Proceedings of DiSS '01 Disfluency in Spontaneous Speech,
ISCA Tutorial and Research Workshop, University of Edinburgh,
August 29–31, 2001, pp. 5–8.

Prolongations: A dark horse in the disfluency stable

Robert Eklund†‡

† Telia Research AB, Sweden
‡ NLPLab, Dept. of Computer and Information Science, Linköping University, Sweden
Robert.H.Eklund@telia.se

Abstract

This paper studies a specific type of disfluency, *viz*. segment prolongation (PR), i.e., the "stretching out" of speech sounds as a means of hesitation. It is shown that the occurrence of PRs varies as a function of phone type, position in the word, lexical factors and word class, and that PRs are subject to phonotactic constraints in Swedish. A comparison between Swedish and Tok Pisin suggests that there are language-specific traits associated with PR production.

1. Introduction

Studies of disfluency phenomena such as filler words, repetitions, pauses, truncations, insertions, deletions and so on have become common recently.

However, one type of disfluency that has received little attention in the literature is segment prolongation, i.e., the "stretching out" of speech segments. It has been shown that PRs are more common than most other types of disfluencies, outnumbered only by filled pauses (FPs, also called "filler words") and unfilled pauses (silences) [2][4][5]. An adequate description of PRs could provide insight into human speech production, and could also help improve durational modeling for automatic speech recognizers.

Moreover, while cross-linguistic studies have shown that there are great similarities with regard to disfluencies across languages [1][2][5][10], Eklund [2] showed that differences between Swedish and Tok Pisin occur at the PR level.

The objective of this paper is to take a more detailed look at the characteristics of PRs.

2. Method

2.1. Corpora

Data from four Swedish spoken language corpora—names in table below—were analyzed. The corpora were all collected as part of the Spoken Language Translator project [8] at Telia Research AB during the period 1996 through 1998. All dialogs were task-oriented within Air Travel Information Service (ATIS) or toward the booking of general business trips [4]. Summary statistics are shown in Table 1.

Table 1: Summary statistics for the Swedish corpora. UW=Unique Word Forms. WT=Word Tokens. H=Human. "M"="Machine" (i.e., Wizard-of-Oz simulation). M=Machine.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
Type	H"M"H	H"M"	HH	HM	—
No. subjects	49	23	8	16	96
M/F	26/23	16/7	6/2	9/7	57/39
No. Dialogs	84	71	24	69	248
No. Utts.	3,104	1,849	1,698	1,888	8,539
No. UW	570	792	1,157	959	3,478
No. WT	5,565	12,190	9,232	12,047	39,034

2.2. Set-up and subjects

All dialogs were made over a telephone line, and high quality recordings were made to enable acoustic analysis. The subjects were all Telia employees, and were all used to travel bookings.

2.3. Disfluency annotation

The annotation scheme used is described by Eklund [4], and is based on, and similar to, that described by Shriberg [9]. All corpora were labeled by the author.

3. Results

3.1. PR rates

Occurrence of PRs is shown in Table 2.

Table 2: Summary statistics. UW=Unique Word Forms. WT=Word Tokens.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. PRs	101	88	129	179	497
% PRs/UW	17.72%	11.11%	11.15%	18.66%	14.29%
% PRs/WT	1.81%	0.72%	1.40%	1.48%	1.27%

As can be seen, 0.7 to 1.8% of the words include prolonged segments at token level.

3.2. Durational data

The mean duration for all PRs (all data pooled) was 0.289 milliseconds (N=497). The 95% confidence interval was 0.275/0.305. Standard deviation was 0.170.

The bottom end of the durational scale is problematic from the point of view of labeling, since it is difficult to say exactly when a segment is in fact prolonged. Thus, the data were also explored with the lower quartile removed (N=373; cut-off point at 0.175). The mean duration for trimmed PR data was 0.340. The 95% confidence interval was 0.323/0.357. Standard deviation was 0.167.

3.3. PRs vs. FPs

PRs and FPs have something in common that distinguishes them from other disfluency phenomena: they both signal hesitation by means of vocalization and duration (unlike repetitions, truncations, mispronunciations etc.). There is no obvious reason to assume *a priori* that there would be any durational differences between the two types since they both serve the same purpose, i.e., signaling hesitation while still speaking (as opposed to inserted silence).

The mean duration for all FPs pooled was 0.488 (N=1,379). The 95% confidence interval was 0.474/0.501. Standard deviation was 0.255.

A *t*-test showed that FPs were significantly longer than PRs ($p<0.001$; two-tailed; mean difference 0.198). Comparing pooled PR and FP values within the same corpora creates problems with regard to whether or not the variables are to be considered dependent or not. Thus, a Wilcoxon signed ranks test and a Mann-Whitney test were also performed. Both tests showed significance at the $p<0.001$ level. Thus, it would appear safe to conclude that FPs are generally longer than PRs. However, these results need be tested at the individual level before more definite conclusions can be drawn.

3.4. Individual differences

The next thing we looked at was whether or not there were any individual differences with regard to PR production, both *per se* and relative to FP production. The observations are shown in Table 3.

Table 3: Relative frequency of PR and FP usage. Note that the sum of the four categories sometimes exceed the numbers of subjects due to the fact that the same person might appear in two cells when the lower number of comparison is zero. >="More frequent than".

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. subjects	49	23	8	16	96
FPs > PRs	35	21	7	13	76
PRs > FPs	10	2	1	3	16
FPs = PRs	4	—	—	—	4
No FPs	6	1	—	—	7
No PRs	16	3	—	—	19

For most subjects, FPs are more common than PRs. It is also more common to not employ PRs at all than it is to totally lack usage of FPs. However, the bulk of subjects who did not make use of PRs at all occur in WOZ-1. This corpus differs from the others in the way the tasks were presented [4], and contains much shorter dialogs. At least a part of the skewness could be attributed to the particulars of the task.

3.5. PR position in the word

As has been reported for Swedish [2][4], American English [5], and Tok Pisin [4], PRs are not evenly distributed within the word. A breakdown of PR position is shown in Table 4.

Table 4: Phone type and position of prolongations. For each corpus the number and percentages of phone position are given.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. PRs	101	88	129	179	497
No. Segments	24,402	52,157	32,549	96,215	205,323
% PRs/Segments	0.41%	0.17%	0.40%	0.19%	0.24%
% Initial	30/29.7	32/36.4	32/24.8	56/31.3	150/30.2
N/% V	8/26.6	8/25.0	6/18.8	10/17.8	32/21.3
N/% C +son	–/–	6/18.8	18/56.2	14/25.0	38/25.3
N/% C –son	22/73.4	18/56.2	8/25.0	32/57.2	80/53.4
% Medial	13/12.9	16/18.2	25/19.4	28/15.6	82/16.5
N/% V	2/15.4	2/12.5	5/20.0	4/14.2	13/15.8
N/% C +son	3/23.1	–/–	4/16.0	1/3.6	8/9.7
N/% C –son	8/61.5	14/87.5	16/64.0	23/82.2	61/74.5
% Final	58/57.4	40/45.4	72/55.8	95/53.1	265/53.3
N/% V	14/24.1	8/20.0	24/33.3	29/30.5	75/28.2
N/% C +son	31/53.5	28/70.0	37/51.4	57/60.0	153/57.8
N/% C –son	13/22.4	4/10.0	11/15.3	9/9.5	37/14.0

While the previously reported 30–20–50 ratio for initial, medial and final position [2][4][5], respectively, is confirmed, Table 4 shows that this tendency does not hold for all *types* of segments. While non-sonorant consonants are the most commonly prolonged segments in initial and medial position, they are the *least* frequently prolonged segments in final position. This indicates that certain types of segments are preferred in certain positions, which hints at an interaction between position in the word and segment type.

The occurrence of word-medial PRs distinguishes them from FPs, since no word-internal FPs have been encountered in our data, although FPs have been reported between lexical roots inside compounds in Swedish [5] and German [7].

3.6. Segment type

A detailed examination of segment type was then undertaken. PR frequency was normalized for overall segment frequency. Since the corpora were all transcribed according to an orthography-based, phonological scheme, making exact calculations cumbersome, all numbers must be considered approximate. The top five segments are shown in Table 5.

Table 5: Most commonly prolonged segments. For all segments, the number of prolonged segments is given divided by the total number of the same segment in the same corpus, in order to normalize for differences in general segment occurrence.

WOZ-1	WOZ-2	Nymans	Bionic
[f] 1.67%	[f] 0.90%	[f] 1.51%	[f] 2.05%
[k] 0.98%	[n] 0.56%	[s] 1.42%	[n] 1.16%
[s] 0.97%	[s] 0.48%	[n] 1.19%	[k] 1.01%
[l] 0.96%	[k] 0.43%	[o] 0.62%	[l] 0.72%
[n] 0.76%	[i] 0.32%	[l] 0.54%	[s] 0.52%

As is shown, almost all of the twenty segments are continuants. The exception is the stop [k], which often comes from the medial [k] in "klockan" (o'clock), accounting for 27% of the cases. The only vowels in the list are [o] occurring in the preposition "på" (on, for dates), which accounts for 86% of the cases, and [i], from the preposition "i" (in), which accounts for 78% of the cases.

Although segment type proper obviously is of importance, a detailed look at the words in which the prolonged segments appear implies that there is also a lexical factor to consider.

3.7. Open versus closed word classes

Given the observations in 3.6, possible differences between open and closed word classes were studied (using a traditional definition of 'open' and 'closed' word classes). Summary statistics are presented in Table 6.

Table 6: Ratio open/closed word classes. UW=Unique Word Forms WT=Word Tokens.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. Open WT	3,355	6,956	4,045	6,519	20,875
% total no. WT	60.3%	57.1%	43.8%	45.9%	53.5%
No. Open UW	497	695	993	893	3,078
% total no. UW	87.2%	87.8%	85.8%	93.1%	88.5%
No. Closed WT	2,210	5,234	5,187	5,528	18,159
% total no. WT	39.7%	42.9%	56.2%	54.1%	46.5%
No. Closed UW	73	97	164	66	400
% total no. UW	12.8%	12.2%	14.2%	6.9%	11.5%

As can be seen, the distribution is more or less 50/50 between open and closed word classes at token level, and averages one-to-eight at unique forms level.

The number of PRs in open and closed word classes are shown in Table 7.

Table 7: Rate of PRs occurring in words belonging to open/closed word classes (tokens).

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. Open	48	39	43	71	201
%	47.5%	44.3%	33.3%	39.7%	40.4%
No. Closed	53	49	86	108	296
%	52.5%	55.7%	66.7%	60.3%	59.6%

As is shown, there is a slight inclination towards prolonging words belonging to closed words classes. The difference is significant at p=0.145 (Pearson chi-square, two-tailed). Given definitional problems associated with the categories 'open' vs. 'closed', these data should be handled with some caution.

3.8. Domain dependency

We then investigated specifically which open words were prone to prolongation. Judging from the examples above, it seemed that within-domain words were more likely to be prolonged than words outside the domain. Examples of prolonged, within-domain words were "boka" (book/reserve), "hotell", "taxi" (taxi), "resa" (travel/go), "rökfritt" (non-smoking), "billigaste" (cheapest), "hemresa" (return trip), and words for dates, times and locations. The results are shown in Table 8.

Table 8: Rate of PRs occurring in words belonging to open word classes (tokens). Figures are given both for general vocabulary and domain-dependent vocabulary.

	WOZ-1	WOZ-2	Nymans	Bionic	Σ
No. open words w/ PRs	48	39	43	71	201
No. open words w/ PRs in domain	27	31	23	61	142
% open words w/ PRs indomain	56.3%	79.5%	53.5%	85.9%	70.6%

In all corpora, within-domain words are more often prolonged than outside-domain words. Pooling all data, the difference is significant at p<0.001 (Pearson chi-square, two-tailed). However, when pooling only WOZ-1 and Nymans, the difference is not significant (p=0.792, Pearson chi-square, two-tailed). These results might be an artefact of general corpus differences. The tasks were presented differently in WOZ-1 [4], and Nymans was human–human, while WOZ-2 and Bionic were similar both with regard to task details and setup. Consequently, no far-reaching conclusions will be drawn here with regard to the observed differences between the corpora.

3.9. Phonological length

A final issue, not to be bypassed, is that of phonological length, which is distinctive in Swedish. It is also mutually exclusive, which means that all VC syllables come either as V:C or VC: (or VCC). In recent work on dynamic segmental effects associated with focusing in Swedish, Heldner & Strangert [6] show that while focused segments in general are lengthened by an average 25%, short vowels are only marginally—not distinctively—lengthened. This observation is repeated in our data. While long vowels and both long and short consonants are subject to prolongation, there are no instances of prolonged short vowels.

4. A comparison with Tok Pisin

4.1. Tok Pisin corpus

In order to test some of the observations made above, a comparative study was made on available Tok Pisin data. The Tok Pisin corpus (TP) consists of authentic ATIS dialogs, collected on location in Kavieng, Papua New Guinea, during the period December 1999 and January 2000 [3]. TP consists of 39 authentic human–human ATIS dialogs, and was labeled by the author (who is not a native speaker of Tok Pisin). Currently, a total number of 654 utterances and 3,538 words have been transcribed, with a total number of 35 PRs [2].

4.2. Durational data

The mean duration for all PRs was 0.347 (N=35). The 95% confidence interval was 0.287/0.407. Standard deviation was 0.170. There was no significant difference between PR durations in Swedish and Tok Pisin (p=0.055, t-test, two-tailed, equal variances assumed).

4.3. PRs versus FPs

It was shown for Swedish that FPs were significantly longer than PRs. To check whether this holds true for Tok Pisin, the values for FPs in TP were explored. The mean for all FPs was 0.456 (N=80). The 95% confidence interval was 0.401/0.501. Standard deviation was 0.244.

FPs were significantly longer than PRs. A t-test resulted in p=0.018 (two-tailed, equal variances assumed), and a Mann-Whitney test resulted in p=0.008 (two-tailed).

4.4. PR position in the word

The distribution of PRs as a function of position in the word is shown in Table 9.

Table 9: Phone type and position of PRs.

	TP
No. PRs	35
No. Segments	12,840
% PRs / Segments	0.27%
% Initial phone	6/17.1%
% vowel	4/66.8%
% cons +sonorant	1/16.6%
% cons –sonorant	1/16.6%
% Medial phone	—
% Final phone	29/82.9%
% vowel	12/41.4%
% cons +sonorant	13/44.8%
% cons –sonorant	4/13.8%

As is shown, the ratio in TP for initial/medial/final position is roughly 15–0–85, which differs from the distribution reported for Swedish and American English, mentioned above.

4.5. Segment type

The most commonly prolonged segments (normalized for overall segment frequency) in TP were, in descending order: [ŋ] (1.20%); [m] (0.82%); [s] (0.51%); [o] (0.41%); [u] (0.35%). That other segments are prolonged more often in Swedish than in Tok Pisin is perhaps not surprising. What is more striking is that the segments seem to be prolonged for the same reason. The phones [ŋ] and [o] mainly occur in the prepositions "long" (general preposition), pronounced [lɔŋ] or [lɔ]) and "bilong" (stronger-binding preposition, genitive marker, conjunction), pronounced [bilɔŋ] or [blɔ].

4.6. Open vs. closed word classes

Rates of words belonging to open and closed word classes and PR rates are shown in Table 10.

Table 10: Ratio open/closed word classes and PR rates in TP.

	TP
No. Open / % total no. words	1,592/45.0%
No. Closed WT / % total no. words	1,946/55.0%
No. Closed UW / % total no. Closed words	39/2.0%
No. PRs Open / % total no. PRs	6/17.1%
No. PRs Closed / % total no. PRs	29/82.9%

The tendency to prolong words belonging to closed word classes is more marked in TP than in the Swedish data. Out of 35 PRs, 29 occur either in prepositions ("long", "bilong") or in grammatical markers such as "i" (predicate marker), "bai" (future marker) or "ol" (plural marker). Moreover, three of the six prolonged words belonging to open word classes are from the domain. "fe" (fare), "ples" (place) and "tri" (three). Also, two instances of a prolonged transitive suffix "-im" are found in the words "salim" (send) and "sekim" (check). These two could arguably be analyzed as grammatical ('closed') prolongations.

5. Discussion

From a **phonological** point of view, it would seem that all segment types might be prolonged, although there is a tendency towards prolonging continuants.

Looking at **phonological length**, it is striking to find that no short vowels are prolonged in our data. This observation supports the hypothesis that phonology puts constraints on the production of PRs, which receives further support from the observations reported by Heldner & Strangert [6].

Looking at **duration** proper, our data suggest that PRs are shorter then FPs, despite their physiological, acoustic and functional similarities. The observation, if tentative, that FPs generally have longer duration could imply that FPs do have a different "status" and are viewed by the speaker as "words" in their own right. Also, that PRs, unlike FPs, are observed in word-medial position is another trait that implies that PRs and FPs do not have the same status in speech production.

From a **morphological** point of view, the favored position for segment prolongation is word-final, in both Swedish and Tok Pisin. However, the observation that the ratio initial/medial/final position differs between Swedish and Tok Pisin could suggest that PR production could be language-specific, being associated with the morphotactics of a given language.

Stepping up to **full words**, the tendency is that words belonging to closed word classes are more prone to prolongation than words belonging to open words classes.

Within the open words class group, most words with prolonged segments are within the **discourse domain**. This is not surprising, since speakers hesitate before or on items with high cognitive load, i.e. either the preposition or article before a semantically heavy item. However, words inside the domain are more likely to be uttered by many speakers, and are thus prone to over-representation, as compared to words outside the domain.

From a **cross-linguistic** perspective, the comparison with Tok Pisin shows that there are similarities between the languages. While the data in TP exhibits the same tendency to prolong segments in words belonging to closed word classes, there are significant differences at the segmental and distributional levels. This could imply that Tok Pisin speakers hesitate at the same places that Swedish speakers do, but that the hesitation affects other types of segments, given different morphological and phonotactic constraints.

6. Conclusions

In conclusion, the prototypical Swedish PR would be the final segment—preferably a continuant—of a preposition or article, or appear in a domain-dependent word which signals crucial information with regard to the task at hand.

The comparison with Tok Pisin suggests that these observations probably do not hold for all languages, and that more cross-linguistic studies of PRs need be done in order to gain deeper insights with regard to the role and function of segment prolongation in human speech production.

7. Acknowledgements

Thanks to Åsa Wengelin, Jaan Kaja and Eva Lindström for a plethora of comments. Thanks to Michael Kiefte for proofs.

8. References

[1] Den, W. & H. Clark. 2000. Word Repetitions in Japanese Spontaneous Speech. *Proc. ICSLP'00*, Beijing 16–20 October 2000, vol. 1, pp. 58–61.

[2] Eklund, R. 2000. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and Tok Pisin Human–Human ATIS Dialogues. *Proc. ICSLP'00*, Beijing, 16–20 October 2000, vol. 2, pp. 991–994.

[3] Eklund, R. 2000. Wanpela deitabeis long Tok Pisin bilong baim tiket bilong balus. (An ATIS database in Tok Pisin.) Methodological observations with regard to the collection of human–human data. *Proc. Fonetik 2000*, The Swedish Phonetics Conference, May 24–26 2000, University of Skövde, pp. 49–52.

[4] Eklund, R. 1999. A Comparative Study of Disfluencies in Four Swedish Travel Dialogue Corpora. *Proc. Disfluency in Spontaneous Speech Workshop*, Berkeley, California, 1 July 1999, pp. 3–6.

[5] Eklund, R. & E. Shriberg. 1998. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues, *Proc. ICSLP'98*, Sydney, 30 November–5 December 1998, vol. 6, pp. 2631–2634.

[6] Heldner, M. & E. Strangert. Temporal effects of focus in Swedish. Accepted for publication, *Journal of Phonetics*.

[7] Lüngen, H., M. Pampel, G. Drexel, D. Gibbon, F. Althoff & C. Schillo. 1996. Morphology and Speech Technology. *Proc. ACL–SIGPHON Conference*, 28 June 1996, Santa Cruz, pp. 25–30.

[8] Rayner, M., D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). 2000. *The Spoken Language Translator*, Cambridge University Press.

[9] Shriberg, E. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of Berkeley, California.

[10] Tseng, S.-C. 2000. Modelling Speech Repairs in German and Mandarin Chinese Spoken Dialogues. *Proc. COLING 2000*, Saarbrücken, 31 July–4 August 2000, vol. 2, pp. 864–870.