Bielefeld University

PhD Thesis

Hesitations in Spoken Dialogue Systems

Simon Betz

supervised by
Prof. Dr. Petra Wagner
co-supervisor:
Prof. Dr. Joana Cholin

March 17, 2020

for my father

Acknowledgments

Before diving into the exciting world of hesitation synthesis, it is time to give thanks to certain people without whom the existence of this book would not have become reality.

I'd first like to thank the people in academia who helped this project to come to existence. To Petra Wagner for everything you helped me with. I cannot find better words to thank you, as you did so much. To Sina Zarrieß for being an inspiration and companion from the first day I came to Bielefeld. To David Schlangen for the massive support and feedback in my early days of academia. To Joana Cholin for the interest in my topic and for your eagerness to support me, regardless of the hurdles bureaucracy placed in the way. To Annett Jorschik for her patience while explaining statistical concepts to me. Over and over again. To Birte Carlmeyer trying to do the same with the technical side of dialogue systems.

Thanks to the many people from the phonetics workgroup, the dialogue systems group and the applied informatics group I had the honor to work with during my time in Bielefeld: Britta Wrede, Julian Hough, Soledad Lopez, Aleksandra Cwiek, Katharina Nimz, Nataliya Bryhadyr, Antje Hey, Katharina Wendler, Zofia Malisz, Laura de Ruiter, Angelika Hönemann, Valentina Schettino, Andreas Windmann, Caro Kirchhof and Barbara Samlowski.

Honorable mention to two guys from the olden days at Münster University, without whom all this would never has happened: René Schiering, who introduced me to Petra about 10 years ago, and Johannes Hoppe, with whom I took the first steps in phonetic experimentation.

Thanks, of course, to my family, all of you, we had challenges beyond imagination to solve over the past years, nonetheless you have always been there for me. Thank you!

Special thanks to Jana Voße for the magnificient support. To my father, whom I wish better days from the deepest of my heart. To my son, who is the most understanding child I can imagine.

Contents

Preliminary Matter

0.0 Zusammenfassung

Die vorliegende Arbeit befasst sich mit der Entwicklung konversationeller Sprachsynthese. Das Ziel ist die Erstellung eines Modells zur spontanen Insertion von Häsitationsphänomenen in den Output von Dialogsystemen.
Sprechende Maschinen sind längst Teil des Alltags geworden. Smartphones, Navigationssysteme, persönliche virtuelle Assistenten oder Smart-Home-Anwendungen sind in der Lage, auf sprachlichen Input ihrer Nutzer zu reagieren und selbst mittels gesprochener Sprache zu kommunizieren. Dabei ist die Qualität der synthetischen Stimme mittlerweile beachtlich, bisweilen kaum noch von Menschen zu unterscheiden. In dieser Arbeit wird die andere Seite der Medaille betrachtet - die Konversation. Sprachliche Interaktion zwischen Menschen erinnert an einen schnellen, teilweise synchronen oder zeitlich überlappenden Tanz. Sprachliche Interaktion zwischen Mensch und Maschine ist verglichen damit zum gegenwärtigen Zeitpunkt eher ein einfaches Spiel, bei dem ein Ball hin- und her geworfen wird, bei dem zudem nach dem Fangen Zeit benötigt wird, um den Ball zurückzuwerfen. Diese Analogie soll illustrieren, dass die Qualität in der Sprachausgabe der heutigen Systeme immer im Wechselspiel ist mit den interaktiven Qualitäten. In dieser Arbeit wird beleuchtet, dass es möglich ist, Dialogsysteme zu konstruieren, welche die gleiche Interaktionsgeschwindigkeit und Anpassbarkeit an den Tag legen wie es in menschlicher Konversation üblich ist. Dies ist jedoch bisher nicht möglich mit der Stimmqualität moderner kommerzieller Systeme. Diese wiederum werden dem inhärenten Anspruch an Interaktionsfähigkeit nicht gerecht.
Menschliche sprachliche Interaktion ist durch weit mehr als das gesprochene Wort

gekennzeichnet. Besonders die zeitliche Mikrosteuerung und das Management des Rederechts werden subtil kontrolliert durch nichtsprachliche Phänomene. In dieser Arbeit soll es unter diesen vornehmlich um Disfluenzen gehen, genauer, um Häsitationen. Diese werden vom Sprecher mehr oder minder unwillkürlich produziert und äußern sich durch stille Pausen, durch in die Länge gezogene Silben und durch die in der Rhetorik berüchtigten Füller wie "äh" oder "ähm".
Diese Phänomene wurden in klassisch linguistischer Forschung oft negativ betrachtet. Mittlerweile sind sie im Zentrum des Interesses angekommen, da sie viele kommunikative Aufgaben erfüllen: Sie kaufen dem Sprecher Zeit, sie können Attitüden wie Unsicherheit signalisieren und verhindern, dass der Gesprächspartner unterbricht, wenn der Sprecher seine Nachricht noch nicht beendet hat, aber Gründe hat, gerade keine Sprache zu artikulieren.
In den letzten circa 15 Jahren hat die Forschung begonnen, diese Phänomene in Bezug auf künstlich erzeugte Sprache zu untersuchen. Könnte es Sinn ergeben, Smartphones oder Navigationssystemen beizubringen, wie Menschen zu pausieren, zu zögern? Bisher sind die Ergebnisse ambivalent. Es zeigt sich immer wieder, dass die Beeinträchtigung der Sprachausgabequalität, die mit der Einbeziehung spontansprachlicher Phänomene einhergeht, die Evaluierung von konversationellen Dialogsystemen erschwert.
In dieser Arbeit werden demnach drei große Areale bearbeitet:

1. Eine Übersicht über bisher geleistete Forschung im Bereich Disfluenzen und Sprachsynthese. Diese Übersicht wird knapp gehalten und auf die wichtigsten Erkenntnisse beschränkt, da es bereits andere Dissertationen gibt, die einen exzellenten Überblick bieten. (Part I)

2. Grundlagenforschung in ebenjenen Bereichen. Im Verlaufe dieses Dissertationsprojektes wurden Disfluenzen und ihr Potential für die Sprachsynthese von verschiedenen Ausgangspunkten beleuchtet. Dabei hat sich insbesondere das Phänomen der Häsitationslängung als bisher unter-erforschter Gegenstand gezeigt, der großes Potential für konversationelle Synthese verspricht. (Part II)

3. Implementierung der gewonnenen Erkenntnisse anhand eines konversationellen Dialogsystems in einer Smart-Home-Umgebung. Im Zuge dessen wird auf

den aktuellen Stand der Evaluierung solcher Systeme eingegangen, welche ebenso wie viele kontemporäre Systeme ihrer Zeit hinterherhinkt und einige entscheidende Aspekte der Interaktion ausblendet. (Part III)

Eingerahmt wird dieser Hauptteil von einer kurzen Einleitung, einer Zusammenfassung, sowie einem Ausblick auf mögliche zukünftige Forschung in diesem Bereich.

0.1 Disclaimer

0.1.1 Previous, Preliminary and Related Work

Parts of this thesis and the work presented therein have been published in the following articles:

Simon Betz, Jana Vosse, and Petra Wagner. Phone Elasticity in Disfluent Contexts. In Fortschritte der Akustik - DAGA 2017, pages 1462-1464, 2017.

Simon Betz, Jana Vosse, Sina Zarriess, and Petra Wagner. Increasing recall of lengthening detection via semi-automatic classification. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017, Stockholm), pages 1084- 1088, 2017.

Simon Betz and Petra Wagner. Disfluent Lengthening in Spontaneous Speech. In Oliver Jokisch, editor, Elektronische Sprachsignalverarbeitung (ESSV) 2016. TUD Press, 2016.

Simon Betz, Petra Wagner, and David 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), pages 2222-2226, 2015.

Simon Betz, Petra Wagner, and David Schlangen. Modular synthesis of disfluencies for conversational speech systems. In Günther Wirsching, editor, Elektronische Sprachsignalverarbeitung (ESSV) 2015, Studientexte zur Sprachkommunikation. TUD Press, 2015.

Simon Betz, Petra Wagner, and Jana Vosse. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. In Phonetik und Phonologie 12, 2016.

Simon Betz, Sina Zarriess, and Petra Wagner. Synthesized lengthening of function words - The fuzzy boundary between fluency and disfluency. In Proceedings of the International Conference Fluency and Disfluency, 2017.

Simon Betz, Birte Carlmeyer, Petra Wagner and Britta Wrede. Interactive Hesitation Synthesis: Modelling and Evaluation. Multimodal Technologies and Interaction 2.1 (2018): 9.

0.1.2 Pluralis Majestatis

The first person narrator of this thesis uses the plural form "we" in the main body of text. This is not to imply that the author suffers from illusions of grandeur, but rather the opposite: As the list of papers above suggests, many studies leading towards this thesis were accomplished together with many other authors. While the author of this thesis was the driving force behind all of these papers, he would still be ashamed to talk in the first person singular and uses the plural to humbly acknowledge the sheer impossibility of this thesis having come to existence on the back of a single, isolated mind.

0.1.3 Gender

Whenever an example is given in this thesis that is easier to comprehend with a speaker as subject, the speaker will be referred to as "she" for simplicity reasons. Readers might call her "Susan" if they like. Whenever a group of people with mixed or unknown genders is referred to, the pronoun "they" will be used for simplicity and neutrality reasons. He / she and combinations thereof might wrongly suggest a binary distribution of gender. In all our experiments, we provided the option to choose "other" as gender for participants, which has been selected in several studies.

Part I

Theoretical Background

Chapter 1

Introduction

Machines have acquired language and learned to speak. This might be the impression one gets when considering the remarkable improvements in sound quality of electronic devices. There are even anecdotes about people falling in love with the imaginary person behind the voice of their smart phone. While synthetic speech appears like a solved problem given the present output quality, there is another side to this medal that will be the topic of this thesis: conversation.
Conversation between humans differs substantially from the interaction between humans and machines. It is governed by awareness for the interlocutor, by timing constraints, and by the ability to rapidly adapt to changes in the dialogue environment. Human conversation is a complex couple dance as opposed to a simple ball-tossing game of contemporary human-machine interaction. As Clark (1996, p. 3) puts it:

> *Doing things with language is [...] different from the sum of a speaker speaking and a listener listening. It is the joint action that emerges when speakers and listeners [...] perform their individual actions in coordination.*

Nowadays, however, synthetic speech is applied in various fields, and it has entered the realm of everyday life: in public transportation announcements, telephone customer services, mobile phone speech output, or smart home environments, to name only a few. Despite the interactive nature of many of these applications,

speech output remains to be rather static and simple, typically reading out predefined texts or responding with an awkward delay.
Another feature of synthetic speech encountered in these realms is its "fluency", that is, it does not contain the hesitations, reformulations, or fillers which are typical for human spontaneous speech production. Rather, speech output, once generated, is produced in a "ballistic", non-interrupted fashion.
In this book, the notion will be explored that this ballistic delivery is sub-optimal for many scenarios of human–machine interaction, where listeners actually need to process information that is synthetically generated, and that hesitations are a conceivable way to improve interaction quality. This is based on the following premises:

1. **Human speakers have ways to allow for extra time during speech planning and production, in case a problem hinders timely delivery.** Hesitations, such as syllable lengthening, silences and fillers, are one way to do so. It is debatable whether these markers are deployed intentionally or a by-product of speech planning difficulties, but what is clear is their property to temporally extend the speech signal, thus *buying extra time*. *Buying time* serves both the speaker, to remedy or improve production, and the listener, by aiding comprehension via reducing the amount of units of processing per unit of time. Viewed from the other side, speech planning requites cognitive effort and the transmission of information from one to the next level of processing takes time. This time can be managed in dialogue by the use of hesitations.

2. **Communication management is a trade-off between communicative material and means of conveying it, and this is true for dialogue systems as well as for humans interacting.** It has been shown that human listeners are willing and able to accept hesitations produced by machines via speech synthesis. It is thus possible to enable machines to buy time in dialogue using the same strategies as humans.

These premises are based on previous work on hesitations. In chapter 2.1, an overview of these phenomena and theories of speech production will be given, as well as basics on dialogue and the technical implementation thereof. Part II of this

thesis concerns empirical investigations of hesitation phenomena, both in the form of corpus studies of human speech production as well as experiments with synthetic realizations of these phenomena. Part III will then introduce and evaluate a model for inserting hesitations in dialogue systems, based on the empirical investigations in the previous part. Part IV summarizes and concludes this work.

Chapter 2

Form, Function and Modeling of Disfluencies

2.1 A First Grasp of the Concept

The topic of this thesis is utilizing hesitations for synthetic speech and dialogue systems. In this chapter, the theoretical foundations will be summarized, beginning with describing hesitations as a subgroup of disfluencies, followed by a detailed description of the phenomena suitable for speech synthesis. Then, the state of the art of speech synthesis and dialogue systems and the applicability of hesitations therein will be discussed briefly.

Hesitations have mostly been studied as a sub-phenomenon of disfluencies, so a general examination of disfluencies is in order. So far, there is no universally agreed-on definition of speech disfluencies, rather, there is huge terminological overlap and ambiguity as well as general criticism of the term (Lickley, 2015). This overlap is not coincidental - the phenomena related to disfluency are entangled heavily, more or less depending on the view of the respective study or field dealing with it. The first part of this thesis is thus devoted to giving the reader an overview of the phenomena involved and define the working vocabulary for this book.

A general notion is that disfluencies mark a deviation from ideal speech delivery. In classical linguistics they have been viewed as a mismatch between the underly-

ing speech plan and the concrete realization (Chomsky, 1965). The major problem with this view is, that there neither is a definition for ideal speech, or, for that matter, for fluency. Often, ideal speech has been confused with read speech, which arguably is not ideal for all speaking situations. A speaker in a conversation would never, without a very good reason, start to speak as if she was reading. As will be shown in this book, there are positive views on disfluencies that treat them as useful conversational tools rather than problems in speaking. These views inspire the modeling of disfluencies for spoken dialogue systems with the aim in mind to improve human-machine interaction and conversation.

2.2 Introducing the Phenomena

Due to the plethora of terms that have been used to describe disfluencies and related phenomena, the terms employed in this book will be outlined, cf. fig. 2.1. Proceeding hierarchically, from broad to narrow, the term that most research uses is *disfluency*, so it will be used overarchingly for all phenomena discussed in this work.

Below this term, we can make a distinction between *forward- and backward-looking disfluencies* (Ginzburg et al., 2014). The overarching term for the former category shall be *hesitation* and, for the latter, *correction*. Within hesitations, which are the main objects of interest in this thesis, *lengthening*, *silences* and *fillers* will be distinguished. Examples for corrections, which will be only touched upon briefly in this thesis, would be *(mid-word) cutoffs* and *repetitions*.

This is a most simplified view of disfluencies and it does not attempt to reflect the plethora of surface forms they can take. In human communication, disfluencies also occur in clusters, and one could argue that in situations in which the speaker corrects herself, hesitations might serve forward- and backward-looking roles at the same time. Within the scope of this thesis, which is the synthetic production of hesitations, it shall suffice to view hesitations as forward-looking disfluencies that will be deployed to manage the interaction between the system and its user. It is necessary for this thesis to enrich the background with a short description of the historical development in order to get a grasp of the partly vague and overlapping

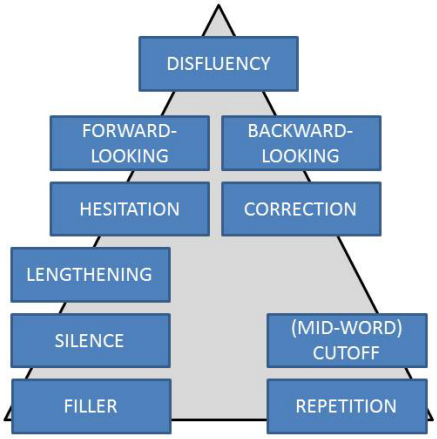

Figure 2.1: *Hierarchy of terms employed in this thesis.*

terminology used in this field, and of the underlying concepts. This overview will be limited to the terms and concepts that are essential for the remainder of this thesis. For a more detailed history of disfluency terminology, please refer to the chapter Robert Eklund (2004) devoted to it in his dissertation.

2.2.1 Disfluency

As Eklund (2004) describes in his dissertation, the term disfluency has been coined by Wendell Johnson and his group in the early 1960s. The group had been studying these phenomena in stutterers for more than a decade and compiled a list of categories and terms that were quickly adopted within the field, cf. Eklund (2004, p. 158). An early attempt on defining the term *disfluency* stems from Eugene Brutten (1963):

> "*Disfluency is defined as interruptions and breaks in the flow of the speech signal.*"

This definition has several implications. First, it appears to not be limited to stuttering as the coiners of the term might have intended; rather, it sounds like a hyperonym, an overarching linguistic concept. Furthermore, it carries the notion of a deviation from some kind of ideal delivery; a recurring concept in disfluency research that is not undebated. Early on there were comments by other researchers

stating that disfluency is not limited to stuttering but might be due to other, totally non-pathological reasons (cf. Neelley (1961)).
Interestingly, a variant spelling exists in various pathological and non-pathological works: *dysfluency*. While sounding similar to disfluency in English, the greek-rooted prefix *dys-* means "abnormal" while *dis-* means "absence of" and is similar to the latin-based prefix *non-*. The term dysfluency is thus only adequate in pathological contexts whereas disfluency or non-fluency resembles the definition by Brutten cited above (Wingate, 1984). Newer research often avoids the term disfluency, as it denotes absence of fluency, with fluency itself being an abstract or undefined concept (cf. sec. 2.4). Equating fluency with ideal delivery or flawlessly read speech cannot account for the characteristics of everyday speech (Lickley, 2015). As will be discussed in section 2.4, a new strand of research emerged that views disfluencies as conversational tools beneficial for both speaker and listener, rather than deviations from ideal delivery. Terms like *own communication management* (Allwood, 1995), *fluency enhancing strategies* (Götz, 2013) and *fluencemes* (Götz, 2013) emerged. These terms are sometimes used to describe the same phenomena traditional research subsumed under the term *disfluency*. The reason for using *disfluency* as an overarching term in this thesis is the fact that it appears to be the most common, most general one. We do not intend to ignore or refute the criticism associated with it, rather, we believe that there will be advantages and disadvantages to every term that could be used. We conceive of it as a neutral term that bundles all related phenomena, it is not intended to convey any negative connotation.

2.2.2 Elements

Throughout the thesis, *elements* will be referred to, in combination with the terms *disfluency* and *hesitation* to denote the individual usages of those in the system that is devised in this thesis. For example, a *disfluency element* could be anything that is used to serve a certain purpose in disfluency synthesis. The same is true for *hesitation elements*, with the limitation that it could only be of use for forward-looking hesitation synthesis.

2.2.3 Hesitation

Hesitation is the central aspect of this work, the very part of disfluency that is to be synthesized later on. This demands a definition: what is hesitation? For this thesis, the working definition shall be: it is anything that temporally extends the delivery of the intended message for whatever reason. A speaker is under pressure to produce and deliver speech rapidly, and will produce hesitations that the listener interprets when smooth message delivery fails. The problem is unavailability of material in any stage of speech production (which will be elaborated on in more detail in section 2.3), and hesitation is the resulting filled or unfilled gap in speech. The temporal extension of the message *buys time* for the speaker to solve problems in speech planning and production, and makes this information accessible to the listener (Brennan and Schober, 2001). The listener, in turn can use this information to infer that the speaker intends to add content and that the unspoken right to speak is not shifting. A short excursion into speaking rights and turns is in order.

Hesitations are closely related to the concept of *buying time* and to the organization of speaking rights in dialogue. The organization of speaking rights in dialogue, also referred to as the *conversational ball*, the *floor* or the *turn*, as well as controversies around which disfluencies are suitable to manage this right to speak will be described in more detail in section 2.5. Since the early stages of disfluency research, the notion is present, yet not universally agreed upon, that hesitations have turn-holding potential: in their foundational study, Maclay and Osgood (1959, p. 41) state:

> "*Let us assume that the speaker is motivated to keep control of the conversational "ball" until he has achieved some sense of completion. He has learned that unfilled intervals of sufficient length are the points at which he has usually lost this control - someone else has leapt into his gap. Therefore, if he pauses long enough to receive the cue of his own silence, he will produce some kind of signal (uhm, uh, or perhaps a repetition of the immediately preceding unit) which says, in effect, "I'm still in control - don't interrupt me!*"" (fillers transliterated)

Beside the reasons for hesitation, there are several possible ways to hesitate, which can be clustered and combined:

1. Silence.

2. Producing content-free speech material.

3. Lengthening speech material.

Silence is a problematic way to hesitate as it yields the risk of losing the conversational floor due to barge-ins, as the interlocutor might infer that the speaker has aborted uttering if the silence lasts too long. Depending on the dialogue situation, a silence might be a sufficient cue. For example, if the dialogue partners have full visual contact, the speaker might provide visual cues that she is still willing to hold the floor, but the more the mode of communication is reduced to speech, the less is silence a sufficient tool to hesitate.
So, when silence fails to buy enough time to resume fluency, the speaker has to produce something. When the problem is unavailability of content then content-free material has to be produced. Unavailability of content can arise for various reasons:

- Retrieval: the speaker has trouble finding the correct word or concept e.g. because it is a rare word, or an object she is describing is vague and ambiguous in e.g. color or shape.

- Change in the dialogue situation: the originally intended message is not relevant anymore, e.g. a station employee explaining to a passenger, why the train has not yet arrived in the very moment it arrives.

- Disturbance: the message has to be paused because of e.g. sudden loud noise or interruption by another speaker.

There are several options of deploying content-free material, which speakers make use of frequently:

- Producing a filler ("*uh*", "*uhm*").

- Producing non-committing material that does not add any information ("*I mean*", "*like*", "*say...*").

- Repeating material that has previously been uttered ("*the situation is ... the situation is ...*").

In this thesis, the focus will be on fillers in the classical sense, but it is a likely extension for future work to expand the filler armory by equipping it with repeats and non-committing words.
The third way of hesitating is lengthening of syllables and phones. Lengthening can be applied to the end of the actual content or to filler material in the middle of the hesitation period. Thus, depending on the dialogue situation, there can be great overlap between lengthening and fillers. In the following sections, the three hesitation elements will be examined in more detail.

2.2.4 Lengthening

Lengthening is a common feature of speech and is in its default form a cue for perceiving phrase boundaries (Peters et al., 2005; Turk and Shattuck-Hufnagel, 2007). A diverse and partly overlapping terminology is associated with this basic type of lengthening, e.g. phrase-final lengthening (Turk and Shattuck-Hufnagel, 2007; Umeda, 1977), utterance-final lengthening (Kohler, 1983), boundary-related lengthening (Turk and Shattuck-Hufnagel, 2007) and prepausal lengthening. The term *prepausal* in this context 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. Umeda (1977), cited in O'Shaughnessy (1995). 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. Peters et al. (2005) 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. Turk and Shattuck-Hufnagel (2007) in their detailed study also focus on lengthening near the edges of intonation phrases rather than near syntactic phrase boundaries. Example 1 illustrates the general concept of phrase-final

lengthening, i.e. the final nasal in the word *station* is longer than it would be in phrase-medial position.

"You have to go to the station: ... then take line 31 to the university."

Example 1: *Phrase-final lengthening.*

Aside from the type described above, there is another form, namely disfluent lengthening. It describes a marked prolongation of one or more phones, resulting in above-average syllable and word duration, cf. Brugos and Shattuck-Hufnagel (2012). This coincides with a local reduction in speech rate that is not expected by the listener, causing an impression of disfluency and hesitation. Phonetically, disfluent lengthening differs from other lengthening in terms of pitch contour: while boundary-related lengthening usually is accompanied by a boundary tone, disfluent lengthening exhibits a flat pitch contour (Shriberg, 2001). As such, it cues the listener that the speaker is still formulating content and thus buys conversational time for the speaker by preventing barge-ins. Doing so, the speaker deploys a less salient disfluency element as e.g. silences or fillers, which are islands in the speech signal, whereas lengthening stretches the message by ongoing phonation. I will argue in this thesis that lengthening is the first level, the starting point of hesitation intervals in speech, the softest measure a speaker can apply to solve problems in speech planning. Example 2 illustrates a hesitation cluster that starts with lengthening.

"You have to go to thee: ... uhm ... bus stop."

Example 2: *Lengthening clustered with other disfluencies: silences and filler.*

In a previous study, we 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 (Betz et al., 2015a). The rarity of these elements in our data is striking. Shriberg (2001) concluded from a large-scale corpus study, that disfluent lengthening *frequently* occurs in otherwise fluent utterances. It raises the suspicion that lengthening without contact to other disfluencies might

be *elusive* in some way. In a later study, we examined the reasons for the rarity of lengthening in the corpus and with the aid of semi-automatic detection, we could show that human annotators frequently miss instances of lengthening, which suggests that it might be a means to buy dialogue time without the listener noticing it (Betz et al., 2017b).

O'Shaughnessy (1995), in a study examining timing and phonetic properties of spontaneous speech, observes different disfluent lengthening-like phenomena, defining fluent speech as containing no hesitation, which is made up of "intrasentential pauses" and "unusual elongation of words". He lists 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)". He further observes 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 I will refer to as standalone lengthening (cf. example 3). Lengthening 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.

"You have to go to thee: bus stop."

Example 3: *Standalone lengthening with shifted vowel quality:* [ðə] *turns to* [ðɪː].

2.2.5 Silence (Silent Pause)

Silences, also often referred to as silent or unfilled pauses in the speech signal might appear trivial on first sight, but there are several issues concerning detection and classification. In his dissertation, Robert Eklund (2004, p. 160) observes:

> "*A problem with unfilled pauses is that they range from the very obvious, like a seconds-long silence in the middle of the word, to hardly noticeable silences between e.g. phrases or even sentences.*"

In figure 2.2, we see a stretch of speech of roughly four seconds that has been annotated for intervals of speech and silence. In order to make these annotations,

some assumptions have to be made. First, we would like to elaborate on the question, what silence actually is. If the answer was "absence of sound" then the first silence in the figure would not be silence, as there is, as the spectrogram reveals, audible noise, here due to inhalation. If the answer was more tailored to the scope of this thesis, it could be "silence is the absence of speech". Under that assumption, both silence intervals in the example figure would qualify as silences. The next question to be asked then would be, how long speech is needed to be absent in order to qualify as silence. This question has frequently been asked in pause research and there is no straightforward answer to it (Lundholm Fors, 2015, p. 40).

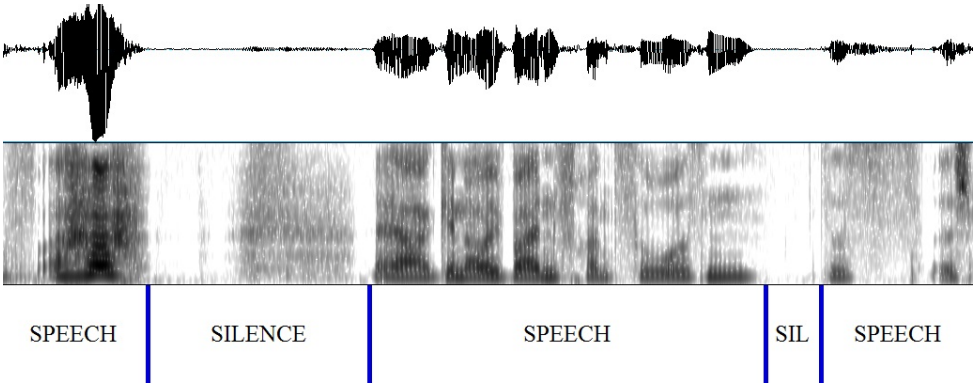

Figure 2.2: *Four seconds of speech with simple speech / silence annotation.*

Studies examining silent hesitation pauses often set thresholds for minimum and maximum pause durations. This has practical reasons, as very short pauses overlap durationally with consonant occlusion silences (Lundholm Fors, 2015; Campione and Véronis, 2002). Very long pauses are untypical of hesitations, as described in section 2.2.3, they bear a great risk of losing the conversational floor and are thus often filled to avoid it. In turn, very long silences can be explicit turn-conceding signals, which places them outside the scope of hesitation research. However, excluding pauses a priori based on duration maliciously impacts results, because they lead to wrong conclusions about durational distributions (Campione and Véronis, 2002). This thesis is interested in hesitant silent pauses, as they are the most common form of hesitation (Eklund, 2004, p. 286), but this poses an additional classification problem; as Eklund (2004, p. 162) puts it:

> "*Silences [...] are cumbersome, since it is hard to tell whether they reflect disfluency or not.*"

There is no way to classify silences as hesitant or not hesitant other than manual annotation and annotator's interpretation. For other phenomena, it is easier to infer relatedness to disfluencies or hesitation. Fillers only occur as hesitation or as part of a disfluent interval in speech, lengthening can be reliably classified into hesitation or accentuation by features such as pitch, word class, syllable position and phone class (Betz et al., 2016), but silences that do not occur in places that contradict expectations from syntax do not provide many features for classification.

One possibility would be to analyze silences that occur within disfluent intervals of speech, clustered with other disfluencies, as that might indicate hesitation function, or at least, disfluency-relatedness. There are corpus studies that dealt with durational parameters of silences, providing mixed results. Clark and Fox Tree (2002) found the type of the preceding filler to influence the duration of the silence after the filler. This was challenged by O'Connell and Kowal (2005) who found post-filler silence duration to vary arbitrarily.

For this work, it is thus necessary to approach silences from the production side. Silences can be deployed by the system as a means to buy time. It does not matter whether the listener perceives it as a hesitation, as long as the system reaches its communicative goal. Viewed this way, it is not necessary to a priori define thresholds for silence duration, as the system and the situation will determine the duration. It is conceivable, however, to set an upper threshold, as a too long silence can be misinterpreted as a signal that the conversational floor is conceded; this will further be elaborated on in the section on turn taking (2.5).

2.2.6 Filler (Filled Pause)

Fillers are probably the most studied and most renowned disfluency element. Rhetoric coaches will train speakers to minimize filler production in their speech because an overdose of fillers is said to evoke an impression of uncertainty, incompetence and dishonesty (Fischer et al., 2017). In a way, this reflects the negative image of disfluencies in classical linguistics, as something erroneous which is best

filtered out or ignored. Newer research tends to see fillers as something useful, sometimes as a cue the listener can interpret, and sometimes to the extent that the speaker produces them intentionally to signal something to the listener. Corley and Stewart (2008, p. 590) neatly summarize the status of fillers in modern research:

> "*Chief among the potentially communicative disfluencies are the so-called* fillers, *such as* um *and* uh*, which (together with prolongations and pauses) mark a hesitation on the part of the speaker.*"

In this chapter, we will consider the form of fillers, and as stated in section 2.2.3, only of fillers in the classical sense, not of closed-class words or repeated material. The communicative potential of fillers that is implied in the quote above will be discussed in more detail in section 2.4 that deals with the function of disfluencies in general, which has the simple reason that a great deal of research on the functional side of disfluencies has been carried out on fillers.
The form of fillers appears to follow a similar, near-universal, pattern in many languages: a centralized vowel and an optional nasal. As a first approach to defining the nature of fillers, we quote Ralph Rose, creator of the website titled "Filled Pause Research Center":

> "*Filled pauses include instances in which speakers utter a syllable which typically consists of a centralized vowel as a nucleus and an optional nasal coda (e.g., in English, uh/um.)*"[1]

In an analysis of repairs in speech, i.e. utterances in which the speaker interrupts herself and then continues, Levelt (1989, p. 483) notes the following about the editing expression *er*, which is similar to the English filler *uh*:

> "*A very special editing term is* er*. It is the most frequently used editing expression, used in 30 percent of all repairs. It is also the only editing expression that is practically universal; it exists, with only minor phonetic variations, in many, if not all languages. The latter should make one suspicious,* er *would be the only universal word. But is* er *a*

[1] Website: filledpause.com/taxonomy, accessed March 21st 2019.

word, or is it rather a neutral sound that is likely to occur under certain speaking conditions?“

So, as the quotes imply, there are at least two surface forms fillers can take, *uh* and *uhm*. Clark and Fox Tree (2002) established the notion that these two forms are indeed seen as two different types of fillers; they disprove the view that *uhm* is a prolonged version of *uh*. Rather both types have their own prolonged forms and are used in complimentary fashion, with *uh* denoting short pauses and *uhm* denoting longer ones.

Based on their findings, Clark and Fox Tree (2002) claim that filler use involve choice and are an intentional signal and should thus be treated as words, a concept referred to as *the signal hypothesis*. This view is challenged by researchers who view fillers and disfluencies rather as by-products of speech production problems that can be interpreted by the listener, but are not actively produced by speakers; e.g. Lickley (2015) explicitly states that Clark and Fox Tree (2002) misinterpreted his results to wrongly conclude active signaling: According to Lickley's own studies (Shriberg and Lickley, 1993), the prosody of fillers makes them *not* stand out from the preceding words, making it unlikely that speakers use fillers for signaling. If they intended to, they would rather place a pitch accent on the filler. Instead, fillers are characterized by a “continuation pitch” that matches the preceding word and the next pitch accent is placed only on the resumption or repair.

The signal hypothesis is connected to the tacit assumption that fillers, like other disfluencies, are used to signal the unavailability of the conversational floor (Maclay and Osgood, 1959). It is, however, unclear if it can be claimed that fillers generally have a floor-holding function. Eklund (2004, p. 241) conducted large-scale corpus studies on Swedish and observed fillers to be more frequent in human-machine than in between-humans dialogue. He reasons that fillers cannot per se be regarded as floor holders, as there is no risk to lose the floor in human-machine communication. Furthermore he concedes that the between-humans task was easier so that the increased amount of fillers in human-machine communication reflects planning difficulties due to increased cognitive load. This would refute the signal hypothesis and clearly position fillers into the vein of by-products of speech production issues. However, it is not clear if the speakers interacting with machines were aware that

there is no risk of losing the floor. It is possible that this condition enhances filler production because humans might infer that silence leads to system errors and make them start over with the task.
To conclude, fillers are probably the most researched disfluency element. They are also the prototypical element that retains the negative image of disfluencies, the folk notion being that fillers are to be avoided in speech. However, regardless of the signal hypothesis and the generally negative reputation of fillers, it is clear that fillers are one type of hesitation that delays the delivery of the message, and this fact can be used by the speaker as well as the listener, which is why we will discuss fillers, their function and their modeling further in later sections of this book.

2.2.7 Cutoff and Repetition

There are other aspects of disfluencies which are worthwhile mentioning, and they differ from the aforementioned lengthening, silence and fillers in the sense that they are not disfluency elements, i.e. not atoms of which disfluencies are composed, but rather symptoms that can regularly be observed in speakers' disfluent intervals. These phenomena are generally referred to as *cutoffs and repetitions*, see example 4 for illustration.

"Take the fir- the, the second elevator to the right."

Example 4: *Mid-word cutoff and repetition.*

Cutoffs (also called *truncations, abandoned words* or *abandoned utterances*) are backward-looking disfluencies, i.e. they appear on the surface only *after* an underlying event that made the speaker abort, and possibly remedy, ongoing production. Cutoffs occur abruptly and are not anticipated, like other disfluencies, by a slowing down of speech. Rather, cutoffs are generally associated with shortening of syllables than lengthening (Shriberg, 2001). Cutoffs can occur mid-word, resulting in word fragments, or between words, resulting in incomplete syntax and interrupted intonation contours (see **repetitions**).
Cutoffs, mid-word as well as between-word, are a very common disfluency phenomenon, although the exact numbers vary strongly by study and language: In

Eklund's corpus studies, *truncations* occur in about 5% of utterances (Eklund, 2004, p. 260). In one of our own studies, 19.5% of standalone disfluencies were mid-word cutoffs and about half of the disfluency clusters examined contained mid-word cutoffs (cf section 3.2). In Bear et al.'s corpus study, they find 60% of repairs containing a cutoff (Bear et al., 1992). These figures show that it is difficult to draw any general conclusion about cutoff frequency, because every study has a different object of study which the numbers relate to. In addition, cutoff rates may vary due to certain factors: they frequently occur in machine-directed speech (Brennan and Schober, 2001) and increase in number when interlocutors are interrupting (Eklund, 2004, p. 260) – in these cases there appears to be no need or no opportunity for the speaker to finish the ongoing word.

Repetitions are any *disfluent* reiterations of previously uttered material in any stretch of the disfluent interval. As Lickley (2015, p. 459) points out, there are fluent repetitions, like recurring digits in phone numbers, or emphatic stress repetitions. However, most repetitions are disfluent and disfluent repetitions are quite frequent, and like other disfluencies tend to have an affinity for function words (Lickley, 2015, p. 460).
As illustrated in example 4: the second instance of "the" replaces the first, in order to restart fluently at a phrase boundary. As can also be inferred from the example, repetitions can also occur in multiple and nested forms, to the point that they have a delaying effect like a filler. In general, repetitions can either be the repair of previous errors, or reparanda, as described in section 2.3.2. It is also conceivable that they can be deployed in an event of hesitation to buy dialogue time by repeating previously uttered material.
Repetition disfluencies are interesting from a phonetics point of view, as the repair words regularly carry a pitch accent whereas the replaced words do not, so speakers signal new or contrasting information in the resumption of fluency (Brennan and Schober, 2001). In this thesis, the focus will be on hesitations, such as lengthening, fillers and silences. A preliminary study reported in chapter 3.3 includes cutoffs for an analysis of disfluency clusters. Repetitions are not covered, however, they are kept in mind for possible elaborations of the hesitation strategy proposed here.

2.3 Models of Speech Production and Disfluencies

In the previous sections, we introduced the surface phenomena, so to speak, the *micro-structure* of disfluencies, that is the main object of investigation of this study. On a higher level, the *macro-structure* of disfluencies can be analyzed. As macro-structure we define the regular structure that is observable in any disfluent interval in speech. This regular structure is inextricably linked to speech production and its modeling. As Chafe (1980, p. 169) noted:

> "*[...] hesitational phenomena can be understood only as natural consequences of the processes which occur during the production of speech. Viewed in that way, they can be seen as contributing important clues to the nature of these processes.*"

In this section, we will briefly discuss basics of speech production models with the prime example being Levelt's blueprint for the speaker (Levelt, 1989), that enabled a theory of disfluencies, most prominently represented in Shriberg's dissertation (Shriberg, 1994). Within the discussion of Levelt's model, we will introduce the concept of *incremental processing* and rules for interruptions, which have direct implications for modeling hesitations for speech synthesis.

2.3.1 Levelt's Blueprint for the Speaker

Willem Levelt created the first full speech production model that is based on the observation of errors and disfluencies in speech corpora (Levelt, 1983, 1984, 1989). Underlying this model is Levelt's observation that corrections (in his terms: repairs) follow a regular structure (Levelt, 1983) (cf. Fig. 2.3). Based on the observations of corrections, Levelt constructed the first speech production model that explains every step "from intention to articulation", which is the title of his 1989 book. Figure 2.4 depicts his "blueprint for the speaker", which follows this pipeline workflow: A message to be conveyed begins as an abstract concept inside the **conceptualizer**, fed by general knowledge accessible to the speaker. It passes a monitoring device, which receives input from the **speech comprehension**

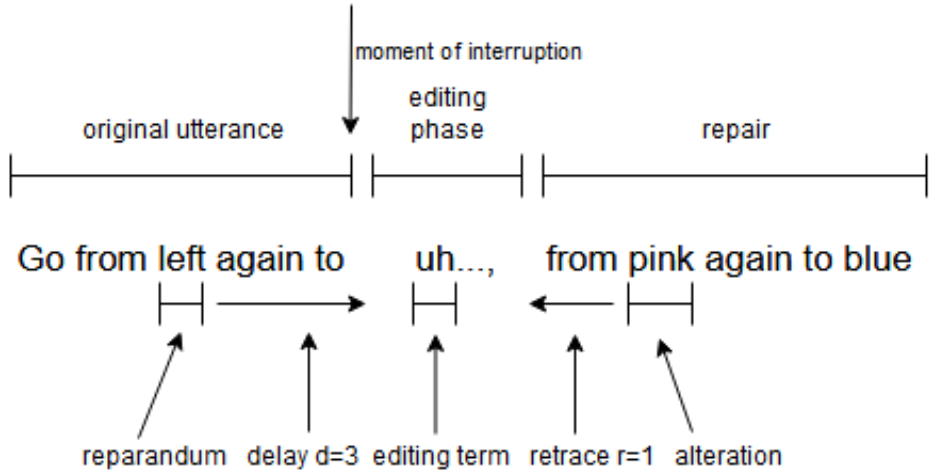

Figure 2.3: *Repair structure (adapted from Levelt 1983).*

system, which attends own speech and speech of interlocutors. The monitoring device can thus revise the message if the speech comprehension system demands it for any reason. The message is passed on in preverbal form to the **formulator**. There, the surface structure of the message is created by phonological encoding on top of grammatical encoding. To do so, lemmas and phonological forms are retrieved from the mental **lexicon**. Internally, the phonetic plan is computed and passed on to the **articulator**, which turns it into overt speech; and to the **speech comprehension system**, which can parse the yet unspoken message like it can parse own speech or speech from interlocutors. At this point, the perceptual loop is complete, as the comprehension system informs the monitor inside the conceptualizer. Crucially, whenever a mismatch between the planned speech and the produced speech is detected by the comprehension system, it will inform the conceptualizer about it where the decision is made whether this mismatch is in need of a remedy, which is when disfluencies arise. The exact shape of the resulting disfluency depends on the timing of detection and the severity of the error, as will be illustrated in the following.

2.3.1.1 Incrementality and Buffering

While Levelt's model is modular and serial in nature, it is important to conceive of the transmissions between the modules as a constant stream. A subsequent mod-

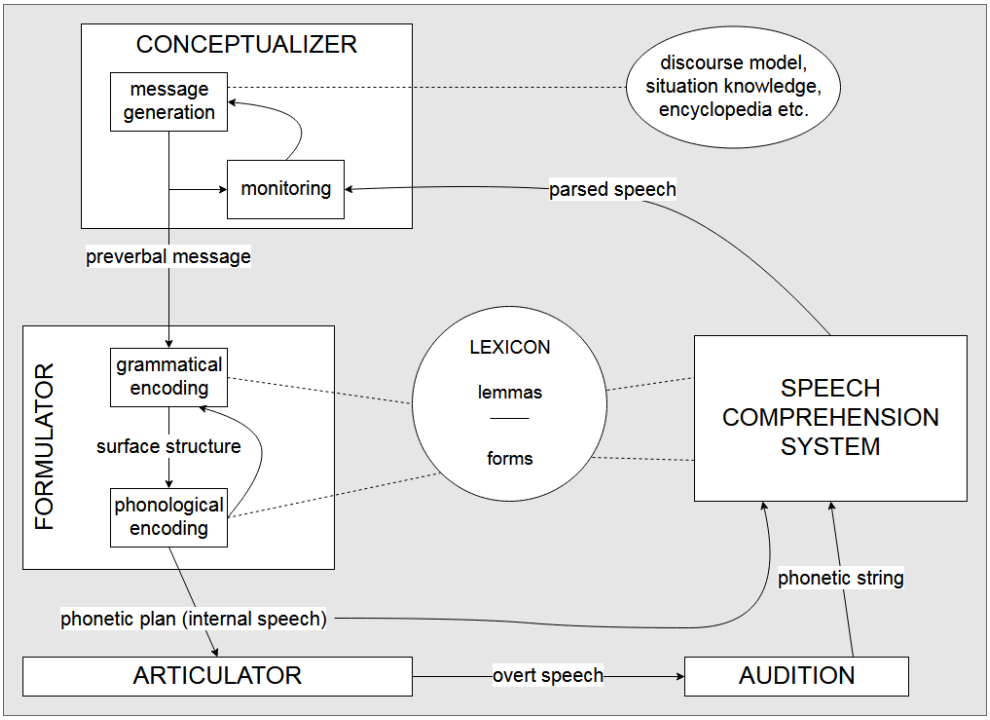

Figure 2.4: *"A blueprint for the speaker" – Levelt's model of speech production.*

ule does not only commence working when the previous module has finished processing, but rather processing is executed incrementally (Kempen and Hoenkamp, 1987). In incremental processing, each subsequent processing step is initiated as soon as sufficient information is available. Levelt (1989, p. 26) calls this *Wundt's principle*:

> "*Each processing component will be triggered into activity by a minimal amount of its characteristic input.*"

Incremental processing applies both to the speakers and their production as well as the listeners and their comprehension. There is evidence that listeners interpret utterances as they unfold over time and do not wait until it is fully uttered (Spivey et al., 2002); which is effective, but sometimes leads to complications: A sentence like "put the apple on the napkin" is syntactically complete, so listeners would be able to interpret it as "napkin" being the goal (Tanenhaus et al., 1995). However, if the speaker continues "...in the box", the listener has to revise and specify her interpretation to the goal being a napkin inside a box. Listeners furthermore

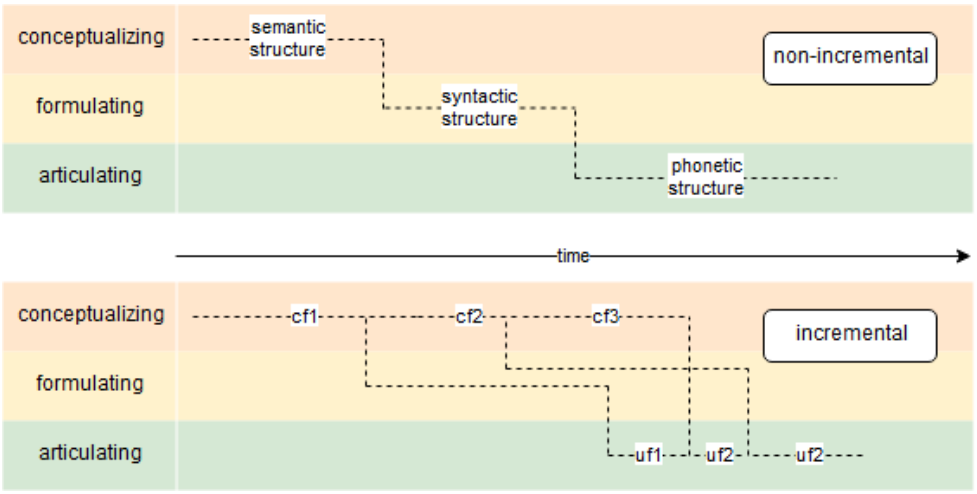

Figure 2.5: *Incremental processing, adapted from Kempen and Hoenkamp (1987). cf = conceptual fragment, uf = utterance fragment. Note that in the incremental version, the order of concept and utterance fragments needs not to be the same.*

anticipate utterance endings. Altmann and Kamide (2007) found that listeners, upon hearing "The boy will eat...", immediately search the scene for edible objects. Listeners are furthermore able to rapidly take up abandoned utterances from the interlocutor and finish them in a syntactically accurate way (Kempen and Hoenkamp, 1987). It is crucial to conceive of Levelt's production model as an incremental one: The formulator can start working based on incomplete, partial, input from the conceptualizer, the articulator based on incomplete input from the formulator, and so on (cf. Fig 2.5). If this was not the case, the rapid interaction speed typical for human communication could not be accounted for.

Incremental processing for speakers requires a lookahead in order to e.g. plan stress patterns to match upcoming phrases, or to cope with language-specific word-order constraints. This in turn requires memory capacities, of which Levelt (1989) assumes three: The *working memory* storing any information relevant to the message that is accessible to the speaker; the *syntactic buffer* residing in the formulator storing the grammatical structure of a planned utterance; and the *articulatory buffer* storing parts of the phonetic plan. These buffers thus absorb asynchronies resulting from the different processing speeds of the modules; the conceptualization and grammatical encoding of a message might be faster than the articulation,

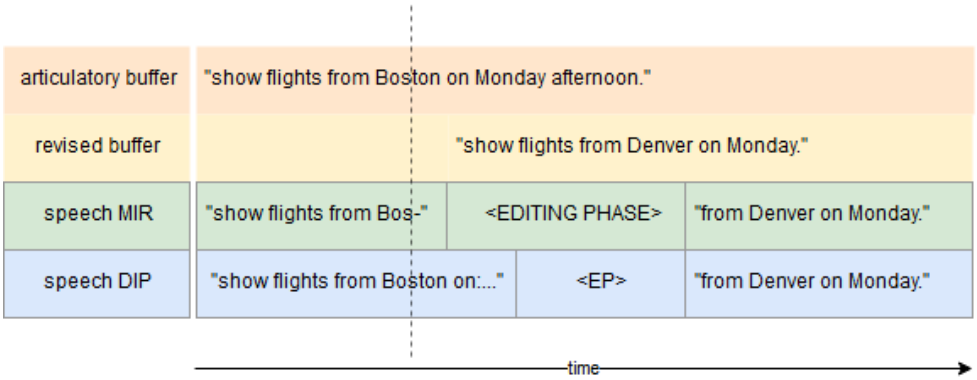

Figure 2.6: *Main Interruption Rule (MIR) and Delayed Interruption for Planning (DIP) exemplified. Speech according to MIR is interrupted as soon as possible after error detection (dashed line). With DIP, speech can be interrupted later, in this example on a function word ending on a nasal which is a frequent target for hesitation lengthening (Betz et al., 2016).*

which is subject to physiological constraints. On the other hand, if a problem is detected at any place in the pipeline, material stored in the buffers can be used to buy time for re-planning.

2.3.1.2 Interruption Rules

Levelt (1983) proposes the Main Interruption Rule, stating that speakers will interrupt their production in order to re-plan as soon as possible after error detection. They will cut off production within a word for critical errors twisting the meaning such as “left” instead of “right” and will interrupt production between words if the error is merely about appropriateness, such as “blank” instead of “white”. This would imply on the one hand that speakers prefer accuracy over fluency, and, on the other hand, that speakers implicitly signal the severity of the error by the positioning of their interruption point. Seyfeddinipur et al. (2008) proposed an alternative hypothesis, namely that speakers prefer fluency over accuracy and rather make excessive use of the material in the articulatory buffer to minimize the time interval from interruption to resumption (i.e. the editing phase in Fig. 2.3), a concept termed the Delayed Interruption for Planning Hypothesis (cf. Fig. 2.6).

For this thesis, it is important to view incremental processing as the basis of

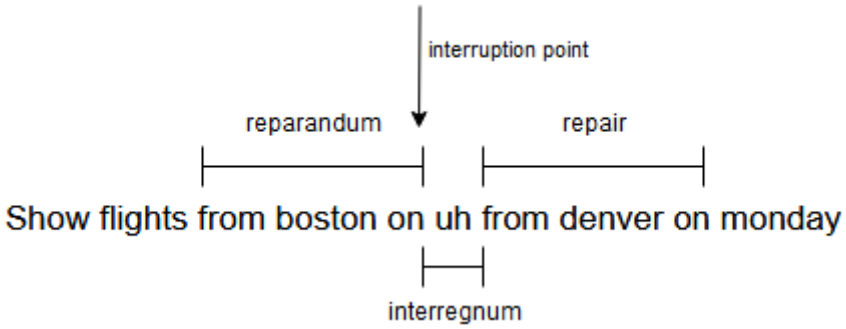

Figure 2.7: *The Shribergian disfluency structure.*

speech production. As will be shown in section 2.6.1, dialogue systems can be equipped with incremental processing capacities as well, which creates a processing pipeline quite similar to human speech production in Levelt's sense. As will be shown in section 6 we conceive of lengthening as the starting point for a hesitation interval, which in turn is initiated at the next best phonological target in the articulatory buffer (metaphorically called so, as speech synthesis does not, in fact, articulate) We thus opt to follow Seyfeddinipur et al. (2008) as an amendment to Levelt's model, which allows for exactly that kind of behavior, as can be seen in Fig. 2.6.

2.3.2 Shriberg's Disfluency Model

In the wake of Levelt's work constituting the first fully fledged model of monitoring, errors and self-repairs, Shriberg (1994) created the first approach to fully model disfluencies. While Levelt (1983) drafted a first sketch to describe repair disfluencies (cf. Fig. 2.3), Shriberg (1994) went one step further and proposed a structural description of disfluencies, which is able to describe any surface form (cf. Fig. 2.7). In terms of structure, this example can be described as a *reparandum*, a stretch of speech the speaker will revoke; followed by the *interruption point* and the *interregnum*, which is a typical place for hesitations to occur, and finally, the *repair*, providing the intended resumption of the utterance. This approach takes up some of the terminology for repair structure by Levelt (1983), using reparandum in a broader sense from error to interruption. Levelt's editing phase is in this

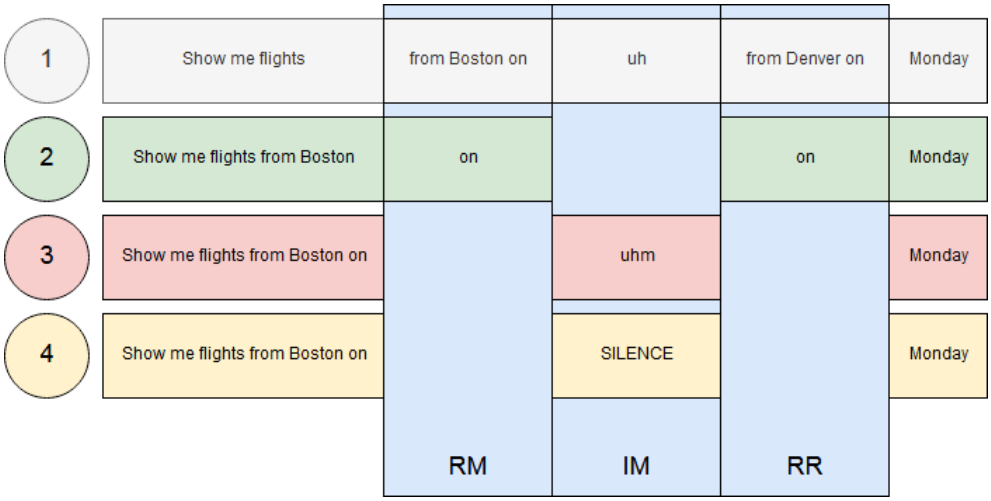

Figure 2.8: *Examples for different disfluencies describable with the Shribergian structure. All regions (RM = reparandum, IM = interregnum, RR = repair) can be conceived as ever-present, but any of these three regions can be empty.*

framework called interregnum to more neutrally denote the time span between interruption and repair, which makes sense: as can be seen in Fig. 2.8, the interregnum can, like any other interval, be empty, which would be an odd feature for something called the editing phase. This structural description does not cover the type of resumption (cf. section 2.3.3) nor the amount of delay or repair as Levelt's account did. The main reason being that it is possible to derive annotation schemes from Shriberg's structure which can be analyzed automatically to derive this kind of information, so there is no need to label it inside the annotation. In general, as will be shown in the following, this structure can be used to describe disfluencies of any level of complexity. The most striking feature is its built-in property to reveal the fluent utterance by removing the reparandum and interregnum (cf. example 5).

"Show flights ~~from Boston on uh~~ from Denver on Monday"

Example 5: *Removal of reparandum and interregnum yields a fluent utterance.*

This structure is capable of describing also more complex and nested forms of disfluencies. These multi-layered cases can better be exemplified in textual form. A number of annotation schemes for disfluencies exist (see Hough et al. (2015)

for an overview) that feature coding for the different regions and the interruption point (cf. example 6).

"Show flights (from Boston on + {uh} from Denver) on Monday"

Example 6: *Annotation scheme by Hough et al. (2015) exemplified on the current example utterance: (marks the beginning of the reparandum. + is the interruption point. { } encapsulate the interregnum.) marks the end of the repair.*

Now consider an utterance with disfluencies inside disfluencies, such as "Show flights to Boston well to from uh Denver on Monday": Using the same annotation scheme, this can be structured as example 7 illustrates:

"Show flights (to Boston + {well} (to + from) {uh} Denver) on Monday"
"Show flights ~~to Boston well to~~ from ~~uh~~ Denver on Monday"

Example 7: *Nested disfluency. "To Boston" is ultimately repaired by "from Denver". In between, "to" is repaired by "from" with an empty interregnum. In addition "uh" occurs as a standalone filler, structurally a disfluency with only an interregnum. Second line shows the fluent utterance when all reparanda and interregna are removed.*

The disfluency structure as described by Shriberg (1994) provides the basis for all analyses of disfluency phenomena that are related to annotation of corpus data and to automatic speech recognition. Especially the ability to infer the lexically fluent utterance by means of excising reparanda and interregna proves very useful: it provides a "sanity check" for annotations, which are only valid if the fluent utterance is deducible; on the technical side, automatic speech recognition can make use of this property to identify the intended meaning.
For this thesis, however, we have to go beyond the Shribergian structure. The goal is synthesis of surface phenomena which are not directly covered; e.g. lengthening disfluencies can be expressed within the Shribergian structure only as phonetic variation applied to any word. This distinguishes lengthening from fillers which would always be identified as (part of) an interregnum. This, however, might not be the whole story. Lengthening, as shown in section 2.2.3, is one way of hesitating and should thus be viewed like fillers or silences, i.e. as a disfluency element in its own right, in the interregnum, which is how lengthening is regarded in this thesis.

2.3.3 Resuming Speech After Disfluencies

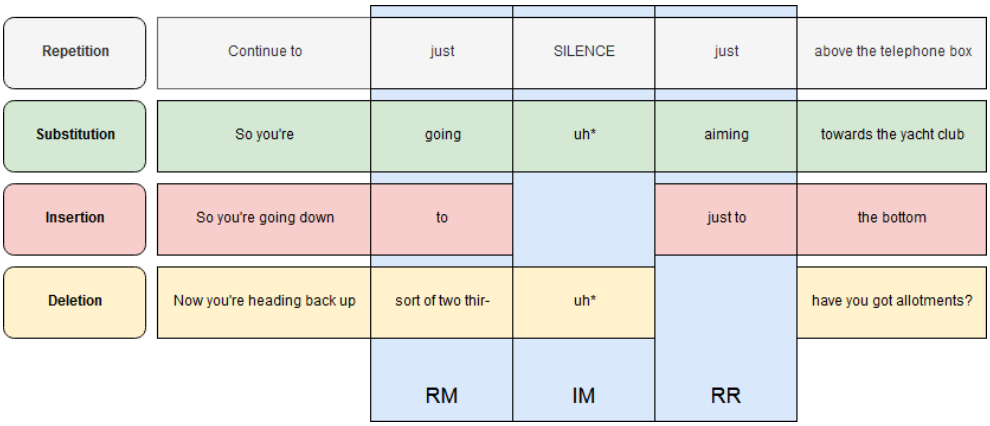

Figure 2.9: *Different ways to continue after a disfluency. Examples adapted from Lickley (2015). Fillers marked * are inserted for illustration purposes. Empty intervals are null time.*

Whenever a disfluency breaks up the structure of the originally intended utterance, there are multiple paths of continuation the speaker can take, if she is given the chance and her turn is not taken. As shown in figure 2.9, there are four general types of surface forms disfluent intervals in speech can take. They can be in general described by the Shribergian structure and their type is named after the way the speaker continues after the disfluency. They are mentioned here for the sake of completeness on the one hand, as these terms are frequently used in disfluency research; and on the other hand to show possible ways a dialogue system could proceed after the insertion of a hesitation.

Repetition. Parts of the original utterance are repeated. In the example, there is one word repeated, but it could also be an entire phrase or a cutoff word which is then repeated fully.

Substitution. Parts of the original utterance are replaced by other lexical items, often a more precise or appropriate one. As with repetitions, the reparandum can consists of word fragments, words or phrases.

Insertion. The utterance is interrupted because something is to be inserted in a position prior tho the interruption point.

Deletion. Parts of the original utterance are revoked entirely and the speaker

continues with entirely different content. This case of disfluency can also break up the syntactic structure that would normally be intact after removing RM and IM.

2.4 Disfluencies as a Conversational Tool

The terminology of fluency versus disfluency seems to imply a dichotomy. In fact, both are hard to define concepts that, on closer inspection, might have more in common than the contrasting word forms suggest. Among the first to suggest that disfluencies might be something useful was Wallace Chafe (1980, p. 170):

> "*I would like to suggest on the contrary that the speaker's chief goal is to get across what he has in mind [...] The speaker is interested in the adequate verbalization of his thoughts. Pauses, false starts, afterthoughts, and repetitions do not hinder that goal, but are steps on the way to achieving it.*"

Following this idea, as described in section 2.3, models of speech production were created that are able to explain the occurrences of disfluencies as a consequence of speech planning and monitoring processes. With this framework established, many studies from the 1990s onward investigated these novel views on disfluencies experimentally. Fox Tree (1995) was among the first to empirically study disfluencies from a psycholinguistic point of view. She comments on the contemporary view on disfluencies as follows:

> "*A common assumption is that [...] hesitations [...] slow understanding. If fluency is every speaker's goal, then disfluency is every listener's nightmare.*"

In her 1995 study she disproves this statement by providing evidence that some disfluencies actually facilitate understanding rather than hindering it. This study is followed by several other attempts to empirically prove the usefulness of disfluencies.

Brennan and Schober (2001) conduct a series of experiments, showing that listeners compensate best for errors in speech when a cutoff word is accompanied by a filler. They elaborate that the form of the filler is not important, rather the

time elapsing aids comprehension. They argue that disfluencies are not superior to fluent speech, but that their presence helps to mitigate the effects of mishaps in speaking. They further state that they do not have any evidence that speakers make such choices deliberately and intentionally.
Clark and Fox Tree (2002) argue that fillers are words and are uttered with specific communicative intentions in mind, namely commenting on the speakers' own performance, enabling the listener to act accordingly. This is one of the strongest claims that disfluencies are not only beneficial, but rather an intentional signal, a claim that has kindled heated debates in the community, in which most researchers formulate rather carefully that it is not clear whether speakers produce hesitations deliberately (cf. e.g. Barr (2001), Brennan and Schober (2001), Corley and Stewart (2008)). Further examining the potential of disfluencies, Arnold et al. (2004) found that disfluencies, in addition to aiding comprehension, also bias the listener towards new rather than given objects, and also unfamiliar and hard to describe objects (Arnold et al., 2007).
The empirical proof of the potential of disfluencies to actually facilitate comprehension rather than hindering it leads to a shifted theoretical view. Some contemporary research undoes the distinction between fluency and disfluency entirely. Götz (2013, p. 8) coins the term *fluenceme*, defining it as follows:

> "*A fluenceme is an abstract and idealized feature of speech that contributes to the production or perception of fluency, whatever its concrete realization may be.*"

Among fluencemes of production she lists phenomena that are traditionally subjects of disfluency research: silences, fillers and repetitions (Götz, 2013, p. 9)[2]. It is quite revealing that the latter two terms are listed under *fluency-enhancement strategies* and *speech management phenomena*, which contrasts the traditional view of these phenomena as *disfluencies*, literally, "things that lack fluency".
While Götz's term *fluenceme* overtly describes hesitation phenomena with a positive term inferring fluency, the idea is not novel. The earlier work of Jens Allwood already worked under similar premises: he avoids the term disfluency altogether

[2]Terms were translated according to the terminology used in this thesis. In Götz's book they are referred to as *unfilled pauses*, *filled pauses* and *repeats*, respectively.

and describes these phenomena as *speech management* (Allwood et al., 1990) and *own communication management* (Allwood, 1995), viewing them as tools to buy time, and to reorganize own utterances.
In the light of this development in research, it comes as no surprise that disfluencies have also received interest from dialogue system designers: Something beneficial for human communication might also work for human-machine communication. This notion will be elaborated on in section 2.6.1.
These are some examples for the modern view on disfluencies, partly characterized by their abandoning of the term itself. Disfluencies are items in the toolbox of the speakers to facilitate their messages. With this notion in mind, the next section gives a brief introduction to turn taking as the key concept of human conversational interaction; then the background part is concluded with a chapter on disfluencies and their potential for conversational speech synthesis and dialogue systems.

2.5 Turn Taking

In conversation, it is not always possible to keep the turn, or the "conversational ball". Following Maclay and Osgood (1959), disfluencies occur in places where other speakers are likely to take over the turn, and where the speaker seeks to secure the right to speak. When synthesizing hesitations in dialogue systems, there must be awareness for the risk of conceding the floor, so a brief examination of disfluencies and turn-taking is in order.
The first seminal study to introduce turn taking was by Sacks et al. (1974) who found that turn-taking in human conversation happens at a rapid pace across languages. Turn changes happen with minor delay of some milliseconds, or even with slight overlap. Planning speech takes time, so the pace with which turn taking happens dictates that the interlocutor must plan responses ahead and uses cues in the speech signal (and, if possible, gestural and facial expression signals) to project upcoming *transition relevant places*, which are theoretical points in dialogue where speaker change is possible (Sacks et al., 1974). Lundholm Fors (2015) suggested that, instead of a point-based approach like transition relevant places, there is *turn change potential*, an ever-present value in dialogue, with different levels of

intensity. Places with high turn-change potential levels will often correspond to transition relevant places, but the potential for turn change is not restricted to certain areas in dialogue.

In fluent speech, or rather, in speech devoid of obvious disfluencies, there are several cues for turn keeping and turn conceding. These cues, in terms of language are mostly related to lexical, syntactic and intonational completeness (Bögels and Torreira, 2015). Complete structures invite interlocutors to take over, whereas incompleteness asks them to wait for the speaker to complete the message. It can be argued, that disfluencies play a crucial role here as they occur often at places of incompleteness.

There is debate how many of these cues need to be present to signal availability or unavailability of the turn. Bögels and Torreira (2015), in line with previous studies suggests that an interplay of many cues is involved and therewith contradicts De Ruiter et al. (2006) who proposed that lexico-syntactic cues are sufficient and intonation cues neither sufficient nor necessary. Gravano and Hirschberg (2009) found that the likelihood of speaker change increases linearly with the number of turn-conceding turns present.

When speakers hesitate, because they ran out of things to say temporarily, they have to take action if they want to keep the turn. Otherwise, the impression of completeness arises and the interlocutor takes over. Following this notion, it has been proposed that hesitations like fillers are devices a speaker could deploy to avoid losing the floor prematurely (Maclay and Osgood, 1959). Eklund (2004) suggested that the phonetic properties of lengthening make it a better candidate as the ongoing vocalization of the current utterance chunk makes it more obvious than a filler that the floor is not up for grabs.

This idea receives support from turn-taking research. Gravano and Hirschberg (2009) found that "final lengthening is more prominent" in turn-medial than in turn-final position. They confirm the notion that a flat pitch contour signals turn-keeping, whereas all other pitch contours signal turn-conceding, for example rising intonation indicating a question, falling intonation signaling completeness of a phrase. As will be shown later, flat pitch contours are also typical for hesitations like fillers and hesitation lengthening (as opposed to emphatic lengthening), which highlights the innate turn-keeping potential of hesitations.

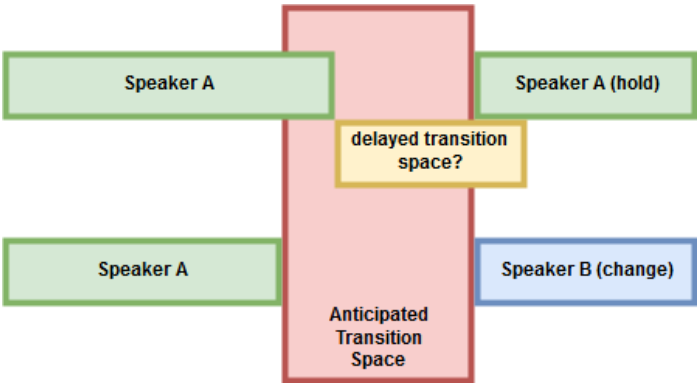

Figure 2.10: *Zellers' transition space model.*

Margaret Zellers elaborates on the idea of lengthening being a turn-keeping cue and proposes a turn-taking model (cf. Fig. 2.10) with a transition space. This space (which is actually a time frame) corresponds to a transition relevant place or to a place with high turn change potential. This space can be intruded by a lengthening of the current utterance, this intrusion delaying the transition space, thus signaling the unavailability of the turn to the listener.[3]

Later in this work, we discuss lengthening as a subtle means to buy time for the speaker, which is very unmarked and frequently missed by the listener. It is an interesting topic for future research to determine if lengthening can be a floor-holding device at the same time, which linguistic parameters determine that, and if it holds true for synthetic speech as well.

2.6 Disfluency Synthesis

2.6.1 Disfluencies in Spoken Dialogue Systems

Dialogue systems are programs that communicate with users in text and/or speech form. They are generally distinguished into task-oriented dialogue agents and chatbots. The latter are designed for extensive conversations, for entertainment or practical application, traditionally in text form. The former are designed to interact with the user in a limited domain in short task-oriented conversations, for

[3]Based on a talk by Margaret Zellers and personal communication at Bielefeld University in January 2019.

example to give directions or control home appliances. Well-known present-day examples would be Siri, Alexa, or Google Home. These current task-oriented dialogue systems are based on speech input and output. The scope of application used to be limited to small domains, but the range of interactions became more like spoken conversation between humans as more computational power and better speech synthesis became available. One major shortcoming of these systems remains to be their lack of adaptivity that stands in contrast to many of their fields of application.

Modern state-of-the-art dialogue systems produce adequate responses to user requests, but they do so in a rather static fashion: the response is delivered in one fully specified, potentially lengthy, or semantically complex utterance. This typically leads to response times that are much higher than the usual promptness of turn-taking observed in human communication (Sacks et al., 1974; Heldner and Edlund, 2010). Also, there currently exists no satisfactory solution as to how systems handle interruptions during the delivery, such as the user refining the request or the user being distracted: some systems can be halted mid-utterance, but as of now, these interruptions do not work authentically like in human communication (Wester et al., 2017). It could thus be stated that many of these systems are less interactive than they should be.

Addressing the adaptivity and interactivity issue, a strand of research evolved that aims to develop *conversational dialogue systems* that are capable of *talking* instead of merely *reading* out pre-defined responses. One key feature on the way to more interactivity between system and user is incrementality.[4] Human dialogue does not work like a ball-tossing game, but rather simultaneously: Responses are planned while the interlocutor is speaking (cf. sec. 2.3.1.1). It can be shown that limited-domain dialogue systems can make use of incremental processing to achieve human-like interaction speed (Skantze and Schlangen, 2009). This gain in velocity helps dialogue systems to respond faster and to be able to react to external changes (attention shifts, barge-ins etc.), making the interaction conversational, as opposed to an exchange of monologues.

[4]In this study, the focus is on incremental spoken dialogue systems. It is worthwhile noting that it was recently demonstrated that an interactive system capable of handling interruptions can be built without incremental processing (Wester et al., 2017).

Hesitations are a useful feature for incremental spoken dialogue systems. On the one hand, these systems might need to buy time for re-planning and can deploy hesitations while doing so. On the other hand, the incrementality enables the system to hesitate immediately and flexibly. To develop conversational dialogue systems, various approaches have been proposed, with incremental processing, with various forms and functions of hesitation and with both incrementality and hesitations. While our focus in this work revolves around time buying, we will briefly discuss other functions of disfluencies in dialogue systems here, in order to create awareness for potential side-effects and alternative interpretations that can result from the insertion of hesitations.

Adell et al. (2010) enriched a non-incremental speech synthesis system used for translation with the capability to produce fillers, taking into account prosodic modifications observed in human filler production. Skantze and Hjalmarsson (2013) built an incremental system based on a general, abstract model for incremental processing (Schlangen and Skantze, 2011) that employs turn-initial hesitations ("eh", "well", "wait a minute") to buy time to generate a response (or in this case, time for the wizard to type the answer). This system exploits the fact that hesitations do not commit content to the conversation. They can literally be used as fillers to bridge gaps in dialogue. Building upon the incremental processing model (Schlangen and Skantze, 2011), further studies with incremental processing and hesitations have been conducted: Kousidis et al. (2014) conducted an experiment in a driving simulator, during which a virtual assistant told the driver about appointments on that day. It was shown that a system that hesitates by means of silences whenever a difficult situation occurs improves both the participants' driving performance as well as their recall of information presented during the task. Carlmeyer et al. (2016a,b) used hesitations (phonetically realized as silences) as a user-oriented strategy, based on observations of the human interaction partner. They investigated the effect of self-interruptions as a strategy to regain the visual attention of distracted users in a smart-home setting with a virtual agent. They showed that insertion of silence whenever the attention of the users shifts away has a positive effect of the attention of the user, but at the cost of less positive subjective ratings. In a similar scenario, the authors showed that incremental information presentation leads to a better task performance (Chromik et al., 2017). Whereas

the authors were able to show that listener-oriented insertion of hesitations (realized as silences) has a positive effect on the interaction, the self-interrupting agent was perceived less friendly in all three studies.
Bohus and Horvitz (2014) used hesitations in human–robot interaction as a disengagement strategy. A directions-giving robot produces lexical hesitations ("so...", "let's see...") after its own speaking turns to bridge the awkward silence during which the user has to decide whether she wants to continue the interaction or not. Interestingly, this usage of hesitations is contrary to many other studies that highlight the usefulness of hesitations to gain attention and to *continue* interacting.
All systems presented here reported positive effects on the interactivity. Not all systems evaluated speech synthesis quality, but those that do report negative effects. This hints at a shortcoming, namely a trade-off between interactivity and sound quality that is a key issue for current and future research in this field. Evaluation of speech synthesis is not a trivial issue, and neither is the evaluation of synthesis with special features such as disfluencies. In the following section, we will briefly discuss issues to be considered for the remainder of this work.

2.6.2 Synthesis Evaluation

When it comes to evaluating speech synthesis, the researcher faces a problem: there is a trade-off between field-testing systems and testing them in a lab. Field tests are necessary to prove the ecological validity and applicability of a system, while lab tests are needed for fine-grained tuning of individual components of the system (Wagner and Betz, 2017; Möller, 2017). This dilemma between field and lab testing is a general one in speech research and other research areas: the lab ensures control over the testing variables, without being able to make predictions about the real-world applicability. In speech synthesis evaluation, the general evaluation paradigm is lab-testing of isolated sentences asking listeners for ratings of intelligibility and naturalness (Wagner and Betz, 2017). It can be argued that this does not reflect the manifold application areas for speech synthesis. Further, naturalness appears to be an ill-defined concept in this matter (Dall et al., 2014b). It is understood as how close to a human voice a speech synthesis system is rated, but a human voice is not necessarily a good baseline for testing. It is question-

able how human-like voices should be at all: a well-known phenomenon related to this question is the uncanny valley (Mori, 1970), reporting negative impressions of robots that are too strongly resembling humans where actually some features of artificiality are expected. This might translate to speech synthesis research, where every application scenario is different and not every scenario involves human-like speech producing entities. In recent years, claims have arisen to move speech synthesis evaluation towards more interactive paradigms rather than testing them in the void (King, 2017; Wagner and Betz, 2017; Betz et al., 2018).

Synthesis capable of producing disfluencies is a special subgroup of speech synthesis systems. Thus, it comes as no surprise that there is no global evaluation standard for these systems, given that general synthesis evaluation lags a bit behind its time. Rather, each study investigating aspects of disfluency synthesis provides an evaluation paradigm of its own. Many follow the standard synthesis evaluation paradigms such as MOS scales (eg. Adell et al. (2010); Betz et al. (2015b)), or dialogue system evaluation methods like task performance (e.g. Carlmeyer et al. (2016a,b) or try to introduce methods novel to speech synthesis evaluation, like game scores, focus groups or task efficiency (e.g. Betz et al. (2017c); Wester et al. (2017); Betz et al. (2018) respectively).

Many of the aforementioned disfluency synthesis evaluation methods have shortcomings. Asking explicitly for feedback on the disfluencies produced (as in e.g. Adell et al. (2010)) might prime users to respond in a way they believe the authors intend and give ratings biased in favor of disfluent systems (Dall et al., 2014a). Asking for feedback without explicitly mentioning the disfluencies, as for example done in one of our own studies (Betz et al., 2015b) which will be discussed in more detail in section 3.3, yields the opposite problem: results will not be biased towards disfluencies, but feedback on overall sound quality might or might not reflect the influence of disfluencies. The same is true for performance-based measures, which are likely telling about the system as a whole, but cannot directly be taken as an indicator of disfluency synthesis quality.

Moreover, while there has been successful attempts of equipping dialogue systems with hesitation capabilities (cf. sec. 2.6.1) there are other studies challenging the notion of positive effects of hesitation in synthetic speech. Carlson et al. (2006) found that listeners are able to perceive hesitation in synthetic speech. They at-

tribute most impact to temporal patterns that deviate from expectations. In their experiment these were silent pauses combined with lengthening of the consonant occlusion at positions that were not predictable from syntax. They further noted that pitch and creak might play supporting roles and that synthesis should not rely on singular cues for hesitation only. Dall et al. (2014a) investigate whether the psycholinguistic results that suggest a positive effect of fillers on information processing on the side of the listeners are replicable and transferable to speech synthesis. They find that the positive effects disfluencies have on reaction time in studies with human speech are lost when synthesized, and attribute this mainly to the difficulty to produce synthetic fillers with the original vowel sound.
This highlights a need for further investigations, both in terms of evaluation paradigms for conversational speech synthesis and dialogue systems as well as on the impact of synthesized hesitations themselves. In the next section, we conclude the background part by giving an overview over different types of speech synthesis and their capability to realize disfluencies.

2.6.3 Types of Speech Synthesis

There are several technical approaches to speech synthesis. Here, the most important ones will be summarized and their applicability for conversational speech synthesis and phenomena such as disfluencies discussed:

- USS: unit selection synthesis
- SPSS: statistical parametric speech synthesis
- SPSS-HMM: SPSS with Hidden-Markov-Models and regression trees
- SPSS-DNN: SPSS with deep neural networks

To very briefly summarize the underlying architectures, USS makes use of large databases of speech recorded by a single speaker. The database is controlled in a way that any possible segment combinations and transitions are featured multiple times. It thus needs carefully selected data of speech that matches the desired speaking style.
USS has been the state of the art in many commercial systems for a long time

(Taylor, 2009). The sound output quality is relatively high, although clearly distinguishable from human speech: its major drawback is the lack of flexibility and adaptivity. It is in general not possible to produce an alternative speaking style without recording an entire new database. Furthermore it is not possible for USS systems to quickly adapt to changes in the dialogue environment.
When it comes to disfluencies, USS runs into problems: while it is generally not problematic to insert silences (by simply pausing the system) or filler words, it is difficult to produce lengthening or fillers, because the phone duration and vowel qualities which are necessary for that are typically not included in the database (Dall et al., 2014a).
The advent of parametric systems from the late 1990s onward challenged the dominance of USS systems (Tokuda et al., 2013): SPSS systems solve the flexibility issues USS systems have. SPSS uses simple machine-learning techniques to train a set of rules on top of which parameters of output speech can be produced and modified. Usually these are based on Hidden Markov Models (HMMs) and regression trees. While being more suitable for adaptive systems and requiring less data size, the sound quality is generally lower than in USS systems. Additionally, the training corpora underlying the SPSS systems are the same type as in USS systems, "cleaned" of disfluencies, as disfluencies serve no apparent purpose for the basic applications of these systems. This in turn implies that, despite adaptivity, filler synthesis is still problematic, as the neutral vowel sounds which are typical of fillers are not part of the training set (Dall et al., 2014a, 2016)
As the last decade saw a huge increase in computational power, computation-intense new synthesis methods have arisen, based on deep neural networks that can, based on big data input model speech output at tremendous quality (Van Den Oord et al., 2016; Arik et al., 2017; Wang et al., 2017, 2018). These approaches replace parts of, or the entire classical synthesis architecture by DNNs, leading to a massive increase in output quality. Their training data needs not to be cleaned of spontaneous speech phenomena, so synthesis of hesitations is not a problem, however, at the time being, no research has been published that investigates the applicability of these new systems for conversational and interactive speech synthesis. WaveNet is stated to be rather slow on the way of achieving good audio (Wang et al., 2017). Tacotron is developed by Google Inc. which has immense

computational power at their disposal, without which, according to external testing and replication, a system like Tacotron takes several weeks of training before being applicable.[5] This means that the greatest achievements in terms of audio quality are at the moment –unsurprisingly– only reached by commercial systems and not open for free research and replication.

The speech synthesis itself can only utter content as fast as the underlying speech planning allows for. It has been shown that incremental processing, combined with SPSS synthesis can interact as rapidly as human turn-taking functions, however with the usual shortcomings of sound-quality (cf. section 2.6.1). Given enough computational power, Wester et al. (2017) showed that better sounding synthesis systems can be bestowed human-like interaction speed as well, by planning multiple utterance threads simultaneously between which can be switched. This is to date the only known example of a system that can facilitate interruptions at human-like turn-taking speed that does not make use of incremental processing. It is up for future research to see if that approach can be extended and used as a replacement for incremental processing in conversational systems.

A promising contemporary approach that appears able to combine many advantages of different approaches is the Merlin toolkit (Wu et al., 2016). It is an open-source example of an SPSS system that replaced the HMM component by DNNs and is thus capable of harnessing the dynamic properties of parametrical synthesis and maintaining a high sound quality at the same time. To summarize, each type of speech synthesis yields different advantages and disadvantages (see table 2.1 for an overview). For the goal of this work, an SPSS approach is most suitable, as we are aiming for maximal flexibility. The best approach in the vacuum would be the Merlin toolkit (Wu et al., 2016) with its combination of SPSS and DNN, but it was released after the first studies had already been carried out. So in this work, studies we describe subsequently, will be based on MaryTTS (Schroeder and Trouvain, 2003), which is a simple, open-source synthesis software which can be easily adapted to our needs and which is implemented in inProTK, a toolkit for incremental processing in dialogue systems (Baumann and Schlangen, 2012).

We have now established a background of disfluencies and their technical ap-

[5]Information found on a research blog, accessed May 3rd 2019: https://medium.com/@rajanieprabha/tacotron-2-implementation-and-experiments-832695b1c86e

Speech Synthesis Types				
Potential	USS	HMM	Open-Source DNN	Commercial DNN
sound quality	+	-	+	++
style variation	-	+	+	++
live adaptivity	-	+	+	-
data sleekness	-	+	+	-

Table 2.1: Potentials, advantages, disadvantages of different synthesis types.

plicability and realizability for dialogue systems. On this basis, the next part of this thesis will be devoted to empirical investigations leading towards a model of hesitation insertion for incremental spoken dialogue systems.

Part II

Empirical Investigations

Chapter 3

Disfluency Basics for Conversational Speech Synthesis

3.1 Introduction

This chapter marks the beginning of the second major area of this work, which is empirical investigations of disfluencies. The ultimate goal of this thesis is creating a model for hesitation insertion into speech synthesis and spoken dialogue systems. This empirical part provides foundational research on several basics of hesitations, disfluencies, their synthesis, and human interactions with these synthesized phenomena. In this chapter we start out by investigating the possibilities of synthesizing disfluencies using open-source speech synthesis software. Based on this, we conduct an in-depth analysis of hesitation lengthening (chapter 4), which appears to be a fruitful candidate to fulfill our needs in dialogue system development (as will be shown in section 3.3).

3.1.1 Scope

Having outlined disfluencies and their potential for spoken dialogue systems, this part will pave the way for a model of hesitations for dialogue systems. The scope of this model is limited, for the simple reason that its focus is the practical application, meaning, in this case, synthetic production. It is not intended to provide a general descriptive model, that effort has already been undertaken by other

researchers (cf. section 2.3). It is thus not our intent to be able to produce *any* conceivable form of hesitation, but to equip dialogue systems with a basic model to produce *plausible* hesitations. This model is intended to predict the optimal starting point and duration of a disfluency interval as well as the shape of hesitations therein. In this, the model is a novel contribution to the field of conversational dialogue systems. The model is designed specifically for incremental spoken dialogue systems (cf. section 2.6.1), utilizing the metaphorical articulatory buffer, i.e. the readily planned but not yet synthesized parts of the utterance to facilitate human-like micro-timing in speech delivery (cf. section 2.3.1.1).

3.1.2 Aim

The aim is to enable incremental spoken dialogue systems with the capability to react to a generic external event with entering hesitation mode for as long as either the model allows or until the external event is over. This will be facilitated by deploying lengthenings, silences and fillers based on rules that are derived from corpus analyses and experimental investigations. We will argue that a sophisticated modeling and insertion of these elements is sufficient to create elaborated synthetic hesitations.

3.2 Corpus Study 1: Human Disfluencies

3.2.1 Method

As a starting point for hesitation and disfluency synthesis research, we conducted a small-scale corpus study to analyze real-world occurrences of disfluencies. For this, we used a part of a spontaneous speech corpus that has been compiled at Bielefeld University. It features dialogues of students furnishing an imaginary apartment, eliciting, among other spontaneous speech phenomena, a large amount of disfluencies due to the high cognitive engagement of the interlocutors (Kousidis et al., 2013). The corpus is hand-labeled by annotators trained to look for disfluencies. We used data of two speakers, who were familiar to each other, totaling in 27 minutes of dialogue.

The focus of this study is on frequencies, distributions and duration of four surface forms of disfluency, which occur in forward-looking (hesitation) situations as well as in backward-looking ones: lengthening, mid-word cutoffs, silences and fillers. The elements were selected because they are common phenomena in spontaneous speech. This study is intended as a general starting point for disfluency synthesis, which is why we analyze not only hesitation phenomena, but also elements typically occurring in corrections like cutoffs.
Lengthening is a more difficult to grasp phenomenon compared to other disfluencies. To call into memory from the background chapter: lengthening describes a marked prolongation of one or more phones, resulting in above-average syllable and word duration, which coincides with a local reduction in speech rate that is not expected by the listener, causing an impression of disfluency and hesitation. There is consequently no hard definition of how long lengthening needs to be in order to be perceived as hesitation, which is why we analyze in this study how lengthening differs from slow speech. We therefore investigated the locus of syllabic lengthening in terms of phone duration and additionally checked for influences of speech rate on lengthening extent. In order to address these issues, we supplemented the durational analysis by also measuring the surrounding local speech tempo as indicated by syllable duration. We obtained the duration for the three preceding and the three following syllables, where available. Based on this corpus study, we then conduct a perception test in which synthetic realizations of these elements will be investigated (section 3.3).

3.2.2 Results

3.2.2.1 Frequency and Distribution

As a first step, we looked at the frequencies and distributions of the disfluency elements in question. Within this part of the corpus, 77% of these disfluency elements occur in standalone fashion, the other 23% of the occurrences were in clusters, cf. table 3.1. While this is a very small-scale approach, and the insights gathered here cannot be more than tendencies, it is interesting to observe that silences are the most frequent disfluency element, which is in line with previous studies. Fillers are less in number than initially expected, with counts below

Standalone disfluencies (77%)			**Disfluency clusters (23%)**		
Type	**Count**	**%**	**Type**	**Count**	**%**
Lengthening	15	7.7	Len. + cutoff	3	5.2
Cutoff	38	19.5	Len. + silence	7	12.1
Silence	110	56.4	Len. + filler	4	6.9
Filler	32	16.4	Cutoff + silence	12	20.7
			Cutoff + filler	7	12.1
			Unaccounted	25	43.1
Total	**195**	**100**	**Total**	**58**	**100**

Table 3.1: *Frequencies and distributions of disfluency elements.*

those of cutoffs. Lengthenings are remarkably rare given the assumption that many disfluencies are introduced by this element, which is a matter that will be investigated further.

3.2.2.2 Duration

As a next step, we looked at the duration of the individual disfluency elements. As summarized in fig. 3.1, Lengthening is the element with the longest duration on average, frequently with values around 500 ms and, more rarely, with about 800 ms. Cutoffs usually span a much shorter time, mostly between 150 and 300 ms, however with some outliers ranging up to 670 ms. Silences 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 are more restricted in range, between 170 and 380 ms, with occasional longer instances of up to 630 ms. The key finding of this analysis is that there is great durational variation. These first insights demand several follow-up investigations. The lengthening duration has been measured on the syllable level, which will vary depending on the types of phones contained. The cutoff can happen at any point in the word, which furthermore dictates a high degree of variability. The fillers in this sample have not been distinguished into types ("uh" and "uhm") fillers and the silences have

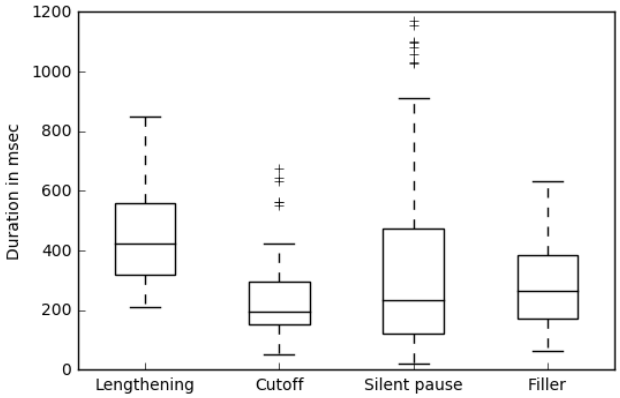

Figure 3.1: *Duration and variation of disfluency elements.*

not yet been regarded in connection with their syntagmatic embedding.

As a last general aspect of duration analysis, we checked if duration of elements in a cluster differed from duration of standalone elements. A linear regression analysis found no such relationship ($F(1, 318) = 1.921, R^2 = 0.006, p = 0.17$). We detected strong inter-speaker variability. For speaker 1, syllable lengthening was significantly higher when the lengthening occurred alone ($F(1, 11) = 10.85, R^2 = 0.5, p = 0.007$). For speaker 2, Word fragments occurring clustered with other disfluency elements were significantly longer ($F(1, 11) = 6.476, R^2 = 0.37, p = 0.02$).

3.2.2.3 Lengthening Versus Slow Speech

We checked the surroundings of long syllables in order to determine how much these syllables differ in duration from the average local speech rate, which shall for this study be defined as the syllable duration of the three preceding through the three subsequent syllables. As can be seen in fig. 3.2, the syllables in question (labeled 0) are significantly longer than their immediately surrounding ones, suggesting that these are indeed cases of hesitation lengthening that saliently stand out from the surrounding and are not merely stretches of slow speech.

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-lengthened syllables span between 50 and

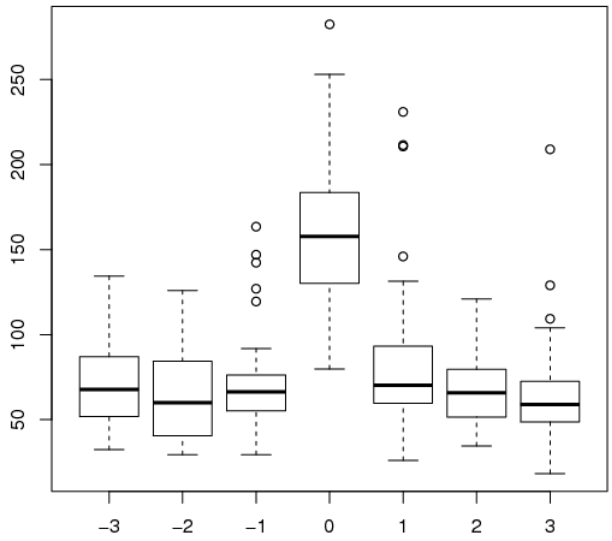

Figure 3.2: *Duration of lengthened syllable and its three surrounding syllables in ms divided by number of phones.*

70 ms in duration, occasionally stretching to about 140 ms. Phones in lengthened 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 duration of the preceding syllables had any predictive value for the extent of the lengthening, but this was not the case. Speaking rate appears to have no influence on this factor: lengthening 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 lengthened one has no influence on the extent of the duration of the following one ($F(1, 33) = 0.057, R^2 = 0.0017, p = 0.81$).

3.2.3 Discussion & Summary

This first exploration into human production of disfluencies yielded several insights that will be investigated further in subsequent sections. In general, disfluencies

exhibit a tendency of occurring in standalone fashion. When they occur in clusters, they do not differ in duration compared to standalone occurrence. In general, however, there is great durational variation within the individual disfluency elements as well as between speakers.
In this sample, silences were the most frequent element, followed by cutoff words and fillers. The remarkably rare lengthening seems to occur and end abruptly, without slowing down speech before or after. This has several implications for the testing of disfluencies in synthetic speech: Introducing too much variation into stimuli bears the risk of having to many variables to be able to control for. Introducing too few variation might create an overly artificial impression. Testing clusters of disfluencies is on the one hand unproblematic, as the individual elements behave, at least duration-wise, not differently from their usual occurrence. On the other hand, clusters are rare in human communication which questions their potential acceptability in human and synthetic speech.

3.3 Experimental Study 1: Modular Disfluency Synthesis

3.3.1 Introduction

As summarized in sec. 2.6.1, several approaches have been undertaken to synthesize disfluencies for various applications. Mostly these approaches insert a single disfluency phenomenon, such as silences or fillers. In human communication, however, there is a wide range of disfluencies that can be deployed. While it is not realistic to cover the entire range in a spoken dialogue system, it is still desirable to be able to produce more than one disfluency element to be able to more dynamically react to demands in the dialogue situation: lengthening might work well to introduce a hesitation interval, but it might not be as well suited to sustain it for a longer time. Longer intervals of hesitation might be better realized by silences or fillers.
In this first empirical study, we thus aim to test a wider range of disfluencies, using lengthening, cutoffs, silences and fillers in any possible combination, to test users'

acceptance of these different phenomena and of their combinability. They are modeled based on observations of human production, as described in the previous chapter. The aim of this first experiment is to explore how realizable individual disfluency elements are using open-source speech synthesis software.

3.3.2 Method

As a first exploratory step into the potential of disfluency synthesis, we synthesized utterances and added several disfluency elements to investigate the questions whether it is technically possible with open-source software to create plausibly sounding disfluencies, whether they are combinable or not and how users perceive and rate them.

The following disfluency elements were selected: lengthening, mid-word cutoffs and pauses. Lengthening and cutoffs could be either absent or present, coded binarily with 0 or 1 respectively, cf. Fig. 3.3. Pauses have three levels, encompassing both silent (1) and filled pauses (2) when present, and coded 0 when absent.

In order to minimize external influence on the testing, we used utterances from the DreamApartment corpus (Kousidis et al., 2013) and generated close-copy synthesis stimuli thereof. The stimuli were synthesized using Mary TTS (Schroeder and Trouvain, 2003) and a Python script was used to transpose the original pitch contour to the XML file from which Mary TTS synthesizes the audio. This ensures a higher voice quality compared to unmodified Mary TTS output, to control for quality judgments of disfluency synthesis really reflecting disfluency quality and not synthesis quality in general.

The criterion for utterance selection was the presence of a mid-word cutoff. We expect cutoff words to be the most salient and therefore most noticeable disfluency element, which is why we wanted to copy real-word occurrences thereof and not jeopardize the experimental setup by placing the cutoff at an arbitrary place that has unforeseeable influence on the quality. Table 3.2 shows the four utterances that were selected for re-synthesis.

In terms of duration, we opted against variation and for controllability: silences and fillers were inserted with fixed durations of 500ms, which is inside their normal duration range. Lengthening was inserted by adding 500ms of duration onto

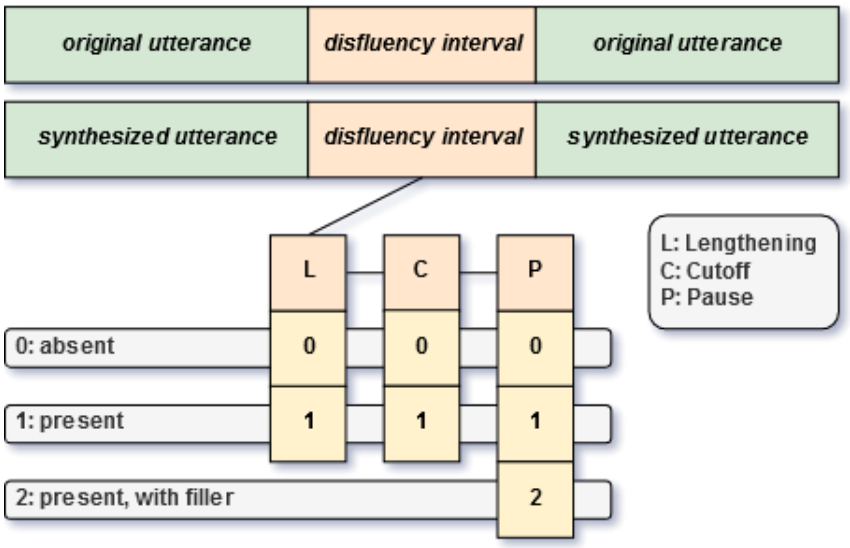

Figure 3.3: *Modular disfluency insertion architecture.*

the duration predicted by the synthesis system. Cutoffs retained their duration from the original instances in the corpus. We reasoned that for this small-scale perception test with an intended number of $n = 30$ participants, a smaller number of variables was regarded as preferable, even though the durational variation of human speech is not represented adequately this way.

Each of these utterances was synthesized in all twelve possible combinations the modular model as depicted in fig. 3.3 allows for; eleven disfluent combinations and a fluent configuration with the original cutoff excised. In table 3.3, utterance B is listed in all twelve possible realizations. The order of the elements is fixed. Every element can optionally be left out, but the order of occurrence cannot change. The configuration is coded with the abbreviation and levels shown in fig. 3.3. A fluent utterance would have all levels of lengthening, cutoff and pause set to 0 (absent), yielding a coding of $L_0C_0P_0$. The original utterances, containing a cutoff as their only disfluency, would be coded $L_0C_1P_0$. The stimuli were presented to 32 participants in a quiet room on a PC with headphones via the PRAAT MFC environment (Boersma and Weenink, 2014). Participants were then asked to provide overall quality judgments on a 5-point Likert scale ("How do you consider the quality of this synthetic speech?" with scale ends captioned "very good" and "very bad"). 15 of the subjects were female and 17 were male. They were between 23

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.

Table 3.2: *Utterances in orthographic representation and English translation*

$L_0C_0P_0$	Dann lassen wir mal die Einzelheiten einfach weg.
$L_0C_0P_1$	Dann [.] lassen wir mal die Einzelheiten einfach weg.
$L_0C_0P_2$	Dann [ähm] lassen wir mal die Einzelheiten einfach weg.
$L_0C_1P_0$	Dann ma- lassen wir mal die Einzelheiten einfach weg.
$L_0C_1P_1$	Dann ma- [.] lassen wir mal die Einzelheiten einfach weg.
$L_0C_1P_2$	Dann ma- [ähm] lassen wir mal die Einzelheiten einfach weg.
$L_1C_0P_0$	Da:n:n lassen wir mal die Einzelheiten einfach weg.
$L_1C_0P_1$	Da:n:n [.] lassen wir mal die Einzelheiten einfach weg.
$L_1C_0P_2$	Da:n:n [ähm] lassen wir mal die Einzelheiten einfach weg.
$L_1C_1P_0$	Da:n:n ma- lassen wir mal die Einzelheiten einfach weg.
$L_1C_1P_1$	Da:n:n ma- [.] lassen wir mal die Einzelheiten einfach weg.
$L_1C_1P_2$	Da:n:n ma- [ähm] lassen wir mal die Einzelheiten einfach weg.

Table 3.3: *Stimulus B in all twelve configurations.*

and 39 years old, most of them monolingual native speakers of German, and most of them university students or graduates. Data of non-native or non-academic user groups showed no significant differences from the other results, so their data were pooled. No participant reported any hearing impairment. Participants were asked for their familiarity 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.

3.3.3 Results

Analyses of Variance (ANOVA) were deployed to identify influences of the main factors lengthening, cutoff and pause. Lengthenings do not appear to have any

effect ($F(1) = 0.009, p = 0.923$), but cutoffs ($F(1) = 13.37, p < 0.001$) and pauses ($F(2) = 46.74, p¡0.001$) do. The interaction between cutoff 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 salient disfluency condition, i.e. conditions in any combination containing C_1 get lower mean results than the same conditions with C_0 instead of C_1 ($p < 0.001$). The same holds true for P_2 which performs worse than P_1 ($p < 0.001$) and P_0 ($p < 0.001$). P_1 yields almost the same mean result as P_0 (3.35 and 3.32 respectively).
First glances at the data reveal that stimulus D gets lower ratings than any other stimulus. For this reason, the ANOVA and TukeyHSD were repeated for stimuli A, B and C only to check for effects of the dis-preferred 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 are significantly lower on stimulus D in the comparison of C_0 and C_1, regardless of lengthening.
A look at the overall means reveals that stimuli with the $L_0C_0P_0$ 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 score of responses. However, the results are scattered widely, so there is no statistically significant correlation ($cc = -0.09$). There are three configurations that get (slightly, but not significantly) higher means than the fluent configuration: $L_0C_0P_1$ ($\mu = 3.53$), $L_1C_0P_1$ ($\mu = 3.53$), $L_1C_0P_0$ ($\mu = 3.45$) as opposed to $L_0C_0P_0$ ($\mu = 3.41$).

3.3.4 Discussion & Summary

The results, while exploratory and with a limited number of participants and a very simplified way of evaluation, exhibit some interesting tendencies. It seems that in general, simple and non-obtrusive disfluencies are preferred; the more disfluencies there are in a stimulus, the worse the user feedback gets. This correlates to the finding that the more disfluencies there are in a cluster, the rarer this specific cluster occurs (Betz et al., 2015a). One might even go so far as to comment

on the traditional linguistics view versus the view of "disfluencies as a tool" here: when utterances are severely disfluent, which in this case is defined by containing multiple different disfluency elements, then they carry a negative image. Positive perspectives on disfluencies seem only justified when the degree of disfluency is in such a level that it aids either speaker or listener in processing.

As expected, the fluent baseline is preferred over most disfluent stimuli, but surprisingly, not over all. Silent pauses and lengthenings fare unexpectedly well and some configurations even get higher user ratings than the fluent baseline. These elements deserve closer inspection in follow-up studies, as they promise to buy valuable dialogue time without being detrimental for perceived sound quality, which is a strong finding for synthesized disfluencies. One of the premises of this work is that hesitations can be used to buy time, and this result is a first hint towards how to best facilitate this.

On the other hand, it is to question, how to deal with fillers and cutoffs. They negatively impact the perceived quality but are still deemed useful to facilitate corrections and prevent barge-ins. Communication management is a trade-off between content and means of conveying it, and this is no different in synthetic speech. It is to be investigated if this trade-off can be tipped to the system's favor when cutoffs and fillers are produced with a better quality.

The key reason for the lack of quality in this study appears to be the lack of variation. Participants heard the same filler with the same duration and with only little prosodic adaptation in every stimulus with P_2 in it. It is very conceivable that introducing variation could improve the perceived quality significantly. Studying the phonetic micro-structure already hinted at the importance of variation (Betz et al., 2015a), but it was not included in this experiment in order to keep the amount of variables in reasonable dimensions. For future implementations, at least durational variation needs to be introduced to the disfluency insertion model.

Chapter 4

In-Depth Investigation of Hesitation Lengthening

One important result of the foundational studies described in the previous chapter is the fact that hesitation lengthening can be synthesized with a remarkable quality. Even with simple synthesis methods, it received ratings higher than the fluent baseline, hinting at the possibility that disfluencies need not be detrimental at all. However, as the perception test presented in the previous chapter was rather exploratory, more research is needed to shed light on this issue. It is desirable to find out if studies that focus on lengthening can confirm the tendencies obtained from the perception test. If the findings can be confirmed, lengthening might be the most versatile tool for conversational speech synthesis. Therefore, this chapter provides an excursion into the details of hesitation lengthening.

4.1 A Search Tool to Aid Lengthening Detection

The basic way to analyze lengthening is based on annotators' 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 Hough et al. (2015), tested on and applied to, e.g., the DreamApartment Corpus (Kousidis et al., 2013). This at the same time hints at the shortcoming of this method: When there is a huge corpus that has not been

enriched with lengthening labels upon annotation, it is an unreasonably high effort to manually parse the entire corpus again. Another potential shortcoming of the perception-based method is the suspected elusiveness of the subject. If lengthening is as subtle as the preliminary studies (cf. section 3.3) suggest, then many instances of highly deviant phone duration might escape the annotator's attention. It is thus desirable to explore the potential of machine-aided lengthening detection and annotation. In this chapter, we introduce a search tool that provides a pre-filtering of potentially lengthening-relevant phones in corpus data based on a simple z-score threshold. For brevity's sake, the tool will be called "detector" in the remainder of this chapter.

Lengthening has been studied in great detail with respect to its occurrence near phrase boundaries, e.g. Peters et al. (2005); Turk and Shattuck-Hufnagel (2007). It has also received attention in connection with other disfluencies, e.g. Adell et al. (2008); Peters et al. (2005). In its standalone form, a hesitation that manifests itself only by means of lengthening has received less detailed phonetic attention: O'Shaughnessy (1995) noted it to be an abrupt slowing down of speech, Adell et al. (2008); Betz et al. (2015a) and Goto et al. (1999) found it occuring at arbitrary places not predictable from data.

4.1.1 Method

The data at hand is the GECO corpus (Schweitzer and Lewandowski, 2013) of spontaneous German speech with phonemic annotation. It contains a total of 92 dialogues with about 30 minutes duration each. 46 female speakers were recorded in a monomodal condition (no visual contact, only speech as means of communication) and a multimodal condition (full visual contact). For each speaker, duration and z-normalized duration is calculated for each phone. As no utterance boundaries are available, stretches between two silences (IPUs = interpausal units) are marked as the unit of analysis. Fillers are also marked in the annotation, so the data set can be enriched with some phrase-positional information: For each phone, the distance from silences, from IPU beginnings and from fillers is given.
The data set was then modified, excluding all stops and [h] phones because, according to Peters et al. (2005), lengthening cannot be realized on these sounds.

Traditionally, only stops are seen as not prolongable in German. In other languages, this might be different, cf. Betz et al. (2017a). We decided to follow Peters et al. (2005) in the exclusion of [h] for two reasons: On first inspection, many errors in the data were caused by forced alignment errors on this phone. Furthermore, it is unlikely that a prolonged [h] would result in high-quality synthetic speech.
Following the inspection of the z-score filtering output of the detector, we repeat the analysis on a narrowed-down dataset excluding predictable distractions in the data, such as utterance-final and filler-preceding phones, which are frequent loci of lengthening in the same duration range as hesitation lengthening. We investigate whether we are able to replicate the findings of other studies with entirely spontaneous German speech with regard to phrase-final lengthening. In terms of spread of the lengthening, we analyze where it occurs and how many phones are affected. Furthermore, we repeat the analysis of phrase-final lengthening, but only with instances preceding fillers to check whether the results differ or stay the same preceding fillers compared to preceding boundaries. Finally, based on these analyses, we investigate whether disfluent lengthening is detectable based on a filtering procedure that takes into account more features besides z-scored duration.

4.1.2 Results

4.1.2.1 Z-Scored Duration

Figure 4.1 shows the distribution of z-scored phone duration of all speakers. 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 high degree of variation in spontaneous speech. It can further be seen that to the negative side there is a defined lower boundary, reflecting the natural limits of shortening (Klatt, 1973, 1976). In other words: syllables cannot be infinitely shortened or compressed, whereas it is theoretically possible to infinitely lengthen syllables (within the limits of physiology). The interesting areas for lengthening are between 2 and 4, where 3.4% of the tokens occur and which clearly exceed the boundaries of natural variance and are thus suspected to con-

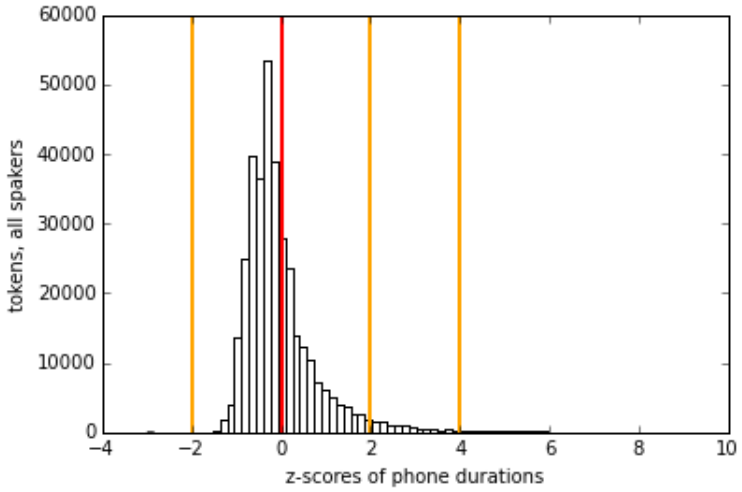

Figure 4.1: *Duration distribution of prolongable phones*

tain disfluencies and probably other types of marked lengthening. Furthermore, the area of z-scores greater than 4 needs to be put to close scrutiny, where we expect errors as well as extreme lengthening. This demands close-ups in the analysis. In the following, we investigate what the cause is for z-scores between 2 and 4, what causes the outliers of z-scores > 4, and, consequently, what pure z-scores can tell about disfluent lengthening.

We inspected the phones with z-scores in the aforementioned interesting ranges and draw the following first conclusions: 95% of the z-scores between 2 and 4 are caused by phrase-final lengthening. Disfluent lengthening makes up only a small part of this. 98% of z-scores greater than 4 are due to force-alignment errors in the original annotation. There are only very few cases of disfluent lengthening per speaker with z-scores this high. This leads to the conclusion that z-scores alone are not sufficient to 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% of lengthening flagged for z-deviation is attributable to boundaries.

4.1.2.2 Boundary-Related Lengthening

The assumptions about phrase-final lengthening in German can be confirmed with entirely spontaneous 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. We expect an increase in phone duration immediately preceding the boundary and we investigate whether the increase affects more phones than the immediately preceding ones. As can be seen in Figure 4.2, the last phone preceding an intonation phrase boundary is clearly longer than the preceding one, between all other phones, maximally a slight increase is notable. 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.
The impression that the main duration increase occurs on the last phone only can be verified statistically. 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 data set 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 4.2 suggests, the increase to the last phone differs significantly ($p < 0.001$) from the increase to the penultimate phone for all sets. No significances were found between other positions.

4.1.3 Summary

In this section we introduced a search tool, or filtering method, to aid the detection of disfluent lengthening. It will be referred to as the "detector" for the remainder of this thesis. With the detector it becomes possible to process large amounts of data without annotation. At the moment, we have not yet performed an evaluation of detector performance against human annotation, which will be addressed in the next section. In order to analyze the nature of truly disfluent lengthening, it appears to be necessary to not only apply z-score filtering, but also tune the filter to exclude positions close to boundaries. It is conceivable that more features

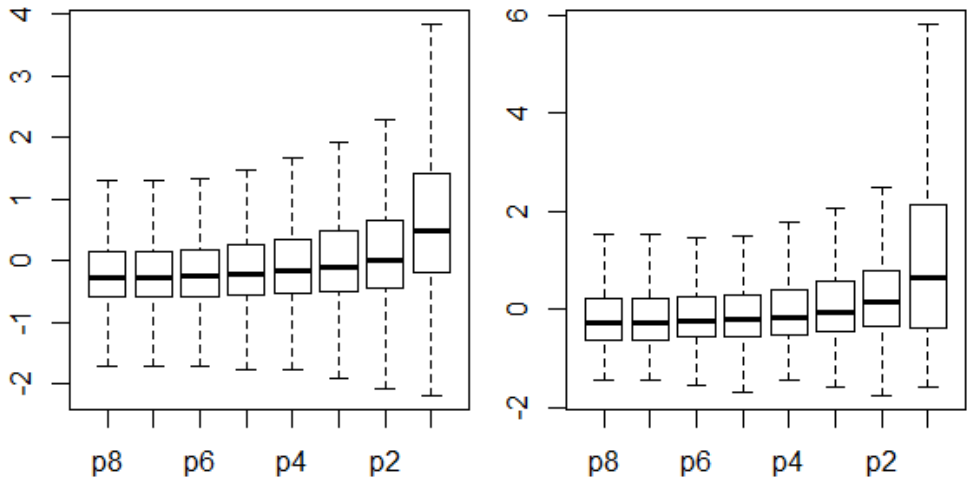

Figure 4.2: *Z-normalized duration increase of phones preceding intonation phrase boundaries (left panel) and preceding fillers (right panel) p8 is 8 phones before, p1 is last phone before boundary.*

have to be taken into account in order to further improve the search tool. In the following section, we evaluate the tool by applying it to corpora with disfluency markup and explore which further features have to be taken into account for its improvement.

4.2 Corpus Study 2: Detector Evaluation

In the previous section, a search tool for detecting lengthening has been presented, which, while functioning satisfactorily, posed some questions to be addressed. In this study, the detector is applied to a corpus that already features disfluency annotation (DUEL-dreamapartment; henceforth: DUEL (Kousidis et al., 2013)). This way, it will be possible to compare, on the one hand, detector versus human performance and, on the other hand, the two different corpora, DUEL, and GECO from the previous study.

In a part of the GECO corpus, with approximately 22 hours of speech, the detector output 750 instances of disfluent lengthening, corresponding to 0.57 instances per minute. The human-labeled instances of lengthening in the DUEL corpus added up to 114 in total in 4.5 hours, or 0.42 instances per minute.

As both corpora consist of spontaneous speech, and the latter features speakers who are highly involved in a collaborative task that was specially designed to elicit disfluencies, it is surprising that the DUEL corpus exhibits a lower occurrence rate of lengthening. The rates would have been expected to be at least equal, if not even higher in the DUEL corpus.
Furthermore, given that lengthening is supposed to be one of the most frequent disfluencies in spontaneous speech (Eklund, 2001), the detected frequencies appear to be surprisingly low. We therefore hypothesize that both strategies of detecting disfluent lengthening, i.e. human annotation and semi-automatic detection, have their shortcomings. The detector can only find instances above its z-threshold and humans can only find instances above an unknown hearing threshold. (An attempt to find such a threshold is described in section 4.4.) We therefore evaluate human versus detector performance and provide insights on how to make both human and automatic lengthening detection more efficient.
There are two different annotations available for the DUEL corpus: One created by humans with the instruction to label on-the-fly any disfluency-related phenomenon they encounter, and one by humans assisted by the detector as a metaphorical magnifying glass highlighting candidates of hesitant lengthening. These can be compared in a subsequent evaluation.

4.2.1 Method

We apply the detector to the duel corpus and compare the number of lengthening instances marked by (a) the detector alone (b) the original annotators and (c) by new annotators with aid of the detector. The method in (c) is expected to yield the most precise results, as possible misses of the original annotators can be spotted and false positives of the detector can be eliminated by the new annotators. We consider the combined set of annotated and semi-automatically detected lengthening as the ground truth for assessing precision and recall. It is important to note that this ground truth is an approximation as there might be lengthening missed by the detector due to the z-threshold, and by human annotators due to the elusiveness of lengthening.

As a last step, we analyze potential improvements of the detector by putting the false positives of its output to close scrutiny, in order to identify recurring types of features giving rise to erroneous output.

4.2.2 Results

4.2.2.1 Counts, Precision, Recall

We performed fine-grained comparisons between the two sets of annotations in order to identify shortcomings or advantages of the various methods. While the frequencies of other disfluency labels are constant within the DUEL corpus, lengthening labels are almost completely absent in the second half of the corpus, see tables 4.1 & 4.2. We thus limit the comparison of human versus detector annotations on the part containing lengthening labels.
The total number of disfluency-related lengthening found in the entire corpus, with human and semi-automatic detection combined is 431 in 4.5 hours of speech, or 1.6 per minute. As expected, this rate is higher than in the GECO corpus (0.57 per minute). The rate remains constant throughout all files, so that the anomaly in lengthening label frequency has to be ascribed to the annotation process, probably as the result of fatigue or a change of the annotator.

Tables 4.1 & 4.3 reveal two main findings: Human annotators miss more than

Type	**Count**	**Percentage**
Detector only	140	59.9
Human only	45	19.2
Detector+Human	49	20.9
Total	234	100.0

Table 4.1: *Detected lengthening instances in the first half of the DUEL corpus.*

Type	**Count**	**Percentage**
Detector only	191	96.4
Human only	6	3.6
Total	197	100.0

Table 4.2: *Detected lengthening instances in the second half of the DUEL corpus.*

half of the instances of disfluent lengthening, with a recall of 40%. The use of semi-automatic detection increases the recall to more than 80%. It comes at a cost, though, as precision drops from 82% in human annotation to 28.8% in semi-automatic detection.

Annotator	**Precision**	**Recall**
Detector	28.8	80.8
Human	82.0	40.1

Table 4.3: *Precision and recall*

4.2.2.2 False Positives

In order to improve the detector, we thoroughly analyzed its false positives. As summarized in Table 4.4, forced-alignment errors are responsible for more than half of the false positives reducing the precision of the semi-automatic approach. The second largest portion of false positives is due to laughter or laughed speech. These intervals add noise to the signal, making it impossible for forced-alignment tools to correctly identify phone boundaries. For this reason, some corpora, such as DUEL (Hough et al., 2016), feature laughter markup. This information can be used to pre-filter the data in future work to increase precision. The remaining 20% of false positives are overhead that is avoidable if the corpus annotation allows for it.

Type	**Count**	**Percentage**
Forced-alignment error	655	58.2
Laughter	219	19.5
Accentuation	94	8.3
Phrase-final	82	7.3
Backchannel	69	6.1
Other	7	0.6
Total	1126	100.0

Table 4.4: *False positives types*

4.2.3 Discussion & Summary

In this study we evaluated the detector against human annotators and conclude that the most precise results can be obtained with human annotation aided by the detector. The detector alone at the time being lacks linguistic knowledge and thus outputs instances of lengthening of any sort, which must then be filtered for the instances desired by the researcher. On the other hand, there is evidence for the idea that was first put forward in section 3.3, that lengthening is an elusive and hard-to-spot phenomenon: A large portion of lengthening that was found with the aid of the detector escaped the expert annotator beforehand. The next step is to improve the performance of the detector based on this evaluation in order to make it a more valuable tool for lengthening analysis.
Lengthening is not only used in hesitation, but also in accentuation, phrase-finality and backchanneling. It is possible that features such as word class or pitch movement distinguish disfluent lengthening from accentuation. In corpora with utterance or speaker turn markup, it could be possible to identify and exclude phrase-final lengthening, which would, however, also exclude some instances that are disfluent *and* phrase-final. Backchannels, as "islands" of one speaker in the other speaker's turn can be detected and excluded if the corpus is annotated accordingly. To sum up, there are ways to increase precision and reduce overhead, given the the corpus data has the required features:

- A corrected phonemic annotation could increase precision by up to 58%.
- Laughter markup that allows for pre-filtering: 19%
- Speaker turn markup to exclude backchannels: 6%

The remaining 15% lack of precision due to accentuation and phrase-finality can only be avoided if the data gathered so far is sufficient to train a classifier to perform the distinction between disfluent and non-disfluent lengthening. This was tested in Betz et al. (2017b), which yielded a minor increase in performance above the majority baseline, but the data at hand was too sparse to draw conclusions about the classifier performance.
At this point we have established that lengthening is an elusive phenomenon, that

escapes even trained annotators. This makes it particularly interesting for our purpose of creating a hesitating dialogue system, because it appears to be able to bridge gaps (and thus *buy time*) in dialogue without the user noticing it. This leaves us with the paradoxical situation of wanting to synthesize elements of speech that will ideally not be actively perceived by the listener.
At the same time, we have put our detector to the test and can conclude that it is a valuable assistant for discovering hesitation lengthening in speech corpus data. In the next section, we will use the detector to extract more features of lengthening from data, which is required for adequately including it into speech synthesis and dialogue systems later.

4.3 Corpus Study 3: Lengthening Features

In this section, we use the detector presented in the previous sections to extract lengthening features from speech corpus data. Up to this point, we have discovered that lengthening is a versatile feature of spontaneous speech, which appears to have the potential to fulfill a central role in hesitating dialogue systems. In order to adequately model and synthesize hesitation, we analyze in this study where lengthening manifests itself on the word level, the syllable level and the phone level.

4.3.1 Method

This study is based on the GECO corpus Schweitzer and Lewandowski (2013), a phonemically annotated corpus of spontaneous German speech. We used the first half of it, the "monomodal" condition, in which speakers had no visual contact. This part was chosen as the absence of visual contact has been found to have emphasizing effect on disfluencies (e.g. Belz and Reichel (2015)). One file had to be omitted due to technical issues, yielding 43 files, each containing 30 minutes of speech, totaling in approximately 22 hours of speech. Speakers are female students from southern Germany engaged in free dialogue.
From these corpus data, we extract instances of lengthening remote from phrase boundaries using the detector described in section 4.1. The resulting tokens are

hand-labeled by two annotators in order to (a) identify hesitation lengthening and (b) explore potential other types of lengthening, such as lengthening for emphasis and accentuation. Inter-annotator agreement is checked via Cohen's kappa scores in order to ensure that the hypothetical categories of lengthening can be identified reliably.
Building on that annotation procedure, we analyze the lengthening tokens with regard to the word class they belong to. We count the frequencies of open-class and closed-class words (function and content words) and use chi-square tests to compare distributions. Zooming in further, we investigate in which syllable position lengthening occurs and which phones it affects. Doing so, we compare hesitation lengthening to accentuation lengthening, which appears to be the second major type of lengthening that is not associated with syntactically motivated boundary-related lengthening. Distributions are tested with the chi-square test and a generalized linear mixed model is fit to test if lengthening type can be expressed as a function of syllable position.

4.3.2 Results

4.3.2.1 Tokens

2.800 tokens of lengthening were extracted from the corpus data at hand. These tokens were hand-labeled by two annotators and fall mainly into three categories: (1) *Disfluent lengthening*, (2) *accentual lengthening*, and (3) *forced-alignment errors*. In total, 1.000 tokens of lengthening, 75% of them disfluent and 25% accent, were extracted from the first half of the corpus. 1.800 tokens were discarded because of grave forced-alignment errors, or for reasons such as the lengthening being phrase-final and neither disfluent nor emphatic. About 500 of the remaining 1.000 lengthening tokens still contain minor boundary errors, that are corrected for future analyses, but are not severe enough to discard the tokens.

4.3.2.2 Inter-Annotator Agreement

The two annotators labeled the output phones according to the three main categories. Inter-annotator agreement was tested on a subset of 13 files of the corpus,

Word	*Translation*	*Frequency*
und	and	61
die	the	35
so	so	27
dann	then	23
in	in	22
ich	I	19
das	the	16
ist	is	16
irgendwie	somehow	15
weil	because	14

Table 4.5: *10 most frequent words lengthened for disfluency*

after a training phase based on four different files from the same corpus. We tested three aspects of agreement: whether annotators agree on a token being disfluent lengthening, whether annotators agree on a token being accentuation lengthening and whether accentuation and disfluent lengthening are confounded by annotators. It turns out that annotators show relatively low agreement in terms of classifying a token as disfluent ($\kappa = 0.433$) or as accentuation ($\kappa = 0.298$) but accentuation and disfluency lengthening are **never** confounded. More specifically: the tokens on which the annotators disagree are labeled as "false positive output from detector" by one of the annotators. When either annotator decided for either disfluency or accentuation, the other annotator never chose the opposite category.

4.3.2.3 Word Classes

As noted by O'Shaughnessy (1995), lenghtening occurs mainly on function words, or closed-class words, such as determiners, prepositions and conjunctions.[1]
This is confirmed by our data: we examined word frequencies of the 755 examined disfluencies and table 4.5 lists the 10 most frequent disfluent words. The same picture extends downward. Apart from auxiliary forms of *sein* "to be", there are

[1]The distinction into function and content words is highly debated in linguistics and open-class and closed-class are possibly more neutral terms to describe this. For simplicity, we call them function and content words. By this we mean to provide only a rough and exploratory distinction into words that add content to the message and words that rather manage and structure information in the message.

	Function words	Content words	Σ
Freq > 1	508 (67.3%)	32 (4.2%)	540 (71.5%)
Freq = 1	74 (9.8%)	141 (18.7%)	215 (28.5%)
Σ	582 (77.1%)	173 (22.9%)	755 (100%)

Table 4.6: *Function and content word distribution and word frequencies within disfluent lengthening*

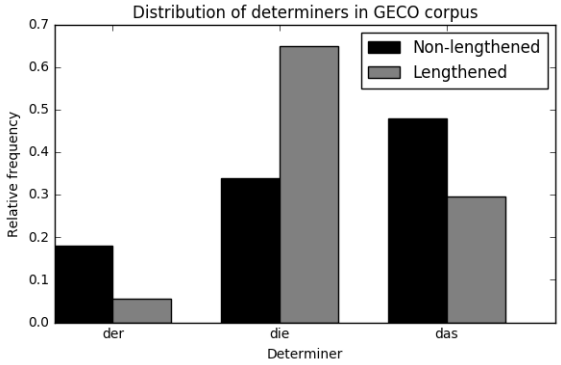

Figure 4.3: *Distribution of relative determiner frequency defined as percentage of lengthened der, die, das in the total number of lengthened der + die + das. Same for non-lengthened.*

no nouns, verbs or adjectives in the top 41 ranks, or in the top 59% of disfluent words.

A preliminary word class tagging was performed, alongside with a split of the data in words that occur only once in the sample and words that occur more often, cf. table 4.6. This reveals that function words make up most of the targets for disfluent lengthening (77.1%) and that most function words occur more than once in the sample (67.3%). To the contrary, most of the content words in the sample occur exactly once.

The distribution of the determiners (that can also serve as relative pronouns in German) of different gender is extremely skewed. As can be seen in table 4.5, the female (*die*, 35) and neutral (*das*, 16) determiners are quite frequent, while there are only three tokens of the male one, *der*. As can be seen in Fig. 4.3, this distribution differs significantly from the overall distribution of words in the same corpus ($\chi^2 = 23.578$, $p < 0.001$).

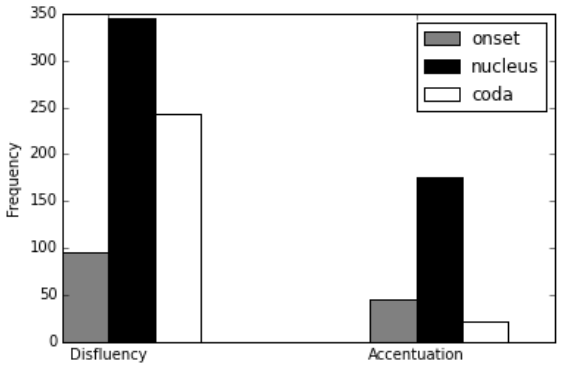

Figure 4.4: *Syllable positions of lengthened phones.*

4.3.2.4 Syllable Positions and Phone Classes

As can be seen in figure: (cf. fig. 4.4), disfluent lengthening is, as opposed to accentuation lengthening, distributed across nucleus and coda positions and to a much lesser extent, the onset. Accentuation lengthening almost exclusively occurs in the nucleus. A chi-square test of syllable position distribution of lengthening tokens (onset, nucleus, coda) in accentuation and disfluent lengthening found the distributions to be significantly different ($\chi^2 = 20.064$, $p < 0.001$). In order to further analyze the difference with a mixed models approach with a binomial factor *syllable position*, another chi-square test was run with only the positions *nucleus* and *coda* considered, as these appeared to be the most impactful factor levels from visual inspection. This shows that the distribution of nucleus and coda lengthening over disfluent and accentuation type is significant as well ($\chi^2 = 17.623, p < 0.001$). For further analysis a generalized linear mixed model was fit using R's lme4 package (Bates et al., 2015), with *syllable position* as the dependent variable, *lengthening type* and *word class* as fixed factors and random slopes for *speakers*. This confirms the significant influence of lengthening type on the position in the syllable ($\beta = -0.6440, SE = 0.1633, z = -3.943, p < 0.001$).

In the following, we turn towards the phone types within the syllables. The position of the lengthening within the syllable might be determined by the phones the syllable consists of. As can be seen in fig. 4.5, the syllable position of the lengthen-

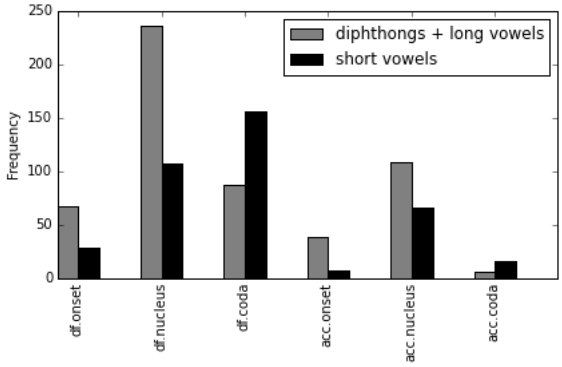

Figure 4.5: *Phone type of vowel nuclei for disfluent (df) and accentuation (acc) lengthening by syllable position of the lengthening.*

ing is related to the nucleus vowel being short or long.[2] If the nucleus of a syllable in which hesitation lengthening occurs is a long vowel, the lengthening will most likely be placed there. If, however, the nucleus of that syllable is a short vowel, the lengthening will fall on the coda instead, if the coda contains prolongable phones. Statistical analysis underpins these findings. Chi-square tests were used to test conservatively if the distributions of syllable positions of lengthening differed respective of vowel nuclei being long or short. They differ significantly, for several configurations tested:

- Full data, onset, nucleus, coda: $\chi^2 = 172.44$, $p < 0.001$
- Full data, nucleus, coda: $\chi^2 = 144.6$, $p < 0.001$
- Accentuation only, onset, nucleus coda: $\chi^2 = 72.83$, $p < 0.001$
- Accentuation only, nucleus coda: $\chi^2 = 51.792$, $p < 0.001$
- Disfluent only, onset, nucleus coda: $\chi^2 = 109.07$, $p < 0.001$
- Disfluent only, nucleus coda: $\chi^2 = 99.258$, $p < 0.001$

This enables a close-up using another generalized linear mixed model, with syllable position as the dependent variable, vowel nucleus type, lengthening type and a

[2] As diphthongs are inherently long, they are grouped here with long vowels.

exploratory distinction of the word being an open-class ("content") or closed-class ("function") word, as well as an interaction of lengthening type and word class as fixed factors and random slopes for speakers. The vowel nucleus type has a significant influence on the position of the lengthening, ($\beta = -2.2358, SE = 0.1990, z = -11.238, p < 0.001$), as has the type of the lengthening, ($\beta = -1.1137, SE = 0.2584, z = -4.310, p < 0.001$). Word class and the interaction between word class and lengthening type do not have a significant influence on syllable position.

4.3.3 Discussion

The tokens that we extracted from the data is comprised of about one quarter accentuation and three quarters disfluent lengthening, after erroneous tokens had been discarded. Many of the tokens still contained minor boundary errors that could be manually corrected before analysis. This reveals that even where the detector outputs the desired material, forced-alignment shortcomings emerge. We suspect that the unusually high length of these phones troubles the language models the forced alignment works with.

It is not straightforward to identify and analyze lengthening. Even on targeted listening, annotators were not consistently able to classify any given instance of lengthening, highlighting that this is a very elusive phenomenon. However, it has become evident that there are at least two categories of lengthening remote from syntactic boundaries, disfluent and accentuation, as there never is disagreement on assigning lengthening to either of these categories. It is up for further studies to determine how long or how salient a lengthening has to be in order to be clearly classified as disfluent, One such endeavor is described in section 4.4.

Previous studies can be confirmed in the sense of hesitation lengthening preferably manifesting itself on function words. In addition, the fact that the great amount of lengthened content words occur only once in our data hints to an interpretation that a random target for hesitant lengthening is likely to be chosen when no suitable function word is available in the articulatory buffer, whereas the frequently recurring function words are the unmarked targets for hesitation lengthening.

Among the different words, conjunctions represent the default word class linking two parts of an utterance. It thus makes sense for speakers to hesitate at this

point, in order to facilitate speech planning for the remainder of the utterance and to signal increased cognitive load to the listener, who can in turn infer that it is not the conjunction which is causing the trouble, but the material that is about to follow.

It is not surprising that the frequently occurring determiners and relative pronouns are also frequent bearers of hesitation lengthening. However, it comes as a surprise that the gender distribution within these words is extremely skewed. It can only be assumed that the long vowel in the open syllable of *die* is easiest or most suitable to sustain, whereas the diphthong in *der* might be less so. This analysis is based on the word level, without distinguishing the word class into pronoun and determiner. It could be a topic for future studies to close-up on this issue.

The observation reported in section 4.1 that hesitation begins with lengthening and has no apparent pre-planning beforehand is supported by this analysis. Disfluent lengthening manifests itself not only in the syllable nuclei but also to a considerable extent in the coda. In contrast, accent related lengthening manifests itself almost exclusively in nuclei. For validation purposes, the same analysis was re-run on the DreamApartment corpus (Kousidis et al., 2013), yielding exactly the same results. To sum up, disfluent lengthening differs substantially from other types of lengthening in terms of a predominance of coda lengthening.

In fluent speech, speakers plan beforehand where they place their accent, so it is likely for them to choose vowel nuclei for realizing these accents. In case of disfluencies, speakers often may not have the chance to time the "perfect phonotactic moment" to hesitate and so they may resort to coda positions. One reason for doing so might be the vowel quality of the nucleus: a short vowel might be less *elastic* than a sonorant coda, making the coda a preferable target to sustain speech production.[3] This could mean that speakers, when they spontaneously have to find the best spot for placing a hesitation, they rather choose an elastic sonorant in the coda than a short vowel nucleus. For accentual lengthening in the nucleus, the vowel types are quite evenly distributed. Accentuation lengthening in the coda is rare, but even so, there is a slight majority of short vowels in the nucleus, which can serve as an explanation for the coda preference.

[3]The concept of elasticity will be further elaborated on in section 4.5.

4.3.4 Summary

In this section we conducted a corpus study, putting to use the detector presented in earlier sections. We were thus able to extract and analyze a considerable amount of lengthening tokens. Summing up, we found 750 instances of hesitation lengthening in 22 hours of spontaneous German speech. We further found that the identification of lengthening is not straightforward, but that two classes of lengthening, hesitation and accentuation clearly exist. In terms of phonetic detail, hesitation lengthening differs from accentuation in that it can occur in long vowel nuclei as well as in sonorant codas, whereas accentuation lengthening mainly occurs in the nuclei. Another difference between the two classes is the observation that hesitation lengthening prefers closed-class, or function words, which makes sense given the function of accentuation lengthening - it is unlikely for a speaker to stress function words, apart from very specific contexts (such as emphasizing a co-occurrence of events by accentuating the conjunction *and*.)

4.4 Experimental Study 2: Searching for a Lengthening Threshold

So far, we have dealt with real-world occurrences of hesitation lengthening and we hypothesize that lengthening can be a potential core component for a hesitation synthesis strategy. In the first experimental study, we heuristically approached disfluency synthesis and found simple lengthening of a fixed duration to be a well-performing element of spontaneous speech synthesis. It is not yet known, however, how much synthetic lengthening is acceptable and how lengthening influences the user's interaction with the system. To address these questions, we set up a computer-based game environment in which users are given instructions with different extents of hesitation in order to test how users perform under different disfluency conditions.

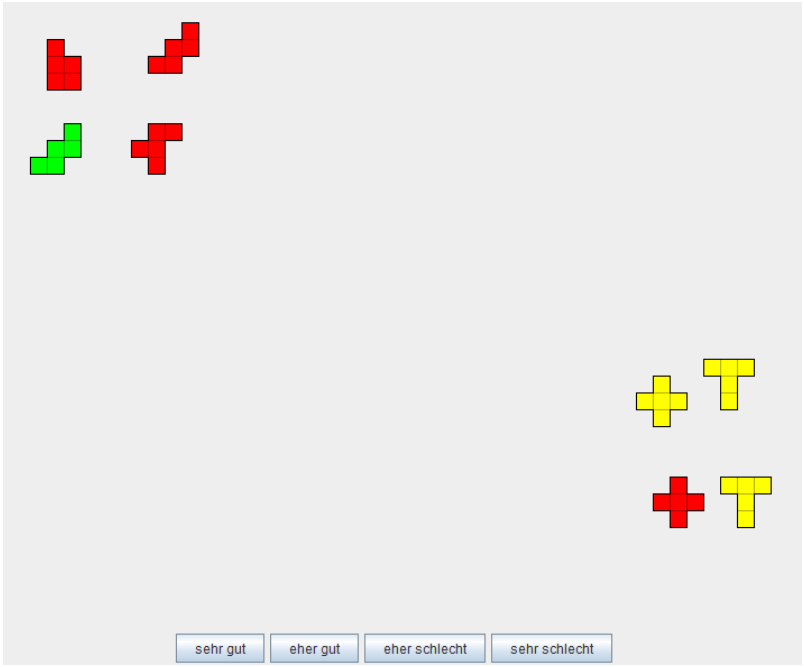

Figure 4.6: *Game scene with sound quality feedback buttons: very good, rather good, rather poor, very poor.*

4.4.1 Method

We designed a perception test to evaluate sound quality of lengthening. This test is embedded in a simple GUI-based game, in which users are asked by a synthetic voice to move around Pentomino pieces on a computer screen (cf. figure 4.6). After each stimulus, participants have to click one of the four quality feedback buttons that constitute a 4-point MOS-scale to proceed. The scale is deliberately set to an even number in order to avoid participants clicking in the middle to proceed faster. In addition to the ratings, we measured *relative task completion time* and checked for influences of lengthening extent. To control for the different sentence lengths, we calculated the time span from beginning of audio until the drop of the Pentomino piece divided by sentence duration.

4.4.1.1 Stimulus Design

Previous studies suggest that lengthening mainly occurs on function words (Shriberg, 1994; O'Shaughnessy, 1995), and that German determiners, conjunctions and pro-

nouns are frequent targets for lengthening 4.3.
In this study, we test synthetic lengthening of function words in different degrees of lengthening with 400, 600, 800, 1000, 1200 and 1400 ms duration of the target word. The degrees are chosen to represent common duration ranges of lengthening in human communication and the step sizes were determined to have a sufficiently high, yet not too complex number of different factors for analysis. The target words are German monosyllables *(der, die, das, und, dann, ihn)* selected because of their high frequency of occurrence and syllable-type balancing, i.e. containing different types of vowel nuclei and sonorant, plosive and empty codas. This balancing does not control for the inherently different duration of the words. E.g. the word *dann* in the 600-ms-condition might appear less stretched than the word *die* in the same condition. This, however had no effect on the results.
The duration for each segment in the target words is determined by applying the duration model based on the elasticity hypothesis (Campbell and Isard, 1991), means and standard deviations for each phone are extracted from the GECO corpus (Schweitzer and Lewandowski, 2013). Each target word is embedded in a different carrier sentence and is located at the junction of two phrases that instruct the user to drag and drop Pentomino pieces (cf. table 4.7). The instructions follow a fixed order of `[<pick up a piece> <conjunction phrase> <move it onto another piece>]` (cf. table 4.7). The resulting six sentences were synthesized in seven different configurations:

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

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

4.4.1.2 Stimulus Presentation

Each trial is composed of a game scene with a Pentomino board (cf. Fig. 4.6) with a corresponding audio instruction to pick up a piece and move it. Participants

#	Stimulus
1	*Nimm das rote Kreuz **und** lege es zum gelben Winkel.*
2	*Die grüne Treppe, **die** muss rüber zum blauen Balken.*
3	*Der gelbe Winkel, **der** muss rüber zum roten Balken.*
4	*Das blaue Kreuz, **das** muss rüber zur grünen Treppe.*
5	*Nimm die rote Treppe, **dann** lege sie zum gelben Kreuz.*
6	*Nimm den grünen Balken **und** lege ihn zum blauen Winkel.*

Table 4.7: *Stimuli with target conjunction in boldface.*

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

4.4.1.3 Participants

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

4.4.2 Results

We used R (R Core Team, 2015) with the lme4 package (Bates et al., 2015) to conduct a linear mixed effects analysis of the influence of lengthening extent on user ratings.[4] As fixed effect, we chose *extent of lengthening*, as random effects we chose

[4]This paper inspired the analysis method and should not pass unnoticed like lengthening: Winter (2013)

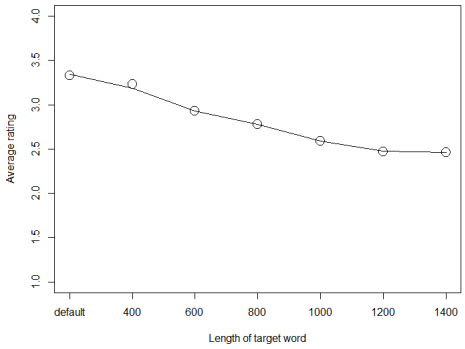

Figure 4.7: *User feedback with respect to word length. 4=good, 1=bad*

intercepts for *stimuli* and *participants*, as well as by-stimulus and by-participant random slopes for the effect of lengthening extent, to control for ideosyncrasies of the participants and stimuli. Visual inspection of the residuals did not reveal any obvious deviations of homoscedasticity or normality.
The analysis shows that regardless of stimulus and participant, lengthening extent influences user ratings ($t(743) = -6.855$), each increment lowering the average rating score by about 0.18 ± 0.027 (standard errors), on a scale where 4 corresponds to the best and 1 to the worst rating.
Using the same mixed models approach as above for relative task completion time (time span from beginning of audio until drop of Pentomino piece, divided by stimulus duration), we found that lengthening also significantly lowers relative task completion time ($t(743) = -4.296$), indicating that participants are not confused by the lengthening, but rather use the extra time to complete the task.

4.4.3 Discussion

As can be seen in Fig. 4.7, stimuli get good overall feedback and the ratings decline very slowly as lengthening increases, reaching a sustained trough at 1200ms. On the one hand, this leads to the assumption that even relatively long lengthening is a valid strategy for spoken dialogue systems. On the other hand, it suggests that lengthening should ideally be kept low to maintain highest-possible quality.
Analyses of the interaction speed support this assumption (cf. Fig. 4.8): Users use

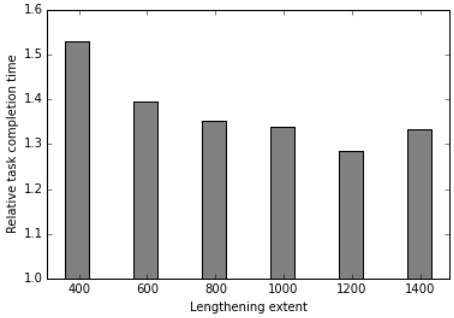

Figure 4.8: *Relative task completion time (divided by stimulus duration) over the different lengthening conditions*

the extra time granted by lengthening to solve the task – they get faster relative to sentence duration as lengthening increases, but appear to get distracted by extreme lengthening, when they appear to slow down again (although the slowdown is not significant).
Despite the lack of evidence for lengthening $> 1200ms$, it could be assumed that there is a turning point in synthesis quality around 1200 ms: In terms of ratings, users do not differentiate anymore; in terms of task completion times, users need more time.

4.5 Lengthening and Phone Elasticity

This study aims at the last piece of information required to synthesize lengthening properly, that is, how to distribute the duration increase resulting from the slowdown of speech over the individual sound segments of the syllable. The idea of using a segment's elasticity to predict its duration is first introduced by Campbell and Isard (1991). Here, elasticity refers to a segment's flexibility in duration: a highly elastic segment, such as a long vowel or a liquid, can be lengthened or compressed to a great extent, whereas less elastic elements, such as short vowels or stops, have a smaller range of possible duration configurations.
The following equation adapted from Campbell and Isard (1991) represents the duration of a bi-segmental syllable as composed of the mean duration plus k times

the standard deviation of each of its segments.

$$syldur_{2\ segs} = (\mu_{seg1} + k * \sigma_{seg1}) + (\mu_{seg2} + k * \sigma_{seg2}) \tag{4.1}$$

Here, every bracket term calculates the duration of a syllable's segment and represents its measure of elasticity. k is a factor within this measure of elasticity and is segment-independent. That means, k is constant within a single syllable, but may have different values for different syllables. From a more formal point of view, k describes the average z-Score of the phonemes within a syllable. That is, k delivers information about the average normalized distance of the syllable's segments from their mean duration.

$$K_{syll} = \frac{(dur_{syll} - \Sigma_{i=1}^{N}\mu_i)}{\Sigma_{i=1}^{N}\sigma_i} \tag{4.2}$$

It has to be stated that the elasticity hypothesis in its 'strong' form, as depicted in (4.1), is not tailored to represent special contexts. Campbell and Isard (1991) note that for such environments, specialized 'weak' forms of the hypothesis have to be constructed:

> "*Weaker forms of the elasticity hypothesis would state that statistics have to be gathered separately for syllables in different positions in the sentence (e.g. finally versus non-finally), for segments in different parts of the syllable (e.g. for those in the onset and rhyme), and in different phonetic contexts (e.g. for vowels before voiced and unvoiced stops).*"

Based on the context-sensitivity of conversational phenomena such as disfluent lengthening, we consider a context-sensitive approach as promising and develop a weak elasticity hypothesis specified for disfluent lengthening. With this, we aim to provide a more accurate segment duration prediction for hesitation synthesis.

4.5.1 Method

In order to evaluate the prediction accuracy of the strong, baseline form and the weak, specialized form of the hypothesis, we first compile a subset of the GECO corpus (Schweitzer and Lewandowski, 2013). This subset contains all syllables of the corpus that have been labeled as containing disfluently lengthened segments.

Second, we use the baseline elasticity hypothesis (4.1) to predict the segmental durations of the subset. The total syllable duration and the means and standard deviations for each segment based on the full corpus serve as input. The equation can then be solved for k (4.2) and each segment's duration can be calculated.
Third, we repeat this process, but this time with the specialized elasticity hypothesis, which differs only in one aspect: the means and standard deviations are now based on the subset rather than the full corpus.
Finally, we repeat the last step, but this time we compare predictions to instantiations in a different corpus.
To evaluate this approach, we calculate root mean square errors (RMSE) between the measured duration instances and the two types of predicted duration instances. The question is whether this specialized form of the hypothesis outperforms the baseline. It is based on the same *kind* of data it is supposed to predict; however the formula uses abstracted mean values to predict concrete values, which is not to be confused with a machine-learning technique trained on the same data it is to predict. If the specialized form outperforms the baseline, future work on this matter can use small specialized corpora which are easier to handle and cheaper to compile than large, general corpora. Molloy and Isard (1998) introduced the concept of *k-deviation*, which is a measure of a syllable's deviation from the elasticity hypothesis. It is defined as the root mean square z-score deviation from the syllable's k-score. This deviation is calculated over all phones within the syllable. Thus, a syllable's perfect fit to the elasticity hypothesis would result in zero k-deviation, which is always the case if the present syllable is monophonic. (Equation (4.3) is taken from Molloy and Isard (1998)).

$$Kdev_{syll} = \sqrt{\frac{\Sigma_{i=1}^{N}(K_{syll} - Z_i)^2}{N}} \tag{4.3}$$

The approach of Molloy and Isard (1998) is based on their observation that syllables do not exactly fit the elasticity hypothesis in reality. The implication that all phones within a syllable have the exact same z-score, and therefore the syllable k-score is identical to all phone z-scores within the syllable, could not be confirmed. The concept of k-deviation has the following implications for the investigation at hand:

First, we follow Molloy and Isard (1998) and take the RMSE as a measure of deviation. As we are primarily interested in deviation regarding concrete segment duration, our formula differs from the k-deviation in terms of the applied variables:

$$Ddev_{seg} = \sqrt{\frac{\Sigma_{i=1}^{N}(D_{seg} - D_{seg\,pred})^2}{N}} \tag{4.4}$$

Using the actual (D_{seg}) and the predicted ($D_{seg\,pred}$) syllable duration, we calculate the root mean square duration deviation from the true syllable duration.
Second, the gold standard for our hypothesis testing should not be a perfect fit of a syllable into the elasticity hypothesis, as this does not correspond to natural language. Hence, a small RMSE is to be aimed at, but it has to be considered that a result of zero cannot be attained within natural speech.

4.5.2 Results

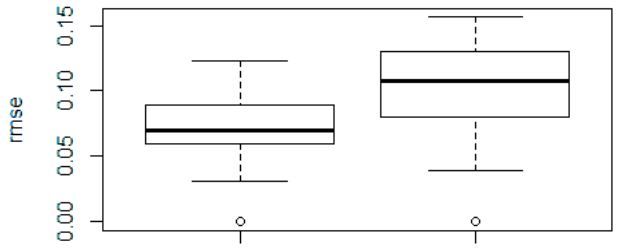

Figure 4.9: *Deviation from observed duration is smaller when predictions are based on disfluent (left box) compared to baseline (right box) data.*

For each of the 750 syllables in the data set that contain a phone with disfluent lengthening, we compare each of its phones' duration to the duration predicted using the disfluency-based form of the hypothesis and the baseline form. As can be seen in figure 4.9, the disfluency-based form of the elasticity hypothesis exhibits a lower root mean square error, i.e. it differs less from the observed duration distributions.
Not only the disfluent phones are predicted more accurately, also each non-lengthened phone is predicted more accurately using the disfluent form. The performance of

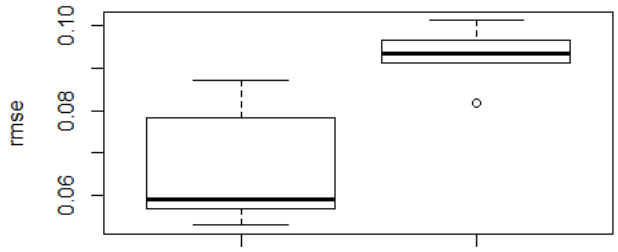

Figure 4.10: *Preliminary tests on a different corpus with the same underlying means yield similar results: disfluency data (left box) exhibit smaller deviation than baseline (right box).*

the disfluent form is significantly better than the baseline form: paired $t(30) = -6.7, p < 0.001$.

4.5.3 Discussion

It is not too surprising that predictions based on a specialized dataset outperform those based on an average dataset. However, the important finding in this study is that a disfluent context is another special context to be taken into account and that it yields significantly more accurate predictions than the baseline prediction, even when sentence position, syllable position and phonetic context are not taken into account individually.

For modeling disfluencies, and probably other conversational speech elements, it appears thus adequate to compile small-scale corpora that are rich in conversational speech phenomena, which can be elicited specifically; such as hesitation phenomena that can be elicited by delaying the flow of information a speaker has to present to a listener. This is interesting in the sense that no cost-intensive gathering of big speech corpora is necessary to improve conversational features of spoken dialogue systems. As in this study, existing large-scale corpora can be used, however, to create sub-sets of the conversational speech element in question.

4.6 Empirical Investigations Summary

In this part, we started out by analyzing human production of disfluencies and mimicking these phenomena with speech synthesis. These first studies proved that it is technically possible to synthesize disfluencies for use in conversational spoken dialogue systems. The subsequent chapter dealt with the question how to insert disfluencies into synthetic speech output. Based on the observation that many human hesitation intervals start with lengthening of material in the buffer, and based on the fact that synthesized lengthening yields very good results in terms of user feedback, we conducted an in-depth investigation of lengthening in order to determine where in an utterance to commence hesitation via lengthening. In order to address this question, we first introduced a search tool, the detector, to help finding instances of this elusive phenomenon in speech. Its elusiveness was highlighted by the application of the detector to an already annotated speech corpus, revealing a large amount of missed instances by trained annotators. The detector was then used to extract instances of hesitation lengthening from 22 hours of spontaneous speech. This yielded several insights. First, there is at least one other distinct type of lengthening that occurs remote from boundaries, namely accentuation lengthening. It behaves similar to hesitation lengthening in terms of duration, but differs in almost every other aspect. Hesitation lengthening prefers function words, accentuation content words. Hesitation can occur ad-hoc in either nucleus or coda (and to a lesser extent: also in onsets) whereas accentuation lengthening exhibits a much higher frequency of occurrence in the nucleus. This seems to be governed by the phoneme inventory. The most preferable phone classes for hesitation lengthening are long vowels or sonorants. Our analyses showed that whenever the vowel nucleus at the locus of a hesitation lengthening is short, the lengthening falls on the sonorant coda. This information helps us to determine the entry point for hesitation in speech synthesis.

The next question was, how long to sustain speech via lengthening before resorting to other measures. To answer this, we conducted a study with a simple GUI-based game and analyzed users' interactions with a system that step-wise increases the amount of hesitation lengthening. We found that lengthening has a positive effect on the interaction, but, alas, failed to find a clear-cut threshold at which length-

ening becomes detrimental. It remains thus hypothetical whether this threshold actually exists. It is clear, however that there are limits in human production as well as in speech synthesis regarding how long a phone can be sustained. We therefore decided to base our lengthening extent on the elasticity hypothesis, which has in this chapter been tested specifically for disfluent contexts.

In the next part, we apply the insights gathered so far by equipping a dialogue system with the ability to hesitate. We will start out by sketching a *hesitation strategy* which serves as a blueprint for an algorithm that determines entry point, composition and extent of a hesitation cluster.

Part III

Implementation and Evaluation

Chapter 5

Hesitation Insertion Strategy for Spoken Dialogue Systems

Based on the insights gained from the studies outlined in the previous chapters, we sketch a draft for a hesitation insertion strategy that can be evoked while a dialogue system is speaking, and that determines the best entry point given an event of hesitation and the best temporal extension of a hesitation. The strategy is centered on lengthening, which has been identified as a reasonable starting point in human as well as machine hesitations (cf. Part II). Human speakers can prolong material in their articulatory buffer, that has been planned, but not yet uttered. Incremental spoken dialogue systems can proceed likewise as they also have a buffer that stores planned utterance chunks before synthesizing. As the analysis of lengthening in human speech suggests, this potentially buys enough time to regain fluency. If this is not the case, other hesitations can be deployed. The realization of the individual hesitations is governed by the insights gathered from the empirical studies described in Part II. This section is a walk through the details of the strategy, which is based on a *hesitation algorithm* that determines when and where which hesitation elements will be spliced into the speech synthesis stream. This strategy can be regarded as the basis of a model for hesitation insertion in incremental spoken dialogue systems. The aim of the strategy proposed here is to buy as much time as possible for the speaker, by lengthening words in the metaphorical articulatory buffer and by inserting silences. Only in severe cases,

where more time is needed, other measures such as fillers will be employed (cf. Figure 5.1). This reflects human speech planning (cf. Section 2.3) and is a feasible way to technically realize on-the-fly hesitation insertion.
The strategy depicted in Figure 5.1 can be summarized as follows:
While an event of hesitation is active, execute the following steps:

1. Apply lengthening to best target.

2. Insert first silence.

3. Insert filler.

4. Insert second silence.

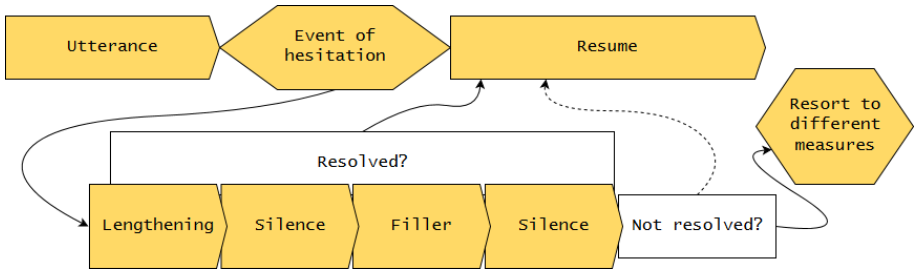

Figure 5.1: *Hesitation insertion strategy.*

When the hesitation ends during any of these steps, the original speech plan is resumed. If all steps have been run through without the event of hesitation ending, the system has to resort to different measures. What these measures are has to be defined externally and will be discussed later in this chapter. In the following, we walk through the individual steps in more detail. For an example of hesitation insertion into an utterance, see Figure 5.2.

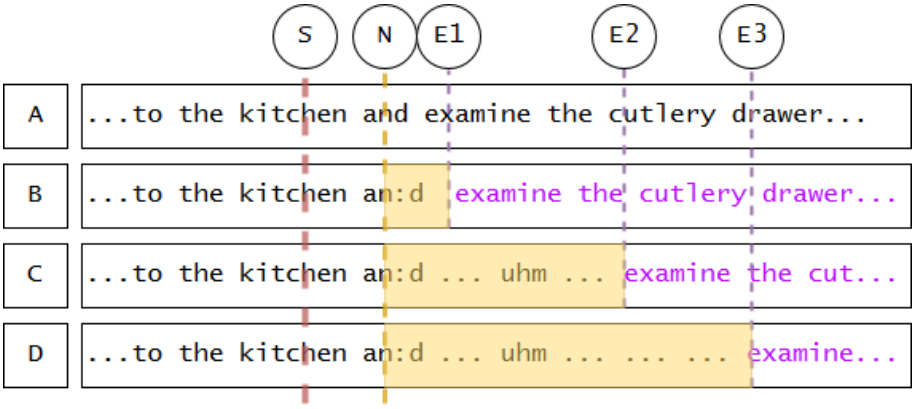

Figure 5.2: *Examples of insertable hesitations. S = Starting point (external event triggering hesitation mode); N = Entry point (best upcoming target segment, in this case nasal sound in function word); E1–E3 = End points (external events triggering end of hesitation mode, after which originally intended utterance A is resumed); B = utterance with short hesitation, resumption (in purple) starting at E1; C = longer hesitation with lengthening, silence, filler, silence; D = same as C, but with a long second silence (when the strategy is "after the loop, remain silent until an external event triggers the end of hesitation mode".) Hesitation mode intervals are highlighted in yellow.*

5.1 Algorithm Walk-Through

In this section, we will explain step-by-step the individual components of the hesitation algorithm depicted in Fig. 5.1.

"Event of hesitation." There are various reasons for hesitating. Any of these reasons could be accounted for in a dialogue system. It could also be a wizard-of-oz setting, where there is a "start" and a "stop" button to delimit the event. The following walk-through is to be understood as "steps that are taken while a generic, externally defined event of hesitation is active. As soon as the hesitation is terminated externally, the system will leave the loop as soon as possible."

1. "Apply lengthening to best target." Hesitation lengthening does not occur arbitrarily. Given the concept of the articulatory buffer, speakers start hesitating as soon as possible, which means, at the next appropriate syllable. Several linguistic and phonetic factors determine which syllable that is, and how much that syllable can be stretched in duration. The lengthening con-

tinues until the phone has been stretched to its maximum, or until hesitation mode ends, whichever occurs first.

2. "Insert first silence." If the lengthening has not bought enough time to resolve the event of hesitation, silence can be added. Following the suggestion of a standard maximum silence of 1 second in conversation (Jefferson, 1989), this silence will continue for maximally 1000 ms, or until hesitation mode ends. For a more elaborate analysis of pauses and their duration, see Lundholm Fors (2015).

3. "Insert filler." If the previous steps did not buy enough time, as a more severe measure of hesitation, fillers ("uhm") can be added. Short fillers ("uh") denote minor pauses and are thus not adequate for long hesitation loops (Clark, 2002).

4. "Insert second silence." If after the filler the hesitation mode is still not resolved, a second silence can be added to buy more time, with the same rules as the first silence. This is based on the observation that fillers are regularly followed by silent intervals: Clark and Fox Tree (2002) suggested that the filler type ("uh" or "uhm") predicts the extent of the following silence; this view is challenged by O'Connell and Kowal (2005), who found that post-filler silences vary arbitrarily in duration. Our assumption is that in a dialogue system, the duration of the post-filler silence is not arbitrary but governed by the duration of the external hesitation event. A cutoff threshold can be inserted, e.g. following the standard maximum silence suggestion.

5. "Resort to different measures." Systems need a strategy to continue when the above steps do not suffice to buy enough time to resolve the event of hesitation. Some conceivable examples of how a system could proceed are as follows:

    - Wait for hesitation event to end.
    - Re-enter the loop or parts of it to buy more time.
    - Repeat parts of previously uttered speech to buy more time.

- Resume own speech plan if possible, despite the event of hesitation not being over.

In our implementation, which will be described in the next chapter, we opt to simply wait for the hesitation mode to end. This is the case when the systems dialogue management component signals the synthesis to proceed. This simple way was chosen as it has not been investigated yet how to properly repeat parts of the loop or of previously uttered material. This is an aspect for future studies which is out of the scope of this work.

In the following chapter, this hesitation insertion strategy will be implemented into a dialogue system residing in a smart-home environment for a first practical test and evaluation.

Chapter 6

Implementation into an Interactive Smart-Home Setting

In the following, we describe how the individual concepts of the model described in section 5 are implemented in a prototype hesitation module. The module is part of a virtual agent's dialogue system embedded in a smart home setting. The hesitation module is fully capable of producing hesitations, but as there were technical issues with the production of fillers (cf. section 6.1.2.4), the evaluation is twofold: (1) we evaluate a system without fillers, that hesitates only by means of lengthening and silence, (2) we evaluate the system with fillers in an exploratory manner, which is still functional, but suffers from bad synthesis quality. The study is conducted with German synthetic stimuli in a smart-home environment, in which users are engaged in an item-retrieval task.
In addition to the interaction study, we conduct a parallel analysis of the speech synthesis output only, using crowdsourcing to collect MOS-based ratings (cf. section 6.3). This is done in order to assess the notoriously difficult evaluation of speech synthesis systems (Wagner and Betz, 2017). By collecting traditional MOS-based data, we aim to compare them to the results gathered in interaction and draw conclusions from differences arising, as it is, as of now, entirely unknown whether interactive evaluation approaches generalize to non-interactive listening tests. It is therefore crucial for speech synthesis and disfluency evaluation in general to investigate differences in evaluation methods.

6.1 Implementation

In this section, we briefly describe the technical environment we implement our strategy in, and then continue to describe step-by-step how the algorithm steps introduced in section 5.1 are implemented into our system.

6.1.1 Technical Implementation

The hesitation module is integrated into an existing incremental spoken dialogue system (Carlmeyer et al., 2014), which uses a toolkit for incremental dialogue processing (Baumann and Schlangen, 2012) and MaryTTS (Schroeder and Trouvain, 2003) as a speech synthesis back-end. The dialogue system is part of a virtual agent that resides in a smart-home environment as an avatar on a screen which can interact with users via speech Wrede et al. (2017). Additionally, the agent has access to behavioral information of the user detected via cameras and sensors in the smart-home.

6.1.2 Implementing the Algorithm

6.1.2.1 Event of Hesitation

In this study, an event of hesitation is triggered when a user stops to maintain eye-contact with the virtual agent, and the event lasts as long as the users' gaze is shifted away. We deploy hesitations as a user-oriented strategy, as a response to visual attention shifts (cf. Carlmeyer et al. (2016b)). The goal is to assist users in their task by only giving them information while they are paying attention.

6.1.2.2 Different Measures

The above definition for hesitation events also governs the strategy for resuming fluency. In this case, it is simply waiting for the hesitation to end, i.e., the user restoring eye-contact. This can result in unfortunate situations in which users misinterpret the silence as a system error and do not bother to look back. For future work, a more elaborate strategy is desirable. A threshold after which the machine breaks the silence would be conceivable, or a cue to regain attention. The perfect

solution would be a module of its own that manages resumptions after events of hesitations, using information about continuations in human communication (cf. section 2.3.3).

6.1.2.3 Lengthening

In our hesitation strategy, lengthening is the starting point for hesitations. The appropriate target syllable is selected from the words in the buffer. We included a lookahead with a five-word limit, in order for the hesitation not to start too late after an attention shift. That means that the best target is selected from the upcoming words, but no later than five words after the trigger. Based on the preference hierarchy for lengthening targets described in Section 5, our system iterates over the buffer, searching for the optimal syllable (i.e., a nasal in a function word), increasing the tolerance for less appropriate targets with each iteration, i.e. when no nasal in a function word is available, it searches for a long vowel or diphthong in a function word; if that is not successful either, it searches for other prolongable sounds, such as fricatives or short vowels in a function word; if no suitable material in a function word is available, it searches for a nasal sound in a content word; this way the search trickles down, and in the "worst case" a fricative or short vowel in a content word is lengthened. In turn, this means that lengthening is always applicable, as there is no five-word stretch in German (or, probably, in any language) consisting solely of unprolongable consonants; it is just not always possible to insert the optimal type of lengthening.
The duration of lengthening is inferred from mean duration values resulting from previous corpus studies, from which a so-called stretch factor is deduced. This factor is calculated by generating Gaussian random numbers with the mean duration and standard deviation for each phoneme. The highest number from 10,000 samples is selected and divided by the mean duration. This factor reflects how much a given phoneme needs to be stretched in duration to achieve its average maximum.

6.1.2.4 Fillers

Due to technical problems, fillers are not included in our main study and could only be tested in an exploratory fashion. These problems are issues related to the incremental processing component, namely (1) audible clicks around the filler and (2) omitted segments and words around the filler. These lead to the impression that the system is having issues with speech production rather than elegantly hesitating.
Four participants were recorded in a condition with fillers, which showed that the insertion works in terms of variable duration and placement of fillers, but the audio artifacts have too much impact on the sound quality. As will be described in Section 6.2.2, we explored the usability of data with this preliminary "full hesitation" version, but most participants were recorded in a "reduced" version containing only lengthening and silence, cf. Section 6.1.2.6.

6.1.2.5 Silences

As fillers are left out, the main study operates with only the first silence of the hesitation strategy. In general, it is designed to last 1000 ms. In our implementation, the duration is variable as we wait for the user to re-focus. Systems with a different fluency-resumption strategy could make use of a threshold, after which other measures are taken. In this study, the system waits until the user restores eye-contact. (In the exploratory condition with fillers, the first silence lasts for 1000 ms and the second silence lasts until the users re-focus.)

6.1.2.6 Reduced Hesitation Model

Given the shortcomings of the synthetic fillers, we conducted the main study with a reduced model, that only employs lengthening and one following silence. As is the nature of lengthening, the duration increase with the stretch factors derived from corpus data is barely audible. This might create the misleading impression that this reduced model only hesitates by means of silence, which is dispreferred based on the results of previous studies (Carlmeyer et al., 2016a,b; Chromik et al., 2017). Therefore, we increased the lengthening extent by 50% to ensure participants are able to perceive it. This implementation, while missing the filler aspect, is fully

functional and capable of dynamically inserting hesitations, only in a less diverse form than originally intended.

6.1.2.7 Paradox Evaluation

The fact that the final study design needed an extra 50% duration increase transplanted onto the carefully extracted durational parameters shows one aspect of evaluation difficulty regarding the phenomenon of hesitation lengthening and its synthesis. We desire to create a barely noticeable hesitation, but something that cannot be perceived by listeners can only be evaluated indirectly: effects on task performance can be observed, but direct feedback on sound quality is not very revealing when the object of evaluation is not consciously perceived.

6.2 Experimental Study 3: Item Retrieval Task

To evaluate the effect of hesitation in human–agent interaction, we conducted an interaction study in *the Cognitive Service Robotics Apartment* (Wrede et al., 2017). The apartment consists of three rooms (kitchen, living room and hallway) which are equipped with various sensors for visual tracking and recording.
The strategy for hesitation synthesis described in Section 5 is evaluated by means of a task in which the participants have to perform a memorization and item retrieval task. A virtual agent provides a background story and instructs the participants to look for hidden treats and candy at seven different places in the apartment. The dialogue system underlying the virtual agent is implemented in two different versions: one *baseline* condition without hesitations or adaptations of any sort, and a *hesitating* condition that monitors participant's attention shifts via gaze tracking and that enters hesitation mode whenever participants look away from the virtual agent.
Our hypotheses for this experiment are: (H1) Memory task performance benefits from the presence of hesitations: as described in section 4.4, users of hesitating dialogue systems can make effective use of the extra time granted by hesitations. (H2) Presence of hesitations influences user ratings of perceived synthesis quality (undirected): the study reported in section 3.3 found different hesitation elements

to evoke different user ratings. Lengthenings and silences fared very well, whereas fillers did not. We thus expect an influence of unknown direction of the presence of hesitation. (H3) There is no negative impact of the presence or absence of hesitation on the system's likability: Carlmeyer et al. (2016a) found that self-interrupting virtual agents that deploy only silences are perceived negatively, we hypothesize that this effect is mitigated by a more elaborated hesitation strategy which chooses the best phonetic entry point and deploys lengthening before the interruption.

6.2.1 Method

We use a between-subjects design, i.e., each participant interacts with the system in either the baseline condition or in the hesitation condition. Before the main study starts, participants are asked to fill out a declaration of consent to be recorded. In addition, they must complete a short memory test, in which they are presented a pre-constructed audio file containing ten words produced by a synthetic voice. The voice is MaryTTS's German female Hidden Markov Model (HMM) voice with no further modification. The words are German nouns that fall into five categories (professions, food, sports, buildings, cities), with two in each category. Each participant is presented with the same words and order of words. They are then asked to recite aloud as many of the words as they can remember. The resulting *memory test score* (percentage of items memorized out of a number of 10 items in total) is surveyed with a checklist for later comparison to the recall rates in the main study, in order to calculate task efficiency (i.e., how well did participants perform relative to their memory capacity attained at the day of the study).
The main study is set in the kitchen and living room of the smart home. As a platform we use the simulation of the anthropomorphic head Flobi (Lütkebohle et al., 2010) displayed on a screen in the kitchen area of the smart apartment (cf. Figure 6.1). The agent is able to detect faces and estimate the current visual focus of attention of the human interaction partner via a gaze-tracking tool (Schillingmann and Nagai, 2015) with a web-cam installed on top of the screen.
As soon as a participant appears in front of Flobi, it starts talking (cf. Figure 6.1).

It first introduces itself and the apartment and then instructs participants about the task they are to perform: Each participant is asked to search for treats that have allegedly been hidden in various places in the apartment (cf. Figure 6.2). The agent lists all potential hiding places, asking the participant to memorize and later investigate these. The task is embedded in a story about construction workers that have just left the apartment and caused confusion in the agent's sensors, due to the dust they stirred.[1] This creates a plausible pre-text for the agent to list all possible hiding places for the participant later to remember, with the hint that it is not sure whether it got all places correctly. The full instruction text can be found in Appendix A.
During the instruction phase, there is an intentional distraction at three fixed points in time. This is included to ensure some degree of distraction and gaze shift for each participant, as this is what triggers the hesitations. The distractions are:

1. **Visual distraction.** Lighting up a door handle in the participants' field of vision.

2. **Audiovisual distraction.** The experimenter entering the room to insert a code for later use in the questionnaire.

3. **Audio distraction.** A music beat being played for two seconds.

The distractions are of the same type and at the same time for each participant in both conditions.
As soon as the agent has finished the instruction, the participants start investigating the possible hiding places. The interaction is monitored audiovisually in an adjacent room. Participants are asked to name aloud every hiding place they are going to examine. This is necessary for the subsequent video analyses in which every retrieved item can thus be classified as found by chance or found by memorizing. Additionally, items that were named, but not retrieved (e.g., due to a stuck door) can be classified as memorized. The number of items memorized comprises the *finding rate*.
After the interaction, participants rate their overall impression of speech synthesis

[1] There was actual visible construction work in the apartment at the time of the study, which inspired this narrative.

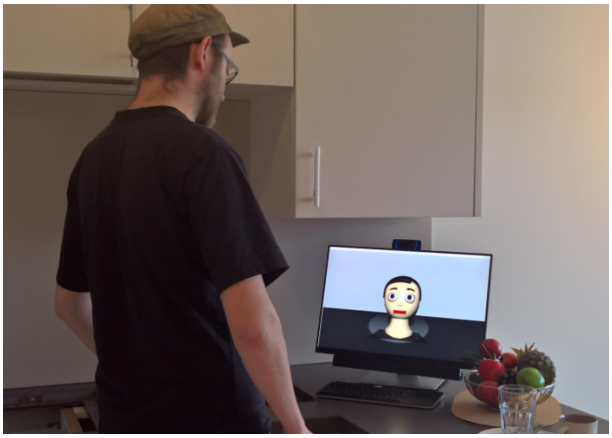

Figure 6.1: *Person being instructed by virtual agent on a screen.*

quality on a 5-point MOS scale. This scale was chosen for maximum comparability with traditional MOS-based synthesis evaluations. In addition, they filled out a questionnaire assessing their subjective impression of the system quality on 24 dimensions using 7-point Likert scales (based on the Godspeed questionnaire (Bartneck et al., 2009)). Additionally, demographic data and previous experiences with robotic systems, the agent Flobi, and speech synthesis systems in general were surveyed. Finally, participants were asked one question in a follow-up interview regarding the interaction, namely, if they felt that the agent adapted to their behavior in any way. All participants received monetary compensation.
The entire interaction was recorded via four cameras mounted on the ceiling of the apartment. In addition, various system events for later analysis are collected (for further information about this process refer to Holthaus et al. (2016)).

The collected data were entered into a generalized linear model (glm) with *finding rate* as dependent variable, *hesitation condition* as fixed factor, and *memory test score*, *gender* and *age* as control variables. To include individual memory performance in participants' retrieval performance, we calculated an efficiency measure: $efficiency = \frac{MemoryScore(\%)}{FindingRate(\%)}$. This is to take into account the users' individual memory capacities and to normalize results accordingly. As efficiency scores are not normally distributed, we used a Mann–Whitney U test to check for effects on *efficiency* by *hesitation condition*. The same test was then used to analyze users' feedback on synthesis quality with regard to the *hesitation condition*.

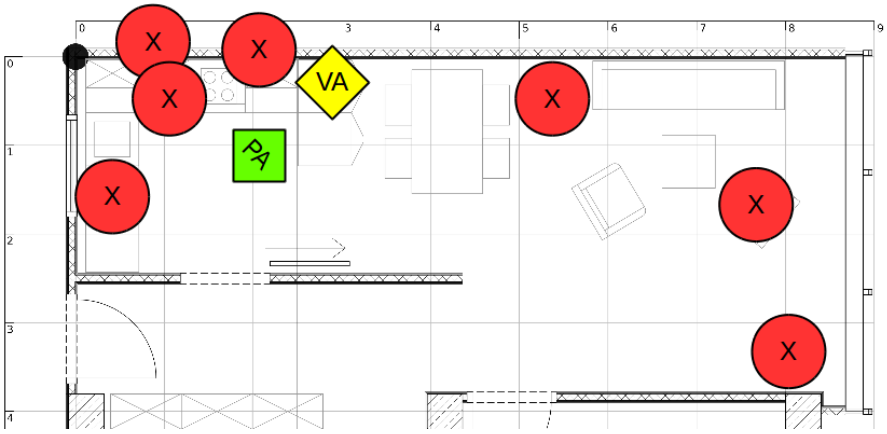

Figure 6.2: *A 2D map of the smart home environment. (X) denotes hiding places of treats, (VA) the position of the screen with the virtual agent, and (PA) the initial position of the participant.*

To evaluate the questionnaires regarding the user's perception of the agent, based on Bartneck et al. (2009), the responses are grouped into five key concepts (*anthropomorphism, animacy, likeability, perceived intelligence* and *safety*). Using Shapiro–Wilk and Bartlett tests, we found the data of all five concepts to be normally distributed and to show equal variances, qualifying the data for a t-test to examine differences of user feedback on key concepts for the two hesitation conditions. This study was devised to test $n = 40$ participants in total, 20 in the baseline and 20 in the hesitation condition. Due to some no-shows, we recorded 37 trials with 24 female and 13 male participants in total. The data of two participants had to be excluded from the analysis because their language competence did not suffice to follow the instructions correctly. Overall, 17 participants interacted with the baseline system, and 14 with the hesitation system. These 31 trials provide the core for our analysis. In addition, four participants were recorded in the full hesitation condition for exploratory purposes, cf. Section 6.1.

Participants were recruited on the university campus and via campus-related social media. Every participant that registered took part in the study, there were no special requirements, apart from functional vision and hearing, basic knowledge of German, and no or little experience with robotic systems, virtual agents, or speech systems in general. Mean age was 24.6 ($SD = 4.2$).

As we have at our disposal four recordings with the full hesitation condition (cf.

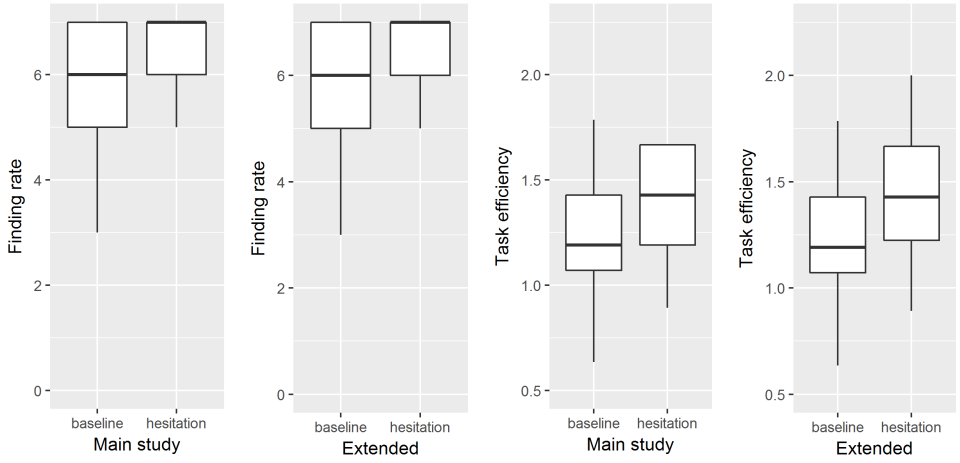

Figure 6.3: *Task performance and efficiency.*

Section 6.1), we repeated the analyses for finding rate and efficiency with the same 17 trials for the baseline condition and with all 18 hesitation trials combined as the hesitation condition. This is an **exploratory extension of the analysis** in order to come closer to the $n = 40$ participants we aimed for. The four additional stimuli have functional hesitations, but severe sound quality issues (cf. Section 6.1.2.4). For that reason, we do not include them in the main study as we cannot control the effects this has on the interaction.

6.2.2 Results and Discussion

Finding Rate. On average, the number of items found is higher in the hesitation condition ($M = 6.36, SD = 0.84$) than in the baseline condition ($M = 5.71, SD = 1.21$), (cf. Figure 6.3, left panel). The glm analysis shows that the effect is not significant ($\beta = 0.8, SE = 0.44, z = 1.84, p = 0.065$). **Efficiency** increases in the hesitation condition ($M = 1.22, SD = 0.3$) compared to the baseline ($M = 1.5, SD = 0.58$), (cf. Figure 6.3, third panel from the left). The Mann–Whitney U test shows no significant effect of *hesitation condition* on *efficiency* ($W = 79, p = 0.11$)

Subjective Speech Synthesis Quality. On average, using a 5-point MOS scale (1 = "very bad", 5 = "very good") users rate synthesis quality worse in the hesitation condition ($M = 1.36, SD = 0.84$) compared to the baseline condition

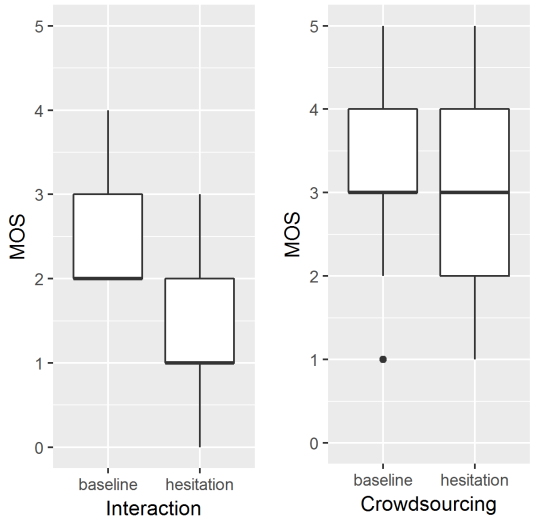

Figure 6.4: *A 5-point mean opinion score (MOS) scale user feedback on synthesis quality.*

($M = 2.53, SD = 0.62$), cf. Figure 6.4, left panel. The Mann–Whitney U test shows that there is a significant effect of *hesitation condition* on users' perception of synthesis quality ($W = 203, p = 0.0004$).

Subjective Rating of the Agent We conducted t-tests for an effect of *hesitation condition* on each subjective ratings of the five key concepts *anthropomorphism, animacy, likeability, perceived intelligence*, and *safety*. The factor *hesitation condition* had no significant influence on any of the user feedback regarding these concepts, cf. Figure 6.5. Aside from the questionnaire results, participants had been encouraged to give free-text feedback in a comments box in the questionnaire, and they had been asked regarding their perception of adaptivity after the study. In previous studies, a system that employed silence rather than hesitation to adapt to participant's level of attention had increased the attention of distracted users, but was perceived as less likable (Carlmeyer et al., 2016b) and rude (Carlmeyer et al., 2016a). This effect appears to be lost in this study, as participants reported that they rather liked the system, which is also reflected in the questionnaire data in both conditions (cf. Figure 6.5).

Regarding the adaptivity, most participants did not report anything in the baseline condition; some participants had the impression that the agent followed their gaze (which is not the case, but the agent looks into the directions of the places

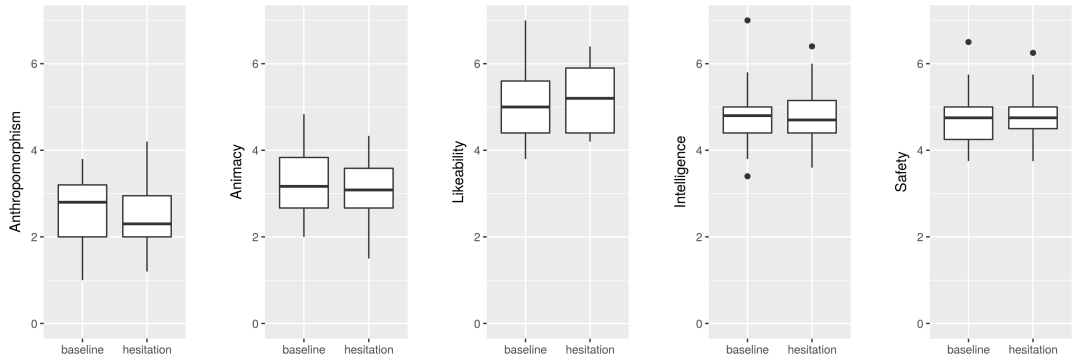

Figure 6.5: *Subjective ratings for the five key concepts.*

he talks about, and users are likely to look in the same direction). In the hesitation condition, many participants noticed the hesitations, but could not figure out what triggered them. Some reported that they like this feature as it grants more time for searching, but most others were put off by the disfluent delivery: In total we have negative sound quality feedback from 13 out of 18 participants that were recorded in the hesitation conditions. In the following interview, however, the notion was rather that the adaptivity is positive and promising for the future, given improvements in the technical realization.

Exploratory Extension of Analysis. We repeated the analysis with the four hesitation stimuli with fillers of bad synthesis quality. The effect on finding rate does not reach significance, however by a very small margin ($\beta = 1.03, SE = 0.53, z = 1.96, p = 0.0504$). The effect on efficiency is significant, when all trials are considered ($W = 83.5, p = 0.02$), (cf. Figure 6.3). This suggests that there is an effect of hesitations on task performance that needs to be considered for future work.

6.2.3 Summary

The results gathered in this preliminary testing of the hesitation model followed the expected directions. Speech synthesis quality suffers from the presence of hesitation, but task performance appears to benefit from it. The evaluation of subjective ratings on the five key concepts as well as qualitative evaluation of user

feedback suggest that the hesitation algorithm tested in this study is acceptable. Thus, for the first study, we can state that H1 and H3 can be accepted for now, and with respect to H2, the results suggest a negative impact of hesitations on users' perceptions of synthesis quality.

6.3 Experimental Study 4: Crowdsourcing-Based Evaluation

In order to assess the quality of the hesitation synthesis in a non-interactive setting, we conducted a parallel online crowdsourcing study. In this evaluation, we used a more traditional approach to speech synthesis evaluation, namely a classic MOS-scale rating task without any interaction between participants and the system. This is done in order to shed light on our underlying assumption that an interactive approach to synthesis evaluation indeed may lead to different conclusions with respect to synthesis quality. As of now, nothing is known about how, and if, interactive and offline evaluation methods produce comparable results. Our main hypothesis for this experiment is undirected, i.e., we do expect a different outcome in terms of speech synthesis quality to that achieved in experiment 3. We do not make any claims about the direction of this hypothesis, as the non-interactive setting may have unforeseeable effects. So far, our only expectation is that the result will differ from the interaction study.

6.3.1 Method

Participants listened to a series of 14 synthetic audio stimuli and rated them individually for their overall quality on a 5-point MOS scale (1 = "very bad", 5 = "very good"). Participants were recruited using mailing lists and social media, and the evaluation builds on a web-based crowdsourcing approach. The listening test was set up using the platform PERCY (Draxler, 2014), specially designed for online audio-based perception studies. Unlike experiment 1, but very much like standard MOS-based synthesis evaluations, participants rated the synthesis quality of each individual stimulus. The participants were not compensated for their participation.

For maximal comparison with the interaction study, we again chose a between-subjects design with the single controlled independent variable *hesitation condition*, which has the two levels, *hesitation* and *baseline*. That is, participants listened to either stimuli containing hesitations only, or to stimuli not containing any hesitations. This may create different stimuli for our two experiments, as in the interactive study, the presence, absence, and length of a hesitation was determined by the participant's individual behavior, and was not necessarily present or absent in each stimulus. Demographic data and information about the output device and individual listening situation is surveyed as well, but not analyzed further.

Before the actual listening tests, participants received some background information of what was being tested (a synthetic voice for usage in an intelligent apartment). They also received some instructions on the procedure of the experiment, i.e., how to use the scale and how long the experiment was likely to last. In both conditions, participants were presented with 14 stimuli which were based upon the text input given to the virtual agent in experiment 3. That way, participants get the same background story (and text) as in the first experiment. Stimuli are divided into 6 introductory, 7 instructive, and 1 concluding utterance. They are presented in the same order for each participant, to generate a coherent story and to ensure maximal similarity with experiment 3. In the baseline condition (non-hesitation), the stimuli are produced with MaryTTS's female German HMM voice, with no further modification. For the hesitation condition, lengthening and silent pauses are woven into each stimulus. In the instructive stimuli, the silent pauses are set to 2000 ms, while in all other stimuli, silences are set to 1000 ms. This difference in duration is motivated by experiment 3, which by design leads to longer pause intervals in the instructions, because participants tend to look around the apartment when possible hiding places are mentioned, these gaze shifts triggering the hesitation mode. Lengthening is applied to syllables preceding the silence with the same durational parameters as in the first study. A list of the stimuli used in this experiment can be found in Appendix B.

The collected data were entered into a linear mixed effects model with *MOS ratings* as the dependent variable, *hesitation condition* as the fixed factor, and *stimulus*, *gender*, and *age* as random factors (random intercepts). This model was compared to a less complex model, leaving out the fixed factor *hesitation condition* using a

likelihood ratio test. All statistical tests were carried out in R, using the R-package *lme4*.

6.3.2 Results and Discussion

We collected ratings from 44 participants (29 female, 15 male) with an age range between 18 and 46 years (mean age: 24.5). With one exception, all participants reported to have entered school in Germany, so we expect them to have a native competence in German. No participant reported any hearing difficulties. Most participants were raised in the vicinity of Bielefeld, and a few in Bavaria. The listening tests typically lasted less than 5 minutes, including the time needed to provide demographic background data. For subsequent analyses, we pooled all participants' data, independent of listening situation, and including one participant who reported to have entered school out of Germany, as the fact that she managed to follow the instructions is an indicator of a sufficiently high competence in German.

On average, MOS-ratings were slightly higher in the baseline condition ($M = 3.28, SD = 0.93$) as compared to the hesitation condition ($M = 2.96, SD = 0.93$) (cf. Figure 6.4). In the linear mixed effects regression (LMER) model containing the fixed factor *hesitation*, the absence of hesitation has a slightly positive, but no significant effect on MOS-ratings ($\beta = 0.31, SE = 0.18, t = 1.78, p = 0.08$). This lack of an effect is further confirmed by the model comparison (likelihood ratio test between models with and without the factor *hesitation*), which does not reveal a significant difference either.

These results are perhaps surprising insofar, as there were reasonable numbers of participants for both conditions (> 20), the test gave listeners a chance to rate each stimulus without being distracted by an ancillary task as in experiment 3, and since participants were confronted with hesitations in each stimulus in the *hesitation condition*. Still, it can only be concluded that even though there is a tendency for stimuli to be rated as slightly less pleasant when hesitations are present, this detrimental effect is not perceived to be significantly strong by listeners in the classic non-interactive approach to speech synthesis evaluation. Of course, most MOS-type analyses rely on within-subjects designs. It is possible,

that participants would have given the stimuli-containing hesitations lower ratings when given a chance for a direct comparison with a stimulus not containing hesitations. However, our aim was to test the influence of an interactive task on speech synthesis ratings. A within-subjects approach would have made such a comparison impossible.

6.4 General Discussion

We tested an incremental spoken dialogue system that is capable of inserting lengthening and silent pauses as a means of hesitation whenever it is required. The experimental results suggest that hesitations are a useful and viable strategy in interaction with users, as they increase task efficiency. Of special interest in this study is the feedback on speech synthesis quality. In addition to the interaction study, we conducted a parallel crowdsourcing experiment with comparable stimuli in order to compare ratings gathered within and without interactive settings. Regarding evaluations in dialogue system and speech synthesis research, we made several observations. Firstly, in dialogue system evaluation, the speech synthesis quality is often not assessed. Secondly, in speech synthesis evaluation, user ratings are surveyed in MOS-based questionnaires regarding stimuli presented without interaction with the system. The results gathered in this study support a claim that has often been reported in the speech synthesis community, which is that the non-interactive evaluation of speech synthesis assesses aspects of synthesis quality that differ from those gathered in interactive settings. Even if it could be guaranteed that what is being assessed really is the "pure" synthesis quality, then it is unclear what to do with this information. Speech synthesis is not used in the void, and there is always some application or interaction associated with it. The problem is not limited to speech synthesis evaluation and human-machine communication, but extends to human conversation. Even in the lab, the circumstances are controllable, but any conversation remains an interaction that cannot be ruled out of the evaluation. As noted by Sacks et al. (1974, p. 699) in their seminal study:

> "*To begin with, a problem for research on actual conversation is that it is always 'situated' – always comes out of, and is part of, some real*

sets of circumstances of its participants."

Our interaction study highlights this notion. As can be seen in Figure 6.4, there are two main differences between MOS-ratings after interaction and after the non-interactive crowdsourcing evaluation. First, stimuli are generally rated better without prior interaction, and second, the presence of hesitation only makes a significant difference in the interaction study. The reason for this discrepancy lies in the nature of the two experimental settings. The crowdsourcing experiment uses neatly pre-constructed stimuli. The interaction study adapts and enhances the stimuli on the fly with spontaneous speech phenomena. The latter will cause artifacts that detriment the synthesis quality, which will be noticed by users and reflected in their feedback. This is the general problem with synthesis evaluation—experimental results from MOS-based questionnaires are not the same as those gathered in interaction studies (and, while being closer to in-the-wild application, interaction studies are still not the reality of application).

It is furthermore possible that the different results for our two experiments may simply be due to the fact that in experiment 3, participants give one score to evaluate the general impression of synthesis, while in experiment 4, participants rate each utterance individually. This was done to truthfully emulate typical MOS-type evaluations, so the main conclusion still holds–we cannot generalize from MOS-type studies to the perceived quality in interactive settings.

Turning to the other objectives of this study, we will now discuss what our evaluation results tell us about the actual system that we tested. It is in general satisfying that there is a tendency towards more task performance and efficiency. The detrimental effect observed for synthesis quality, in turn, highlights the need for improvement.

In general, it shows that the main goal of this thesis, the creation of a hesitation insertion strategy for incremental spoken dialogue systems has been accomplished. We built a working prototype that is able to produce hesitations on-the-fly whenever necessary. The triggers for entering hesitation mode can be adapted, making the system re-usable and transferable for future applications. Relating back to the premises of this thesis, we can state that successful time-buying is possible with synthesized lengthening and silence, only the synthesis quality of fillers is in need

of improvement. With regard to the second premise, it could be shown that users are willing to accept hesitating machines, as despite synthesis quality is rated negatively, users report positive experiences with the hesitating virtual agent.
The fact that some of the effects can be attributed to the technical realization of our hesitation model yielding some audible artifacts gives rise to the question if a simpler strategy could not have achieved the same thing. It may appear unnecessary to develop and implement a complex model that yields technical problems that could have been avoided by simply being silent. In previous studies (Carlmeyer et al., 2016b; Chromik et al., 2017), it was found that strategies that use only silence as a means of hesitation increase visual attention and task performance, but are perceived as rude and less friendly. This is an effect that we cannot observe in our study–the presence of hesitation has no detrimental or beneficial effect on perceived friendliness. Also, feedback gathered in the comments section of the questionnaire and in the short interview after the study suggests that participants assess the adaptive strategy of the system positively, despite the fact that many are rather put off by the disfluent speech delivery. This suggests that the general approach to overtly indicate system hesitation is a promising extension for (virtual) agents' dialogue systems, and doing so with more sophisticated methods than only being silent is credited by users.
To conclude, given some technical improvements, we expect the hesitation model to have future application, which is an objective to explore in follow-up studies. The established strategies of speech synthesis evaluation itself also need to be improved; synthesis designed for interaction needs to be evaluated in interaction.

Part IV

Conclusion

Chapter 7

Summary, Conclusion & Outlook

In this dissertation, we have examined hesitations in human communication and within the framework of incremental spoken dialogue systems. We view the phenomena subsumed under the term *hesitation* as something positive and useful for dialogue. The major premise underlying this work was the concept of *buying time* being an essential tool for managing interaction between speakers or interactants. Time is a valuable commodity in dialogue, both for the speaker and the listener. Disfluencies, especially hesitations are elements of conversation that buy time by bridging gaps of content, which can occur for various reasons, in human communication as well as in human-machine communication. This cues to refer back to the second major premise, namely *machines hesitating like humans are acceptable for human listeners*. It has been shown that humans readily interact with hesitating machines and that this capability of dialogue behavior needs not be perceived negatively. The most frequently asked question in this context is why machines should be able to hesitate. When this research commenced, it was feasible to use hesitations to bridge gaps resulting from processing time. With the advent of greatly increased computational power, this reason has dwindled in importance. However, hesitations have been proven useful to manage attention and information delivery in dialogue, which is a core aspect of communication between humans as well as between humans and machines. There are numerous issues to be addressed regarding the synthetic realization of hesitation elements, but we were able to prove that machines in general can be equipped with hesitation capabilities.

Models of speech production suggest that monitoring mechanisms within the human speech production apparatus are able to rapidly detect mismatches between planned speech and its acoustic realization. It is suggested to use pre-planned but not-yet-uttered speech material in the articulatory buffer for hesitation purposes, i.e. continuing to articulate a path of previously planned speech, in spite of the knowledge that this path is a dead end. Only when this material runs low, hesitations will be deployed, starting out by lengthening material from the buffer, and producing more salient and severe hesitations when the correct resumption fails to become available in time. This human behavior is mimicked in a simplified form in the hesitation insertion strategy for speech synthesis and dialogue systems presented in this thesis. The strategy itself is a very small selection of molecular elements out of the vast amount of possible surface forms of disfluencies. We showed that a dynamic insertion of lengthening, silences (and fillers) results in plausible hesitation behavior, which has a positive effect on task efficiency in humans interacting with a virtual agent and which does not detriment likability of the system compared to other studies deploying simpler hesitation strategies.
The core part of the model is lengthening. In the studies presented here, it could be shown that lengthening does not behave like other disfluencies. Lengthening instances dodge perception, even of trained annotators, and they can be synthesized with remarkable quality. Consequently, lengthening is a means to temporally extend the speech signal, thus buying extra time. When this process is not noticed by the listener, then the time is virtually *bought for free*: as was shown in this thesis, other means of hesitation, such as fillers, can have a negative effect - the user not noticing the hesitating system buying time is thus a desideratum. This leads to the paradoxical situation to desire synthesizing something that users won't be able to perceive, in order to achieve minimally invasive time-buying for dialogue systems. This subtle and elusive phenomenon is a very promising candidate for future work on conversational speech synthesis that endeavors to create machines capable of talking to users instead of reading out text. Furthermore, on a different level, the applicability of this lengthening-centered model has implications for research in psycholinguistics. It suggests that Levelt's concept of starting hesitations by prolonging buffer material is not only a theorem for human speech production, but is also practically applicable for conversational speech synthesis.

When the project of equipping dialogue systems with hesitation started five years ago, we were confronted with many questions. A lot of these questions have been answered in this thesis, but inevitably, new questions arose. Parts of this thesis are built upon such follow-up questions, such as the deep focus on lengthening, which only came into existence after the first empirical study. Within the studies on lengthening, more questions arose, such as whether there is a duration maximum threshold for lengthening, or how hesitation lengthening is related to the elasticity hypothesis. These questions have been answered and the answers have been implemented into the virtual agent for the final study. However, many other questions remain out of scope for this thesis and thus unanswered. While the body of text constituting this thesis is running towards its end, we will close with an outlook to some of the questions which we hope to be able to answer in the time to come.

The first question regards the technical realization and the speech synthesis software itself. The final implementation of the hesitation strategy was done with the open-source tools for incremental speech synthesis, MaryTTS and InproTK. Time moves swiftly and during the roughly five years this project encompassed, new commercial systems have revolutionized the market and new open-source tools tapping on the potential of *deep learning* have become available that enable synthesis of spontaneous speech phenomena at higher quality. To our knowledge, there is no incremental speech synthesis software of this high quality, but it also has been demonstrated that incrementality is not mandatory to build interactive systems (Wester et al., 2017). Many insights gained in this dissertation are very foundational and can be implemented into any synthesis system. The main objective is finding a system that is open-source, high-quality and interactive. At the moment (2019), *Merlin* (Wu et al., 2016) seems a suitable candidate, and follow-up studies with this system are currently being developed by the author of this thesis.
Another question that inevitably arises when working in this field regards evaluation. There are paradigms to evaluate dialogue systems (usually asking for feedback on key concepts), and there are paradigms for evaluating speech synthesis (usually collecting Mean Opinion Scores). In this thesis, however, we did not

build a new speech synthesis system, nor did we create a novel dialogue system. We rather did something in between: we adapted speech synthesis to be able to function in highly interactive dialogue systems. Of course we can evaluate the resulting dialogue system or ask people for opinions about the synthesis. In fact we did exactly that. But the thing we are actually interested in is somewhere in between: how suitable is this adaptive speech synthesis with hesitations for interactive dialogue systems? This is something which is hard to come by with current evaluation paradigms. In this thesis, we therefore strongly advocate rethinking the field of synthesis evaluation and move towards interactive evaluation for interactive synthesis.

As of now, we can only muse how to achieve this. In general, there are two possible starting points: one can either use the dialogue system evaluation to infer something for speech synthesis quality, or one can make offline evaluations more interactive. There is no obvious way to get precise first-hand user feedback on synthesis quality from an interaction study, as the interaction cannot be interrupted in between to ask for feedback. Neither can task performance measures from the study be used to directly infer the impact of the speech synthesis. One conceivable option would be to have external evaluators review the recorded interactions and give feedback on the synthesis quality every given time interval. If the stimuli that participants have to rate would be embedded in small-scale interactive scenarios, interactive measures like reaction time, task completion time, or task performance in general could be surveyed in addition to the MOS feedback, helping to analyze and interpret the results. Preliminary tests with relative task completion time for instructive stimuli in connection with MOS-feedback were explored in section 4.4. Speech synthesis evaluation as of now is an unsolved problem. Speech synthesis does not exist without interaction, thus it makes no sense to evaluate it without. If any given speech synthesis system achieved good MOS scale ratings, it would at least be necessary to test the system in interaction to see if the results can be justified. Non-interactive MOS-based evaluation, however, maximally reflects the opinion of a user testing it in a disembodied way without the application it may be designed for. This may suffice for general evaluation purposes like overall intelligibility, but the challenge remains to evaluate the quality of *conversational*

synthesis in interaction. We dipped into novel ways of evaluating speech synthesis and combined traditional questionnaire-based methods with interactive measurements. Much work is to be done in order to optimize gathering of evaluation results in interactive scenarios, but it is undisputed that without interaction, evaluation is incomplete. The next step in hesitation synthesis evaluation on our agenda is a mouse-tracking paradigm that monitors users' task and gaming behavior exposed to hesitant instructions.

Another subject that was touched in this thesis, but not dealt with in depth is the *signal hypothesis*. The disfluency community has witnessed heated debates over the question whether disfluencies are an intentional signal by the speaker, or merely a symptom that is interpreted by the listener. There is a tendency to the latter, but most research on these lines relied heavily on fillers. Having dealt a lot with lengthening over the course of this project, we want to explore how synthesized lengthening behaves in these circumstances. Can lengthening signal anything? Can listeners "read" lengthening like other disfluencies? Our hypothesis is that the notion should not be "disfluencies are useful", but rather "different disfluency elements do different interesting things".

Other hesitation elements also demand future attention. One valuable contribution to the disfluency community would be an investigation of the notion of *standard maximum silence*. Our heuristic for silence insertion worked, and it is unsure if it is worth the effort investigating the standard maximum silence for minor refinements. But from a general research perspective it is desirable to shed light on the question if there really is such a thing, a threshold after which silence in dialogue becomes unbearable. So far only one study (on English) exists (Jefferson, 1989). It would be interesting to try and replicate that study for different languages.

Lengthening has done a lot of work for us in the past five years, and it being the workhorse of hesitation synthesis, it is honored with the final open question, namely how do lengthening and other hesitation elements like fillers interact? It needs to be investigated in more detail how regularly lengthening precedes fillers. Based on first exploratory inspections, contrary to some claims in the literature (e.g. Adell et al. (2008)), they do not appear to introduce every filler, but rather fillers of the *uhm* type. It would be an important strategic information for further elaboration of the hesitation synthesis model to know of regularities, e.g. length-

ening only preceding certain types or extents of fillers.
With this, we leave the reader pondering questions and undertake no further attempts of lengthening this thesis.

Part V

Appendix

Appendix A

Stimulus Text for Smart-Home Study

Reading pauses are indicated by their duration (in seconds) in brackets.

Hallo, schön, dass du an dieser Studie teilnimmst. (0.5) Mein Name ist Flobi und ich bin dein virtueller Ansprechpartner in dieser Studie. (0.5) Durch die Sensoren in diesen Räumen bin ich über alles informiert, was hier geschieht. Ich werde dir heute ein wenig über dieses Apartment erzählen, und dann habe ich eine kleine Aufgabe für dich. Du könntest mir nämlich beim Suchen helfen; hier sind eben ein paar Sachen verloren gegangen. (0.5) Aber zunächst zu der Umgebung. (1.0) Du bist hier in einem intelligenten Apartment, das mit einer Menge Technik ausgestattet ist. (0.5) Ein paar Beispiele: (0.5) Zu deiner linken Seite siehst du die Küche. Sie ist voll funktionsfähig. An den Schränken sind an manchen Türgriffen Lichter angebracht. Diese signalisieren dir in blau, dass ich dir dort etwas zeigen möchte. Sobald der Schrank geöffnet wird, werden sie grün, und wenn er wieder geschlossen wird, geht das Licht wieder aus. (1.0) Rechts von dir ist das Wohnzimmer. (0.5) Wie du siehst, hängt dort ein großer Bildschirm rechts an der Wand. (0.5) Der Tisch, der dort steht, ist interaktiv. Man kann sich auf ihm einen Plan von dieser Wohnung anschauen. Außerdem können der Tisch und auch der Bildschirm genutzt werden, um bei einer Besprechung, Präsentatzionen und Videos zu zeigen. (1.0) Das Apartment ist auch mit einer Reihe von Kameras und Sensoren

ausgestattet, die größtenteils an der Decke hängen. Du wirst gleich noch die Gelegenheit bekommen, dich hier etwas umzusehen. Ich möchte dir nämlich jetzt von der Aufgabe erzählen. (1.0) Du sollst mir helfen, ein paar Dinge wieder zu finden. (0.5) Vor etwa einer Stunde ist hier folgendes passiert: (0.5) Einige Handwerker waren hier im Appartment und haben die Küche umgebaut. (0.5) Dabei wurde jede Menge Staub aufgewirbelt, was meine Sensoren beeinträchtigt hat. (0.5) Währenddessen haben andere Leute Sachen hier im Appartment versteckt. (0.5) Ich vermute, es handelt sich dabei um die Münzen und die Schokolade, die ich vorher auf dem Tisch gesehen habe. (0.5) Ich konnte wegen des Staubs leider nicht genau erkennen, wo die Sachen versteckt wurden, aber ich werde dir alles erzählen, was ich weiß. (0.5) Dann kannst du versuchen, soviel wiederzufinden wie möglich. (0.5) Pass jetzt gut auf. Ich sage dir was ich gesehen habe. (0.5) Versuch, dir alles zu merken! (0.5) Danach muss ich mich einmal neu starten, um meine Sensoren wieder klar zu kriegen. (0.5) Wenn ich neu gestartet bin, kannst du dich auf die Suche begeben. (1.0) Jemand hat die Waschmaschine bedient und das Waschpulverfach geöffnet. Da würde ich später auf jeden Fall mal nachsehen! (1.0) Und ich habe gesehen, wie jemand zur Pflanze im Wohnzimmer gegangen ist und etwas am Blumentopf gemacht hat. Da solltest du auch mal nachsehen! (1.0) Danach hat jemand die Beschteckschublade geöffnet und hat dort rumgewühlt. Vielleicht ist da etwas versteckt! (1.0) Und dann habe ich beobachtet dass jemand den Schrank über der Mikrowelle aufgemacht hat. Schau doch da mal rein! (1.0) Dann wurde einer der Stühle im Wohnzimmer bewegt. Die solltest du auch mal untersuchen. (1.0) Irgend etwas ist mit den Kaffeetassen auf dem Tisch im Wohnzimmer passiert. Da könnte auch etwas versteckt sein. (1.0) Zu guter Letzt war noch jemand am Beschteckfach der Spülmaschine. Ich weiß nicht, ob es mit der Sache zu tun hat, aber schau gleich mal nach. (1.0) So, ich starte mich neu und bin in wenigen Sekunden zurück. (5.0) So, jetzt bin ich auch wieder bereit. (0.5) Jetzt kommt dein Part. (0.5) Schau in beliebiger Reihenfolge an den Orten nach, die ich dir genannt habe. (0.5) Bevor du an einem Ort nachsiehst, sag mir bitte einmal den Namen des Ortes. (0.5) Alles, was du findest, darfst du behalten.

Appendix B

Stimuli for Crowdsourcing Study

The following stimuli are used for the crowdsourcing experiment described in Section 6.3. Lengthened syllables are indicated by appended colons. Pauses are indicated by seconds in brackets. Lengthening durations are determined as described in Section 6.1.2.3. Stimuli for the baseline condition are the same, except without lengthenings and pauses.

Introduction

1. "Hallo, schön, dass du an: (1.0) dieser Studie teilnimmst."
2. "Ich werde dir heute ein wenig über dieses Apartment erzählen, un:d (1.0) dann habe ich eine kleine Aufgabe für dich."
3. "Du könntest mir nämlich beim Suchen helfen. Hier sind eben ein paa:r (1.0) Sachen verloren gegangen."
4. "Einige Handwerker waren hier im Apartment un:d (1.0) haben die Küche umgebaut."
5. "Ich konnte wegen des Staubs leider nicht genau erkennen, wo die: (1.0) Sachen versteckt wurden."

Instruction

1. "Jemand hat die Waschmaschine bedient un:d (2.0) das Waschpulverfach geöffnet."

2. “Und ich habe gesehen, wie jemand zur Pflanze im Wohnzimmer gegangen ist, un:d (2.0) etwas am Blumentopf gemacht hat.”

3. “Danach hat jemand die Beschteckschublade geöffnet un:d (2.0) hat dort rumgewühlt.”

4. “Und dann habe ich beobachtet dass jemand den Schrank über der: (2.0) Mikrowelle aufgemacht hat.”

5. “Dann wurde einer der Stühle im: (2.0) Wohnzimmer bewegt.”

6. “Irgend etwas ist mit den Kaffeetassen auf dem Tisch im: (2.0) Wohnzimmer passiert.”

7. “Zu guter Letzt war noch jemand am Beschteckfach der: (2.0) Spülmaschine.”

Conclusions

1. “Schau in beliebiger Reihenfolge an: (1.0) den Orten nach, die ich dir genannt habe.”

Bibliography

Adell, J., Bonafonte, A., and Escudero-Mancebo, D. (2008). On the generation of synthetic disfluent speech: Local prosodic modifications caused by the insertion of editing terms. In *Proceedings of Interspeech*, pages 2278–2281.

Adell, J., Bonafonte, A., and Escudero-Mancebo, D. (2010). Modelling filled pauses prosody to synthesise disfluent speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4810–4813.

Allwood, J. (1995). Reasons for management in spoken dialogue. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, 142:241–241.

Allwood, J., Nivre, J., and Ahlsén, E. (1990). Speech management—on the non-written life of speech. *Nordic Journal of Linguistics*, 13(1):3–48.

Altmann, G. T. and Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of memory and language*, 57(4):502–518.

Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., et al. (2017). Deep voice: Real-time neural text-to-speech. *arXiv preprint arXiv:1702.07825*.

Arnold, J. E., Kam, C. L. H., and Tanenhaus, M. K. (2007). 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*, 33(5):914.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., and Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological science*, 15(9):578–582.

Barr, D. J. (2001). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. *Oralité et gestualité: Interactions et comportements multimodaux dans la communication*, pages 597–600.

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Baumann, T. and Schlangen, D. (2012). The inprotk 2012 release. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, SDCTD '12, pages 29–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bear, J., Dowding, J., and Shriberg, E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics.

Belz, M. and Reichel, U. D. (2015). Pitch characteristics of filled pauses in spontaneous speech.

Betz, S., Carlmeyer, B., Wagner, P., and Wrede, B. (2018). Interactive hesitation synthesis: modelling and evaluation. *Multimodal Technologies and Interaction*, 2(1).

Betz, S., Eklund, R., and Wagner, P. (2017a). Prolongation in German. In Eklund, R. and Rose, R., editors, *Proceedings of DiSS 2017, Disfluency in Spontaneous Speech*, volume 58, pages 13–16.

Betz, S., Voße, J., Zarrieß, S., and Wagner, P. (2017b). Increasing recall of lengthening detection via semi-automatic classification. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017, Stockholm)*, pages 1084–1088.

Betz, S., Wagner, P., and Schlangen, D. (2015a). 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)*, pages 2222–2226.

Betz, S., Wagner, P., and Schlangen, D. (2015b). Modular synthesis of disfluencies for conversational speech systems. In Wirsching, G., editor, *Elektronische Sprachsignalverarbeitung (ESSV) 2015*, Studientexte zur Sprachkommunikation, pages 128–134. TUD Press.

Betz, S., Wagner, P., and Vosse, J. (2016). Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. In *Phonetik und Phonologie 12*, pages 19–23.

Betz, S., Zarrieß, S., and Wagner, P. (2017c). Synthesized lengthening of function words - The fuzzy boundary between fluency and disfluency. In Degand, L., editor, *Proceedings of the International Conference Fluency and Disfluency*, pages 15–19.

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

Bögels, S. and Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Bohus, D. and Horvitz, E. (2014). Managing human-robot engagement with forecasts and... um... hesitations. In *Proc. of the 16th International Conference on Multimodal Interaction*, pages 2–9, New York, USA. ACM.

Brennan, S. E. and Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

Brugos, A. and Shattuck-Hufnagel, S. (2012). A proposal for labelling prosodic disfluencies in tobi. In *Poster presented at Advancing Prosodic Transcription for Spoken Language Science and Technology*.

Brutten, E. J. (1963). Palmar sweat investigation of disfluency and expectancy adaptation. *Journal of Speech, Language, and Hearing Research*, 6(1):40–48.

Campbell, W. N. and Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, 19(1):37–47.

Campione, E. and Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Speech prosody 2002, international conference*, pages 199–202.

Carlmeyer, B., Schlangen, D., and Wrede, B. (2014). Towards closed feedback loops in hri: Integrating inprotk and pamini. In *Proceedings of the 2014 Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction*, ICMI-MMRWHRI '14, pages 1–6. ACM.

Carlmeyer, B., Schlangen, D., and Wrede, B. (2016a). Exploring self-interruptions as a strategy for regaining the attention of distracted users. In *Proceedings of the 1st Workshop on Embodied Interaction with Smart Environments - EISE '16*. Association for Computing Machinery (ACM).

Carlmeyer, B., Schlangen, D., and Wrede, B. (2016b). "Look at Me!": Self-Interruptions as Attention Booster? In *Proceedings of the Fourth International Conference on Human Agent Interaction - HAI '16*, pages 221–224. Association for Computing Machinery (ACM).

Carlson, R., Gustafson, K., and Strangert, E. (2006). Cues for hesitation in speech synthesis. In *Ninth International Conference on Spoken Language Processing*, pages 1300–1303.

Chafe, W. (1980). Some reasons for hesitating. *Temporal variables in speech: Studies in Honour of Frieda Goldman-Eisler*, pages 169–180.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press.

Chromik, M., Carlmeyer, B., and Wrede, B. (2017). Ready for the Next Step?: Investigating the Effect of Incremental Information Presentation in an Object Fetching Task. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pages 95–96. Association for Computing Machinery (ACM).

Clark, H. (2002). Speaking in time. *Speech Communication 36*.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Clark, H. H. and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Corley, M. and Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602.

Dall, R., Tomalin, M., and Wester, M. (2016). Synthesising Filled Pauses: Representation and Datamixing. In *Proc. SSW9*, pages 7–13, Cupertino, CA, USA.

Dall, R., Wester, M., and Corley, M. (2014a). The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech. In *Proc. Interspeech*, pages 56–60.

Dall, R., Yamagishi, J., and King, S. (2014b). Rating naturalness in speech synthesis: The effect of style and expectation. *Proceedings Speech Prosody, Dublin, Ireland*, pages 1012–1016.

De Ruiter, J.-P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Draxler, C. (2014). Online experiments with the percy software framework - experiences and some early results. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 235–240, Reykjavik, Iceland. European Language Resources Association (ELRA).

Eklund, R. (2001). Prolongations: A dark horse in the disfluency stable. In Core, M. G., editor, *Disfluency in Spontaneous Speech (DiSS '01)*, pages 5–8, Edinburgh, Scotland.

Eklund, R. (2004). *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Linköping University Electronic Press.

Fischer, K., Niebuhr, O., Novák-Tót, E., and Jensen, L. C. (2017). Strahlt die negative reputation von häsitationsmarkern auf ihre sprecher aus? In *Proc. 43rd Annual Meeting of the German Acoustical Society (DAGA), Kiel, Germany*, pages 1450–1453.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6):709–738.

Ginzburg, J., Fernández, R., and Schlangen, D. (2014). Dysfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, pages 9:1–9:64.

Goto, M., Itou, K., and Hayamizu, S. (1999). A real-time filled pause detection system for spontaneous speech recognition. In *Eurospeech*, pages 227–230.

Götz, S. (2013). *Fluency in native and nonnative English speech*, volume 53. John Benjamins Publishing.

Gravano, A. and Hirschberg, J. (2009). Turn-yielding cues in task-oriented dialogue. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 253–261. Association for Computational Linguistics.

Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555 – 568.

Holthaus, P., Leichsenring, C., Bernotat, J., Richter, V., Pohling, M., Carlmeyer, B., Köster, N., zu Borgsen, S. M., Zorn, R., Schiffhauer, B., Engelmann, K. F., Lier, F., Schulz, S., Cimiano, P., Eyssel, F., Hermann, T., Kummert, F., Schlangen, D., Wachsmuth, S., Wagner, P., Wrede, B., and Wrede, S. (2016).

How to address smart homes with a social robot? a multi-modal corpus of user interactions with an intelligent environment. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3440–3446, Paris, France. European Language Resources Association.

Hough, J., de Ruiter, L., Betz, S., and Schlangen, D. (2015). Disfluency and laughter annotation in a light-weight dialogue mark-up protocol.

Hough, J., Tian, Y., de Ruiter, L., Betz, S., Schlangen, D., and Ginzburg, J. (2016). DUEL: A Multi-lingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter. In *10th edition of the Language Resources and Evaluation Conference*, pages 1784–1788.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a'standard maximum'silence of approximately one second in conversation.

Kempen, G. and Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive science*, 11(2):201–258.

King, S. (2017). What speech synthesis can do for you (and what you can do for speech synthesis). In *Proceedings of the 18th International Congress of the Phonetic Sciences (ICPhS 2015)*, pages 249–253.

Klatt, D. H. (1973). Interaction between two factors that influence vowel duration. *The Journal of the Acoustical Society of America*, 54(4):1102–1104.

Klatt, D. H. (1976). Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59(5):1208–1221.

Kohler, K. J. (1983). Prosodic boundary signals in german. *Phonetica*, 40(2):89–134.

Kousidis, S., Kennington, C., Baumann, T., Buschmeier, H., Kopp, S., and Schlangen, D. (2014). Situationally Aware In-Car Information Presentation Using Incremental Speech Generation: Safer, and More Effective. In *Proceedings of the EACL 2014 Workshop on Dialogue in Motion*, pages 68–72.

Kousidis, S., Pfeiffer, T., and Schlangen, D. (2013). Mint.tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora. In *Proceedings of Interspeech*, pages 2649–2653.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Levelt, W. J. (1984). Spontaneous self-repairs in speech: Processes and representations. pages 105–115.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. MIT Press.

Lickley, R. J. (2015). *Fluency and Disfluency*, pages 445–474. John Wiley & Sons, Inc.

Lundholm Fors, K. (2015). *Production and Perception of Pauses in Speech*. PhD thesis, University of Gothenburg.

Lütkebohle, I., Hegel, F., Schulz, S., Hackel, M., Wrede, B., Wachsmuth, S., and Sagerer, G. (2010). The bielefeld anthropomorphic robot head "flobi". In *2010 IEEE International Conference on Robotics and Automation*, pages 3384–3391. IEEE.

Maclay, H. and Osgood, C. E. (1959). Hesitation phenomena in spontaneous english speech. *Word*, 15:19–44.

Möller, S. (2017). *Quality Engineering: Qualität kommunikationstechnischer Systeme*. Springer-Verlag.

Molloy, L. and Isard, S. (1998). Suprasegmental duration modelling with elastic constraints in automatic speech recognition.

Mori, M. (1970). Bukimi no tani: the uncanny valley. *Energy*, 7:33–35.

Neelley, J. N. (1961). A study of the speech behavior of stutterers and nonstutterers under normal and delayed auditory feedback. *The Journal of speech and hearing disorders*, page 63.

O'Connell, D. C. and Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, 34(6):555–576.

O'Shaughnessy, D. (1995). Timing patterns in fluent and disfluent spontaneous speech. In *International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995*, volume 1, pages 600–603. IEEE.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735.

Schillingmann, L. and Nagai, Y. (2015). Yet another gaze detector: An embodied calibration free system for the icub robot. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 8–13.

Schlangen, D. and Skantze, G. (2011). A general, abstract model of incremental dialogue processing. *Dialogue and Discourse*, 2(1):83–111.

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

Schweitzer, A. and Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)*, pages 525–529.

Seyfeddinipur, M., Kita, S., and Indefrey, P. (2008). How speakers interrupt themselves in managing problems in speaking: Evidence from self-repairs. *Cognition*, 108(3):837–842.

Shriberg, E. (1994). Preliminaries to a theory of speech disfluencies. *Ph D. thesis University of California*.

Shriberg, E. (2001). To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Shriberg, E. E. and Lickley, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, 50(3):172–179.

Skantze, G. and Hjalmarsson, A. (2013). Towards incremental speech generation in conversational systems. *Computer Speech and Language 27*, pages 243–262.

Skantze, G. and Schlangen, D. (2009). Incremental Dialogue Processing in a Micro-Domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 745–753.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., and Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive psychology*, 45(4):447–481.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Taylor, P. (2009). *Text-to-Speech Synthesis:*. Cambridge University Press.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K. (2013). Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, 101(5):1234–1252.

Turk, A. E. and Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35(4):445–472.

Umeda, N. (1977). Consonant duration in american english. *The Journal of the Acoustical Society of America*, 61(3):846–858.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*.

Wagner, P. and Betz, S. (2017). Speech Synthesis Evaluation – Realizing a Social Turn. In *Tagungsband Elektronische Sprachsignalverarbeitung (ESSV)*, page 167–172.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*.

Wester, M., Braude, D. A., Potard, B., Aylett, M., and Shaw, F. (2017). Real-time reactive speech synthesis: Incorporating interruptions. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017, Stockholm)*, pages 3996–4000.

Wingate, M. E. (1984). Fluency, disfluency, dysfluency, and stuttering. *Journal of Fluency Disorders*, 9(2):163–168.

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications.*

Wrede, S., Leichsenring, C., Holthaus, P., Hermann, T., and Wachsmuth, S. (2017). The Cognitive Service Robotics Apartment: A Versatile Environment for Human-Machine Interaction Research. *KI - Kuenstliche Intelligenz (Special Issue Smart Environments)*, pages 299–304.

Wu, Z., Watts, O., and King, S. (2016). Merlin: An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*, pages 202–207.