PHONETIC EVIDENCE FOR TWO TYPES OF DISFLUENCY

Jixing Li, Sam Tilsen

Department of Linguistics, Cornell University
jl2939@cornell.edu, tilsen@cornell.edu

ABSTRACT

Disfluency, such as pause (silences), 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 [2, 5]. According to one model of speech production [8], there are two types of disfluency: disfluency at the internal planning stage (e.g., word-retrieval difficulties), and disfluency at the external monitoring stage (e.g., self-correction of speech errors). The current study provides phonetic evidence for the two types of disfluency by examining word durations before different types of disfluency in the Switchboard corpus [6]. The results showed only a marginal increase in the durations of words before cutoffs, but a large increase in the durations of words before repetitions, silences and filled pauses, suggesting internal processing difficulty before non-cutoff disfluency, but not before cutoff disfluency.

Keywords: disfluency, duration, self-monitoring, Switchboard

1. INTRODUCTION

Human speech is rarely fully fluent. Disfluency, such as pause (silences), filled pause ('um', 'uh'), repetition ('the the', 'and and') and cutoff word ('e[ven]-', 'h[ow]-'), etc. occurs regularly at a rate of 6 to 10 per 100 words [2, 5]. According to current models of disfluency [7, 9], an interruption in the flow of fluent speech signals the speaker's detection and attempted correction of a problem in language production. The problem is usually immediately repaired after its articulation, as in cutoff disfluency (e.g., "here is a hori[zontal]– a vertical line"). The problem could also be a difficulty in retrieving a word, resulting in repetitions, silences and filled pauses [9]. The present study provides phonetic evidence for the distinction between the two types of disfluencies by examining the durations of four words before cutoffs, repetitions, silences, and filled pauses in the Switchboard corpus [6]. The results showed only a marginal increase in the word durations before cutoffs, but a large increase in the word durations before repetitions, silences, and filled pauses. The two different durational patterns suggest different mechanisms underlying external and internal disfluencies, consistent with Levelt's ([8]) "double perceptual loop" theory of self-monitoring.

2. HYPOTHESES

2.1. Self-monitoring of disfluency

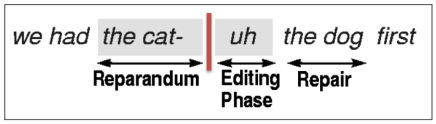
Levelt's ([8]) "double perceptual loop" theory of self-monitoring involves an external loop for perception of self-produced overt speech and an internal loop for perception of internal speech. A problem can be detected via the internal loop and "covertly repaired" before it is articulated. For example, in the sentence "here is a – uh vertical line", the speaker might be about to say "horizontal", but he detected the error after generating the phonetic plan for "horizontal" and repaired it covertly before articulation. Alternatively, the speaker might have a problem retrieving the word "vertical", either because of difficulty in retrieving the lemma or retrieving the lexeme [3]. This problem occurs because of a failure to generate a phonetic plan, resulting in a hesitation period filled with repetitions (e.g., "here is a – a vertical line"), filled pauses or silences ("here is a – SIL vertical line") [4].

A problem could also occur at the external loop of speech production after the speaker hears their own production. Since self-monitoring at the internal loop requires optimal attentional conditions, which is not always given, the speaker may be too late to intercept an error. The result is an "overt repair" at the external speech, like cutoffs (e.g., "here is a hori[zontal]– a vertical line"). Upon the detection of error, either internally or externally, the speaker will immediately halts (within an estimated constant latency of 200 ms) further formulation of the present utterance. This is called the "Main Interruption Rule" [7]. The Main Interruption Rule was supposed to hold for both internal and external speech, although the temporal estimate for interruption latency was only based on examination of overt repairs.

2.2. Structure of disfluency

Levelt ([7]) divided a typical disfluency into three components: the reparandum, the editing phase, and the repair. The reparandum is the troublesome item before the interruption of fluent speech; the (optional) editing phase is the hesitation period (e.g., 'um's and 'uh's, repetitions and silences); the repair is the resumption of fluency consisting of making the repair proper (see Figure 1, the red line indicates the interruption point).

Figure 1: Levelt's model of disfluency structure (from Shriberg ([9]))



Although Levelt's model of disfluency structure has been widely accepted, it is not always the case that a disfluency fits into the reparandum-repair frame. Disfluencies caused by word-retrieving difficulties, for example, do not have an overt reparandum to be repaired: in the sentence "we had – uh the dog first", there is no clear error before the editing phase "uh", and it is not clear either what constitutes the repair proper. The notion of an "interruption point" is also unclear in disfluencies due to word-retrieval difficulties: the speaker may have detected the problem several words before the hesitation period, thus the so-called "interruption point" before the hesitation is not the actual time where a fluent flow of speech is interrupted.

The present study avoided the terms of reparandum and repair, and used the term "surface interruption point" to refer to the interruption point identified in the previous literature [7, 9]. We consider a disfluency as consists of a fluent stretch before the surface interruption point, a disfluent stretch (usually the editing phase), and the resumption of fluency. The words before the surface interruption point are marked as as "-1" (the first word preceding interruption), "-2" (the second word preceding interruption), and so forth; the words in the disfluent stretch are marked as "0" , and the words in the resumption of fluency are marked as "1", "2", etc. (see Table 1).

Table 1: The interruption point for different disfluencies

CUT:			"one	thing	that	w[as]	–		that	they ... "
REP:	"she	had	used	a	walker	for	–		for	quite ..."
FP:	"she	lived	in	an	apartment	and	–	um	that	was ..."
SIL:		"place	my	mother	in	a	–	SIL	nursing	home..."
			-4	**-3**	**-2**	**-1**		**0**	**1**	**2**

2.3. The present study

Shriberg ([9]) speculated that words preceding the interruption point of a disfluency are lengthened except for overt repairs, such as cutoffs. The rationale behind the hypothesis 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 at the external loop. Thus there should be no abnormal durational pattern before the onset of the cutoff word. For the non-cutoff disfluencies due to word-retrieving difficulties, the problem occurs prior to formulating the phonetic plan for the troublesome item cannot be generated. Therefore, lengthening of the words in the current articulatory buffer may be a strategy to buy time for retrieving the troublesome item, and if all words in the articulatory buffering has been used, a hesitation period would follow until an item is finally retrieved. As a result, we should see an increase in the normalized duration (difference between the raw duration and the expected duration) from word -4 to -1 before repetitions, silences and filled pauses. The four-word window provides an estimate for the onset of the word-retrieving difficulty, i.e., the position where a significant increase in duration starts.

It is also hypothesized that the lengthening of words -4 to -1 is positively correlated with the duration of the following disfluent stretch. If the hesitation period is used for retrieving the troublesome item, then a more serious word-retrieving problem would result in a longer period of hesitation, and possibly a larger increase in duration before the hesitation at word -1 to -4 because the speaker need to buy more time.

3. METHOD

3.1. Annotation of disfluency

Following Levelt's model of disfluent structure, a disfluent token in Switchboard was annotated as containing a reparandum and a repair (including the "editing phase" in Levelt's model). However, as discussed in Section 2.2, not all disfluencies fit into a reparandum-repair frame, so the original annotation of disfluencies left out a number of disfluent words that are not followed by a repair. There is also no information on the types of disfluency. We therefore re-annotated disfluency in Switchboard based on the following markers:

1. A silence transcribed as SIL at the phone level;
2. A word with a part-of-speech tag of 'UH' (interjection, e.g., 'uh', 'um', 'well', 'like');
3. A word string (<=4 words) followed by a same

word string;
4. A cutoff word (tagged with '...]-' in the orthography);

Since many of the disfluent stretches consist of more than one type of disfluency (e.g., "uh people who SIL uh ran SIL"), we narrowed down the four types of disfluency under the following restrictions:

1. Silence contains only a silence, and is not preceded by a cutoff words;
2. Filled pause contains only "um" and "uh", and is not preceded by a cutoff words;
3. Repetition contains only one repeated word, and is not preceded by a cutoff words;
4. Cutoff contains one cutoff word.

3.2. Normalization of duration

To normalize the raw durations of words at -1 to -4 positions, we first calculated their expected durations based on a linear mixed effects model for durations of fluent words at other positions. The fixed effects include the number of the syllables in the word (from 1-7 syllables), and the natural logarithm of the token frequency of the word in Switchboard; the random effect is the speaker [1]. The formula used for the linear mixed effects model is:

$$(1) \quad exp\_dur \sim 1 + nsylb + freq + (1|spkr)$$

The normalized duration are the difference between the raw durations and the expected durations:

$$(2) \quad norm\_dur = raw\_dur - exp\_dur$$

4. RESULTS

4.1. General description of the disfluency data

The current annotation of disfluency yields 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 [2, 5].

4.2. Duration of words before different types of disfluency

The normalized durations of words before different types of disfluency showed very different patterns for cutoffs and the other three types of disfluency. For cutoffs, a marginal lengthening was observed from -4 to -2 positions. One-sample *t*-tests revealed them to be significantly different from 0 at $p < .001$ level (see Figure 2(a)), but a one-way ANOVA suggested no significant effect of word position (-4 to -2) on normalized duration at $p < .001$ level (($F(2) = 4.89$, $p = 0.01$)). The normalized durations for words -1 to -4 before repetitions, silences and filled pauses are similar: there was a small increases from -4 to -3, a medium increase at -2, and a great increase at -1. One-sample *t*-tests revealed all the normalized durations to be significantly different from 0 at $p < .001$ level (see Figure 2(b-d)). One-way ANOVAs also suggested significant effects of word position on normalized duration before repetitions, silences and filled pauses ($F(3) = 88.9$, $p < .001$; $F(3) = 1721.2$, $p < .001$; $F(3) = 64.9$, $p < .001$, respectively). Figure 3 showed the boxplots for the normalized durations of -1 and -4 before the four types of disfluencies.

Figure 2: One sample t-tests for normalized duration of word -1 to -4 before different types of disfluency

(a) Cutoffs

	-4	-3	-2	-1
N	869	873	876	759
M	29.9	23.9	39.2	-137.2
SD	0.1	0.1	0.1	0.1
t	8.4	7.5	10.5	-36.2
p	< .001	< .001	< .001	< .001

(b) Repetitions

	-4	-3	-2	-1
N	1451	1456	1439	1449
M	27.1	31.4	77.9	95.7
SD	0.1	0.1	0.1	0.1
t	10.4	12	27	36.5
p	< .001	< .001	< .001	< .001

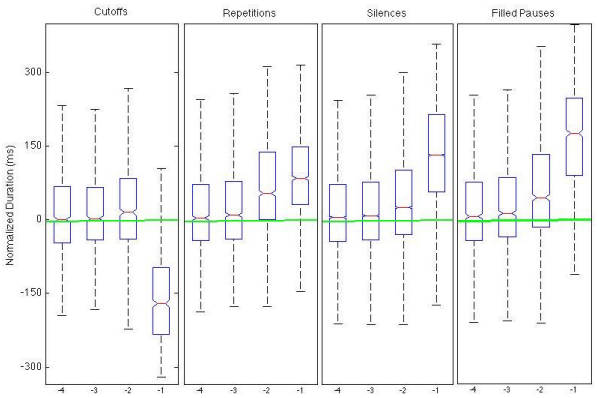
(c) Silences

	-4	-3	-2	-1
N	25108	25114	25007	23817
M	25.7	29.5	48	146.6
SD	0.1	0.1	0.1	0.1
t	41.3	46.4	70.9	194.5
p	< .001	< .001	< .001	< .001

(d) Filled Pauses

	-4	-3	-2	-1
N	1129	1129	1107	1057
M	27.1	37.4	67.8	170.9
SD	0.1	0.1	0.1	0.1
t	9	12.3	20	49.4
p	< .001	< .001	< .001	< .001

Figure 3: Normalized duration of word -1 to -4 before different types of disfluencies



4.3. Duration of words before cutoff and non-cutoff disfluency

Repetitions, silences and filled pauses were grouped as non-cutoffs to compare with the cutoffs. There is a small increase in the normalized durations of words before non-cutoffs from -4 to -2, and a large

increase at -1 (see Figure 4). One-sample t-tests revealed them to be significantly different from 0 at $p < .001$ level (see Figure 5), and a one-way ANOVA suggested a significant effect of word position on normalized duration (($F(3) = 1875.3$, $p < .001$)).

Figure 4: Normalized duration of word -1 to -4 before cutoffs and non-cutoffs

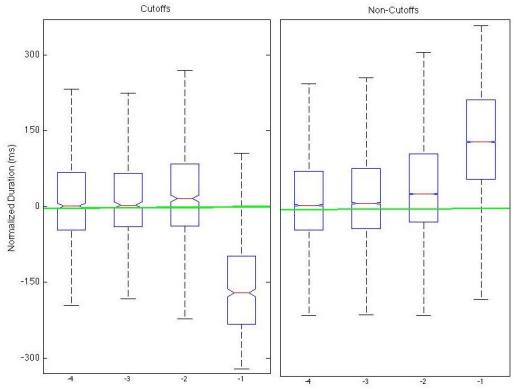


Figure 5: One sample t-tests for normalized duration of word -1 to -4 before non-cutoffs

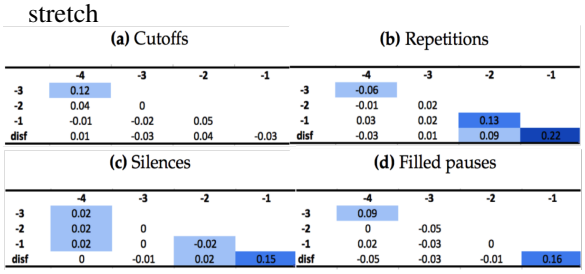
	-4	-3	-2	-1
N	27688	27699	27553	26323
M	25.9	29.9	50.3	144.8
SD	0.1	0.1	0.1	0.1
t	43.6	49.4	77.6	202.5
p	< .001	< .001	< .001	< .001

A 2x3 ANOVA (2 disfluency types (cutoffs, non-cutoffs) and 3 word positions (-4, -3, -2)) revealed a main effects of word position ($F(2, 23392) = 29.1$, $p < .001$). Post-hoc multiple comparison tests suggested that the normalized duration for word -2 before cutoffs and non-cutoffs are significantly different ($p < .001$), while the the durational differences for word -3 and -4 before cutoffs and non-cutoffs are not significant.

4.4. Correlation matrix

Correlation matrix of normalized durations of word -1 to -4 and log duration of disfluent stretch are shown in Figure 6. Weak correlations were found between the normalized duration of word -1 and the normalized duration of the repeated word ($r = 0.22$, $n = 1449$, $p < .001$), the log duration of the silence ($r = 0.15$, $n = 23817$, $p < .001$), and the log duration of the filled pause ($r = 0.16$, $n = 1057$, $p < .001$). No correlation was found between the normalized duration of word -1 and the log duration of the following disfluent stretch ($r = -0.03$, $n = 759$, $p = 0.46$).

Figure 6: Correlation matrix of normalized duration of word -1 to -4 and log duration of disfluent stretch



5. DISCUSSION

The present study examined the normalized durations of four words before the surface interruption point of cutoffs, repetitions, silences and filled pauses. The results suggested a marginal increase in the durations of words before cutoffs and a large increase in the durations of words before the other three types of disfluency. The differences in pre-interruption lengthening between cutoff disfluencies and other patterns of disfluency supported Shriberg's ([9]) hypothesis that lengthening effect occurs before hesitations due to word-retrieving difficulties, but not before detection of overt speech errors. It also provides phonetic evidence for Levelt's ([8]) double-loop theory of self-monitoring, which suggests that detection of error can occur both at the internal loop before the articulation of the error and the external loop after the articulation of the error.

Weak correlations between the durations of word -1 and the following disfluent stretches, supporting the hypothesis that a more serious processing problem would predict a larger lengthening effect and a longer duration of disfluency. No correlation was found between the durations of word -1 and the following disfluent stretches, further supporting the distinction between cutoffs and non-cutoffs.

The current classification of disfluency is only based on the occurrence of a cutoff word, a repeated word, a silent period ("SIL") and a filled pause ("um" and "uh"), disfluencies of syntactic and semantic nature without these overt signals are not examined. In addition, the silence and filled pause disfluencies are only considered as due to word-retrieval difficulties that occur at the internal loop of speech production, but they may as well occur after external speech errors, patterning the cutoff disfluencies (e.g., "here is a horizontal – uh/SIL a vertical line"). A syntactic-semantic metric should be developed in future to include the syntactically abnormal sentences in the disfluent groups, and to exclude the heterogeneity of silence and filled pause disfluencies.

6. REFERENCES

[1] Babel, M. 2009. *Phonetic and social selectivity in speech accommodation*. PhD thesis University of California; Berkeley.

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

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

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

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

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

[7] Levelt, M. 1983. Monitoring and self-repair in speech. *Cognition* 14, 41–104.

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

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

[1] We considered the same speaker in different conversations as different variables (so we have 1282 (641*2) speaker variables, not the actual number of speakers), because speaker may talk differently with a different conversation partner (see [1] for phonetic convergence)