MODULAR SYNTHESIS OF DISFLUENCIES FOR CONVERSATIONAL SPEECH SYSTEMS

Simon Betz, Petra Wagner, David Schlangen

Bielefeld University
Faculty of Linguistics and Literary Studies
Phonetics and Phonology Workgroup, Dialogue Systems Group
simon.betz@uni-bielefeld.de

Kurzfassung: It has been shown that dialogue systems benefit from incremental architectures to produce fast responses and to interact with the interlocutor in a more human-like way. The advantage of quick responses yields the disadvantage of running out of things to say for a while. In such occasions, humans tend to produce disfluencies as a listener-oriented strategy to signal the ongoing production process and to buy time for finalizing the turn. Introducing disfluency capabilities into a speech synthesis module of a dialogue system may therefore be a straightforward strategy towards conversational speech systems.

Disfluencies are a very complex matter, they can take various chaining and nested forms in human communication. We do not attempt to equip our system with the full range of possible disfluent time-buying strategies found in human interaction. For a first perceptual evaluation of the most suitable synthetic disfluency strategy to be integrated into the dialogue system, we focus on three structural factors that are able to cover a wide range of attested disfluency patterns: lengthening, word cutoffs and pauses. This leads to several different configurations a disfluent sentence can take. Sentences from a spontaneous speech corpus were resynthesized in all possible configurations using Mary TTS. In order to identify euphone configurations, these stimuli were then presented to test subjects in a perception test.

1 Introduction

It has been shown that dialogue systems benefit from incremental architectures to produce fast responses and interact with the interlocutor in a more human-like way. The advantage of quick responses yields the (also very human-like) disadvantage of running out of things to say for a while [4]. In such occasions, humans tend to produce disfluencies as a listener-oriented strategy to signal the ongoing production process and to buy time for finalizing the turn. Introducing disfluency capabilities into a speech synthesis module of a dialogue system may therefore be a straightforward strategy towards both an improved naturalness and a prevention of barge-ins. It would also be a step beyond the scope of reading tasks typical for traditional TTS systems [10] and towards more flexible and adaptive conversational speech systems. In this paper, we are describing the first steps towards devising a full dialogue system capable to produce acceptable disfluencies.

Disfluencies are a very complex matter, they can take various chaining and nested forms in human communication (see [8] for an overview). They serve an important role in time management in dialogue in order to structure turns and signal boundaries [5]. However, not many attempts to synthesize disfluencies have been undertaken so far, with notable exceptions like [1]

	Lengthening	Fragment	Pause	
FLD	+	-	+	
	-	+	+	BLD

Tabelle 1 - Assumed distribution of disfluency phenomena within FLD and BLD

and [9], among others. From a dialogue system's perspective, two groups of disfluencies can be distinguished: forward-looking (FLD) and backward-looking (BLD) ones. FLDs are useful when the system detects that there will be trouble and needs to stall for extra time. BLDs can serve as repair strategies when the system has uttered an erroneous item and has to interrupt and correct itself.

Our long-term goal is to construct a dialogue system capable of conversational speech. This study is to provide a first glance at the implementation of disfluencies from a phonetic point of view: We want to find out (a) if the universal structure of disfluencies can be boiled down to a modular architecture for constructing versatile and euphone disfluencies, and (b) which types or configurations of disfluencies are acceptable in synthetic speech output as evaluated via a perception experiment.

2 Methods

We do not attempt to equip our system with the full range of possible disfluent time-buying strategies found in human interaction. For a first perceptual evaluation of the most suitable synthetic disfluency strategy to be integrated into the dialogue system, we focus on three structural factors that are able to cover a wide range of attested disfluency patterns: Pre-disfluency lengthening (L), word fragments due to cutoffs (F), and pauses (P). While the former two are binary, being either absent (0) or present (1) the latter factor has three parameters: (0) absent, (1) present but silent, (2) containing a filler. This leads to twelve potentially different configurations a sentence can take, eleven disfluent ones plus the quasi-fluent one with every factor set to 0. This way we hope to construct a handy, yet versatile framework for generating and synthesizing disfluencies. Their diversity in nature cannot be fully accounted for, but if the tri-factorial approach described above proves fruitful, the diversity of the system output can be strongly increased. Aside from the insights expected from the evaluation experiment, it is promising to have a modular architecture for building disfluencies, with three factors leading to twelve configurations per sentence (cf. table 4).

We assume that not all of the twelve configurations will be acceptable in every context. In terms of FLD and BLD, we assume that only the factor pause is present in both strategies, whereas lengthening and word fragments will only feature in FLD and BLD respectively. It is unlikely that factors lengthening and fragment co-occur in natural speech, thus a lower user rating is expected for the configurations that feature both. The factor pause, however, could prove a powerful allround tool which works in both directions as a hesitation marker, so we expect it to perform well or at least not to prove detrimental. Given these presumed limitations in distributability, we hypothesize that only some combinations of disfluencies are valid and acceptable: A tendency that highly disfluent sentences are dispreferred is likely to emerge, as the same would be expected of natural speech.

2.1 Stimuli

In order to model the twelve different configurations for a perception experiment, stimuli were built based on four spontaneously produced disfluent sentences taken from from the dreamapartment corpus, a corpus of highly involved task oriented dialogues [6] using MaryTTS [7]. In

ID	Sentence
A	Also dann hab ich fünfund- ne gar nicht, dann hab ich vierzig Quadratmeter.
	Well then I've got twentyfi- no wait, then I've got forty square meters.
B	Dann ma- lassn wir mal die Einzelheiten einfach weg.
	Then w- then we'll simply leave out the details.
C	Dann würde ich sagen ein f- Zimmer für dich, eins für mich.
	Then I'd say one f- room for you, one for me.
D	Ja ist doch alles bisher so re- ach nein, das Wohnzimmer war ja L-förmig.
	Yeah so far everything is rec- ah, no, wait, the living room was L-shaped.

Tabelle 2 - The stimuli sentences in orthographic representation and english translation

Factor	Description
L0	no Lengthening
L1	Lengthening present
F0	no Word Fragment
F1	Word Fragment present
P0	no Pause
P1	Silent Pause present
P2	Filled Pause present

Tabelle 3 - Factors for disfluencies

order to provide a high sound quality and naturalness, we extracted the original pitch curve and the durational parameters of each phone and transferred them to the MaryXML input file using a python script.[3] This way we ensure that quality judgements for disfluencies are less biased by overall synthesis quality. The central interest while selecting speech material for this study was to be as close to naturally occuring disfluencies as possible and to have to invent as little as possible. The transfer of original prosodic information provides a very good sound quality. For that reason, we do not want to invent speech material that is not present in the original utterance, as that would imply that we would be obliged to invent that new material's prosody too, which would either result in high workload or in poor quality. This made us depart a little from the full Shribergian model of disfluencies ([8], see above), which would have included four different ways of continuing after a disfluency. As we did not want to invent three alternative sentence endings, we stuck to the phonetics around the interruption point. We chose to select as original input utterances that featured a mid-word cutoff, since it involves a bit of inventing when we tried to artificially insert a cutoff. Leaving out the fragment resulting from a cutoff, however, poses not much of a problem, as it is conceived as an optional element in disfluency structure. This leaves lengthening and pauses to be considered. For lengthening, the last syllable before the disfluency was extended in duration with MaryTTS. For silent pauses, a break was inserted. The durational parameters used for this are mean values obtained from the corpus. For filled pauses, a german Filler *ähm* produced by a speaker in the corpus was reused, optionally with pitch adjustment if the rest of the stimulus required it.

These four utterances were re-synthesized according to the twelve disfluency strategies using MaryTTS while preserving pitch and segmental duration values of the original speakers, resulting in 48 stimuli in total. These stimuli were rated by test subjects according to their situation-specific level of acceptability on a 5-point Likert scale.

[3]The function of the script is quite limited to resynthesizing utterances. If that is exactly what you want to do, feel free to contact corresponding author for more information.

L0F0P0	Dann lassen wir mal die Einzelheiten einfach weg.
L0F0P1	Dann [.] lassen wir mal die Einzelheiten einfach weg.
L0F0P2	Dann [ähm] lassen wir mal die Einzelheiten einfach weg.
L0F1P0	Dann ma- lassen wir mal die Einzelheiten einfach weg.
L0F1P1	Dann ma- [.] lassen wir mal die Einzelheiten einfach weg.
L0F1P2	Dann ma- [ähm] lassen wir mal die Einzelheiten einfach weg.
L1F0P0	Da:n:n lassen wir mal die Einzelheiten einfach weg.
L1F0P1	Da:n:n [.] lassen wir mal die Einzelheiten einfach weg.
L1F0P2	Da:n:n [ähm] lassen wir mal die Einzelheiten einfach weg.
L1F1P0	Da:n:n ma- lassen wir mal die Einzelheiten einfach weg.
L1F1P1	Da:n:n ma- [.] lassen wir mal die Einzelheiten einfach weg.
L1F1P2	Da:n:n ma- [ähm] lassen wir mal die Einzelheiten einfach weg.

Tabelle 4 - Stimulus B in all twelve configurations, original configuration was L0F1P0

2.2 Test Subjects

At the time this paper was written, 32 Test subjects had participated in the experiment. The experiment is planned to be continued with another 30 subjects, so the results presented here are to be understood as tendencies and the paper as work-in-progress. Subjects were presented the stimuli via the PRAAT MFC [3] environment in a quiet room on a pc with headphones. 15 of the subjects were female and 17 were male. They were between 23 and 39 years old, most of them monolingual native speakers of German, and most of them university students or gradutates. Data of non-natives or non-academics showed no significant differences from the other result, so these data were treated as normal. No subject reported any hearing impairment. Subjects were asked for their familarity with synthetic voices and were asked to report what devices, if any, they use, and how regularly they do so. As there was no significant difference in the results regarding this factor, it will not be considered later on.

3 Results

We used Analyses of Variance (ANOVA) to identify influences of the main factors lengthening, fragment and pause. Lengthening does not appear to have any effect ($F(1) = 0.009$, $p = 0.923$), but Fragment ($F(1) = 13.37$, $p < 0.001$) and Pause ($F(2) = 46.74$, $p < 0.001$) do. The interaction between fragment and lengthening reaches borderline significance ($p = 0.071$) when all four stimuli are compared. A TukeyHSD post-hoc test was conducted to find out which configurations of stimuli yielded significant results in comparison. It showed that the mean of responses decreased upon the transition into a more disfluent condition, i.e. conditions that feature F1 get lower mean results than the same conditions with F0 instead of F1 ($p < 0.001$) The same holds true for P2 which performs worse than P1 ($p < 0.001$) and P0 ($p < 0.001$). P1 gets almost the same mean result as P0 (3.35 and 3.32 respectively). First glances at the data reveal that stimulus D gets lower ratings than any other sentence. For this reason, the ANOVA and TukeyHSD were repeated for stimuli A, B and C only to check for effects of the dispreferred stimulus. The overall results remain the same, but the borderline interaction between fragment and lengthening is lost ($p = 0.2$). The TukeyHSD reveals that results lower significantly on stimulus D in the comparison of F0 and F1, regardless of lengthening. Looking at the overall means reveals that stimuli with the L0F0P0 condition, the quasi-fluent one, get slightly higher scores than others, yet not significantly higher ones ($p = 0.45$). There is a tendency that the higher the number of disfluencies in a stimulus, the lower the mean of responses. As to be seen in Fig.1, the results are scattered widely, so there is no statistically significant correlation ($cc = -0.09$). There are

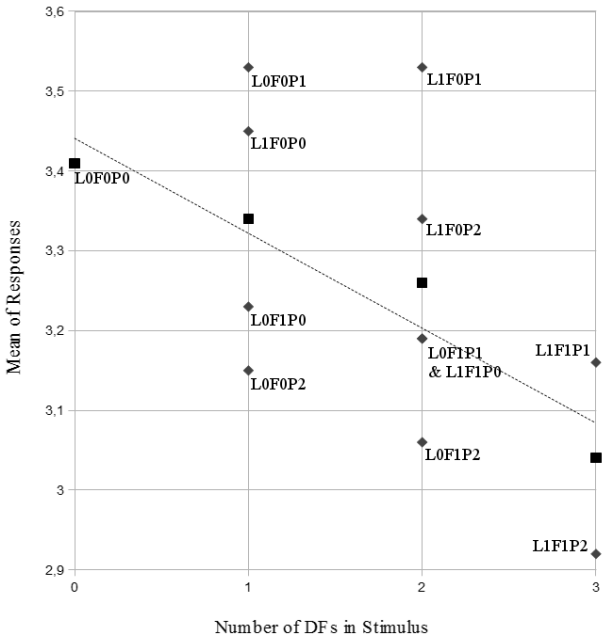

Abbildung 1 - Means of responses per configuration. Squares are the means for a group of configurations with regard to the number of disfluencies they contain. Line illustrates tendency of the means.

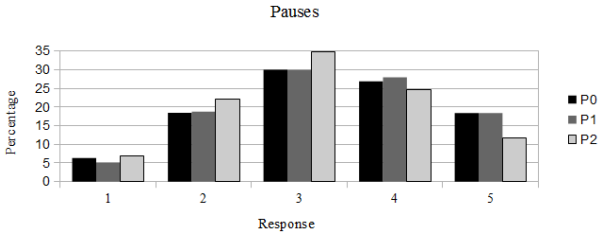

Abbildung 2 - Distribution of responses to different pause categories

three configurations that get (slightly, but not significantly) higher means than the fluent configuration: L0F0P1 (μ = 3.53), L1F0P1 (μ = 3.53), L1F0P0 (μ = 3.45) as opposed to L0F0P0 (μ = 3.41)

4 Discussion

The results show a tendency that simplicity and unobtrusiveness is preferred. Sentences with no pauses or with silent pauses get better ratings than those with filled pauses. As can be seen in Fig.2, filled pauses (P2) outrank the other pause categories in the lower response categories, while the other pauses score higher in the above-average categories. The same tendency is found with fragments. As can be seen in Fig.3, Stimuli with no fragments dominate in the two above-average response categories, while those where fragments are present score higher in the lower response areas. Lengthening, on the contrary, does not appear to have any significant effect. Interestingly enough, this does not mean that the quasi-fluent configurations (with no lengthening, no fragment and no pause) are preferred, as they do not perform significantly better than the others. There is, however, a tendency that shows that overall quality decreases the more disfluencies are introduced into a stimulus (Fig.1). Within this, there is another interesting aspect: There are three disfluent configurations that get even slightly higher means than the fluent one

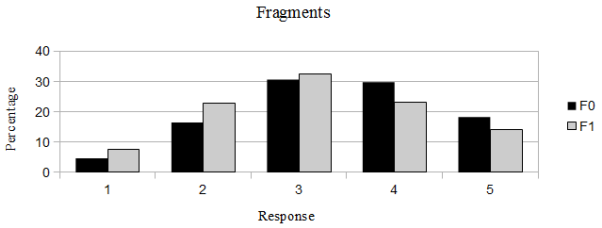

Abbildung 3 - Distribution of responses to different fragment categories

(L1F0P0, μ = 3.45; L0F0P1, L1F0P1, μ = 3.53). Those configurations have in common that they feature either lengthening or silent pause or both. Lengthening appeared not to be significant in the analyses, which might be a hint that it gets high scores due to unobtrusiveness. Also for silent pauses the point could be made that it is an unobtrusive disfluency which is acceptable in speech synthesis. These could be very useful findings for future research if dialogue systems could freely insert lengthenings and silent pauses to stall for processing time. One remaining task would be to examine to what extent durations of disfluencies can be stretched before they become detrimental for perceived quality.

Another point to discuss is why complexity is dispreferred and what the implications of this are for future work. Filled pauses did not perform badly, only in the framework of this study they appeared less natural or euphone than just silences. There are some explanations for this. First, there was no variation within the fillers. They were adjusted to fit the prosody of the surrounding stimulus, but at the end of the day, it was the same *ähm*. We suspect that introducing more variability into the filler could improve the performance. This variability can be purely durational, that is, a lengthening within the filler, which occurs frequently in human conversation, but was for reasons of comparability omitted from this study. Filled pauses can behave a bit differently from silent ones [2] and fulfill an additional role in conversation: the prevention of barge-ins. So even if filled pauses were to ever perform weaker than silent ones, they still would be interesting for future work. After all, communication management is a tradeoff between content and means of conveying it.

Basically the same is true for fragments. The difference is, though, that here we cannot blame architectural constraints since the fragments were copied from natural speech. Again, they did not perform terribly, but we suspect that these kinds of interruptions are not quite expected from a conversational agent. On the other hand it could be worthwhile considering that maybe cutoffs are easy to handle within synthesis. If invented cutoffs at arbitrary points yielded similar results, this could be a very powerful feat for facilitating corrections.

A further result to ponder is the lengthening and fragment interaction. It is quite surprising that a configuration which sounds so unnatural does not yield significantly negative results. As illustrated in table 1, we expected that these factors would not be freely combinable. If they are, it becomes even easier to satisfy the needs of FLD and BLD in various dialogical contexts. It is however possible that the continuation of this study sheds a new light on these tendencies. Maybe lenghtening plays more of a supportive role in disfluency structure. The demand for durational variability in the filler design could be a hint that there is more than one place where lengthening can be. Maybe lengthening could be better conceived of as a parameter stretching over one or more elements of the disfluency rather than only being an unobtrusive disfluency element. Our modular architecture should be updated taking these findings into account. It is good to have pre-disfluent lengthening as a module in our repertoire, but it should not be limited to that.

What comes as no surprise at all is the fact that not all stimuli are rated equal, in the sense that one sentence sounded worse than the others. In this study, it was the sentence where the original

speakers discussed the L-shapedness of a room. As it turns out, MaryTTS is not very useful for pronouncing individual letters. It would have been an option to take any German word with [El] as initial syllable and cut out the unnecessary parts in post-hoc processing. Fragments are tricky for future research. On the one hand, it appears possible to interrupt ongoing production at arbitrary phones, on the other hand, it yields lower-quality output. It is important to note, that in a full system, the stop will have to be executed in the synthesis output and not in planning, for the same reason as above: When MaryTTS is forced to synthesize unknown word fragments, it will falter quite frequently.

5 Conclusion

We succeeded in phonetically modelling a variable set of disfluencies using a modular design that can insert eleven different disfluencies into any given sentence. The disfluent stimuli produced this way fared not significantly worse than the fluent ones, some even got unexpectedly good user feedback. The transfer of prosodic information yielded a high quality of synthetic speech, making it desirable to improve future language generation models in terms of prosody. Our design is capable of covering both forward- and backward-looking disfluency strategies, making it a versatile tool for supporting dialogue system's dynamic output in the future.

Literatur

[1] ADELL, J., A. BONAFONTE und 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] ADELL, J., A. BONAFONTE und D. ESCUDERO-MANCEBO: *Modelling Filled Pauses Prosody to Synthesise Disfluent Speech*. 2010.

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

[4] BUSS, O. und D. SCHLANGEN: *DIUM – An Incremental Dialogue Manager That Can Produce Self-Corrections.*. In: *Proceedings of Semdial*, 2011.

[5] CLARK, H.: *Speaking in Time*. Speech Communication 36, 2002.

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

[7] SCHROEDER, M. und 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.

[8] SHRIBERG, E.: *Preliminaries to a Theory of Speech Disfluencies*. Ph D. thesis University of California, 1990.

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

[10] WAGNER, P.: *(What is) the contribution of phonetics to contemporary speech synthesis (?)*. In: *Systemtheorie. Signalverarbeitung. Sprachtechnologie. Rüdiger Hoffmann zum 65. Geburtstag. Studientexte zur Sprachkommunikation. Vol Band 68.*, 2013.