Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data

Simon Betz[1,2], *Jana Voße*[1,3], *Petra Wagner*[1,2]

[1]Bielefeld University, Phonetics and Phonology Workgroup
[2]Bielefeld University, Center of Excellence Cognitive Interaction Technology (CITEC)
[3]University of Gothenburg, Department of Philosophy, Linguistics and Theory of Science
simon.betz@uni-bielefeld.de

Abstract

Our overarching research project explores the usability of disfluencies in incremental spoken dialogue systems. This endeavor requires basic phonetic research on disfluencies in spontaneous speech corpora as to define strategies for synthesizing disfluencies in a meaningful way. In this paper, our current research focus lies in an investigation of disfluency-related lengthening as a promising time-buying strategy in synthesized dialogue [1][2]. We base our analyses on the results of a search tool aiming to automatically detect lengthening in spontaneous speech corpora occurring without adjacency to phrase boundaries or other disfluencies, i.e. standalone lengthening phenomena. We analyzed disfluency-related lengthening in the "monomodal" half of the GECO corpus [3], with regard to their context, word class, syllable position and phone type. We then postulate a disfluency insertion strategy for synthetic speech that prioritizes lengthening phenomena based on the results obtained in our study.

Keywords: Disfluencies, Hesitation, Lengthening, Spontaneous Speech, Speech Synthesis

1. Introduction

Disfluencies have become increasingly popular from a speech synthesis perspective [4][5][1]. Especially incremental spoken dialogue systems, that plan and prepare their responses while the interlocutor is speaking, are promising areas of their application [5][1]. One of the reasons for this development is that conversational speech phenomena such as disfluencies can buy valuable time to retrieve content, to facilitate the production of corrections and to signal complexity to the listener.

Disfluencies are manifold in structure and the terminology used to describe them is often ambiguous and varies depending on publication date and perspective. In general, we use the terminology established by [6] and [7] to describe the overarching macro-structure of disfluencies, and refer to the phonetic correlates in the speech signal, such as silent pauses, fillers, or lengthened words, as disfluency elements [1].

In this study, we focus on one particular disfluency element, namely *standalone lengthening*, which we define as a stretch of unexpectedly high segmental duration in an utterance that features no other disfluency elements. For a start, any elastic phone (i.e. one that is prolongable) in any syllable or word can carry the lengthening. However, we hypothesize that there are restrictions as to where lengthening manifests itself. To detect regularities of disfluent lengthening in German is one aim of this study. Are certain word classes, syllable types or phone types preferred?

Lengthening in general appears to be capable of buying valuable dialogue time without being detrimental for synthesis quality [1]. Lengthening occurs by default toward the end of syntactic and intonational phrases. Additionally, overt hesitations containing fillers such as "uhm" are regularly preceded by lengthening [2][8][9]. Standalone lengthening has gotten little to no attention so far, but our position is that if lengthening is both capable of buying time and can do so without being detrimental to synthesis quality, then it is worthwhile considering the synthesis of standalone lengthening. In order to do so, we examine in this study tokens of standalone lengthening extracted from human dialogue data.

We propose a general strategy for the synthesis of hesitation that does reflect human speech planning as described by [10], cited in [8] and provides a good testing environment for standalone lengthening:

1. **Lengthen if possible**
2. Silent pause if issue not solved
3. Insert filler if issue still not solved

When a speaker or dialogue system senses an upcoming production issue, such as end of available, pre-planned speech material, or the anticipation of upcoming complex information that needs more processing time, lengthening is applied if the articulatory buffer still contains suitable material [8]. If lengthening cannot be suitably applied or the planning issue has not been solved during the insertion of lengthening, measures with more severe acoustic impact, such as the insertion of silences or fillers can be taken. On the other hand, if the lengthening successfully bought enough time to solve the issue, fluency can be resumed, resulting in a standalone lengthening on the surface signal.

We hypothesize that standalone lengthening does not occur at arbitrary places and that certain rules have to be paid attention to when synthesizing them. In previous work, we conducted a corpus study based on spontaneous conversational German speech and automatically filtered out standalone lengthening [2]. For this study, the output of this search is annotated and analyzed with regard to its surrounding, word class, syllable position and phone type, thus providing an empirical basis for modeling synthetic hesitation.

2. Methods

2.1. Corpus Data and Lengthening Extraction

This study is based on the GECO corpus [3], a phonemically annotated corpus of spontaneous German speech. We used the first half of it, the "monomodal" condition, where speakers had

no visual contact. One file had to be omitted due to technical issues, yielding 43 files each containing 30 minutes of speech.

The method presented in [2] searches phonemically annotated corpora for places of markedly high phone duration of a z-score of 3 or more, that are not followed by fillers, silences or utterance endings, i.e. noticeable "standalone lengthenings" that are not caused by phrase finality. Z-scores were calculated per phone type and per speaker.

2.800 tokens of lengthening were extracted from this part of the corpus. These tokens fall mainly into three categories: (1) *Disfluent lengthening*, (2) *accentual lengthening*, and (3) *forced-alignment errors*. All tokens were hand-labeled by two annotators for further analyses.

2.2. Inter-annotator Agreement

The two annotators labeled the output phones according to the three main categories. Inter-labeler agreement was tested on a subset of 13 files of the corpus, after a training phase based on four different files from the same corpus.

Agreement was calculated on three categories. The most important one is the distinction between accentuation and disfluencies, where annotators agreed in 98.8% of cases. Furthermore, it was checked how many instances of accentuation or disfluency were only labeled by one annotator, i.e. instances where the other annotator labeled nothing. 92.2% of disfluencies were labeled by both annotators as well as 89.8% of accentuations.

It appears straightforward for listeners to identify disfluency and accent related lengthening. Agreement on the distinction between disfluency and accent related lengthening is also very high, yet it can be seen that not all instances of lengthening can clearly be defined as being of one type or the other. Overall it can be claimed that inter-annotator agreement is sufficient to base further analyses on these annotations.

2.3. Token Frequencies and Errors

In total, 1.000 tokens of lengthening, 75% of them disfluent and 25% accent, were extracted from the first half of the corpus. 1.800 tokens were discarded because of grave forced-alignment errors, or for reasons such as the lengthening being phrase-final and neither disfluent nor emphatic.

About 500 of the remaining 1.000 lengthening tokens still contain minor boundary errors, that are corrected for future analyses, but are not severe enough to discard the tokens.[1] This reveals that even where the search tool outputs the material we're after, forced-alignment shortcomings emerge. We suspect that the unusually high length of these phones troubles the language models the forced alignment works with.

3. Results and Discussion

3.1. Is there "pre-lengthening lengthening"?

Phrase-finality and disfluencies like filled pauses are regularly preceded by lengthening that extends and gradually increases over several phones [2]. The standalone lengthening examined here lacks such a feature. As can be seen in figure 1, no systematic durational variation can be observed in any phones preceding standalone lengthening. Normalized duration means cluster around the mean (0), only the outliers directly before the disfluent lengthening (position -1) hint at a slight increase. Note that due to our filtering method, the phones that follow the -1 position have a z-score of 3 and more, which is a drastic increase from position -1. Pairwise t-tests were conducted on all pairs of adjacent positions, yielding no significant results ($p > 0.1$), thus supporting the hypothesis that there is no systematic increase in duration preceding a standalone lengthening disfluency.

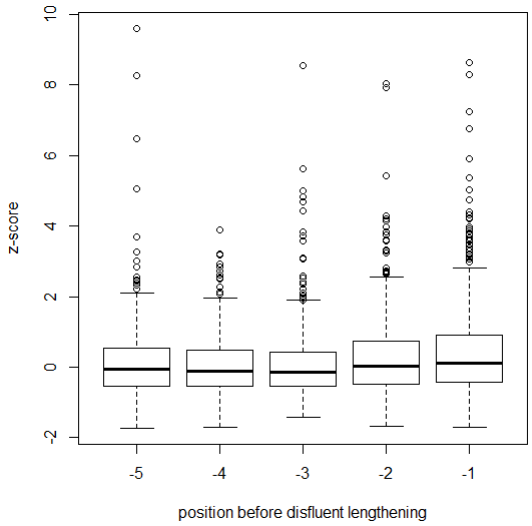

Figure 1: Normalized phone durations preceding lengthening. Positions are indicated in phones relative to position of lengthened phone ($= 0$).

This further supports the hypothesis that lengthening is the first signal of hesitation, i.e. the primary measure that speakers employ before using silent or filled pauses. These lengthening-only hesitations are not introduced by a slowing down of speech rate. Rather, they *are* the slowing down - but in case of successful time-buying, they appear without any further surface disfluency element following. The cases examined here are very likely ones where speakers are able to resume fluency after the lengthening.

3.2. Syllable positions and phone classes

The observation that hesitation begins with lengthening and has no apparent pre-planning beforehand is supported by the fact that disfluent lengthening manifests itself not only in the syllable nuclei but also to a considerable extent in the coda. In contrast, accent related lengthening manifests itself almost exclusively in nuclei (cf. table 2).

In fluent speech, speakers plan beforehand where they place their accent, so it is likely for them to choose vowel nuclei. In case of disfluencies, speakers often do not have the chance to time the "perfect phonotactic moment" to hesitate and resort to coda positions. One reason for doing so might be the vowel quality of the nucleus.

As can be seen in figure 3, the syllable position of the lengthening is related to the nucleus vowel being short, long or a diphthong. If disfluent lengthening occurs in the nucleus, it has a tendency to be realized on long vowels. Much more striking is that when disfluent lengthening happens in the coda, the preceding vowel is likely to be short. This could mean that

[1] The 1800 tokens that were discarded for example were tokens labeled as /a:/ but were an entirely different phone. The 500 erroneous tokens that we kept in were ones that contain the right phone, but the boundaries are dislocated by $<$ 20 ms.

Disfluent Word	*English Transl.*	*Frequency*
und	and	61
die	the	35
so	so	27
dann	then	23
in	in	22
ich	I	19
das	the	16
ist	is	16
irgendwie	somehow	15
weil	because	14

Table 1: 10 most frequent words lengthened for disfluency

Function words	Content words
582 (77%)	173 (23%)
Total of words with freq. > 1	With freq. = 1
540 (71.5%)	215 (28.5%)
Content words with freq. > 1	With freq = 1
32 (5.9%)	141 (94.1%)

Table 2: Function and content word distribution within disfluencies

speakers, when they spontaneously have to find the best spot for placing a hesitation, they rather choose an elastic sonorant in the coda than a short vowel nucleus. For accentual lengthening in the nucleus, the vowel types are quite evenly distributed. Accentuation lengthening in the coda is rare, but even so, there is a slight majority of short vowels.

3.3. Word classes

3.3.1. Function words and content words

As noted by [11], lenghtening occurs mainly on function words, such as determiners, prepositions and conjunctions. This is confirmed by our data: we examined word frequencies of the 755 examined disfluencies and table 1 lists the 10 most frequent disfluent words. The same picture extends downward. Apart from auxiliary forms of *sein* "to be", there are no nouns, verbs or adjectives in the top 41 ranks, or in the top 59% of disfluent words. A preliminary word class-tagging was performed, showing that function words add up to 77% of the disfluencies. 28.5% of the words occur only once, and 81.5% of the content words fall into that region. To put it differently, 94.1% of the words that occur only once are content words, while only 5.9% of the more frequent words are content words (cf. table 2). It appears that hesitation indeed preferably manifests itself on function words. The fact that the great amount of lengthened content words occur only once in our data hints to an interpretation that a random target for hesitant lengthening is likely to be chosen, when no suitable function word is available in the articulatory buffer.

3.3.2. Conjunctions

The by far most frequent word on which disfluent lengthening occurs is the conjunction *und* "and". Conjunctions represent the default word class linking two parts of an utterance, so it makes sense for speakers hesitate at this point, in order to facilitate speech planning for the remainder of the utterance and to signal increased cognitive load to the listener, who can in turn infer that it is not the conjunction which is causing the trouble, but

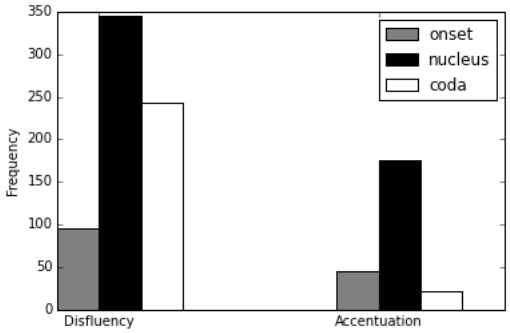

Figure 2: Syllable positions of lengthened phones

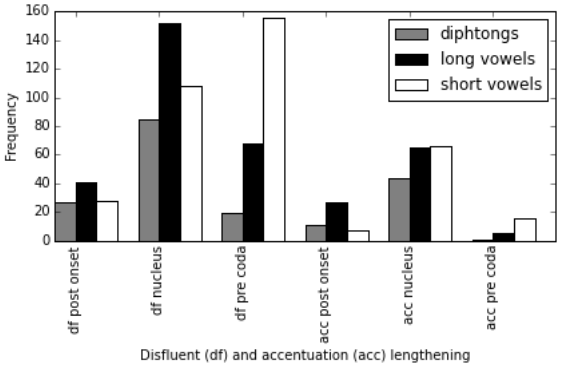

Figure 3: Vowel qualities related to syllable position of lengthening

the material that is about to follow.

3.3.3. Determiners

Quite remarkably, the distribution of the determiners of different gender is extremely skewed. As can be seen in table 1, the female (*die*, 35) and neutral (*das*, 16) determiners are quite frequent, while there are only three tokens of the male one, *der*. German word frequency studies predict these three words to be equally frequent. It can only be assumed that the long vowel in the open syllable of *die* is easiest or most suitable to sustain, whereas the diphtong in *der* might be less so.

4. Conclusions and Outlook

Our study set out to characterize naturally occuring standalone lengthenings in conversational speech as a blueprint for modeling hesitation in synthetic speech as a strategy for "buying time". Our reasoning based on the hypothesis that an unsystematic synthesis strategy to "lengthen anything anywhere whenever needed" may be detrimental for synthesis quality if natural conversational lengthening is characterized by a more specialized pattern, such as centering on function words and containing cues to differentiate between hestiation, accentuation and phrase-final lengthening. Our analyses strengthen this assumption, as annotators were consistently able to differentiate between accentual and disfluent lengthening and we assume that the annotator's ability to do so is at least partly due to the differ-

ent distributions of the two types of lengthenings with respect to phonotactic position, phone type and word class. Of course, other acoustic cues such as accent related pitch excursions may play an additional role and the examination of these cues will be future work.

At the moment, we cannot draw any conclusions with respect to listeners' ability to differentiate between phrase final and disfluency-related lengthening phenomena. For the time being, we assume that many of the lengthenings caused by disfluencies are interpreted as indicating phrase-finality. Many disfluency-related lengthenings occurred together with conjunctions, which can be seen as optimal syntactic position for placing an intonation phrase boundary. Our evidence thus points to a speaker strategy aiming to synchronize hesitation-related lengthening and places of naturally occurring phrase final lengthening. Still, speakers are not always able to match hesitations with such "ideal positions". From a synthesis perspective, it will be of future interest to find out whether hesitation-related lengthening interrupts the prosodic structure of the ongoing intonation phrase which is later resumed, or whether it initiates a new intonation phrase.

To conclude, we postulate that in order to model disfluencies in the synthetic conversational speech, a more sophisticated routine than random lengthening has to be developed. From the insights gained here, the following sequence of steps appears reasonable in order to determine the suitable place for lengthening insertion:

1. Is a function word available in the buffer, preferably a conjunction or determiner?
2. If yes, apply lenghtening[2] on long vowel nucleus of the final syllable.
3. If yes, but nucleus has no long vowel or diphthong, but coda contains a sonorant, lengthen coda instead.
4. If no, apply lengthening as described in the steps before to last syllable of last content word in the buffer.
5. If none of the above locations are available, don't lengthen but proceed to next step in disfluency insertion (silent pause)

[2]The extent of the lengthening will be determined on a follow-up study that tests the acceptability of various lengthening extents with respect to phone elasticity.

5. Bibliographie

[1] S. Betz, P. Wagner, and D. Schlangen, "Micro-structure of disfluencies: Basics for conversational speech synthesis," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015, Dresden)*, 2015, pp. 2222–2226.

[2] S. Betz and P. Wagner, "Disfluent Lengthening in Spontaneous Speech," in *Elektronische Sprachsignalverarbeitung (ESSV) 2016*, O. Jokisch, Ed. TUD Press, 2016.

[3] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)*, pp. 525–529.

[4] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "Modelling filled pauses prosody to synthesise disfluent speech," 2010.

[5] G. Skantze and A. Hjalmarsson, "Towards incremental speech generation in conversational systems," *Computer Speech and Language 27*, 2013.

[6] W. J. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

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

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

[9] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "On the generation of synthetic disfluent speech: Local prosodic modifications caused by the insertion of editing terms," in *Proceedings of Interspeech*, 2008.

[10] E. Shriberg, "Toerrrr'is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, no. 1, pp. 153–164, 2001.

[11] D. O'Shaughnessy, "Timing patterns in fluent and disfluent spontaneous speech," in *International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995*, vol. 1. IEEE, 1995, pp. 600–603.