INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The greennn tree – lengthening position influences uncertainty perception

Simon Betz[1], *Sina Zarrieß*[1], *Éva Székely*[2], *Petra Wagner*[1]

[1]Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany
[2]Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
simon.betz@uni-bielefeld.de

Abstract

Synthetic speech can be used to express uncertainty in dialogue systems by means of hesitation. If a phrase like "Next to the green tree" is uttered in a hesitant way, that is, containing lengthening, silences, and fillers, the listener can infer that the speaker is not certain about the concepts referred to. However, we do not know anything about the referential domain of the uncertainty; if only a particular word in this sentence would be uttered hesitantly, e.g. "the greee:n tree", the listener could infer that the uncertainty refers to the color in the statement, but not to the object. In this study, we show that the domain of the uncertainty is controllable. We conducted an experiment in which color words in sentences like "search for the green tree" were lengthened in two different positions: word onsets or final consonants, and participants were asked to rate the uncertainty regarding color and object. The results show that initial lengthening is predominantly associated with uncertainty about the word itself, whereas final lengthening is primarily associated with the following object. These findings enable dialogue system developers to finely control the attitudinal display of uncertainty, adding nuances beyond the lexical content to message delivery.

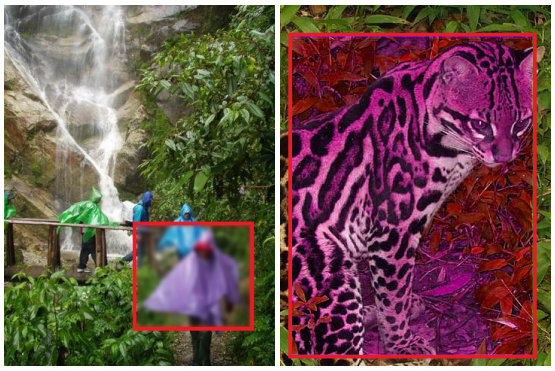
Index Terms: lengthening, hesitation, speech synthesis, attitudinal synthesis, uncertainty

1. Introduction

1.1. Uncertainty in dialogue (systems)

The ability to communicate uncertainty is an important aspect of grounding in human interaction [1], and is often considered essential for legibility and transparency in human-machine interaction as well [2, 3]. Recent work on conversational speech and synthesis investigated how various lexical and prosodic cues contribute to the perception of uncertainty. The presence of hesitations like filled pauses or lengthened syllables have been shown to be indicative of uncertainty in both English [4, 5] and German [6]. While acoustic features appear to be more important than lexical features [7], a limitation of studies investigating uncertainty in recordings of natural speech is that the lexical and acoustic parameters cannot be separated and their impact on the perception of uncertainty cannot be studied independently. Several methods of eliciting varying levels uncertainty in spontaneous speech have been proposed [8, 7]. Using synthesized speech allows for a more detailed investigation of the different factors influencing the perception of uncertainty [9, 10, 11]. [12] found that decreased vocal effort, use of filled pauses, and lengthening of function words increase the degree of perceived uncertainty of synthesised utterances.

Recent studies have proposed a relatively detailed approach to linguistic *cues* signaling uncertainty. However, the underlying notion of *uncertainty* itself, as an aspect of an utterance's meaning, remains surprisingly vague and is notoriously problematic to pin down and annotate [13]. In some studies, uncertainty is investigated as an emotional or cognitive state [13] related to the speaker's commitment or beliefs. Here, uncertainty typically applies to the propositional meaning of an utterance as a whole (i.e. a speaker is generally uncertain about the truth of an utterance) and can be treated as a scale (from perfectly certain to absolutely uncertain). Other studies have looked at interpretation of disfluencies and hesitations in situated dialogue, where utterances refer to objects in a visual environment. In these works, hesitations have been investigated as cues of formulation effort, e.g. when an object is difficult to describe [14] or is likely to be confused with a different object [15]. Here, uncertainty can be seen as referring to a specific domain, i.e. signalling a formulation problem for a particular semantic property of the referred target. For instance, in the seminal study by [14], hesitations are found to express uncertainty with respect to the category of the referent in visual scenes that contains easy-to-describe and difficult-to-describe objects.



(a) (b)

Figure 1: *Two images (shown as examples in the warm-up phases of our experiment) illustrating different domains of uncertainty: object category (a) and color (b).*

To date, surprisingly little is known as to how and whether the domain of the uncertainty can be varied by manipulating the realization of the respective linguistic cues. For realistic objects in a visual environment, varying properties could be considered as difficult to describe: e.g. its color, shape, its name or category, as illustrated in Figure 1. A dialogue system that describes these objects to a human listener might face these different types of uncertainty, cf. [16]. Hence, even in relatively short utterances describing these visual objects, various positions and realizations of uncertainty cues and various interpretations of the uncertainty domain are possible. Our aim is to show that these interpretations shift as a function of hesitation position in the speech signal.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2572

1.2. Lengthening premises

Lengthening in speech occurs for various reasons, among others, because of accentuation, upcoming phrase boundaries or hesitation. Depending on the underlying reason, the locus of the lengthening within the affected word might change: accentuation lengthening always occurs on the vowel nucleus of a word, whereas hesitation lengthening is less restricted in terms of occurrence, frequently manifesting itself in different syllable positions, such as the onset or coda [17]. Hesitation, as one reason for the occurrence of lengthening, can itself have several underlying reasons, such as changes in the dialogue situation, lexical retrieval issues, or uncertainty. In this study we test if placing lengthening in initial or final positions within a word influences the perception of the uncertainty.

1.3. Hypotheses

In this study we test two hypotheses:

1. Lengthening the onset and nucleus of a word is interpreted as uncertainty about the semantic property expressed by the word itself.
2. Lengthening the coda of a word is interpreted as uncertainty about the semantic property of the next content word in the utterance.

These hypotheses are driven by the desire to deploy hesitations and to control signaling uncertainty in dialogue systems. Previous studies revealed that hesitation lengthening predominantly manifests itself on function words such as conjunctions or determiners, but there might be a communicative desire to hesitate in different locations [18]. Whereas hesitating on a function word suggests that a possible uncertainty is likely to be centered on the next content word, it is unclear what effect hesitation on a content word itself has. In our work, we assume that it makes a difference *where in a given word* hesitation lengthening is placed.

2. Methods

2.1. Setup

The hypotheses are investigated by means of a perception test, conducted via the crowdsourcing platform PERCY [19]. Participants were recruited via social media and mailing lists. 64 volunteers from all parts of Germany, with a majority from the north-western part, participated in the experiment (24 male and 40 female; age range= 18–51, mean = 29.59, median = 25). The participants were required to be familiar with the German language, and to ensure they had a suitably quiet environment. Users self-reported the input and audio devices they used as well as their current environment and native language.

Participants were asked specifically to feedback on two freely adjustable sliders with no preset value how certain or uncertain they perceive the system to be regarding two different referential domains, in this case, color and object (cf. Figure 2).

This perception test was preceded by a page with instructions and a cover story about an artificial agent than can automatically describe images and is capable of expressing its own uncertainty. The participants were further told that the system had been given a set of images that were blurred or had modified coloration which are difficult to describe. For illustration, five blurred and five color-modified images were shown (cf. Figures 1).

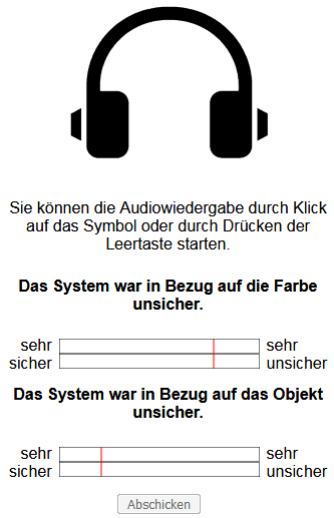


Figure 2: *Experiment interface. Engl.: "System was uncertain about color." "System was uncertain about object." Slider labels: "very certain" (left), "very uncertain" (right).*

Table 1: *Words used for the stimuli*

color	*trans.*	**object**	*trans.*	**distract.**	*trans.*
rot	*red*	Wagen	*car*	Kater	*tomcat*
blau	*blue*	Ball	*ball*	Kamel	*camel*
grün	*green*	Jacke	*jacket*	Auto	*car*
braun	*brown*	Schild	*sign*	Dose	*box*
schwarz	*black*	Buch	*book*	Pferd	*horse*
weiss	*white*	Tisch	*table*	Kette	*chain*

2.2. Stimuli

The stimuli presented to users were short German sentences following a template of "look for the <color> <object>". This grammatical template was chosen so that all color words would get the inflectional ending {-en}, regardless of the grammatical gender of the object word. This is useful for applying lengthening, which is frequently realized on nasal sounds like [n] [17]. This yields sentences as shown in example (1).

(1) Suche nach dem grünen Baum
(look for the green tree)

Six color words and six object words were used throughout the experiment and were synthesized in every possible combination for each of the three conditions: **baseline**, **initial lengthening** and **final lengthening**. The color words were selected to feature at least one synthetically prolongable sound before the vowel nucleus, i.e., a non-plosive sound. In addition to these $3x36$ stimuli, 36 distractor stimuli were constructed, six different object words were combined with the same color words, yielding $4x36 = 144$ stimuli in total. The words used are summarized in Table 1.

The stimuli were generated with Mary TTS [20] using a male German hsmm voice. For the baseline condition, no further modification was applied. For the lengthened conditions and distractors, the XML generated by Mary TTS was modified before generating the audio (cf. section 2.2.1).

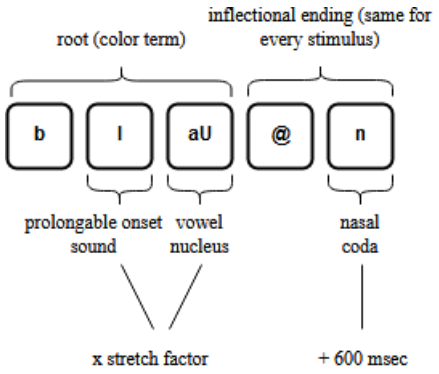


Figure 3: *Duration modification for initial and final lengthening exemplified on the color term "blauen" (blue + ending).*

2.2.1. Duration modifications for initial and final lengthening

For this experiment, our stimuli need to be lengthened enough to ensure that they can be perceived as expressing hesitations. At the same time, we have to ensure that the duration increase does not exceed 800 msec, as previous studies suggest that longer synthetic prolongations might have uncontrollable impacts on perception [21].

Another limiting factor is the phone itself. Different phones have different inherent elasticity [22, 23], which has direct implications for parametric speech synthesis: prolonging phones to a degree exceeding their elasticity may result in degraded and buzzy sound quality.

To apply duration modification based on phones' elasticity, we calculate a *stretch factor*, by which the duration predicted by Mary TTS can be multiplied to achieve *natural maxima*: for each phone, we obtain its mean duration and standard deviation from a corpus of spontaneous German speech [24]. Based on these, we predicted 10,000 concrete duration instances using the random.normal(mean,sdev) function of the python package numpy. Out of these, the highest value was selected and divided by the mean duration, yielding the *stretch factor*.

Initial lengthening is realized on several phones, in case of our stimuli on [r], [l], [v] respectively (due to the choice of words, cf. Table 1). We opt to split the lengthening between the vowel nucleus and the prolongable phone immediately preceding it. Multiplying both phones' duration with their stretch factor yields total duration increases of about 500 msec for the initial lengthening condition (cf. Fig. 3 for an illustration of lengthening distribution within words).

The *natural maximum* approach has also been tested for the final lengthening, which yielded unsatisfactory results, in the sense that the lengthening was barely perceivable. Final lengthening is a common phenomenon in speech and might thus require a higher amount of lengthening to be perceivable as hesitation, which is why we opted to only lengthen the coda by a fixed amount of 600 msec. Nasal sounds like [n] have a high inherent elasticity, so this procedure does not impact sound quality negatively.

2.2.2. Distractors

To prevent participants from learning patterns, 36 distractor stimuli were included with different object words and different hesitations. Instead of lengthening, they featured a filler ("ähm", (engl.: *uhm*)) embedded in 400 msec of silence, both before and after. These hesitation clusters were put before the color word in the one half and before the object word in the other half of cases.

2.2.3. Block design

To avoid fatigue in our volunteer participants in a tedious online task, we split the stimuli into three groups of 48 stimuli each, selected so that every word and condition is featured in every group. Each participant is then assigned randomly to one of these groups. This way the experiment takes about 10 minutes to complete for each participant.

2.3. Statistical analyses

We fitted two generalized linear mixed effects models with color and object uncertainty rating as the dependent variables each. The ratings on a 0-99 scale were log-transformed to satisfy normal distribution needs. As fixed factors, we used the type of lengthening split into contrasts between initial and no lengthening, and final and no lengthening respectively. As fixed effects, we also included *noise* (set to "yes" if participants did not use headphones or were in public, based on self-reports), *mobile* (set to "yes" if participants used a tablet or smartphone as input device) and native language (set to either "German" or "other"). Comparisons between full and reduced models indicated no influence of these.

As random factors, we considered participant, color word and object word. For both object and color rating we stepwisely excluded the random slopes and intercepts that accounted for the least variance and tested each different model against the previous one using ANOVAs until a difference became significant. The last model before reaching significance was considered the best fit.

3. Results

As can be seen in Fig. 4, there are clear differences in the way users perceive uncertainty. Initial lengthening is predominantly associated with color uncertainty, i.e., with the word it occurs in. Final lengthening is associated with both color and object uncertainty, i.e., with the word the lengthening is realized on, together with the following one. Both conditions are in contrast with the no-lengthening baseline condition, which is not associated with any uncertainty. The distractors with fillers appear clearly distributed into color uncertainty for fillers occurring before the color word and object uncertainty for fillers occurring before the object word. Both lengthening conditions, initial and final, differ significantly from the baseline for both color and object uncertainty (cf. Table 2).

Table 2: *Model output for fixed factors*

color uncertainty: initial vs. baseline $SE = 0.089$, $df = 47.37$, $t = -27$, $p < 0.001$
color uncertainty: final vs. baseline $SE = 0.138$, $df = 63.70$, $t = 9.78$, $p < 0.001$
object uncertainty: initial vs. baseline $SE = 0.086$, $df = 38.83$, $t = -5.3$, $p < 0.001$
object uncertainty: final vs. baseline $SE = 0.143$, $df = 65.28$, $t = 11.3$, $p < 0.001$

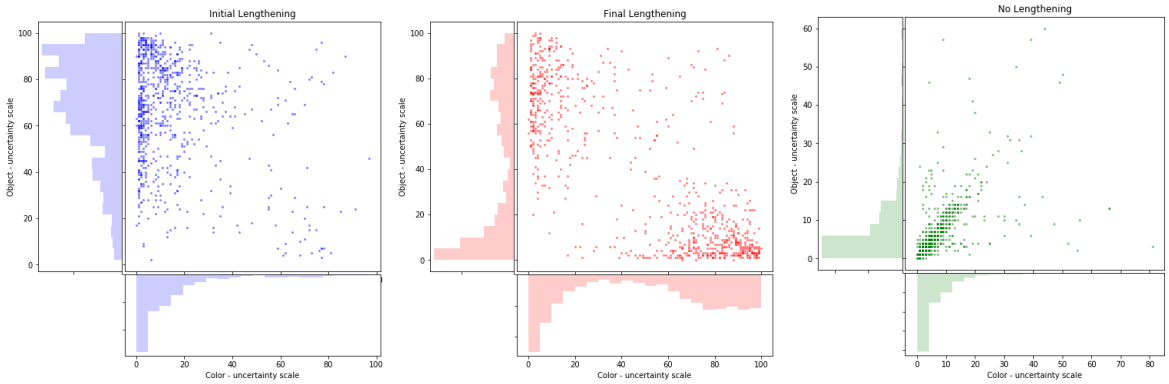


Figure 4: *Uncertainty ratings for the different conditions.*

4. Discussion

The results show that the domain of uncertainty does not entirely coincide with the location of its phonetic marking. There can be many possible aspects in an utterance that there is uncertainty about, and paralinguistic marking of uncertainty may or may not happen on the uncertainty reference itself.

Word-final lengthening has a wider scope, as it is interpreted as both signaling uncertainty about the word it occurs in as well as about the word that follows. It is possible that this effect arises due to the fact that final lengthening is a very frequent phenomenon at the crossroads of speech and language, commonly marking phrase-endings or hesitations, so that there may not always be a clear-cut way to infer its intended prosodic function, especially in spontaneous speech.

Compared to final lengthening, initial lengthening is quite rare in spontaneous interactions [25], and its communicative function seems to be strongly associated with the word it occurs on. This can be interpreted as follows: if speakers do not hesitate at the lesser marked, word final lengthening position, they communicate that the reason for hesitation is in the very word. If it was later in the utterance, speakers could resort to final lengthening or other hesitation markers such as silences or fillers between words.

Due to our limited experimental environment, and the circumstance that we only investigated the perception of *uncertainty*, we cannot conclude that word-initial lengthening is *limited* to the expression of uncertainty, and does not fulfill further communicative functions. However, we can confidently conclude that the expression of uncertainty is one of the functions of initial lengthening.

An exploratory examination of the distractors provides interesting additional insights: The uncertainty is regularly perceived to be centered on the word that *follows* the hesitation cluster of silence and filler, cf. Fig. 5. This points to a general hypothesis that the perception of uncertainty is associated with the first concept uttered *after any kind of hesitation*.

There are some limitations that are to be taken into account when interpreting the results. First, this experiment was conducted with German synthetic speech data, and it is subject of future studies to determine whether the findings are translatable to other languages.Second, this experiment only explicitly asks for opinions on uncertainty, so we do not know if further notions are conveyed by hesitation in this setting. Third, the audio stimuli are presented in a disembodied way, without visual stimuli containing the references that need to be resolved. The latter

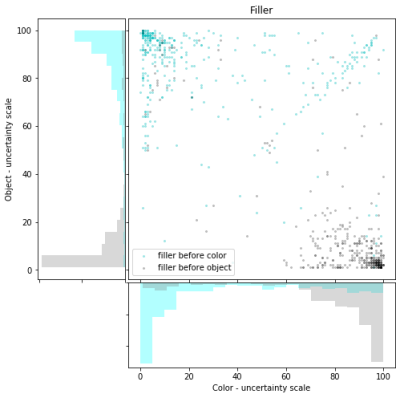


Figure 5: *Uncertainty ratings for distractor stimuli*

point, however, also yields interesting implications, namely that synthetic speech can be used to convey uncertainty without any other contextual cues present.

5. Conclusions

We can confirm the hypotheses underlying this experiment: Lengthening in initial position of a word is interpreted as uncertainty about the semantic domain represented by the word itself. Lengthening in final position within the word is interpreted as uncertainty regarding the semantic domain represented by the following content word. The latter is to be interpreted carefully though, as word-final lengthening appears to be associated with uncertainty in general, but it contrasts with initial lengthening which is clearly associated only with the word it occurs in. These findings have implications for future development of dialogue systems, which could tap on the potential of signaling uncertainty about certain dialogue topics. When designing future experiments on synthetic speech perception, it is fundamental to keep in mind that word-final lengthening is associated with a plethora of functions, while word-initial lengthening appears to be more narrow in scope.

6. Acknowledgements

Thanks to Annett Jorschick, Marin Schröer and Christoph Draxler for their valuable advice and support during this study!

7. References

[1] H. H. Clark, *Using Language*. Cambridge: Cambridge University Press, 1996.

[2] J. Hough and D. Schlangen, "It's not what you do, it's how you do it: Grounding uncertainty for a simple robot," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 274–282.

[3] C. Liu, R. Fang, and J. Y. Chai, "Towards mediating shared perceptual basis in situated dialogue," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2012, pp. 140–149.

[4] H. Pon-Barry and S. M. Shieber, "Recognizing uncertainty in speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 251753, 2011.

[5] V. L. Smith and H. H. Clark, "On the course of answering questions," *Journal of Memory and Language*, vol. 32, no. 1, 1993.

[6] T. Schrank and B. Schuppler, "Automatic detection of uncertainty in spontaneous german dialogue," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] G. Romigh, C. Rothwell, B. Greenwell, and M. Newman, "Modeling uncertainty in spontaneous speech: Lexical and acoustic features," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3401–3401, 2016.

[8] H. Pon-Barry, S. M. Shieber, and N. S. Longenbaugh, "Eliciting and annotating uncertainty in spoken language," 2014.

[9] E. Lasarcyk, C. Wollermann, B. Schröder, and U. Schade, "On the modelling of prosodic cues in synthetic speech–what are the effects on perceived uncertainty and naturalness?" in *Proc. of NLPCS*, 2013.

[10] A. Hönemann and P. Wagner, "Synthesizing Attitudes in German," in *Proc. of The Australasian International Conference on Speech Science and Techonology*, 2016.

[11] C. Lai, "What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue," in *Proc. Interspeech*, 2010.

[12] É. Székely, J. Mendelson, and J. Gustafson, "Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies." in *INTERSPEECH*, 2017, pp. 804–808.

[13] H. Pon-Barry, S. Shieber, and N. Longenbaugh, "Eliciting and annotating uncertainty in spoken language," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 1978–1983.

[14] J. E. Arnold, C. L. H. Kam, and M. K. Tanenhaus, "If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 33, no. 5, p. 914, 2007.

[15] S. E. Brennan and M. F. Schober, "How listeners compensate for disfluencies in spontaneous speech," *Journal of Memory and Language*, vol. 44, no. 2, pp. 274–296, 2001.

[16] S. Zarrieß and D. Schlangen, "Easy things first: Installments improve referring expression generation for objects in photographs," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 610–620. [Online]. Available: http://www.aclweb.org/anthology/P16-1058

[17] 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.

[18] S. Betz, B. Carlmeyer, P. Wagner, and B. Wrede, "Interactive hesitation synthesis: modelling and evaluation," *Multimodal Technologies and Interaction*, vol. 2, no. 1, p. 9, 2018.

[19] C. Draxler, "Online experiments with the percy software framework - experiences and some early results," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[20] M. Schroeder and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching." *International Journal of Speech Technology, 6:365-377.*, 2003.

[21] 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.

[22] S. Betz, J. Voße, and P. Wagner, "Phone elasticity in disfluent contexts," in *Tagungsband der 43. Jahrestagung fr Akustik*, march 2017, pp. 1462–1464.

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

[24] 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.

[25] S. Betz, R. Eklund, and P. Wagner, "Prolongation in German," in *Proceedings of DiSS 2017, Disfluency in Spontaneous Speech*, R. Eklund and R. Rose, Eds., vol. 58, no. 1, 2017, pp. 13–16.