Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis

Simon Betz[1] [2], *Petra Wagner*[1], *David Schlangen*[2],

[1]Bielefeld University, Phonetics and Phonology Workgroup
[2]Bielefeld University, Dialogue Systems Group
simon.betz@uni-bielefeld.de

Abstract

Incremental dialogue systems can produce fast responses and can interact in a human-like fashion. However, these systems occasionally produce erroneous material or run out of things to say. Humans in such situations use disfluencies to remedy their ongoing production and signal this to the listener. We devised a new model for inserting disfluencies into synthesis and evaluated this approach in a perception test. It showed that lengthenings and silent pauses can be built for speech synthesis with low effort and high output quality. Synthesized word fragments and filled pauses, while potentially useful in incremental dialogue systems, appear more difficult to handle for listeners. While we were able to get consistently high ratings for certain types of disfluencies, the need for more basic research on their micro structure became apparent in order to be able to synthesize the fine phonetic detail of disfluencies. For this, we analysed corpus data with regard to distributional and durational aspects of lengthenings, word fragments and pauses. Based on these natural speaking strategies, we explored further to what extent speech can be delayed using disfluency strategies, and how to handle difficult disfluency elements by determining the appropriate amount of durational variation applicable.

Index Terms: speech synthesis, disfluencies, spontaneous speech, dialogue systems, incrementality

1. Introduction

Disfluencies have been studied in depth since the seminal works of [1] and [2]. They occur frequently in everyday speech, about 6% of words uttered contain a disfluency [3]. The term covers any kind of deviation from the ideal, perceivable as hesitation and expressable in the proposed structure of reparandum, editing phase and repair. It is a structural phenomenon that brings along phonetic correlates, which we will call *disfluency elements* to distinguish them from the general term and to imply that a structural disfluency acoustically consists of smaller parts. They are no longer viewed as speech errors but as solutions to errors in speech planning. As communicative cues they signal troubles in delivery to the listener and aid in comprehending the intended grammatic and semantic structure. Disfluencies follow a predictable pattern of erroneous material to be revoked, an editing phase and a repair phase delivering the intended content. In general, two kinds of disfluencies can be distinguished: forward looking (FLD) and backward looking (BLD) ones, the first type anticipating trouble and delaying the production of speech, the latter one detecting an already-uttered error and interrupting in order to repair [4]. This structural predictability is advantageous for applications such as speech synthesis [5], [6] or speech recognition [7], where they grew increasingly popular recently. Incremental systems, that produce speech while receiving input [6], are prone to errors or to run out of things to say. Disfluencies nicely fit these articulatory niches, bridging the gap in a human-like way.

Most research on disfluencies has focused on the macro-level, like [1] and [2], aiming to describe their general structure. This study digs a little deeper, into the underlying micro-structure and into the phonetic detail, providing insights about the basics required for conversational speech synthesis. The micro-structure can be understood as the phonology and phonotactics of disfluencies, addressing the questions: What elements does a disfluency consist of and how can these elements be syntactically combined? In terms of phonetic detail, we analyze durational properties in the micro-structure. These investigations are crucial for a better understanding of the finer features of disfluencies and their acoustical modeling for later use in conversational speech synthesis systems. In order to provide a framework for description and analysis of the micro-structure of disfluencies, we devised a modular model for synthetic disfluencies, based on [2]. We tested this model in a perception test where we asked for user feedback on sound quality of disfluencies (cf. [8] for preliminary work). Taking that as a starting point, we investigated real-world disfluencies in a corpus of spontaneous German speech [9] to refine our model and provide basic insights for constructing conversational speech systems.

2. LFP: A Simplified Modular Model for the Micro-Structure of Disfluencies

Consider an utterance that could occur in real speech, as in the following example:

(1) I will go tomorrow

upon production, the speech plan might change, yielding:

(2) I will go tom- {F uhm} on tuesday

In terms of macro-structure, this utterance can be segmented into reparandum, the erroneous (tom-) that is to be replaced, followed by the optional editing phase, here occupied by the filler {F uhm}, and the repair, the material that replaces the reparandum, in this case “on tuesday” [1], [2]. Our micro-structural approach is interested in a simplified acoustic realisation of the macro-structure. Imagine an incremental system was uttering (1). Then, while producing the output, it is informed that the speech plan has changed and that tomorrow is no longer the right target. Two scenarios, here exemplified in pseudo-code, are possible:

```
(3) Scenario FLD - This information arrives timely
    - Slow down production by lengthening
    - If no new speech plan is available, add silent pause
    - If still no new plan, insert filler
```

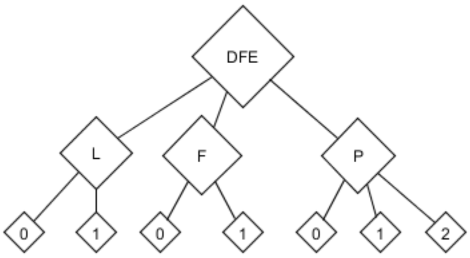

Figure 1: *Disfluency Elements (L)engthening, (F)ragment and (P)ause, and their parameters: 0=absent, 1=present, 2=with filler*

(4) Scenario BLD - The information arrives late
- interrupt immediately (for example at "tom-")
- if no new plan is available, pause, add filler...

Either scenario buys some valuable miliseconds, maybe entire seconds, to let generation deliver new content. We hypothesize that these three micro-structural elements (and combinations thereof)

- Pre-disfluent syllable **(L)**engthening,
- Cutoffs leading to word **(F)**ragments
- Silent or filled **(P)**auses.

provide synthesis systems with a variety of options to be disfluent in a well-sounding way. We tested this hypothesis in a perception experiment [8]. For that, we built a simple modular model which can be used to insert a disfluency into any given utterance, following a fixed order of L first, F second, P third. Each element could be present or absent, P could also get the parameter *with filler*, see Figure 1 for reference.

The fixed order of LFP allows for evaluating each element on its own in the perception test. We do not attempt to cover the entire range of shapes disfluencies can take in real speech. The LFP model is a simple framework for constructing variable disfluencies on the one hand, and for evaluation of particular disfluency elements on the other.

2.1. Experimental Setup

As stimulus material, we took four disfluent utterances from a spontaneous speech corpus [9] that featured a word fragment and no other disfluency elements. Consider the following example:

(5) Dann ma- lassen wir mal die Einzelheiten einfach weg
Then jus- let's just leave out the details

In terms of LFP structure, this example would be coded as L0 F1 P0, or in short notation 010, meaning that it contains no lenghtening, a word fragment and no pause. Cutting out the fragment would yield the fluent utterance "Dann lassen wir mal die Einzelheiten einfach weg". Using the LFP model, we can generate 12 different configurations of each stimulus, from the fluent one (000) to a very disfluent one that features all elements possible (112). The next example shows the configuration 102, with lengthening (Dann:), without fragment and with a filled pause({F ähm}). Synthesis was performed using Mary TTS [10]. In order to satisfy the needs of incremental synthesis, we have to use a hmm voice, despite using a unit-selection voice would yield higher sound quality.

(6) Dann: {F ähm} lassen wir mal die Einzelheiten einfach weg
Then: {F uhm} let's just leave out the details

The resulting stimuli were presented to participants via the Praat MFC environment [11] in random order with one repetition. First they had to finish a training phase in which 24 random stimuli were presented in order for the participants to calibrate their expectations. The training stimuli were re-used later, the results of the training phase were not included in the analysis. In the test phase, they were asked to assign an intuitive overall quality feedback on a 1 to 5 MOS scale that was presented on a screen. Responses of 32 participants, each of whom rated 96 Stimuli, were collected and analyzed with regard to the influence of the individual disfluency elements on the responses.

2.2. Results Summary

We conducted an ANOVA which showed that, in general, lengthening does not influence user feedback significantly (F(1) = 0.009, p = 0.923), but Fragment (F(1) = 13.37, $p < 0.001$) and Pause (F(2) = 46.74, $p < 0.001$) do. Post-hoc analyses revealed that stimuli with fragments fared significantly worse than those without, and that the same holds true for stimuli with filled pauses compared to those with silent or no pauses (cf. [8] for more details). An interesting tendency that emerged was, that some disfluent configurations (featuring silent pauses and lengthenings) get slightly, yet not significantly, higher mean results than the fluent baseline, which contradicts our expectation that disfluent stimuli should fare worse than fluent ones.

2.3. Discussion and Conclusions

The experiment served as an orientation in terms of the capabilities of disfluencies in incremental speech synthesis. We drew the following conclusions, linked closely to research questions to be adressed to the corpus, which will be taken up in the discussion section:

1. Lengthenings and silent pauses are unobtrusive in synthesized speech. To what extent can their duration be increased without becoming detrimental for perceived quality? What is their durational variability in real-world spontaneous speech?
2. Fragments are dispreferred. We hypothesize that this is not only the case with synthetic speech, but also in human speech. However, fragments due to word cutoffs provide a great flexibility for facilitating self-repairs. Can we use insights from the corpus study to be able to produce cutoffs with reasonable quality when no other option is available?
3. Filled pauses are detrimental for synthesis quality. They do however serve an important role in dialogue, namely the prevention of barge-ins. How can we improve their quality to be able to tap on their potential for our application? Can we learn from the corpus how they behave in reality?

3. Disfluency Elements in a Corpus of Spontaneous German Speech

The empirical research is based on the *Traumappartment* corpus [9], which consists of nine dyadic conversations in which the speakers imagine and describe and collaboratively plan the appartment of their dreams. Two dialogues totalling 27 minutes

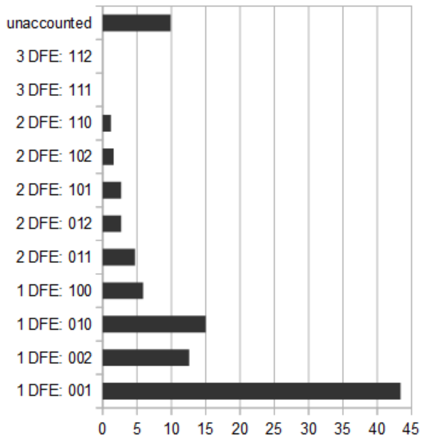

Figure 2: *Frequency percentage of disfluency element configurations in the corpus. Total number of configurations: 253.*

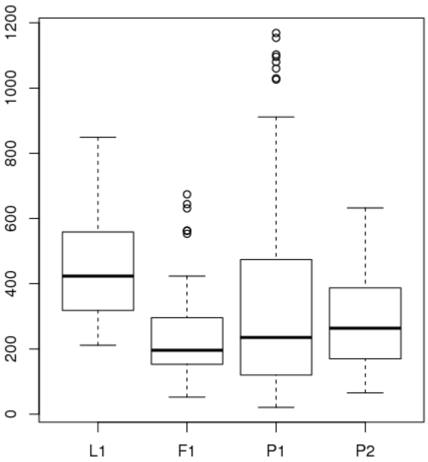

Figure 3: *Durational distribution of disfluency elements in ms.*

of speech were used for this study. Disfluencies that can be expressed via the LFP model cover about 3% of the total dialogue duration. If number of words correlates to duration, we can assume that, based on [3], about half of the speech perceived as disfluent can be expressed via a combination of the three disfluency elements LFP. Disfluencies not expressable in this framework are mostly repairs that are disfluent in structure, but are facilitated without the aid of hesitations to delay, or cutoffs to rescue production.

3.1. Frequencies and Distributions

Figure 2 shows the frequency of each possible combination of disfluency elements. In addition, the frequency of combinations our model cannot account for is given. In total, 253 tokens featuring one or more disfluency elements were counted. As can be seen, there are about 10% of configurations that can not be expressed within LFP. These either contain more than three elements or elements in different order than LFP. A very frequent combination within these (40%) is a double pause, i.e. a filled pause with adjacent silence.

Most frequent among the configurations that can be expressed are those that feature only one disfluency element. Standalone silent pauses (001) account for almost half of the total number of configurations (43.4%). The other configurations that feature exactly one disfluency element add up to another 35.5% of configurations. 13% of the configurations in the corpus consisted of exactly two elements. Standalone disfluency elements which serve as a forward-looking hesitation appear to be the most important ones in spontaneous speech. The production of clustered configurations is less frequent.

We checked if duration of elements in a cluster differed from duration of standalone elements. A linear regression analysis found no such interaction (F(1,318) = 1.921, R^2 = 0.006, p = 0.17). What was found was a hint towards strong inter-speaker variability. For speaker 1, syllable lengthening was significantly higher when the lengthening occured alone (F(1,11) = 10.85, R^2 = 0.5, p = 0.007). For speaker 2, Word fragments occuring clustered with other disfluency elements were significantly longer (F(1,11) = 6.476, R^2 = 0.37, p = 0.02). This could be good news, as one could hypothesize that many different shapes of conversational speech synthesis will be acceptable due to the variability of real speech.

3.2. Durational Features of Disfluency Elements

First, we looked at the durations of the individual disfluency elements. As summarized in Figure 3, Lenghtening (L1) is the element with the longest duration on average, frequently with values of 500 ms and, more rarely, with about 800 ms. Fragments (F1) usually span a much shorter time, mostly between 150 and 300 ms, however with some outliers ranging up to 670 ms. Silent Pauses (P1) exhibit the greatest degree of variability. Most instances vary in duration between 120 and 470 ms, but outliers with a duration up to 1170 ms are observed. Fillers (P2) are more moderate, ranging mostly between 170 and 380 ms, with occasional longer instances of up to 630 ms.

3.3. Phonetic Detail of Lengthening

Lenghtening is a slightly more complex issue than the other elements. It is unclear what exactly the term means and how long a syllable has to be in order to be perceived as a hesitant disfluency element. We therefore investigated the locus of syllabic lengthening and additionally checked for influences of speech rate on lenghtening extent.

In order to address these issues, we supplemented the durational analysis by also measuring the surrounding local speech rate, roughly following [12], we obtained the durations for the three preceding and the three following syllables, where available.

As can be seen in Figure 4, the annotator's detection of lengthening is quite reliable in the sense that the syllables in question (labelled 0) are indeed significantly longer than their surrounding ones. This figure shows a normalized syllable duration obtained by dividing absolute syllable duration by the number of phones contained. So the values are to be understood as follows: Phones of non-lenghtened syllables span between 50 and 70 ms in duration, occasionally stretching to about 140 ms. Phones in lenghtened syllables vary mostly between 140 and 180 ms, with much higher variability, up to 250 ms. In terms of absolute duration, the majority of lengthened syllables is between 300 and 450 ms long, with variability up to 600 ms. As a general rule, lengthened phones and syllables are roughly twice as long as normal ones, but the high variability indicates that even much more lengthening could be acceptable. We checked if the durations of the preceding syllables had any predictive value for the extent of the lengthening, but this was not the

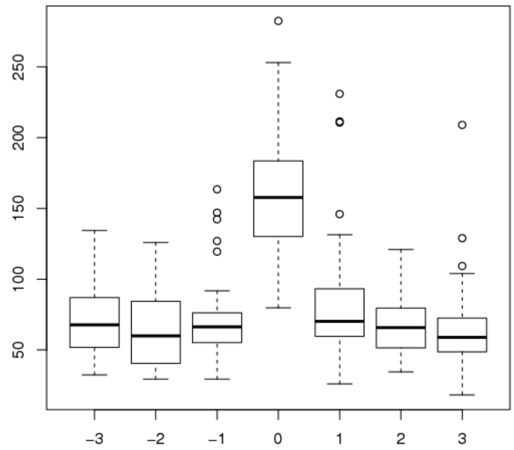

Figure 4: *Duration of lengthened syllable and its three surrounding syllables in ms divided by number of phones.*

case. Speaking rate appears to have no influence on this factor, lenghtening can be twice as long as the preceding syllable, or it can be five times as long. A linear regression analysis confirmed that the duration of the last syllable before the lenghtened one has no influence on the extent of the duration of the following one (F(1 ,33) = 0.057, R^2 = 0.0017, p = 0.81).

4. General Discussion

4.1. Lengthenings and Silent Pauses

We will first discuss the elements L1 and P1 which were perceived as unobtrusive in the experiment. As shown in Figure 3, both elements exhibit a great degree of variability. Their real-world occurences frequently cover long time spans of 500 ms and more. A solitaire silent pause is by far the most frequent element. We assume that their wide range of durational variability is the reason for the high user acceptancy. It is likely that valuably long delays in incremental speech synthesis can be easily built with these elements while keeping a high standard of synthesis quality. Future work could see a perception test that measures the actual extent of stretching applicable. The hypothesis would be that any pause and lengthening duration of 500 ms would be acceptable. A more tolerant model would predict that also 800 ms are unproblematic.

4.2. Fragments due to Word-Cutoffs

Two things are important to note about fragments. First, we have a strong suspicion that they will be dispreferred in human communication as well, so quality judgments of fragments should in future work better be measured against a disfluent human baseline. Second, our approach to fragments in the experiment was too static. We took the fragment from the original utterance and applied no variation to it. But as can be inferred from the empirical analysis, variation is a key factor in human production of disfluencies.

As can be seen in Figure 3, there is great variability in fragments, even if only duration is considered. Durational variation in fragments can mean two things. On the one hand, the cutoff point appears to be arbitrary at any phone boundary and is not governed by structural constraints such as syllable boundaries. So this variation could be merely a reflection of the fact that the words turned into fragments by cutting them off during production have been uttered to a varying degree. Examples from the data that illustrate this would be the fragments “f-” and “Quadratm-” which would yield very different results in durational analysis.

On the other hand, there is another option we have not included so far. Lengthening is more than a stretching of the predisfluent syllable, as it can be seen as durational variation applicable to all other disfluency elements. So it might be necessary to lengthen word fragments too, regardless of their cutoff point, which would account for more durational variation. If we take up the examples above, imagine, the “f-” was lenghened and the “Quadratm-” was not, this might lead to an assimilation in duration despite the very different cutoff point.

Although fragments are dispreferred, they can add flexibility for correction management to the system. With the insights at hand, we hypothesize for future work that cutoffs should only be facilitated if really necessary and that durational variability should be included for fragments, such as the option to lengthen the phone after which the cutoff is to happen.

4.3. Filled Pauses

Variability also is the keyword for the last element in question, P2. An explanatory note first: On the macro-level, filled pauses often feature silence around the filler, as reflected by the double-pause phenomenon reported above. The duration measured here concerns only the disfluency element P2 itself, the filler that is produced.

In the experiment, we did not apply any variation to the fillers, which in the light of the variability seen in Figure 3 might well be the reason for their bad performance. There was a padding of about 100 ms around the fillers we inserted, but also this approach was static. So, in order to synthesize well-sounding fillers, we need to account for the natural variability, be it only the durational one we observed in this study. In addition it could prove fruitful if we applied a wider concept of pauses, encapsulating fillers that can be embedded in or combined with silent pauses directly.

5. Conclusions

We provided some basic research on the micro-structure of disfluencies to improve dialogue synthesis systems. Variability is the key to high sound quality in conversational speech synthesis. We should be able to build a high-quality system with human-like interaction speed by equiping it with spontaneous speech phenomena that are tailored to the demands of incremental processing.

A great deal of the variability encountered appears to arise from differences between speakers. While this makes deriving robust rules difficult, it presents a high degree of freedom in terms of designing speech synthesis. If each speaker produces spontaneous speech elements differently anyway, it is suspectable that system flexibility, surfacing in speech synthesis as variation, will be readily accepted by human listeners.

While forward-looking hesitations like lengthenings and silent pauses already yield good results, work has to be done in order to improve backward-looking correction qualities. With the inclusion of variability discussed above, we hope to be able to supplement our system with the capability to interrupt itself and make the best of it when no other option is available, but which for minor issues prefers delaying strategies.

6. References

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

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

[3] J. E. F. Tree, “The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech,” *Journal of memory and language*, vol. 34, no. 6, pp. 709–738, 1995.

[4] J. Ginzburg, R. Fernández, and D. Schlangen, “Dysfluencies as intra-utterance dialogue moves,” *Semantics and Pragmatics*, 2014.

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

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

[7] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1526–1540, 2006.

[8] S. Betz, P. Wagner, and D. Schlangen, “Modular synthesis of disfluencies for conversational speech systems,” 2015.

[9] S. Kousidis, T. Pfeiffer, and D. Schlangen, “Mint.tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora,” in *Proceedings of Interspeech*, 2013.

[10] 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.

[11] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [computer program]. http://www.praat.org/,” 2014.

[12] H. Pfitzinger, “Local speech rate as a combination of syllable and phone rate,” 1998.