Increasing Recall of Lengthening Detection via Semi-Automatic Classification

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

[1]Bielefeld University, Bielefeld, Germany
[2]CITEC, Bielefeld, Germany
simon.betz@uni-bielefeld.de

Abstract

Lengthening is the ideal hesitation strategy for synthetic speech and dialogue systems: it is unobtrusive and hard to notice, because it occurs frequently in everyday speech before phrase boundaries, in accentuation, and in hesitation. Despite its elusiveness, it allows valuable extra time for computing or information highlighting in incremental spoken dialogue systems. The elusiveness of the matter, however, poses a challenge for extracting lengthening instances from corpus data: we suspect a recall problem, as human annotators might not be able to consistently label lengthening instances. We address this issue by filtering corpus data for instances of lengthening, using a simple classification method, based on a threshold for normalized phone duration. The output is then manually labeled for disfluency. This is compared to an existing, fully manual disfluency annotation, showing that recall is significantly higher with semi-automatic pre-classification. This shows that it is inevitable to use semi-automatic pre-selection to gather enough candidate data points for manual annotation and subsequent lengthening analyses. Also, it is desirable to further increase the performance of the automatic classification. We evaluate in detail human versus semi-automatic annotation and train another classifier on the resulting dataset to check the integrity of the disfluent - non-disfluent distinction.

1. Introduction

Disfluencies, such as repetitions, filler words, silent pauses or word lengthenings are useful in spoken dialogue systems [1]. As in human communication [2][3], they can be used to gain extra time for speech planning or to signal processing delays to the listener. This is especially interesting for incremental systems that generate responses in real-time, where extra computational time is valuable [1]. Consider a dialogue system talking to you:

(1) "Your next train to Hamburg leaves aaa:t 12.03"
(2) "Your next train to Hamburg leaves aaa:t ... uh ... 12.03"

(1) is an example of a short hesitation: the system lengthens one word and then returns to normal speech rate. This allows for extra time without having a negative impact on speech quality: users rate synthesized lengthening disfluencies very positively, probably because it is an unobtrusive, and hard-to-notice phenomenon in speech [4][5].

(2) is an illustration of a more severe hesitation: the system lengthens one word, then adds silent and filled pauses to gain additional extra time. The filled pause furthermore signals the delay to the listener and thus manages dialogue interaction by preventing the loss of the conversational floor [6]. However, it has been proven difficult to synthesize filled pauses of high quality [5][7], which leads to the following heuristic strategy for generating hesitations in speech synthesis:

- Always start with word lengthening.
- Only add fillers if more time is needed.

Disfluencies are ambivalent in their perceptual nature: they can be difficult to notice [8], yet they provide the listener with meta-information [6]. The hesitation examples above illustrate this point: listeners can try to remedy their production issues while lengthening a word, and if they fail to solve their problems, the result will frequently be a longer hesitation lengthening, followed by silences and fillers.

To utilize word lengthening for speech synthesis, corpus studies of this phenomenon are necessary. This poses the following dilemma: If they are so hard to notice, then how can we guarantee finding the relevant instances of hesitation-related lengthening in our data? We argue in the following that there is a recall problem when annotators are to label hesitation lengthening, which can be solved by using semi-automatic detection.

After a general introduction to our data (section 2.1.1) and methods (section 2.1.2), our study aims at making the following points:

- Human annotators, confronted with the task to label disfluency phenomena indeed miss most of the objectively measurable lengthenings present in the data (cf. section 3.1).
- Annotation aided by a simple classifier that works on a normalized duration threshold drastically increases the recall by human annotators and thus, the size of the data set available for lengthening analysis (cf. section 3.1).
- Several factors lowering the precision of the classifier can be eliminated if the corpus data allows for it (cf. section 3.2).
- Classifiers can be trained on the resulting dataset for further enhancement of automatic lengthening detection (cf. section 3.3).

2. Methods

2.1. Preliminaries

2.1.1. Corpus data

In order to model human-like hesitations in the synthesis output of spoken dialogue systems, we need to rely on corpus data that consists of spontaneous speech, has disfluency markup and phone-level annotation. For German, the available spontaneous speech corpora contain either disfluency markup or phone-level annotation. For lengthening analysis, we therefore used two existing corpora modified to fit our needs:

GECO is a large-scale German spontaneous speech corpus with phone-level annotation [9]. It is compiled to analyze convergence in speech, features free dialogues and has no disfluency annotation.

DUEL-Dreamapartment (henceforth: DUEL) is a smaller, more specialized corpus in which speakers collaboratively design the apartment of their dreams in highly engaged interaction. The interaction task is specifically designed to elicit spontaneous speech phenomena such as disfluencies and laughter [10]. DUEL contains disfluency markup, but no phone labels.

2.1.2. A detector based on phone duration

To help detecting hesitation lengthening in phonemically annotated corpora without disfluency markup, we created a semi-automatic search tool (henceforth: detector) [11][12]. It is essentially a simple classifier that calculates the z-normalized duration for every phone in a corpus and flags every phone with a duration exceeding a pre-set z-threshold of $z - score > 3$. The threshold is based on a previous study [11] that suggests the best balance between hits and false positives for a score of 3 or greater. The z-score was calculated per phone and speaker. The flagged output is then manually checked. False positives have to be sorted out and actual lengthenings have to be classified to be disfluent or non-disfluent (e.g. accentuation or phrase-final). In this study we provide a detailed evaluation of the detector, based on the DUEL corpus, cf. section 3. The original version of the detector was built using the GECO corpus, which is used for comparison in this study.

There are several possible levels when analyzing lengthening. [13] showed that it is possible to use word durations and their deviation from modeled expected duration to account for lengthening. The syllable would be another possible level for lengthening analysis, e.g. in [14], we investigated the possibility to predict segmental durations in a frame of disfluently lengthened syllables. We opt for a phone-based approach as we are interested in precise segmental durations for later use in speech synthesis.

2.1.3. Lengthening frequencies in the data

In a part of the GECO corpus, with approximately 22 hours of speech, the detector approach identified 750 instances of disfluent lengthening, corresponding to 0.57 instances per minute. The human-labeled instances of lengthening in the DUEL corpus added up to 114 in total in 4.5 hours, or 0.42 instances per minute.

As both corpora consist of spontaneous speech and the latter being one where speakers are highly involved in a collaborative task especially designed to elicit disfluencies, it is surprising that the DUEL corpus exhibits a lower occurrence rate of lengthenings. We would have expected the rates to be at least equal, if not even higher in the DUEL corpus.

Furthermore, given that lengthenings are supposed to be the third most frequent disfluency in spontaneous speech [3], the detected frequencies appear to be surprisingly low. We therefore hypothesize that both strategies of detecting disfluency lengthenings, i.e. human and detector-based have their shortcomings. The detector can only find instances above its z-threshold and humans can only find instances above a hearing threshold about which we do not know anything. We therefore evaluate human versus detector performance and provide insights on how to make both human and automatic lengthening detection more efficient.

2.2. Data preparation

In a first step, we automatically create a phonemic annotation for the DUEL corpus [10] using forced alignment software [15]. Then we use the detector described in section 2.1.2 to flag all phones with a z-score > 3. The resulting flags are then manually labeled for being disfluent or not. Our annotation decisions are based on a set of criteria taken from previous studies (e.g. [16][12]) to be indicators of hesitant lengthening, such as:

- Is hesitation perceivable from the utterance context?
- Is the phone followed by a silent or filled pause?
- Is the phone in an unaccented position?
- Is the phone in a function word?

The classification into disfluent and non-disfluent lengthening is straightforward. Agreement between two expert annotators was tested in [12] and reaches 98.8%. In this study we explore automatic decision-tree classification, confirming the robustness of these two categories, cf. 3.3. The lengthening labels created by human annotators are then re-checked with the same criteria, to account for the possibility that among these labels are instances that do not qualify as disfluent by our definition.

That way, we have two different annotations available for the DUEL corpus: One created by humans with the instruction to label on-the-fly any disfluency-related phenomenon they encounter, and one by humans assisted by the detector as a metaphorical magnifying glass highlighting candidates of hesitant lengthening. These can be compared in a subsequent evaluation.

We finally train a classifier on the resulting dataset to check the integrity of these distinctions (see section 3.3)

3. Results and evaluation

We performed fine-grained comparisons between the two sets of annotations in order to identify shortcomings or advantages of the various methods.

3.1. Human annotation versus semi-automatic detection

While the frequencies of other disfluency labels are constant within the DUEL corpus, lengthening labels are almost completely absent in the second half of the corpus, see tables 1 & 2. We thus limit the comparison of human versus detector annotations on the part containing lengthening labels.

The total number of disfluency-related lengthening found in the entire corpus, with human and semi-automatic detection combined is 431 in 4.5 hours of speech, or 1.6 per minute. As expected, this rate is higher than in the GECO corpus (0.57 per minute). The rate remains constant throughout all files, so that the anomaly in lengthening label frequency has to be ascribed to the annotators, probably as the result of fatigue or a change of the annotator. We consider the combined set of annotated and

Table 1: *Detected lengthening instances in the first half of the DUEL corpus.*

Type	Count	Percentage
Detector only	140	59.9
Human only	45	19.2
Detector+Human	49	20.9
Total	234	100.0

Table 2: *Detected lengthening instances in the second half of the DUEL corpus.*

Type	Count	Percentage
Detector only	191	96.4
Human only	6	3.6
Total	197	100.0

semi-automatically detected lengthenings as our ground truth for assessing precision and recall. It is important to note that the ground truth is an approximation as there might be lengthenings missed by the detector due to the z-threshold, and by human annotators due to the elusiveness of lengthening.

Tables 1 & 3 reveal two main findings: Human annotators miss more than half of the instances of disfluent lengthening, with a recall of 40%. The use of semi-automatic detection increases the recall to more than 80%. It comes at a cost, though, as precision drops from 82% in human annotation to 28.8% in semi-automatic detection. For a detailed analysis of the reasons for the low precision, see section 3.2.

Table 3: *Precision and recall*

Annotator	Precision	Recall
Detector	28.8	80.8
Human	82.0	40.1

3.2. Detector precision and false positives

Table 4: *False positives types*

Type	Count	Percentage
Forced-alignment error	655	58.2
Laughter	219	19.5
Accentuation	94	8.3
Phrase-final	82	7.3
Backchannel	69	6.1
Other	7	0.6
Total	1126	100.0

As summarized in Table 4, forced-alignment errors are responsible for more than half of the false positives reducing the precision of our semi-automatic approach. That means in turn that the detector is expected to perform well on corpora with manually corrected phonemic annotation, which is a valuable insight for future applications.

The second largest portion of false positives is due to laughter or laughed speech. These intervals add "noise" to the signal, making it impossible for forced-alignment tools to correctly identify phone boundaries. For this reason, some corpora, such as DUEL [10], feature laughter markup. This information can be used to pre-filter the data in future work to increase precision.

The remaining 20% of false positives are overhead that is avoidable if the corpus annotation allows for it. Lengthening is not only used in hesitation, but also in accentuation, phrase-finality and backchanneling. It is possible that features such as word class or pitch movement distinguish disfluent lengthening from accentuation. In corpora with utterance or speaker turn markup, it could be possible to identify and exclude phrase-final lengthening, which would, however, also exclude some instances that are disfluent *and* phrase-final. Backchannels, as "islands" of one speaker in the other speaker's turn can be detected and excluded if the corpus is annotated accordingly.

To sum up, there are ways to increase precision and reduce overhead, given the the corpus data has the required features:

- A corrected phonemic annotation could increase precision by up to 58%.
- Laughter markup that allows for pre-filtering: 19%
- Speaker turn markup to exclude backchannels: 6%

The remaining 15% lack of precision due to accentuation and phrase-finality can only be avoided if the data gathered so far is sufficient to train a classifier to perform the distinction between disfluent and non-disfluent lengthening, cf. 3.3.

3.3. Towards automatic classification

The automatic threshold-based lengthening detection described above is based on a single criterion: normalized phone duration. This simplistic procedure of automatic filtering thus does not account for a range of other criteria necessary for making the distinction between fluent and disfluent lengthenings. We therefore explore whether a classifier for disfluency-related lengthening detection can help us exploring more complex, context-driven patterns related to hesitation. As our current data set is rather limited, however, we focus on learning the distinction between disfluent and non-disfluent lengthenings above our duration threshold. Thus, instead of classifying all phones in a corpus into lengthenings and non-lengthenings, we are aiming for enhancing the threshold-based detector with some more fine-grained decision rules.

Data and Features: For training and testing, we consider the set of 603 phones that were flagged either by our detector or by human annotators (i.e. we do not consider instances resulting from forced alignment errors, but we do include non-disfluent lengthening, such as accentuation.). This set divides into 431 disfluent and 172 non-disfluent instances of lengthened phones, meaning that a simple majority baseline that always predicts disfluent lenghtenings would achieve an accuracy of 71%. We represent the phone instances using the following set of features:

- phone class (fricative, sonorant, diphthong, short vowel, vowel, plosive)
- phone position in word (initial, medium, final)
- phone position in syllable (onset, nucleus, coda)
- coarse-grained part-of-speech (function word, content word)
- pitch (pitch movement, no movement, no pitch)
- word is followed by filled pause (true, false)

Classifier: First, we train a logistic regression classifier[1] and evaluate using 5-fold cross-validation as the data set is fairly small. The classifier achieves an average accuracy of 0.73 (sd: 0.018) and an average F1-score of 0.82 (sd: 0.013). So it outperforms the majority baseline by a small margin, but this trend is not consistent among different splits of the data. In order to

[1]We use available classification libraries from `http://scikit-learn.org`.

```
if phone class = 'diphthong':
  if word position = 'final': True
  if word position = 'medial':
    if pitch = 'no movement': True
    if pitch = 'movement':
      if pre-filler = False: False
      if pre-filler = True: True

if phoneclass = 'fricative':
  if p.o.s. = 'content-word':
    if syl.position = 'coda':
      if word position = 'final': True
      if word position = 'medial': False
    if syl.position = 'onset': True
  if p.o.s. = 'function-word': True

if phoneclass = 'shortvowel':
  if p.o.s. = 'content-word':
    if pre-filler = False:
      if syl.position = 'coda': True
      if syl.position = 'nucleus': False
    if pre-filler = True: True
```

Figure 1: *Some automatically learned rules in the decision tree classifier distinguishing disfluent and non-disfluent lengthening, shown in pseudocode.*

be able to interpret the impact of particular features, we also train a decision tree classifier as it allows for easy inspection of the automatically learned decision rules. We train on 500 randomly sampled training instances and test on the remaining 103 instances. The decision tree classifier achieves an accuracy of 74% on this testset.

Figure 1 shows some decision rules automatically learned by the classifier. The top node of the decision tree is the feature "phone class", meaning that specific decision rules are learned for each type of phone in our data set. This suggests that the small gain in performance could be due to sparsity and that more instances of each phone class are needed to learn more robust decision patterns.

Decision rules: As shown in figure 1, in case of diphthongs, pitch influences the decision as expected: lengthenings are likely to be disfluent when the pitch contour is a plateau. For fricatives and short vowels, non-disfluent lengthenings occur more in content-words than in function words, and in both cases the feature interacts with syllable position. These are patterns that we expect to be scalable to other phone classes, but the dataset at hand is too small. For future work on fully automatic detection, more phones have to be labeled. It could be beneficial to lower the z-score threshold to a value that yields an equal number of disfluent and non-disfluent phones for better training of the decision tree. This, however, can only be done with a corpus with carefully checked phonemic annotation, in order to keep the overhead due to forced-alignment errors in a reasonable dimension.

3.4. Results summary and discussion

We show that the semi-automatic approach drastically increases recall of lengthening detection. Precision can be increased by using pre-filtering methods if the corpus data is structured accordingly. We can use the resulting dataset off a relatively small, but specialized corpus to model basic hesitation lengthenings for speech synthesis. It is possible to train a decision tree classifier on the resulting data set, that reflects the human classification into disfluent and non-disfluent, however, the improvement over the baseline is negligible. In order to improve automatic classification, larger datasets and possibly more features are desirable.

4. General discussion

Lengthening is a subtle phenomenon in speech that is interesting for speech synthesis in incremental dialogue systems, because it is difficult to perceive. It seems odd at first glance to put effort into synthesizing a phenomenon that listeners do not notice - however, hesitation is important for controlling microtiming in dialogue and being able to do so without producing a filled pause (of poor synthesis quality) seems crucial for creating conversational speech synthesis of acceptable quality.

Disfluencies such as filled pauses are known to provide helpful cues for the listener [2][6], but from a production side, this listener orientation and potential listener benefit may be secondary. In the incrementally ongoing language and speech production process, speakers can be expected to lengthen words in the articulatory buffer in order to facilitate a subtle, hardly perceivable remedy, and only if more remedy is needed will the speaker resort to overtly perceivable disfluencies such as noticable hesitations or filled pauses [12][17][13].

Modeling these hesitation strategies for dialogue systems requires an in-depth analysis of lengthening phenomena in corpora of spontaneous speech. In the light of human annotators' limitations in perceiving disfluency-related lengthenings make it inevitable to use machine-aided approaches for annotating disfluencies on a large scale and at a fine level of granularity. The method presented here is suitable to create a corpus upon which a first version of a speech synthesis that is able to produce human-like disfluencies can be modeled. In the future, it is desirable to compile a larger dataset to allow for more sophisticated synthesis modeling and disfluency classification.

5. Acknowledgements

This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

6. References

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

[2] H. Clark, "Speaking in time," *Speech Communication 36*, 2002.

[3] R. Eklund, "Prolongations: A dark horse in the disfluency stable," in *Disfluency in Spontaneous Speech (DiSS '01)*, M. G. Core, Ed., Edinburgh, Scotland, 2001, pp. 5–8. [Online]. Available: http://www.isca-speech.org/archive_open/archive_papers/diss_01/dis1_005.pdf

[4] S. Betz, S. Zarrieß, and P. Wagner, "Synthesized lengthening of function words - The fuzzy boundary between fluency and disfluency," in *Proceedings of the International Conference Fluency and Disfluency*, 2017.

[5] 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.

[6] M. Corley and R. Hartsuiker, "Hesitation in speech can... um... help a listener understand," in *Proceedings of the twenty-fifth meeting of the Cognitive Science Society*, 2003, pp. 276–281.

[7] R. Dall, M. Tomalin, and M. Wester, "Synthesising Filled Pauses: Representation and Datamixing," in *Proc. SSW9*, Cupertino, CA, USA, 2016.

[8] R. J. Lickley and E. G. Bard, "When can listeners detect disfluency in spontaneous speech?" *Language and speech*, vol. 41, no. 2, p. 203, Apr 01 1998.

[9] 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.

[10] J. Hough, Y. Tian, L. de Ruiter, S. Betz, D. Schlangen, and J. Ginzburg, "DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter," in *10th edition of the Language Resources and Evaluation Conference*, 2016.

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

[12] S. Betz, P. Wagner, and J. Vosse, "Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data," in *Phonetik und Phonologie 12*, 2016.

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

[14] S. Betz, J. Voße, and P. Wagner, "Phone Elasticity in Disfluent Contexts," in *Proceedings of 43. Jahrestagung für Akustik*, 2017.

[15] L. Schillingmann, B. Wrede, and E. Belke, "Aligntool." [Online]. Available: https://aiweb.techfak.uni-bielefeld.de/aligntool/aligntoolvm.7z

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

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