. Li and S. Tilsen, "Phonetic evidence for two types of disfluency." in *Proceedings of ICPhS 2015*, Glasgow, UK, August 2015.

[11] H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive Psychology*, vol. 37, pp. 201–242, 1998.

[12] F. Goldman-Eisler, "Speech production and the predictability of words in context." *Quarterly Journal of Experimental Psychology*, vol. 10, pp. 96–106, 1958.

[13] D. Klatt, "Linguistic uses of segmental duration in english: Acoustic and perceptual evidence," *The Journal of the Acoustical Society of America*, vol. 59, pp. 1208–1221, 1976.

[14] M. Brookes, "Voicebox: Speech processing toolbox for matlab." [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[15] H. Clark and J. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, pp. 73–111, 2002.

[16] J. Hale, "A probabilistic earley parser as a psycholinguistic model," in *of the North American C Proceedings of the Second Meeting hapter of the Association for Computational Linguistics*, 2001.

[17] ——, "Uncertainty about the rest of the sentence." *Cognitive Science*, vol. 30, pp. 609–642, 2006.

[18] T. Landauer and S. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological Review*, vol. 104, pp. 211–240, 1997.