DISFLUENT LENGTHENING IN SPONTANEOUS SPEECH

Simon Betz, Petra Wagner

Universität Bielefeld, Phonetics and Phonology Workgroup
simon.betz@uni-bielefeld.de

Abstract: We investigate lengthening in spontaneous speech with the aim in mind to use it as a time-management strategy in incremental spoken dialogue systems. lengthening is a common feature of speech, occurring regularly near the edges of intonation phrases. It behaves similar to disfluencies when it occurs in places remote from phrasal boundaries. Disfluencies have proven useful in incremental spoken dialogue systems that require real-time interaction with human users. Previous studies suggest that lengthening might be a low-cost way for dialogue systems to buy valuable time. This is to be examined in detail, based on a corpus study of lengthening behavior in spontaneous German speech. To aid the analyses, we explore new methods of automatic lengthening detection.

1 Introduction

1.1 Lengthening in fluent speech

Lengthening is a common feature of speech and is in its default form a cue for perceiving phrase boundaries [16][13]. A diverse and partly overlapping terminology is associated with this basic type of lengthening, e.g. phrase-final lengthening [17] [16], utterance-final lengthening [10], boundary-related lengthening [16] and prepausal lengthening [12]. The term prepausal was used to distinguish lengthening in spontaneous speech from that in read speech, because phrase-final lengthening was attributed to read speech only, e.g. [17], cited in [12]. This appears due to the fact that it used to refer to syntactic phrases only. More recent research tends to refer to intonation boundaries when using the term phrase-final lengthening. [13] analyze spontaneous German speech and identify final lengthening as one frequent phonetic cue for phrase boundaries and state that syntactic boundaries are not needed for prosodic boundaries, yet they can co-occur. [16] in their detailed study also focus on lengthening near the edges of intonation phrases rather than near syntactic phrase boundaries.

1.2 Disfluent Lengthening

Aside from the type described above, there is another form, namely disfluent lengthening, which we will refer to as ”standalone”. It describes a marked prolongation of one or more phones, resulting in above-average syllable and word duration, cf. [5]. This coincides with a local reduction in speech rate that is not expected by the listener, causing an impression of disfluency and hesitation. As such, it cues to the listener that the speaker is still formulating content and thus buys conversational time for the speaker by preventing barge-ins and maintaining a higher fluency compared to silent or filled pauses. We argue that lengthening is the first level of hesitation, the softest measure a speaker can apply to solve problems in speech planning.

Most interest in disfluent lengthening comes from conversational speech synthesis research. [8] is an early example, labeling filled pauses and lengthening in Japanese as hesitation phenomena

which are almost identical in function.[1] In line with findings on English and German, they note that these phenomena appear to occur at arbitrary places. [1] describe lengthening as a method of smoothly inserting filled pauses into unit-selection synthesis. Their basis is corpus data where filled pauses are regularly preceded by lengthening. They mention standalone hesitant lengthening which in contrast displays no regularities of occurence. In a previous study, [2] analyzed standalone lengthening and found it to be a rare element in spontaneous speech, that occurs abruptly with no prediction from speech rate, often limited to one syllable.

An example of general phonetic interest in timing of spontaneous speech, [12] observe different disfluent lengthening-like phenomena, defining fluent speech as containing no hesitation, which is made up of ”intrasentential pauses” and ”unusual elongation of words”. They list among the options a speaker has upon reaching a hesitation point ”abruptly slow[ing] down for 1 or 2 syllables (often followed by a pause)” and ”enter[ing] a mode of much slower speech for a few words (often containing pauses)”. They further observe frequent instances of lengthening which are not clearly perceivable as hesitation, but seem like thinking pauses which manifest preferably on function words.

To conclude, there is one form of lengthening which is a disfluency on its own, which we call ”standalone lengthening”. The interest in lengthening from a conversational speech synthesis perspective is not limited to this element, though. Lenghtening also occurs preceding filled pauses. It is assumed that this is due to the fact that filled pauses often create an intonation phrase boundary, which in turn coincides with phrase-final lengthening. To synthesize filled pauses, as [1] noted, lengthening has to be applied. More lengthening occurs in a weaker form on function words. These ”thinking pauses” are not per se disfluencies, but have to be considered when designing time-management strategies for conversational synthesis.

1.3 Lengthening Analysis

The basic way to analyze lengthening is based on perception. This requires the annotator or the researcher to label the entire corpus for instances of lengthening. This is a reasonable method when it is done on the fly, as suggested by the guidelines of [9], applied e.g. to [11]. This implies the shortcoming of this method: When there is a huge corpus that has not been enriched with lenghtening labels upon annotation, it is an unreasonanbly high effort to manually parse the entire corpus again. Another general shortcoming of the perception-based method is the elusiveness of the subject. If lengthening is as subtle as expected, then many instances of highly deviant phone duration will slip through the annotator’s nets. This is reflected by the findings of previous analyses by the authors of this study, who found that places labeled manually for lengthening are reliable, but infrequent [2]: Only 38 instances were found in 27 minutes of speech, or, put differently, with a rate of 1.4 per minute. The instances found are too few to be a foundation for thorough analysis, and it is suspected that it is not due to the data but to the fact that many interesting occurrences simply pass unnoticed. Depending on the structure of the corpus, more specialized methods can be employed. [7] worked with a corpus of Hood German which features an unusually frequent particle, which they used as a target word and applied duration measurement to it. [13] used a corpus where intonation phrase boundaries were labeled. These studies with specialized corpora yield more analyzable instances of lengthening compared to more general spontaneous speech corpora, but are less suited to finding disfluent lengthening.

[1] This is rather extreme from a Germanic-languages point of view as Japanese is one of the languages that can indeed lengthen a prepausal syllable to create a filler-like effect.

2 Motivation

2.1 Spotting standalone lengthening

Lengthening has been studied in great detail with respect to its occurrence near phrase boundaries, e.g. [16, 13]. It has also received attention in connection with other disfluencies, e.g. [1, 13]. The standalone form, a hesitation that manifests itself only by means of lengthening has gotten less detailed phonetic attention: [12] noted it to be an abrupt slowing down of speech, [1, 2] and [8] found it occuring at arbitrary places not predictable from data. We found this type of lengthening to be very rare when using corpora with lengthening labeled based on annotator's perception [2].
Large corpora are obviously necessary to find sufficient tokens of standalone lengthening. These tokens need to be spotted directly from data as it is suspected that annotators miss too many of them. It is yet unclear, whether this be done without the aid of assistive labels as in [13] or [7].

2.2 Buying time in conversational speech synthesis

Time is valuable in dialogue, and no less when related to spoken dialogue systems. Disfluencies are promising strategies for human-like timing in dialogue [15]. Especially lengthening appears an elegant element capable of facilitating micro-level timing adjustments that do not sound as awful as artificial filled pauses: We conducted preliminary studies to evaluate the quality of various disfluency elements included in hmm-based speech synthesis [3] with the following lengthening-related findings:

(a) Stimuli with lengthening get very good user feedback.
(b) Filled pauses are dispreferred, especially when not preceded by lengthening.
(c) The non-application of lengthening variation degrades stimulus quality.

With disfluencies being generally acceptable in dialogue systems, result **(a)** is a very promising hint that valuable time can be 'bought' with stable, or even increased, quality.

2.3 Aim of this study

This study provides basic research on disfluent lengthening. It can be seen as a continuation of [13] who described boundary-related lengthening in German spontaneous speech and also touched the subject of disfluencies with respect to boundaries. We use a new corpus of spontaneous German speech [14] that is phone-level segmented, but has no markup for lengthening.[2] We investigate if lengthening, especially the disfluent type that occurs remote from boundaries, can be detected directly from data. If this proves fruitful, conversational speech synthesis research gains valuable extra information on a highly relevant item.

3 Corpus Study

3.1 Data preparation

The data at hand is the GECO corpus of spontaneous German speech segmented on phone level [14]. As a first step, a data frame table is prepared for each speaker, with a row for each

[2]It is worthwhile noting that [13] used a corpus with strict separation of turns and an appointment-making scenario of students playing roles. The corpus in this study is entirely spontaneous with no turn restrictions.

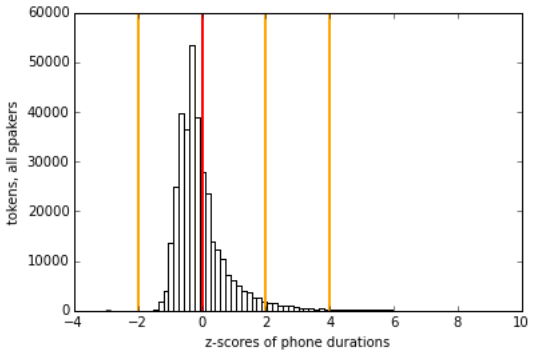

Figure 1 - Duration distribution of prolongable phones

Range	Tokens	Percentage
-1 to 1	622580	84.181
-2 to 2, without -1 to 1	85726	11.591
2 to 4	25115	3.395
greater than 4	6079	0.821
less than -2	71	0.009

Table 1 - Ranges and percentages of z-tokens

phone. For each phone, duration and z-normalized duration is calculated. As no utterance boundaries are available, stretches between two silences (IPUs = interpausal units) are marked as the unit of analysis. Filled pauses are also marked in the annotation, so the data frame can be enriched with some phrase-positional information: For each phone, the distance from silences, from IPU beginnings and from filled pauses is given.

We prepared a parallel dataset where all stops and [h] are excluded because lengthening cannot be realized on these sounds [13] and we report differences in results between these two sets, where applicable.[3]

Figure 1 shows the distribution of z-normalized phone duration of all speakers with only prolongable phones included. The graph with stops included looks exactly the same, just with more tokens. As can be seen, most tokens (95%) cluster in the area of z-scores between -2 and 2. Within these, most tokens (84%) are in the area of -1 to 1, which is the range of their standard deviation. The fact that many tokens have z-scores of up to 2 or -2 reflects the variation in spontaneous speech. It can further be seen that to the minus side there are limits, reflecting the natural limits of shortening, i.e. phones cannot be shorter than a certain minimal duration, whereas extreme lengthening is always possible. The interesting areas for lengthening are between 2 and 4, where still 3.4% of the tokens lie and which are clearly lengthened, and the area of z-scores greater than 4, where we expect errors and extreme lengthening. This demands close-ups in the analysis:

(1) What causes the z-scores between 2 and 4?
(2) What causes the outliers of z-scores grater than 4?
(3) Can pure z-scores tell anything about disfluent lengthening?

3.2 What information can be gained from z-scores only?

Several open-source tools were used to navigate to places with z-scores in interesting ranges for visual inspection [6, 4, 11]. We generated Praat TextGrids with intervals for those phones that have z-scores in the range to examine.

First insights are:

[3]Traditionally, only stops are seen as not prolongable. We also excluded the [h] for two reasons: 1) On first inspection, lots of errors in the data were caused by forced alignment errors on this phone. 2) This phone is not what we want to lengthen for well-sounding synthesis.

(1) The z-scores between 2 and 4 are almost always caused by phrase-final lengthening. Disfluent lengthening makes up only a small part of this.
(2) Most z-scores greater than 4 are due to force-alignment errors in the original annotation, though not exclusively. There are only a few cases of disfluent lengthening per speaker with z-scores this high.
(3) Z-scores alone cannot detect disfluent lengthening. It is suspected that disfluent lengthening is in the same durational range as phrase-final lengthening, but far less frequent. Thus, an estimated 95 per cent of lengthening flagged for z-deviation is attributable to boundaries.

In order to detect disfluent lengthening, we need to apply more filters to the data. In the following section, we redo the analysis but exclude utterance-final, utterance-initial and filled-pause preceding phones. Moreover, we address general questions of interest to lengthening, now that the data is prepared. We address the following questions:

(4) Phrase-final lengthening: Can we replicate the findings of other studies with entirely spontaneous German speech? In terms of spread of the lengthening, where does it occur, how many phones does it affect, does it make a difference if unprolongables are excluded?
(5) Pre-Filler lengthening: If we redo the analysis of (4), do we get the same results? Is the lengthening observed before filled pauses the same as phrase-final lengthening?
(6) Disfluent lengthening: Can it be detected with more sophisticated filtering?

3.3 Lengthening preceding boundaries

3.3.1 Phrase-final lengthening

Phrase-final lengthening can be confirmed with entirely spontaneous German speech. We assume each silence in the corpus to be a cue for phrase-final lengthening, since it marks a intonation phrase boundary, or a syntactic boundary, or often, both. For each silence in the corpus that has at least 8 preceding phones, we analyzed phone duration as the speech production approaches the boundary. As can be seen in Figure 2, the last phone preceding a boundary is clearly longer than the preceding one, between all other phones, maximally a slight increase is notable.
In general, it makes no difference for the analysis whether prolongable phones are excluded or not. Excluding them yields two minor advantages: The picture is a little clearer. As can be seen in Figure 2, there are minimal ups and downs between the phones further away from the boundary. The other advantage is related to the goal of this analysis: We do not want to synthesize lengthened stops, so we ease the analysis by ignoring them.
Earlier studies examined the range of phrase-final lengthening with regard to stress patterns and syllable position [16]. The findings in this study suggest that the last two prolongable phones might often correspond to the same range and could serve as a simplified model for synthesizing phrase-final lengthening.

3.3.2 Lengthening preceding filled pauses

[1] observed filled pauses in their data to be regularly preceded by lengthening. We hypothesize that this has the same reasons as phrase-final lengthening, as filled pauses can be understood as marking intonational boundaries. Tests with our data support this hypothesis.
In the corpus at hand, fillers are labeled on the syllable level. The preceding phones could thus be inspected easily. It is, however, not certain that this approach for really covers all lengthening

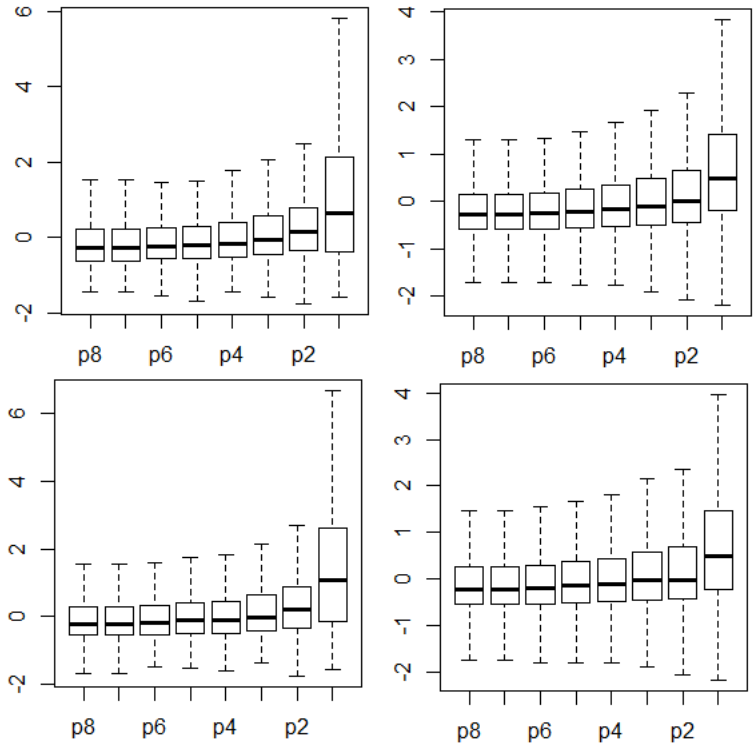

Figure 2 - z-normalized duration increase before boundaries. p8 is 8 phones before, p1 is last phone before boundary. Top: Phrase-final lengthening. Bottom: Pre-filler lengthening. Left: All phones. Right: Stops cut.

in front of fillers, depending on how fillers are defined. Low-content words and some lexemes in German can serve the same purpose and might thus be preceded by lengthening as well.

Our data suggests that lengthening before filled pauses is the same as phrase-final lengthening. This comes not as a surprise, since fillers, disfluent in their nature, cause the utterance to be split into multiple intonation phrases. As can be seen in figure 2, the picture is essentially the same for pre-filler lengthening as for phrase-final lengthening. The range of spread and the behavior of prolongable and unprolongable phones also exhibit the same similarity.

3.3.3 Statistics

For phrase-final and filler-preceding lengthening, we subtracted the column with the penultimate phones, p2 from p1, p3 from p2 etc., to get new columns with measures of increase. We then t-tested the resulting new columns in a pairwise fashion. Since the phrase-final dataset was much bigger than the pre-filler one, we performed the tests on 1018 randomly selected rows from the phrase-final set, so that set sizes were matched. As Figure 2 suggests, the increase to the last phone differs significantly ($p < 0.001$) from the increase to the penultimate phone for all sets, both with stops in and stops cut. For lengthening before fillers there is borderline significance between the increases to the antepenult and the penultimate ($p = 0.04$ with stops, $p = 0.07$ without stops).

3.4 Lengthening in utterance-medial position

For later inclusion into speech synthesis systems, we want information on hesitant lengthening without boundary relation. Since z-scores alone point mostly to pre-pausal lengthenings, we

Disfluency	13
Accentuation / Focus	7
Boundary misplacement	19
Preceding filler or silence	8

Table 2 - Lengthening types & totals in flagged phones

hypothesize that these places of interest can be found with a combination of z-deviation and utterance position remote from pauses.

Finding utterance-medial phones is not trivial with the data at hand, since it does not provide information on utterance boundaries. IPUs may consist of very short feedback signals and very long multi-utterance stretches. We thus enriched the data with indices of distance to silences, fillers, and beginnings of IPUs. We then generated textgrids that have labels for phones with a z-score of more than 3 and are at least two phones away from beginnings, silences and fillers. The distance of two is due to the spread range of lengthening as described in previous sections.

This method yields promising results. A considerable percentage of the flagged phones are the wanted standalone disfluent lengthenings, or rather, are ones that behave in a way that is useful for disfluency synthesis. A thorough analysis of the detection quality is still in progress. Some problems in defining the success rate are discussed below.

4 Evaluation

4.1 Hits

As a first evaluation, the highlighted phones were examined manually in the first audio file of the corpus. They were labeled on an extra tier with lengthening type classification. The main types are summarized in Table 2

An unexpected finding is that overlap between labels is possible. For example there can be errors in the boundary position causing a wrong z-score, the phone nevertheless carrying disfluent lengthening. Or there might be a long phone before an unlabeled intonation phrase boundary which appears to be accentuated for narrative purposes and simultaneously carrying a notion of hesitation.

With a z-score minimum set to 3, 40 phones are marked by the detector. Table 2 gives the count for each type.[4] In the z=3 condition, 13 phones or 32,5% contain a hesitant lengthening. In addition, there are 7 phones, or 17,5% where the lengthening is not hesitant but rather due to focus and accentuation, often in the context of re-telling experience or speech. This could be interesting as well, as it is markedly long speech sounds that still sound natural.

4.2 False Positives

The method will inevitably output false positives as well. The goal is to find a viable balance between hits and false positives, a balance that is expected to shift upon modifying the z-score minimum for analysis. In our sample, 50% percent were false positives, i.e. labels that are not useful at all for our purposes. The greater part of these false positives are boundary position errors. We expect a drastic increase in false positives due to phrase-final lengthening when we reduce the z-minimum, though it is up for investigation where the threshold is.

A task for the future is the search for the z-score that finds the best balance between hits and false positives. With the minimum set to 3 , the detector finds roughly the same number of

[4]Note that the total exceeds 40 due to the possible overlaps.

Z-Minimum:	3.0	2.8	2.6	2.4	2.2	2.0	1.8	1.6	1.4	1.2
Flags:	40	47	52	67	88	111	148	199	274	358

Table 3 - Number of flags created per z-score minimum threshold

hits per minute of speech as human annotators found in another corpus [11]. Maybe this is the threshold that corresponds to ”clearly perceivable lengthening”. As a first outlook, we checked how many phones are flagged when we set the z-score minimum lower (cf. Table 3). For reference, we checked the types of a version with a 2.4 threshold. This yields 67 instead of 40 tokens, but only 4 of the 27 new ones are disfluencies. In addition, 8 of the new ones are accentuation lengthening, which could potentially be interesting. From this first inspection, a 3.0 threshold appears to yield the best balance between hits and false positives, but this has to be confirmed on the other speakers' data.

4.3 Misses

As of now, misses cannot be counted for this corpus because it is not labeled for lengthening. We tried to re-label parts of it with annotators primed to mark lengthening, which caused them to find much more instances than expected from other corpora e.g. [11], with very low agreement. There are only two options of testing this: Either the entire GECO corpus has to be re-labeled by annotators with an instruction manual like [9], which does not focus the attention on lengthening, or the detector has to be tested on corpora that feature labeling. The first option is way too much effort. The point of the detector is to automate tasks exactly like these. The second option yields the problem that there are, to the knowledge of the authors, no corpora of spontaneous speech with lengthening labels that are phone-level segmented, so extra work would also be needed.

4.4 Discussion

Considering that we were dealing with a corpus with about 20.000 phrase-final lengthenings that could distract the analysis, this result is not bad at all. The total number (20) of false positives is comparatively small, in a dimension that is easily manually inspectable, yielding the same total number of points of interest per 30 minutes speech, which is roughly the same compared to [11].

The misses-issue is maybe not too problematic. This method implies a trade-off between fine-grainedness and false positives in the output. Using a lower z-score for filtering is assumed to drastically decrease misses, alongside a drastic increase in false positives. Since false positives are the factor that can really slow down post-processing, we can rather accept a high miss rate as long as a satisfactory number of data points for synthesis is yielded.

5 Conclusion

We found a way to ease analysis of disfluent lengthening. While the method is not fully automatic, it is a helpful filtering, highlighting places of interest among thousands of distractors. We do not claim that all of these places are genuine lengthening, as one might argue that it is not lengthening when it is not perceived as marked. The methodology presented here provides a foundation for further analysis of lengthening and its semantic reasons. For the time being, what counts for us is the fact that we can find phones that are unusually long but do not sound unnatural. Implementing this enables conversational speech synthesis to buy time at a discount - valuable extra seconds without a decrease in quality.

References

[1] ADELL, J., 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.

[2] BETZ, S., 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)*, pp. 2222–2226.

[3] BETZ, S., P. WAGNER and D. SCHLANGEN: *Modular Synthesis of Disfluencies for Conversational Speech Systems*. Proceedings of ESSV, 2015.

[4] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer [Computer program]. http://www.praat.org/*. 2014.

[5] BRUGOS, A. and S. SHATTUCK-HUFNAGEL: *A proposal for labelling prosodic disfluencies in ToBI*. In *Poster presented at Advancing Prosodic Transcription for Spoken Language Science and Technology*, 2012.

[6] BUSCHMEIER, H. and M. WLODARCZAK: *TextGridTools: A TextGrid Processing and Analysis Toolkit for Python*. Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013), pp. 152–157, 2013.

[7] FUCHS, S., J. KRIVOKAPIC and S. JANNEDY: *Prosodic boundaries in German: Final lengthening in spontaneous speech.*. The Journal of the Acoustical Society of America, 127(3):1851–1851, 2010.

[8] GOTO, M., K. ITOU and S. HAYAMIZU: *A real-time filled pause detection system for spontaneous speech recognition.*. In *Eurospeech*, 1999.

[9] HOUGH, J., L. DE RUITER, S. BETZ and D. SCHLANGEN: *Disfluency and Laughter Annotation in a Light-weight Dialogue Mark-up Protocol*, 2015.

[10] KOHLER, K. J.: *Prosodic boundary signals in German*. Phonetica, 40(2):89–134, 1983.

[11] KOUSIDIS, S., T. PFEIFFER and D. SCHLANGEN: *MINT.tools: Tools and Adaptors Supporting Acquisition, Annotation and Analysis of Multimodal Corpora*. In *Proceedings of Interspeech*, 2013.

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

[13] PETERS, B., K. J. KOHLER and T. WESENER: *Phonetische Merkmale prosodischer Phrasierung in deutscher Spontansprache*. Prosodic structures in German spontaneous speech (AIPUK 35a), 35a:143–184, 2005.

[14] SCHWEITZER, A. 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.

[15] SKANTZE, G. and A. HJALMARSSON: *Towards incremental speech generation in conversational systems*. Computer Speech and Language 27, 2013.

[16] TURK, A. E. and S. SHATTUCK-HUFNAGEL: *Multiple targets of phrase-final lengthening in American English words*. Journal of Phonetics, 35(4):445–472, oct 2007.

[17] UMEDA, N.: *Consonant duration in American English*. The Journal of the Acoustical Society of America, 61(3):846–858, 1977.