




Article

Interactive Hesitation Synthesis: Modelling and Evaluation

Simon Betz ^{1,2,3,*} , Birte Carlmeyer ^{1,3,4}, Petra Wagner ^{1,2} and Britta Wrede ^{1,4}

¹ Cluster of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University, 33615 Bielefeld, Germany; b.carlmeyer@uni-bielefeld.de (B.C.); petra.wagner@uni-bielefeld.de (P.W.); bwrede@techfak.uni-bielefeld.de (B.W.)

² Phonetics and Phonology Workgroup, Faculty of Linguistics and Literary Studies, Bielefeld University, 33615 Bielefeld, Germany

³ Dialogue Systems Group, Faculty of Linguistics and Literary Studies, Bielefeld University, 33615 Bielefeld, Germany

⁴ Applied Informatics Group, Faculty of Technology, Bielefeld University, 33615 Bielefeld, Germany

* Correspondence: simon.betz@uni-bielefeld.de; Tel.: +49-521-106-3518

Received: 8 December 2017; Accepted: 26 February 2018; Published: 2 March 2018

Abstract: Conversational spoken dialogue systems that interact with the user rather than merely reading the text can be equipped with hesitations to manage dialogue flow and user attention. Based on a series of empirical studies, we elaborated a hesitation synthesis strategy for dialogue systems, which inserts hesitations of a scalable extent wherever needed in the ongoing utterance. Previously, evaluations of hesitation systems have shown that synthesis quality is affected negatively by hesitations, but that they result in improvements of interaction quality. We argue that due to its conversational nature, hesitation synthesis needs interactive evaluation rather than traditional mean opinion score (MOS)-based questionnaires. To validate this claim, we dually evaluate our system's speech synthesis component, on the one hand, linked to the dialogue system evaluation, and on the other hand, in a traditional MOS way. We are thus able to analyze and discuss differences that arise due to the evaluation methodology. Our results suggest that MOS scales are not sufficient to assess speech synthesis quality, leading to implications for future research that are discussed in this paper. Furthermore, our results indicate that synthetic hesitations are able to increase task performance and that an elaborated hesitation strategy is necessary to avoid likability issues.

Keywords: speech synthesis; evaluation; hesitation; virtual agents; interaction

1. Introduction

1.1. Motivation and Aims of This Study

Synthetic speech is applied in various fields, and it has entered the realm of everyday life, be it in public transportation announcements, telephone customer services, smartphone speech output, or smart home environments. Despite the interactive nature of many of these applications, speech output remains to be rather static, typically reading out pre-defined texts or often responding with an awkward delay.

Also, a special feature of synthetic speech is its “fluency”, i.e., it does not contain the hesitations, reformulations, or filled pauses typical in human spontaneous speech production. Rather, speech output, once generated, is produced in a single, non-interrupted fashion. The study we are presenting in this paper rests on the assumption that this is suboptimal for many human–machine interactions where listeners need to actually process information that is synthetically generated, and where a human speaker would try to deliver the information in a way which is suited to the

listener's attention level to enable him or her to follow and process what is being said (previously explored in [1,2]). In order to test this assumption, we will explore the space of possible improvements of speech synthesis for interactive purposes using synthesized hesitations.

Our assumption rests on the finding that the hesitations produced in spontaneous speech communication are not merely disturbances or "errors" of human speech production, a view first presented by Wallace Chafe [3] (p. 170). Rather, they serve an important role in dialogue—they allow the speaker to have extra time in situations where it is needed, e.g., when searching for the right thing to say, and to signal this to the listener. That way, hesitations help to keep the metaphorical right to speak. It has been shown previously that spoken dialogue systems can utilize hesitations to bridge gaps in dialogue, and to successfully handle interruptions and attention shifts (e.g., [1,2,4]).

In this study, we explore the applicability of an elaborated hesitation synthesis strategy that is based on observations of human hesitations. Upon an event of hesitation, a hierarchical hesitation insertion is triggered that continues "buying time" as long as possible or until it receives a signal for ending the hesitation mode. The start and stop signals for hesitation insertion are inferred from the user's attention—when users look away, the system will enter a hesitation mode until users re-focus. Furthermore, we test a model of optimal hesitation placement. Compared to previous hesitating systems, the approach presented here allows for dynamic hesitation insertion in the middle of an utterance and for flexible, scalable hesitation clusters tailored for hesitation events of various extents.

Due to the intrinsically interactive fashion of our hesitation strategy, its evaluation is not straightforward. While the system as a whole can be evaluated with interactive measures such as task performance, speech synthesis components are usually evaluated using non-interactive measures, in which listeners are asked to rate the quality of synthetic speech, typically individual utterances, using mean opinion scores (MOSs). Despite numerous criticisms of this method, alternatives have seldom been proposed [5–9]. Also, to our knowledge, no previous study actually provides empirical evidence for these critical claims. We therefore test our hesitation synthesis twice. First, we evaluate it in direct connection with the dialogue system evaluation in interaction, and interpret objective measures like task performance and efficiency alongside subjective user ratings of system features such as synthesis quality. Second, we assess the subjective speech synthesis quality in a non-interactive, crowdsourcing-based parallel study that uses the same stimuli. That way, we can compare user task performance and their subjective impression of a system with subjective ratings where the interaction quality is not part of the evaluation strategy. Ultimately, we hope to not only be able to evaluate our own synthesis approach, but also shed light on the issue of what traditional approaches to speech synthesis evaluation actually reveal. We do not claim to present an all-new evaluation paradigm, but to shed light on these under-researched areas for their future improvement.

1.2. Scope of This Study

In this study, we present and discuss a model to dynamically insert hesitations into ongoing speech output of incremental spoken dialogue systems. A preliminary implementation of this model is evaluated in an interaction study: it is fully capable of producing hesitations, but does not yet exploit the full depth of the model. We discuss the current evaluation paradigm and explore new ways of evaluating speech synthesis components of dialogue systems. This study is conceived as foundational research to design and evaluate interactive and conversational spoken dialogue systems. A fully fledged implementation of the hesitation model is subject for follow-up studies.

1.3. Structure of This Paper

In the following chapter, we provide further background for this study. First, we define the term *hesitation* as we use it and give a brief overview of its research history in Section 2.1. We continue with the description of a model of incremental speech production, which serves as a foundation for defining and discussing hesitations, incremental spoken dialogue systems, and synthesis strategies in Section 2.2. In Section 2.3 we continue with a brief introduction to dialogue systems with a special

focus on systems with incremental processing, as this is a crucial prerequisite of our hesitation model. With this foundation set up, we turn towards our model for a hesitation synthesis strategy for incremental spoken dialogue systems based on studies of human speech production in Section 3. In the empirical part of this paper we present two experiments that make up this study. First, we describe the methods, results and discussion of an interaction study in Section 4, continuing with a parallel crowdsourcing-based study for evaluation purposes in Section 5. The main part then concludes with a general discussion of both experiments and their implications in Section 6.

2. Background and Related Work

2.1. Form and Function of Hesitations

Hesitations are lexical and non-lexical elements that delay information delivery in speech. The most common hesitations include fillers, silences, lengthenings, and repetitions, cf. Example 1. Traditionally, hesitations are often associated with disfluencies. In this study we only consider hesitation phenomena. For an excellent overview on the historical entanglement of hesitations and disfluencies, see [10]. For the most influential descriptive work on disfluencies in general, see [11].

Example 1. *“Take thee ... uhm ... the, the red line to the university”. Different surface forms of hesitation: lengthening, silence, filler, repetition.*

It has been noted that hesitations do not only buy time, but that they are a useful strategy for both the speaker and listener to manage the conversation. The authors of [12] suggested that speakers intentionally decide to produce a filler as either “uh” or “uhm”, the former denoting only a small delay, the latter a major problem accompanied by a longer pause in speech. This leads to the assumption that this difference in form is a listener-oriented strategy, a means to ensure that the interlocutor is informed about the current dialogue state and does not attempt to grab the conversational floor too early.

It has further been observed that hesitations, with their property to control information timing in dialogue, are linked to users’ visual attention. This relationship may be bilateral: the author of [13] found that speakers hesitate when the listener is apparently distracted, and the authors of [14,15] found that listeners may heighten attention when a hesitation is made.

While it is highly controversial whether hesitations and disfluencies are produced in order to actively transmit information to the listener (see [16] for an overview), or if it is merely the fact that the listener can do something with the information, we have sufficient evidence that listeners can make active use of the extra time that hesitations grant in dialogue, an effect that is replicable for human–machine interactions, (e.g., [1,2,4,17]).

Shifting the focus back to the speaker, with the aim in mind to adapt speaker strategies for dialogue systems, we encounter several common reasons for hesitating. Speakers might have trouble retrieving the correct or most appropriate item (cf. Example 2). They might run out of things to say before having conveyed the intended message (cf. Example 3). The dialogue situation might change, causing a change in speech plan, that needs time (cf. Example 4).

Example 2. *“The capital of Serbia is ... uhm ... Belgrade.” Difficulty retrieving an infrequent lexical item.*

Example 3. *“There is no direct flight to Sydney ... uhm ... today or tomorrow...”. Travel agent giving information, but the database query takes time.*

Example 4. *“You can take a seat ... in the living room.” Originally, the plan was to offer a seat in the kitchen, but as the interlocutor apparently shifted her attention to the living room during the dialogue, a new speech plan was realized.*

The above three are all fictional examples, but they shed light on the various usages of hesitations. The surface forms might be indefinitely complex for every situation, with any combination of the elements suggested in the introductory Example 1, and equally often will they be minimally complex, consisting of only one of the hesitation elements in question. The challenge in this study will be to model plausible surface forms of hesitations for a dialogue system that can use them on the fly whenever the situation requires it. We hypothesize that due to the manifold forms hesitations can take, listeners will be tolerant and accept a wide range of forms instantiated by a dialogue system.

2.2. Incremental Speech Production

Hesitations are closely related to the way humans speak. When initiating an utterance, speakers have not fully pre-planned what to say and how to say it. Instead, they plan and produce speech *incrementally*, in a piece-meal fashion unfolding over time [18]. Doing so, speakers use and interpret information from interlocutors rapidly and simultaneously formulate their own speech plan ([18,19], summarized in [4]). Despite the lack of a complete speech plan, human speech requires ahead planning of a certain degree. Psycholinguistic studies suggest that speakers plan at least one word, and usually more than one word ahead [4]. Evidence for the concept of *incremental processing* comes from several observable phenomena in spontaneous speech, many of which are closely related to hesitations:

- **Anticipatory speech production errors.** (e.g., “a cuff of coffee”) where parts of the utterance are produced in advance, or metathetically switched around, anticipating upcoming phonemes or syllables.
- **Hesitation-lengthening form in English.** (“Theee:” vs. “the”) The lengthened form has a different vowel quality (i: vs. ə), so the speaker must be aware of upcoming challenges in the speech plan (cf. [20]).
- **Different types of fillers.** (“uh” vs. “uhm”) The former appears to denote minor, the latter major problems in the speech plan, both requiring ahead planning [21].
- **Hesitation occurrence probability.** Hesitations are more likely to occur before longer utterances [22].

Models of incremental speech production inspire the design of incremental spoken dialogue systems, which will be further described in Section 2.3. In this study, we investigate whether human-like features that are typical of incremental processing, such as hesitation phenomena, are suitable for dialogue systems as well. Special attention will be paid to the concept of the articulatory buffer, which provides insights how to commence hesitation in incremental spoken dialogue systems.

The concept of the articulatory buffer was introduced in Levelt’s model of speech production [19] (p. 414) to describe the lookahead of several words that speakers have access to when speaking. It describes temporary storage for words that have been planned, but have not yet been articulated. The content of the buffer can be amended when the speech plan changes. Based on [19,20], Li and Tilsen [23] hypothesize that the material in the articulatory buffer can be lengthened by speakers in order to buy time for solving word retrieval problems. We assume that this might not only be the case for word retrieval issues, but make the proposition that this may hold as a general strategy for phonetically producing hesitations. Based on this assumption, we present a general model for hesitation insertion in conversational dialogue systems, as well as a first, preliminary implementation and evaluation.

2.3. 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 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 is limited to small domains, but the interaction became more like spoken conversation between humans as more computational power and better speech synthesis became available. One major shortcoming of these systems is their lack of adaptivity that contrasts their field of application. They produce adequate responses to user requests, but they do so in a rather static fashion: After some seconds processing time, 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 [24,25]. 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. It could thus be stated that these systems are less interactive than they should be. They perform their tasks, but cannot do anything conversational beyond that.

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 is incrementality. In this study, we explore 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 [8]. As described in Section 2.2, human dialogue does not work like a ball-tossing game, but rather simultaneously: Responses are planned while the interlocutor is speaking. It can be shown that limited-domain dialogue systems can make use of incremental processing to achieve human-like interaction speed [26]. 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 as it is of now.

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 use hesitations to do 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.

The authors of [4] built an incremental system based on a general, abstract model for incremental processing [27] 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. The authors of [28] 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. The authors of [29] 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.

The authors of [1,2] used hesitations (silence) 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 [30]. Whereas the authors were able to show that listener-oriented insertion of hesitations (in this case: silences) has

a positive effect on the interaction, the self-interrupting agent was perceived less friendly in all three studies. The authors of [17] found that hesitation lengthening, for a duration shorter than 800 ms, has a positive effect on users' task performance in a game setting.

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. An off-line evaluation study [31] suggested that different hesitation strategies differ inherently with regard to sound quality: while lengthenings and silences get relatively good user feedback (stimuli with lengthening got even better user feedback than fluent baseline stimuli), fillers and other disfluencies like mid-word cutoffs are dispreferred. The same authors investigated the reasons for the good performance of lengthening and found a somewhat paradoxical situation: the reason for the good ratings of synthetic lengthenings might be that they are barely perceivable. In a follow-up study the authors of [32] showed that even phonetically trained annotators, given the task to label disfluencies, miss up to 80% of lengthening instances that can be identified with a semi-automatic classification. This makes lengthening a promising candidate for application in conversational dialogue systems, as it has the potential to allow for extra time without the user even noticing. Based on the assumption that the underlying reasons for hesitations are similar in dialogue systems and humans, and in the light of the positive effect hesitations have on the interactive capacities of dialogue systems, we will explore a hesitation strategy for dialogue systems that generates a suitable hesitation initiation, overall duration and phonetic structure, and is based on observations of hesitation strategies in conversations among humans. Doing so, we hope to improve our system regarding subjective ratings compared to [1,2,30], by using a smoother hesitation insertion strategy that will not, as we hope, evoke a notion of rudeness.

3. Towards a Hesitation Synthesis Strategy for Incremental Spoken Dialogue Systems

3.1. A Model for Hesitation Insertion in Incremental Spoken Dialogue Systems

Given the insights summarized in Section 2.3, 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. In this section, we walk through the details of the algorithm that can be seen as our general model for hesitation insertion in dialogue systems. In Section 3.2, we give details on how we realized a preliminary implementation for this study.

The aim of the strategy proposed here is to buy as much time as possible for the speaker, by lengthening words in the articulatory buffer and inserting silences. Only in severe cases, where even more time is needed, will other measures such as fillers be employed (cf. Figure 1). This approach is governed by technical constraints. The choice to prioritize lengthening and silence over other hesitations is due to the simple fact that they can be synthesized with better sound quality [31] to account for the fact that sound quality is a weak spot of many incremental systems. Moreover, this strategy is motivated by the general assumption stated in Section 2.2 that suggests that a hesitation is always initiated by lengthening the phonetic material available in the articulatory buffer.

The strategy depicted in Figure 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.

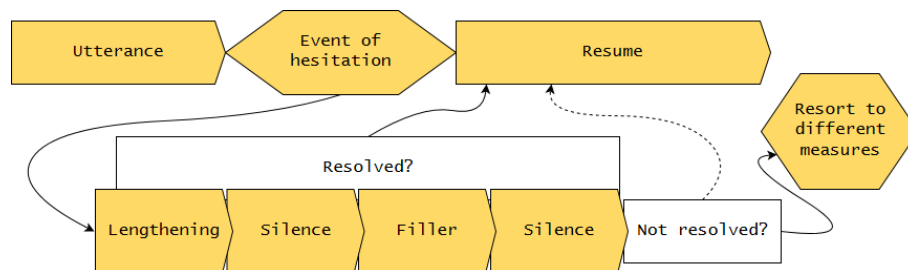


Figure 1. Hesitation insertion strategy.

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, resort to different measures. In the following, we walk through the individual steps in more detail. For an example of hesitation insertion into an utterance, see Figure 2.

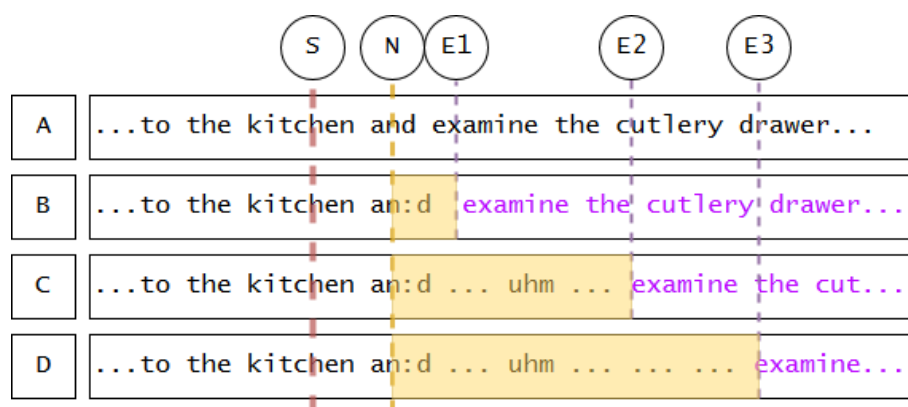


Figure 2. Examples of insertable hesitations. S = Starting point (external event triggering hesitation mode); N = Entry point (best upcoming target segment, in this case nasal sound in function word); E1–E3 = End points (external events triggering end of hesitation mode, after which originally intended 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.

When the event of hesitation is active. As described in Section 2.1, 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. To summarize findings of Refs. [33,34]:
 - Lengthening prefers closed-class (“function”) words.
 - Lengthening prefers, in this order, nasals, long vowels and diphthongs, short vowels, other non-plosive sounds. (The latter is language-specific. In some languages, plosives can be lengthened (e.g., Swedish) in others not (e.g., German).)

- The extent of the lengthening is governed by the elasticity of the phone in question.

The lengthening continues 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 s in conversation [35], this silence will continue for maximally 1000 ms, or until hesitation mode ends. For a more elaborate analysis of pauses and their duration, see [36].
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 [21].
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: the authors of [12] suggested that the filler type (“uh” or “uhm”) predicts the extent of the following silence; this is challenged by the authors of [37] who found that post-filler silences vary arbitrarily in duration.
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. This strategy depends on the architecture. Some examples of how a system could proceed (see Section 6 for a detailed discussion of possible continuation strategies) 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 (cf. Example 1).
 - Resume own speech plan if possible, despite the event of hesitation not being over.

3.2. Implementing the Algorithm

In the following, we describe how the individual concepts of the model described in the previous Section 3 are realized in this preliminary implementation. We present a system that is fully capable of producing hesitations, does not (for the time being) exploit the full depth of the model, as there were audio issues with the insertion of fillers. The system we explore here is therefore a reduced version that hesitates only by means of lengthening and silence, which is sufficient to evaluate the effect of hesitations in dialogue system interaction. The implementation of the full model is scheduled for follow-up studies.

3.2.1. Event of Hesitation

In this study, we define an event of hesitation as the time interval a user does not maintain eye-contact with our virtual agent. This is based on one of the reasons from Section 2.1—change in dialogue environment. We deploy hesitations as a user-oriented strategy (cf. [2]), as a response to visual attention shifts. The goal is to assist users in their task by only giving them information while they are paying attention.

3.2.2. Different Measures

This definition for events of hesitation also governs the strategy for continuation. In this case, it is simply waiting for the hesitation to end, i.e., the user looking back.

3.2.3. Lengthening

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 5-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 5 words after the trigger. Based on the preference hierarchy for lengthening targets described in the previous Section 3, 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.

The duration of the lengthening is inferred from mean duration values from previous corpus studies, from which a so-called stretch factor is deducted. 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.

3.2.4. Fillers

Due to technical problems, fillers are not included in our main study and could only be tested in an exploratory way. The problems are bugs related to incremental processing, namely (1) audible clicks around the filler and (2) omitted segments and words around the filler. This leads to the impression that the system is broken 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. Implementing the full model will be addressed in a follow-up study. As will be described in Section 4.2, we explored the usability of data with this preliminary “full hesitation” version, but most participants were recorded in a “reduced” version with only lengthening and silence, cf. Section 3.2.6.

3.2.5. Silences

As fillers are left out, the main study operates with only the first silence. In the general model, it is designed to last 1000 ms. In our implementation, the duration is variable as we wait for the user to re-focus. (In the exploratory condition with fillers, the first silence lasts for 1000 ms and the second silence lasts until the users re-focus.)

3.2.6. 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 [1,2,30]. Therefore, we increased the lengthening extent by 50% to ensure participants are able to perceive it. For a follow-up study with the full hesitation model with fillers working, this extra duration will be removed again. This implementation, while covering only half of the model, is fully functional and capable of dynamically inserting hesitations, only in a less diverse form than originally intended.

3.2.7. Technical Implementation

From a technical perspective, the hesitation algorithm is integrated as separate module into an existing incremental spoken dialogue system [38], which uses a toolkit for incremental dialogue processing [39] and MaryTTS [40] as a speech synthesis back-end.

4. Experiment 1: Interaction Study

To evaluate the effect of hesitation in human–agent interaction, we conducted an interaction study in the *Cognitive Service Robotics Apartment* <https://cit-ec.de/en/csra> (CSRA) [41]. 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 3 is evaluated by means of a task in which the participants have to perform a memorization task. A virtual agent provides a background story and instructs the participants to look for hidden treats 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 that:

1. We expect memory task performance to benefit from the presence of hesitations.
2. We expect that presence of hesitations influences user ratings of perceived synthesis quality (undirected)
3. We expect no negative impact of the presence or absence of hesitation on the system's likability.

4.1. Methods

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 [40] 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 say 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).

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 [42] (cf. Figure 3) displayed on a screen in the kitchen area of the smart apartment. The agent is able to detect faces and estimate the current visual focus of attention of the human interaction partner [43] with the help of a web-cam installed on top of the screen.

As soon as a participant appears in front of Flobi, it starts talking (cf. Figure 3). 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 4). 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. (There was actual visible construction work in the apartment at the time of the study, which inspired this narrative.) 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. 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) lighting up a door handle in the participants' field of vision (visual distraction); (2) the experimenter entering the room to insert a code for later use in the questionnaire (audiovisual distraction); and (3) a music beat being played for two seconds (audio distraction). 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 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 [44]). 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 [45]).

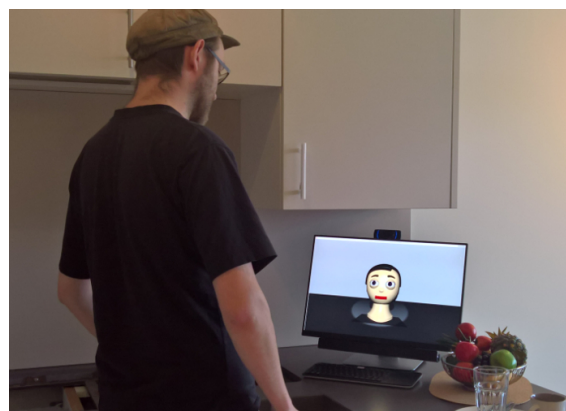


Figure 3. Person being instructed by virtual agent on a screen.

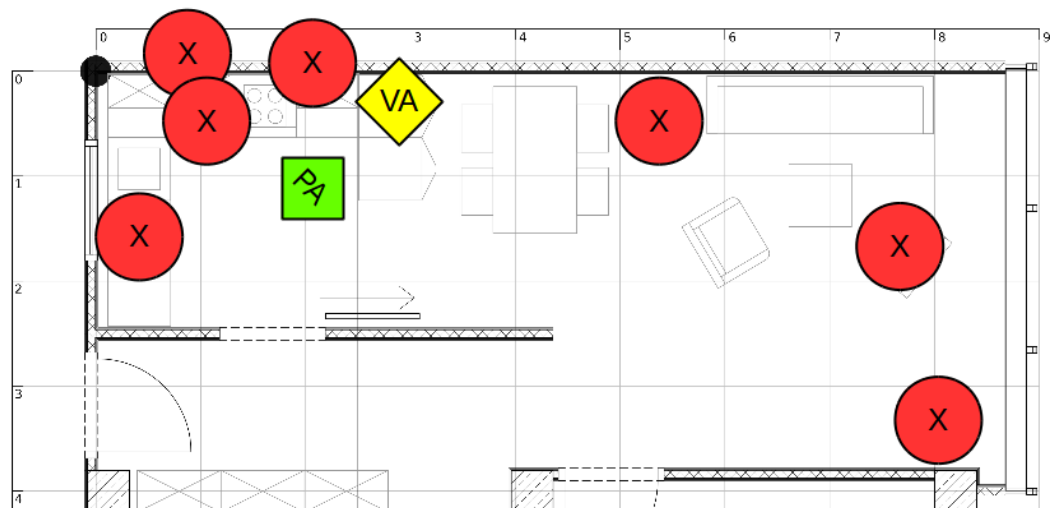


Figure 4. A 2D map of the smart home environment. (X) denotes hiding places of treats, (VA) the position of the screen with the virtual agent, and (PA) the initial position of the participant.

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*.

To evaluate the questionnaires regarding the user's perception of the agent, based on [44], 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 of *key concept* and *hesitation condition*.

4.2. Results and Discussion

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 3.2.

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).

4.2.1. 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 5, left panel). The glm analysis shows that the effect is not significant ($\beta = 0.8$, $SE = 0.44$, $z = 1.84$, $p = 0.065$).

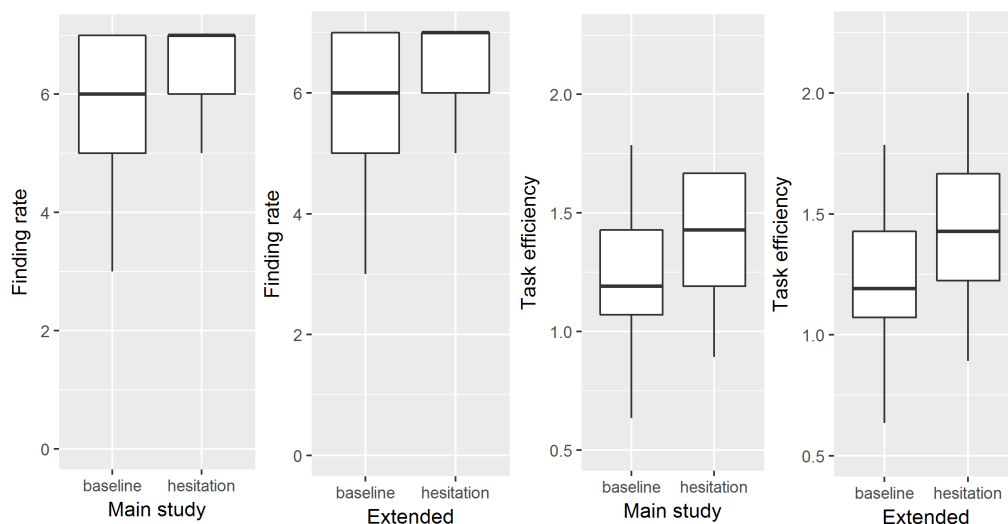


Figure 5. Task performance and efficiency.

4.2.2. Efficiency

Efficiency increases in the hesitation condition ($M = 1.22$, $SD = 0.3$) compared to the baseline ($M = 1.5$, $SD = 0.58$), (cf. Figure 5, third panel from the left). The Mann–Whitney U test shows no significant effect of *hesitation condition* on *efficiency* ($W = 79$, $p = 0.11$).

4.2.3. 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 ($M = 2.53, SD = 0.62$), cf. Figure 6, left panel. The Mann–Whitney U test shows that there is a significant effect of *hesitation condition* on users’ perception of synthesis quality ($W = 203, p = 0.0004$).

4.2.4. 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 feedbacks regarding these concepts, cf. Figure 7.

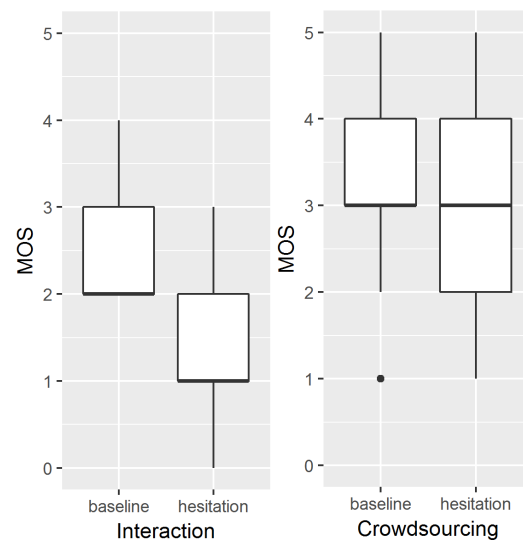


Figure 6. A 5-point mean opinion score (MOS) scale user feedback on synthesis quality.

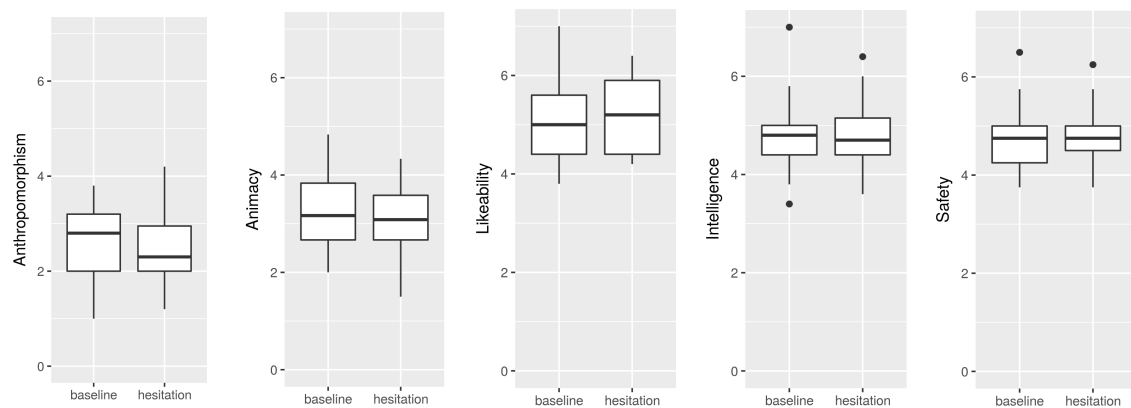


Figure 7. Subjective ratings for the five key concepts.

Aside from the questionnaire results, participants were encouraged to give free-text feedback in a comments box in the questionnaire, and they were 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 increased the attention of distracted users [2], but was perceived as less likable [2] and rude [1]. 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 7).

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

4.2.5. Exploratory Extension of Analysis

As we have at our disposal four recordings with the full hesitation condition (cf. Section 3.2), we re-did 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 3.2.4). For that reason, we do not include them in the main study as we cannot control the effects this has on the interaction. 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 5). This suggests that there is an effect of hesitations on task performance that needs to be considered. We assume that these effects will be confirmed in a follow-up study with the full model implemented and with 20 participants in the baseline condition and 20 in the full hesitation condition.

4.2.6. 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 hypotheses (1) and (3) can be accepted for now, and with respect to hypothesis (2), the results suggest a negative impact of hesitations on users' perceptions of synthesis quality.

5. Experiment 2: Crowdsourcing-Based Evaluation of Hesitation Synthesis

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. 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 1. 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.

5.1. Methods

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 [46], 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 a deviation between 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 1. 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 1. In the baseline condition (non-hesitation), the stimuli are produced with MaryTTS’s [40] female German HMM voice, with no further modification. For the hesitation condition, lengthenings 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 1, 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 hesitation mode. Lengthenings are 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 A.

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* (version 1.1-12).

5.2. Results and Discussion

We collected ratings from 44 participants (29 female, 15 male) with an age range between 18 and 46 years (median: 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 problems. Most participants were raised in the vicinity of Bielefeld, and a few in Bavaria. The listening tests typically lasted less than 5 min, 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). 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 1, 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. 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. Our evaluation is, however, not limited to objective assessments of the system as a whole, rather, we also assessed subjective system ratings via participant feedback.

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.

Our study highlights this point. As can be seen in Figure 6, 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 1, participants give one score to evaluate the general impression of synthesis, while in experiment 2, 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 the perceived quality in interactive settings.

An important issue that arises is how to gather quality measures that do account for the interactive nature of speech synthesis applications. 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. It thus appears more fruitful to enrich offline evaluations. 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 [17].

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. If the system cannot reach the same quality level in interaction due to technical limitations, as observed in this study, then the offline version could serve as a gold standard to be reached in interaction via further development of the system. 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 synthesis in interaction.

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. The fact that some of the effects can be attributed the technical realization of our hesitation model yielding some audible artefacts 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 [2,30], 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. In a follow-up study we will further explore the applicability of our model with some extensions regarding the quality of hesitations in order to minimize the irritating effects reported for this first prototype.

A further open question is the point that our hesitation model contains the placeholder “resort to different measures”. What should be done in the case that the hesitation algorithm has run out of options but the issue is not resolved? In general, any conceivable strategy can be combined with the model presented here. In our study, the strategy we employed was to wait until the user looks back and be silent until then. This can lead to awkward situations in which users do not ever look back because the instructions have ceased. If the chosen strategy is to wait, some threshold for an exit strategy has to be included, after which the system notifies the user that it has more to say, or just continues speaking, depending on the scenario. There are other strategies than can be employed, such as producing non-committing material: repetitions of previous content or filler phrases. It is, however, subject to further investigation, and how to sophisticatedly realize continuation strategies

is out of the scope of this study. Finally, it could be a strategy to simply re-enter the same hesitation algorithm immediately or after a certain time has elapsed (which, again, would require the system designer to define what to do until then). From a technical standpoint, our hesitation model would, once it has run through, evoke a dialogue management module that contains a follow-up strategy that can contain a conceivable continuation plan.

To conclude, given some necessary improvements on the technical side, 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. It is, as of now, one of the greatest challenges for the speech synthesis community to develop and establish evaluation paradigms that allow us to go beyond pure MOS scales.

Acknowledgments: This research was carried out as part of the CITEC Large Scale Project “The Cognitive Service Robotics Apartment as an Ambient Host” (CSRA) and was supported by the Cluster of Excellence Cognitive Interaction Technology ‘CITEC’ (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). We warmly thank our participants. We are also grateful to Christoph Draxler, who invested much of his to set up experiment 2; Timo Baumann and Soledad Lopez-Gambino, who gave advice regarding several issues with the synthesis in InProTK; Monika Chromik and Ayla Canpolat for their massive support during experiment 1; and all the people who invested a lot of their time to set up the CSRA as a platform for interaction research.

Author Contributions: All authors conceived and designed the experiments, and analysed the data jointly. Birte Carlmeier and Simon Betz conducted experiment 1. Simon Betz constructed the stimuli for experiment 2. Petra Wagner conducted experiment 2. Simon Betz, Birte Carlmeier and Petra Wagner wrote the paper.

Conflicts of Interest: The authors do not declare any conflict of interests.

Appendix A. Stimuli for Crowdsourcing Study

The following stimuli are used for the crowdsourcing experiment described in Section 5. Lengthened syllables are indicated by appended colons. Pauses are indicated by seconds in brackets. Lengthening durations are determined as described in Section 3.2.3. 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 paar: (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 Bescheckschublade 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 Bescheckfach der: (2.0) Spülmaschine.”

Conclusions

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

Reference

1. Carlmeyer, B.; Schlangen, D.; Wrede, B. 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, Tokyo, Japan, 16 November 2016; Association for Computing Machinery (ACM): New York, NY, USA, 2016.
2. Carlmeyer, B.; Schlangen, D.; Wrede, B. “Look at Me!”: Self-Interruptions as Attention Booster? In Proceedings of the Fourth International Conference on Human Agent Interaction-HAI '16, Singapore, 4–7 October 2016; Association for Computing Machinery (ACM): New York, NY, USA, 2016.
3. Chafe, W. Some reasons for hesitating. In *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*; Walter de Gruyter: Berlin, Germany, 1980; pp. 169–180.
4. Skantze, G.; Hjalmarsson, A. Towards incremental speech generation in conversational systems. *Comput. Speech Lang.* **2013**, *27*, 243–262.
5. King, S. 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), Glasgow, UK, 10–14 August 2015.
6. Mendelson, J.; Aylett, M. Beyond the Listening Test: An Interactive Approach to TTS Evaluation. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), Stockholm, Sweden, 20–24 August 2017; pp. 249–253.
7. Rosenberg, A.; Ramabhadran, B. Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), Stockholm, Sweden, 20–24 August 2017; pp. 3976–3980.
8. Wester, M.; Braude, D.A.; Potard, B.; Aylett, M.; Shaw, F. Real-Time Reactive Speech Synthesis: Incorporating Interruptions. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017), Stockholm, Sweden, 20–24 August 2017; pp. 3996–4000.
9. Wagner, P.; Betz, S. Speech Synthesis Evaluation—Realizing a Social Turn. In Proceedings of the Tagungsband Elektronische Sprachsignalverarbeitung (ESSV), Saarbrücken, Germany, 15–17 March 2017; pp. 167–172.
10. Eklund, R. Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues. Ph.D. Thesis, Linköping University Electronic Press, Linköping, Sweden, 2004.
11. Shriberg, E. Preliminaries to a Theory of Speech Disfluencies. Ph.D. Thesis, University of California, Berkeley, CA, USA, 1994.
12. Clark, H.H.; Tree, J.E.F. Using uh and um in spontaneous speaking. *Cognition* **2002**, *84*, 73–111.
13. Goodwin, C. *Conversational Organization: Interaction between Speakers and Hearers*; Academic Press: Cambridge, MA, USA, 1981.
14. Tree, J.E.F. Listeners’ uses of um and uh in speech comprehension. *Mem. Cognit.* **2001**, *29*, 320–326.
15. Collard, P. Disfluency and Listeners’ Attention: An Investigation of The Immediate and Lasting Effects of Hesitations in Speech. Ph.D. Thesis, University of Edinburgh, Edinburgh, UK, 2009.
16. Corley, M.; Stewart, O.W. Hesitation disfluencies in spontaneous speech: The meaning of um. *Lang. Linguist. Compass* **2008**, *2*, 589–602.
17. Betz, S.; Zarrieß, S.; Wagner, P. Synthesized lengthening of function words—The fuzzy boundary between fluency and disfluency. In Proceedings of the International Conference Fluency and Disfluency, Louvain-la-Neuve, Belgium, 15–17 February 2017.
18. Kempen, G.; Hoenkamp, E. Incremental sentence generation: Implications for the structure of a syntactic processor. In Proceedings of the 9th Conference on Computational Linguistics-Volume 1, Prague, Czechoslovakia, 5–10 July 1982; Academia Praha: Praha, Czech Republic, 1982; pp. 151–156.
19. Levelt, W.J.M. *Speaking: From Intention to Articulation*; MIT Press: Cambridge, MA, USA, 1989.
20. Shriberg, E. To ‘errrr’ is human: Ecology and acoustics of speech disfluencies. *J. Int. Phon. Assoc.* **2001**, *31*, 153–169.
21. Clark, H. Speaking in Time. *Speech Commun.* **2002**, *36*, 5–13.
22. Shriberg, E. Disfluencies in switchboard. In Proceedings of the International Conference on Spoken Language Processing, Philadelphia, PA, USA, 3–6 October 1996; Volume 96, pp. 11–14.

23. Li, J.; Tilsen, S. Phonetic evidence for two types of disfluency. In Proceedings of the ICPhS, Glasgow, UK, 10–14 August 2015.
24. Sacks, H.; Schegloff, E.A.; Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *Language* **1974**, *50*, 696–735.
25. Heldner, M.; Edlund, J. Pauses, gaps and overlaps in conversations. *J. Phon.* **2010**, *38*, 555–568.
26. Skantze, G.; Schlangen, D. 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), Athens, Greece, 30 March–3 April 2009; pp. 745–753.
27. Schlangen, D.; Skantze, G. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue Discourse* **2011**, *2*, 83–111.
28. Kousidis, S.; Kennington, C.; Baumann, T.; Buschmeier, H.; Kopp, S.; Schlangen, D. 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, Gothenburg, Sweden, 26–27 April 2014; pp. 68–72.
29. Bohus, D.; Horvitz, E. Managing Human-Robot Engagement with Forecasts and... Um... Hesitations. In Proceeding of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; ACM: New York, NY, USA, 2014; pp. 2–9.
30. Chromik, M.; Carlmeyer, B.; Wrede, B. 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, Vienna, Austria, 6–9 March 2017; Association for Computing Machinery (ACM): New York, NY, USA, 2017.
31. Betz, S.; Wagner, P.; Schlangen, D. 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, Germany, 6–10 September 2015; pp. 2222–2226.
32. Betz, S.; Voße, J.; Zarriß, S.; Wagner, P. 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, Sweden, 20–24 August 2017; pp. 1084–1088.
33. Betz, S.; Wagner, P.; Vosse, J. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. In Proceedings of the Phonetik und Phonologie, München, Germany, 26–28 May 2016.
34. Betz, S.; Voße, J.; Wagner, P. Phone Elasticity in Disfluent Contexts. In Proceedings of the Fortschritte der Akustik-DAGA, Kiel, Germany, 6–9 March 2017.
35. Jefferson, G. Preliminary notes on a possible metric which provides for a “standard maximum” silence of approximately one second in conversation. In *Conversation: An Interdisciplinary Perspective*; Roger, D., Bull, P., Eds.; Multilingual Matters: Bristol, UK, 1989.
36. Lundholm Fors, K. Production and Perception of Pauses in Speech. Ph.D. Thesis, University of Gothenburg, Gothenburg, Sweden, 2015.
37. O’Connell, D.C.; Kowal, S. Uh and um revisited: Are they interjections for signaling delay? *J. Psycholinguist. Res.* **2005**, *34*, 555–576.
38. Carlmeyer, B.; Schlangen, D.; Wrede, B. 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, Istanbul, Turkey, 16 November 2014; ACM: New York, NY, USA, 2014; pp. 1–6.
39. Baumann, T.; Schlangen, D. The InproTK 2012 Release. In Proceedings of the NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data, Montreal, QC, Canada, 7 June 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 29–32.
40. Schroeder, M.; Trouvain, J. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *Int. J. Speech Technol.* **2003**, *6*, 365–377.
41. Wrede, S.; Leichenring, C.; Holthaus, P.; Hermann, T.; Wachsmuth, S. The Cognitive Service Robotics Apartment: A Versatile Environment for Human-Machine Interaction Research. *Kuenstliche Intell.* **2017**, *31*, 299–304.
42. Lütkebohle, I.; Hegel, F.; Schulz, S.; Hackel, M.; Wrede, B.; Wachsmuth, S.; Sagerer, G. The Bielefeld Anthropomorphic Robot Head “Flobi”. In Proceedings of the 2010 IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 3–7 May 2010; pp. 3384–3391.

43. Schillingmann, L.; Nagai, Y. Yet another gaze detector: An embodied calibration free system for the iCub robot. In Proceedings of the 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), Seoul, Korea, 3–5 November 2015; pp. 8–13.
44. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *Int. J. Soc. Robot.* **2009**, *1*, 71–81.
45. Holthaus, P.; Leichsenring, C.; Bernotat, J.; Richter, V.; Pohling, M.; Carlmeyer, B.; Köster, N.; Zu Borgsen, S.M.; Zorn, R.; Schiffhauer, B.; et al. 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), Portoroz, Slovenia, 24 May 2016; European Language Resources Association: Paris, France, 2016.
46. Draxler, C. Online Experiments with the Percy Software Framework—Experiences and some Early Results. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; Chair, N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., Eds.; European Language Resources Association (ELRA): Reykjavik, Iceland, 2014.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).