Synthesized lengthening of function words - The fuzzy boundary between fluency and disfluency

Simon Betz[1,2,3], Sina Zarrieß[2,3], Petra Wagner[1,3]
Bielefeld University, Bielefeld, Germany
[1]Phonetics and Phonology Workgroup [2]Dialogue Systems Group [3]CITEC*

I. Introduction & Background

As [1]'s model of speech production suggests, speakers sense upcoming difficulties and can correct them before uttering. A reasonable strategy to bridge resulting gaps is to prolong the words in the articulatory buffer [2]. This often buys enough time to correct the issue, resulting in standalone disfluent lengthening, after which fluency is resumed [3]. In case of more severe difficulties, the lengthening may be followed by other disfluencies such as silent or filled pauses or repetitions. Similar hesitation strategies might be useful in automatic speech production, e.g. for spoken dialogue systems that interact with human users and typically face a variety of challenges in natural language understanding and generation.

Lengthening is an ambivalent phenomenon in speech that seems to be located at the fuzzy boundary between fluency and disfluency. It regularly occurs before phrase boundaries [4][5] and besides constitutes a common hesitation disfluency. Some disfluencies consist of lengthening [3] only, and some lengthenings appear so subtle that they pass unnoticed [6][7].

We assume that these characteristics of lengthening make it a key component in spoken dialogue systems that are capable of producing disfluencies, as they enable to buy a variable amount of time whilst being unobtrusive to the listener [6]. It is not yet known, however, how much synthetic lengthening is acceptable and how lengthening influences the user's interaction with the system. To address these issues, this study tests the effects of step-wise increases of synthesized lengthening on user ratings and interaction speed.

II. Methods

We designed a perception test to evaluate sound quality of lengthening. This test is embedded in a simple game, in which users are asked by a synthetic voice to move around pentomino pieces on a computer screen (figure 1). The instructions follow a fixed order of `[<pick up a piece> <conjunction phrase> <move it onto another piece>]` (cf. sentences in example 1 with the conjunction phrase in boldface). After each stimulus, to proceed, participants have to click one of the four quality feedback buttons that constitute a 4-point MOS-scale.

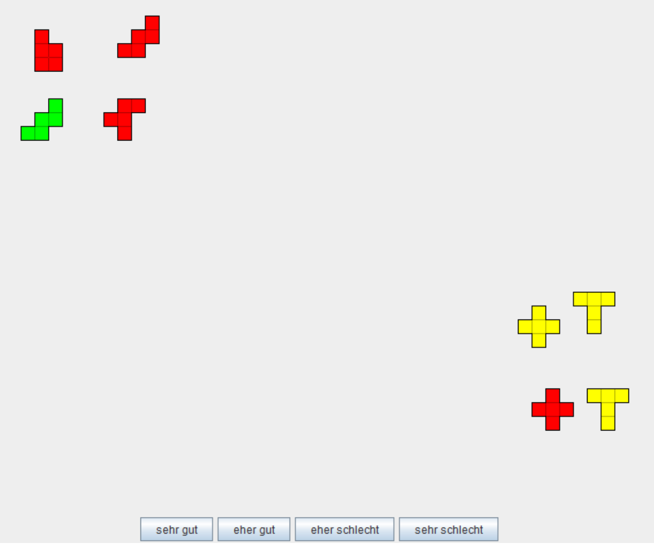

Fig. 1. Game scene with sound quality feedback buttons: very good, rather good, rather poor, very poor.

A. Stimulus design

Previous studies suggest that lengthening mainly occurs on function words [8][9], and that German articles, conjunctions and pronouns are frequent targets for lengthening [3]. For this study we test synthetic lengthening of function words in different degrees of lengthening with 400, 600, 800, 1000, 1200 and 1400 ms duration of the target word. The target words are German monosyllables *(der, die, das, und, dann, ihn)* selected because of their high frequency of occurence and syllable-type balancing.[1] The duration for each segment in the target words is determined by applying the duration model based on the elasticity hypothesis [10], means and standard deviations for each phone are extracted from the

*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).

[1] This balancing does not control for the inherently different duration of the words. E.g. the word *dann* in the 600ms condition might appear less stretched than the word *die* in the same condition. This, however had no effect on the results.

GECO corpus [11].[2] Each target word is embedded in a different carrier sentence and is located at the junction of two phrases that instruct the user to drag and drop pentomino pieces. The resulting six sentences (cf. Example 1) were synthesized in seven different configurations:

- The default configuration (i.e. with all segmental durations as predicted by the synthesizer's language model)
- The six different lengthening configurations (i.e. the same as the default, except that the target word's duration is set to 400, 600, ... 1400 ms.).

Example 1: Sentences (lengthened elements in boldface)
(1) *Nimm das rote Kreuz* ***und*** *lege es zum gelben Winkel.*
(2) *Die grüne Treppe,* ***die*** *muss rüber zum blauen Balken.*
(3) *Der gelbe Winkel,* ***der*** *muss rüber zum roten Balken.*
(4) *Das blaue Kreuz,* ***das*** *muss rüber zur grünen Treppe.*
(5) *Nimm die rote Treppe,* ***dann*** *lege sie zum gelben Kreuz.*
(6) *Nimm den grünen Balken* ***und*** *lege ihn zum blauen Winkel.*

In addition to the resulting 42 stimuli for analysis, we created 56 additional stimuli with different shapes and colors and without lengthening as distractors. Another six different stimuli were created for training the participants.

B. Stimulus presentation

Participants were instructed to act incrementally, i.e. start the task as soon as possible during the instruction and not wait until the voice has finished speaking. Each participant got the same set of 42 stimuli and 56 distractor sentences in a random order. Each session started with a short training phase to get participants used to the task.

C. Participants

23 participants took part in the experiment, all of them were students of Bielefeld University, between 19 and 37 years old (mean age 26.3). Six of the participants (26%) were male, 16 (73%) female and one of other gender. 20 (86%) had German as their mother tongue. 15 (66%) had previous experience with some kind of speech synthesis. None reported impairments of vision or hearing. The participants were paid 3€ for their effort. None of the above mentioned variables (gender, mother tongue, experience with synthesis) had any apparent influence on the results. One participant was excluded from the final analysis, because inspection of their data revealed that they did not proceed incrementally.

III. Results

Following suggestions by [12], we used R [13] with the lme4 package [14] to conduct a linear mixed effects analysis of the influence of lengthening extent on user ratings. As fixed effect, we had lengthening extent. As random effects we had intercepts for stimuli and participants, as well as by-stimulus and by-participant random slopes for the effect of lengthening extent, to control for ideosyncrasies of the participants and stimuli. Visual inspection of the residuals did not reveal any obvious deviations of homoscedasticity or normality.

We found that regardless of stimulus and participant, lengthening extent influences user ratings (t(743) = -6.855), each increment lowering the average rating score by about 0.18 ± 0.027 (standard errors), on a scale where 4 corresponds to the best and 1 to the worst rating.

In addition to the ratings, we measured relative task completion times and checked for influences of lengthening extent. To control for the different sentence lengths, we calculated the time span from beginning of audio until the drop of the pentomino piece divided by sentence duration. Using the same mixed models approach as above, we found that lengthening also significantly lowers relative task completion times (t(743) = -4.296), indicating that participants are not confused by the lengthening, but rather use the extra time to complete the task.

IV. Discussion

As can be seen in Fig. 2, stimuli get good overall feedback and the ratings decline very slowly as lengthening increases, reaching a sustained trough at 1200ms. On the one hand, this leads to the assumption that even relatively long lengthening is a valid strategy for spoken dialogue systems. On the other hand, it suggests that lengthening should ideally be kept low to maintain highest-possible quality. Analyses of the interaction speed support this assumption, cf. Fig. 3. Users use the extra time granted by lengthening to solve the task - they get faster relative to sentence duration as lengthening increases, but appear to get distracted by extreme lengthening, when they appear to slow down again (although the slowdown is not significant).
Even lacking any evidence for lengthenings $> 1200ms$, we take these as indicators for a turning point in synthesis quality around 1200ms: In terms of ratings, users do not differentiate anymore; in terms of task completion times, users need more time.

We furthermore suspect that lengthening is sometimes hard to notice due to its frequency of occurrence and its diversity of functions in everyday speech [6][7]. Summing up, our results raise the question as to the point at which lengthening characterizes a disfluency. In this experiment, we deliberately operationalized lengthening as a means to express hesitation, so it certainly counts as a disfluency from the production perspective. However, we still do not know the exact point (or the exact extent of lengthening) at which listeners start perceiving it as a disfluency. The slow and steady decline of our ratings suggests a fuzzy boundary rather than a clear threshold between “fluent” and “disfluent” lengthening.

V. Conclusion

We showed that synthesized lengthening gets good user feedback and does not negatively impact interaction speed.

[2] For this study, the ”strong” form of the elasticity hypothesis was applied, i.e. general mean durations were used. At the moment, we test the reliability of an elasticity model that is based on disfluent lengthening data only to predict segment durations.

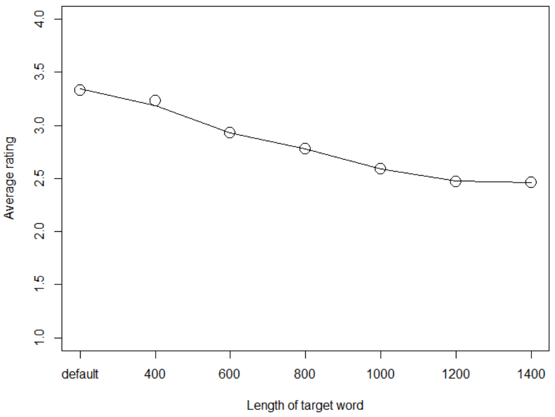

Fig. 2. User feedback with respect to word length. 4=good, 1=bad

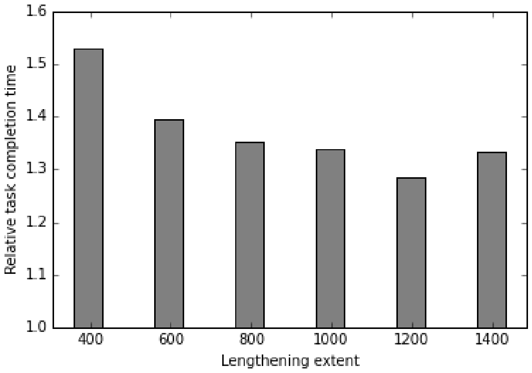

Fig. 3. Relative task completion time (divided by stimulus duration) over the different lengthening conditions

Although this study reveals more of a fuzzy boundary than a clear threshold in lengthening acceptability, ratings and interaction times in the conditions over 1000ms suggest that there is an upper limit to synthetic lengthening. Possible follow-ups could examine the impact of greater lengthening extents to determine whether there is a turning point around 1200ms or whether this is merely an outlier. Lengthening in general appears well suited for disfluency synthesis. It is to be determined if longer hesitations should be covered by lengthening over multiple words or with combinations with other disfluencies such as silent and filled pauses.

References

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

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

[3] 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.

[4] B. Peters, K. J. Kohler, and T. Wesener, "Phonetische Merkmale prosodischer Phrasierung in deutscher Spontansprache," *Prosodic structures in German spontaneous speech (AIPUK 35a)*, vol. 35a, pp. 143–184, 2005.

[5] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, oct 2007.

[6] 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.

[7] 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.

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

[9] 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.

[10] W. N. Campbell and S. D. Isard, "Segment durations in a syllable frame," *Journal of Phonetics*, vol. 19, no. 1, pp. 37–47, 1991.

[11] 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.

[12] B. Winter, *Linear models and linear mixed effects models in R with linguistic applications.*, 2013. [Online]. Available: http://arxiv.org/pdf/1308.5499.pdf

[13] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.