Article

Interactive Hesitation Synthesis and Its Evaluation

Simon Betz [1,2,4]*, Birte Carlmeyer [1,3,4], Petra Wagner [1,2] and Britta Wrede [1,3]

1 Cognitive Interaction Technology Center (CITEC), Bielefeld University
2 Phonetics and Phonology Workgroup, Faculty of Linguistics and Literary Studies, Bielefeld University
3 Applied Informatics Group, Faculty of Technology, Bielefeld University
4 Dialogue Systems Group, Faculty of Linguistics and Literary Studies, Bielefeld University
* Correspondence: simon.betz@uni-bielefeld.de; Tel.: +49-521-106-3518

Abstract: Conversational spoken dialogue systems that interact with the user rather than merely reading text can be equipped with hesitations to manage the dialogue flow and the users' attention. Based on a series of empirical studies, we built an elaborated hesitation synthesis strategy for dialogue systems that inserts hesitations of scalable extent wherever needed in the ongoing utterance. So far, evaluations of hesitating systems have shown that synthesis quality is affected negatively by hesitations, but that there is improvement in interaction quality. We argue that due to its conversational nature, hesitation synthesis needs interactive evaluation rather than traditional MOS-based questionnaires. To prove this point, we dually evaluate our system's speech synthesis component: on the one hand, linked to the dialogue system evaluation, on the other hand, in the traditional MOS way. This way we are 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, which has implications for future research that are discussed in this paper. Furthermore, our results indicate that hesitations work well to increase task performance and that an elaborated 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, the 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 for 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. Rather, they serve an important role in dialogue: They allow the speaker extra time in situations where this is needed, e.g. when searching for the right thing to say, and to signal this to the listener. That way, hesitations help to

Peer-reviewed version available at Multimodal Technologies Interact. 2018, 2, 9; doi:10.3390/mti2010009

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. [1–3].

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 user's attention: When users look away, the system will enter 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 (MOS). Despite numerous criticisms for this method, alternatives have seldom been proposed [4–8]. Also, to our knowledge, there exists no previous study that actually verifies these critical voices. 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 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.

1.2. 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 (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 (2.2). We continue with a brief introduction to dialogue systems with a special focus on systems with incremental processing, which we will work on in this study (2.3). 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 (3). In the empirical part of this paper we present two experiments that make up this study. First, we describe the methods and results of an interaction study (4), continuing with a crowdsourcing-based parallel study for evaluation purposes (5). The experiments sections are each concluded with short discussions. The main study is then concluded with a general discussion of both experiments and their implications (6).

2. Background and Related Work

2.1. A brief introduction to 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. [1]

"Take theee ... uhm ... the, the red line to the university"

Example 1: 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 speaker and listener to manage the conversation. [11] 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 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: [12] found that speakers hesitate when the listener is apparently distracted, and [13,14] found that listeners may heighten attention when a hesitation is uttered.

While it is highly controversial whether hesitations and disfluencies are produced in order to signal something to the listener (see [15] for an overview), or if it is merely the fact that the listener can do something with the information, it can be claimed that listeners can make use of at least the extra time that hesitations grant in dialogue, an effect that is replicable for human-machine interactions, e.g. [1–3,16].

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 the most appropriate item (cf. Example 2). They might run out of things to say before having conveyed the intended message (cf. Example 3). The dialogical situation might change, causing a change in speech plan, that needs time (cf. Example 4).

"The capital of Serbia is ... uhm ... Belgrade."

Example 2: Difficulty retrieving an infrequent lexical item.

"There is no direct flight to Sydney ... uhm ... today or tomorrow..."

Example 3: Travel agent giving information, but the database query takes time.

"You can take a seat ... in the living room."

Example 4: 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

[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 [9]. For the most influential descriptive work on disfluencies in general, see [10].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2017 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at Multimodal Technologies Interact. 2018, 2, 9; doi:10.3390/mti2010009

elements suggested in the introductory Example 1. 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.

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 [17]. Doing so, speakers use and interpret information from interlocutors rapidly and simultaneously formulate their own speech plan([17,18], summarized in [3]). 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 ahead, usually more [3]. Evidence for the concept of *incremental processing* comes from several observable phenomena of spontaneous speech, many of those 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 (iː vs. ə), so the speaker must be aware of upcoming challenges in the speech plan, cf. [19].
- **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 [20].
- **Hesitation occurrence probability.** Hesitations are more likely to occur before longer utterances [21].

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 [18] (p. 414) to describe the lookahead of several words that speakers have access to when speaking. It describes a 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 [18] and [22], 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 in this study a general model for hesitation insertion in conversational dialogue systems.

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 in- and output. The scope of application is limited to small domains, but the interaction has become 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 can only produce static responses, but are incapable of interpreting

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2017 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at Multimodal Technologies Interact. 2018, 2, 9; doi:10.3390/mti2010009

user feedback or handling interruptions. 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.[2] 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 [24].

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.

[3] built an incremental system based on general, abstract model for incremental processing [25] 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. [26] 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. [27] uses hesitations in human-robot interaction as a disengagement strategy. A directions-giving robot produces lexical hesitations ("so...", "let's see...") after 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.

[1,2] use 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 could show that incremental information presentation leads to a better task performance [28]. Whereas the authors could 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. [16] found that hesitation lengthenings, as long as they are shorter than 800ms, have 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, 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 [29] suggests, that different hesitations 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 paradox situation: the reason for the good rating of synthetic lengthening might be that they are barely perceivable. In a follow-up study [30] showed that even corpus annotators with the task to label disfluencies miss up to 80% of lengthening

[2] 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 [7].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2017 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at Multimodal Technologies Interact. 2018, 2, 9; doi:10.3390/mti2010009

instances that can be identified with semi-automatic classification. This makes lengthening a promising candidate for application in conversational dialogue systems.

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 [1,2,28], 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 now propose an elaborated and dynamic hesitation insertion strategy, that can be evoked (1) while a dialogue system is speaking, (2) and that determines the best entry point, given an event of hesitation, (3) 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 the implementation of it 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 is due to the simple fact that they can be synthesized with better sound quality [29], the absence of which is a weakness of many incremental systems. Moreover, this strategy is motivated by the general assumption stated in 2.2 that suggests that a hesitation is always initiated by lengthening the phonetic material available in the articulatory buffer.

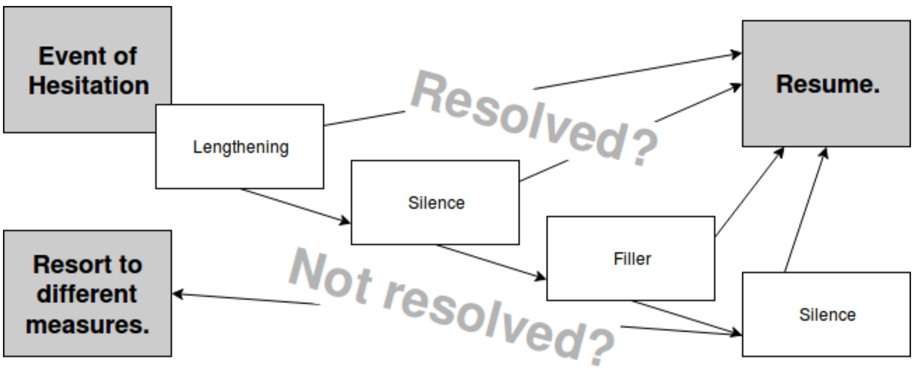

Figure 1. Hesitation insertion strategy

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.

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2017 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at Multimodal Technologies Interact. 2018, 2, 9; doi:10.3390/mti2010009

While 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.

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 [31] and [32]:

   - Lengthening prefers closed-class ("function") words.
   - Lengthening prefers, in this order, nasals, long vowels and diphthongs, short vowels, other non-plosive sounds.[3]
   - 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 second in conversation [33], this silence will continue for maximally 1000ms, or until hesitation mode ends. [4]
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 [20].
4. **Insert second silence.** If after the filler the hesitation mode is still not resolved, a second silence can be added, with the same rules as the first silence.
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 is depending on the architecture. Some examples of how a system could proceed:

   - 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 event of hesitation is not 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 study.

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 to 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] The latter is language-specific. In some languages, plosives can be lengthened (e.g. Swedish) in others not (e.g. German).
[4] For a more elaborated analysis of pauses and their duration, see [34].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2017 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at Multimodal Technologies Interact. 2018, 2, 9; doi:10.3390/mti2010009

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. This factor was additionally multiplied by 1.5 for this study, because, as it is the nature of lengthening, the original duration increase was barely audible.

3.2.4. Fillers

Due to technical problems, fillers are not included in our main study. Four participants were recorded in a condition with fillers, but it became apparent, that the negative impact on sound quality is too great for the time being. This issue will be addressed in future studies. As will be described in section 4.2, we explored the usability of data with this preliminary "full hesitation" version.

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 1000ms. 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 1000ms and the second silence lasts until the users re-focus.)

3.2.6. Technical implementation

From the technical side, the hesitation algorithm is integrated as separate module into an existing incremental spoken dialogue system [35], which uses a toolkit for incremental dialogue processing [36] and MaryTTS [37] as speech synthesis back-end.

4. Experiment 1: interaction study

To evaluate the effect of hesitation in a human-agent interaction, we conducted an interaction study in *the Cognitive Service Robotics Apartment*[5] *(CSRA)* [38]. 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* version without hesitations or adaptations of any sort, and a *hesitating* version 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:

---

[5] https://cit-ec.de/en/csra

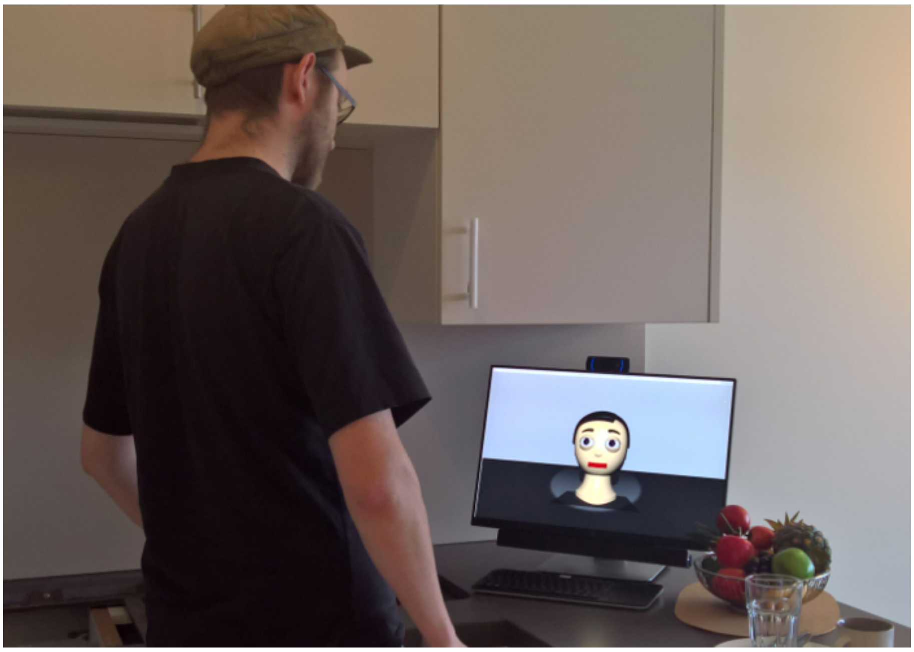

Figure 2. Person being instructed by virtual agent on a screen.

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 [37] German female HHM voice with no further modification. The words are German nouns that fall into five categories (professions, food, sports, buildings, cities), 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* 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 good 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 [39] (cf. Figure 2) displayed on a screen in the kitchen area of the smart apartment. With a web-cam on top of the screen, the agent is able to detect faces and estimate the current visual focus of attention of the human interaction partner [40]. Flobi is also able to show facial expressions and to pay attention to the current focus of discourse by looking at it.

As soon as a participant appears in front of Flobi, it starts talking (cf. figure 2). 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 3). 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

Peer-reviewed version available at Multimodal Technologies Interact. 2018, 2, 9; doi:10.3390/mti2010009

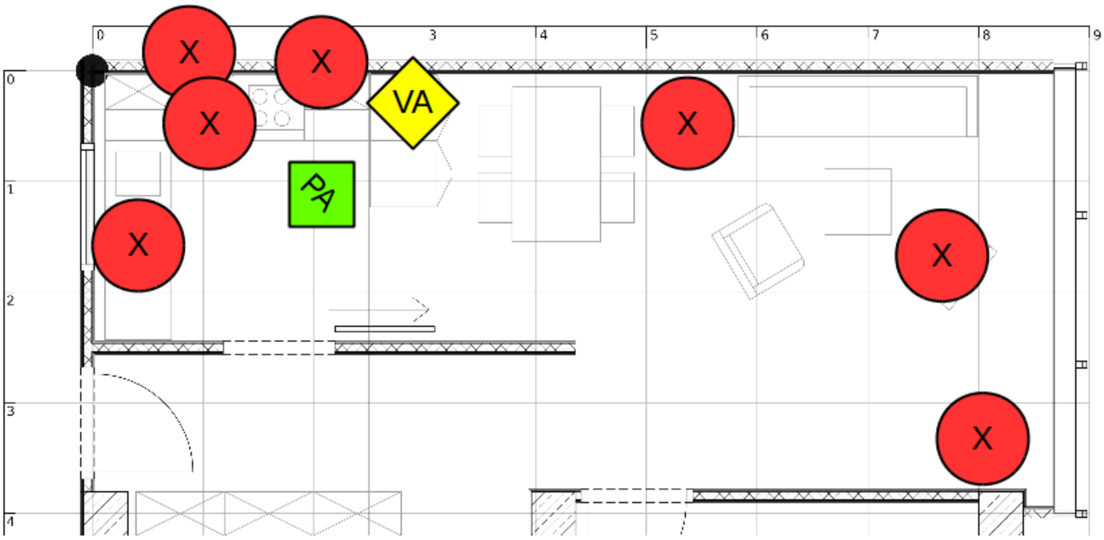

Figure 3. 2D map of the smart-home environment. (X) denotes hiding places of treats, [VA] the position of the screen with the virtual agent, [PA] the initial position of the participant.

left the apartment and caused confusion in the agent's sensors, due to the dust they stirred.[6] 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 intentional audiovisual distraction at three fixed points in time. This is included to ensure some degree of distraction and gaze shift for each participant. The distractions are: (1) Lighting up a door handle in the participants' field of vision, (2) The experimenter entering the room to bring a code for use in the questionnaire, (3) A music beat being played for two seconds.

As soon as the agent has finished the instruction, the participants start investigating the possible hiding places. They are asked to call out each place before looking at it, to ensure that they remember the places and that they do not search the entire place and find things by chance. The interaction is monitored audiovisually in an adjacent room. The number of treats thus retrieved is taken down for each participant as *finding rate*. After the interaction, participants fill out a questionnaire that assesses their subjective impression of the system quality on 24 dimensions using 7-point Likert scales (based on the Godspeed questionnaire [41]), and in which they also rate their impression of speech synthesis quality using a 5-point MOS scale. Additionally, demographic data and previous experiences with robotic systems, the agent Flobi and speech synthesis systems in general are surveyed. Finally, participants are 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 receive monetary compensation.

The entire interaction is 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 [42]).

The collected data were entered into a generalized linear model (glm) with *finding rate* as dependent variable, *hesitation condition* as fixed factor, *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*

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

 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at *Multimodal Technologies Interact.* **2018**, *2*, 9; doi:10.3390/mti2010009

condition. The same test was then used to analyze users' feedback on synthesis quality with regard to *hesitation condition*.

To evaluate the questionnaires regarding the user's perception of the agent, based on [41], the responses are grouped into five key concepts (*anthropomorphism, animacy, likability, 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

We recorded 37 trials with 24 female and 13 male participants in total. Participants were recruited on the university campus and via campus-related social media. Mean age was 24.6 (SD = 4.2). Two participants had to be excluded from the analysis because their language competence did not suffice to follow the instructions correctly. 17 participants interacted with the baseline system (ten female and seven male), and 14 with the hesitation system (ten female and four male). These 31 trials provide the core for our analysis. In addition, four participants (three female and one male) were recorded in the full hesitation condition for exploratory purposes, cf. section 3.2. The participants are balanced with regards to the their prior experience with robotic systems, the virtual agent Flobi (mostly no or very little experience) and speech systems in general (little experience).

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

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 4, 3rd panel from the left). The Mann-Whitney-U test shows no significant effect of *hesitation condition* on *efficiency* ($W = 79, p = 0.11$)

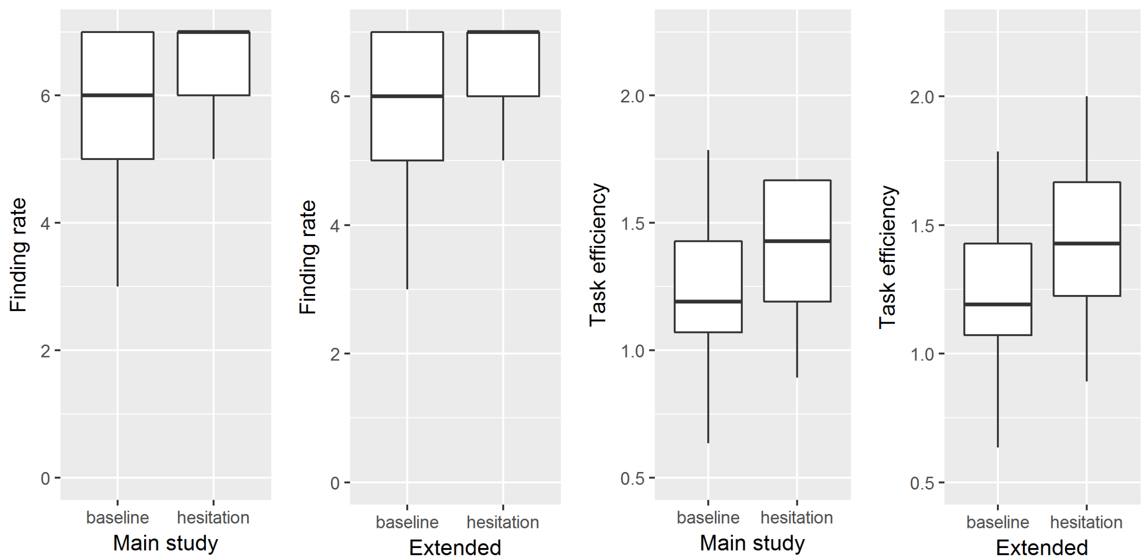

Figure 4. Task performance and efficiency.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2017 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at *Multimodal Technologies Interact.* **2018**, *2*, 9; doi:10.3390/mti2010009

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, likability, perceived intelligence* and *safety*. The factor *hesitation condition* had no significant influence on any of the user feedbacks regarding these concepts, cf. Figure 5. Aside from the questionnaire results, participants were encouraged to give free-text feedback in a

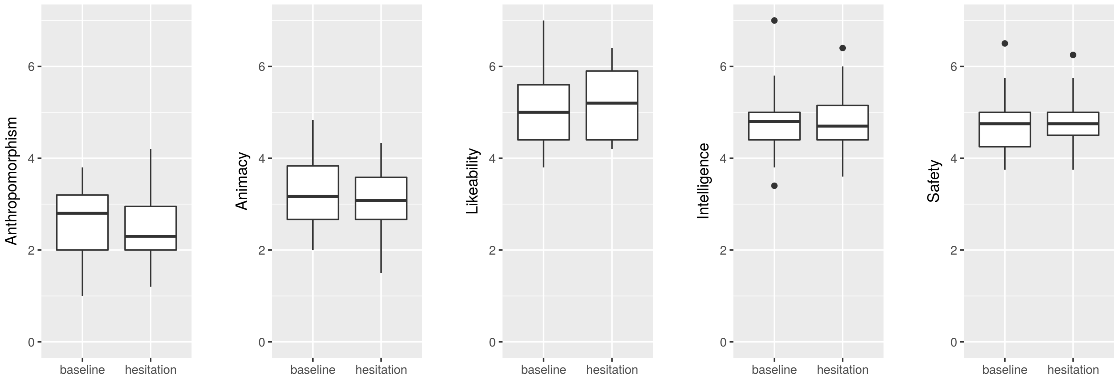

Figure 5. Subjective ratings for the five key concepts.

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 5).

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 triggers 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 the tendencies observed for finding rate and efficiency failed to reach the 0.05 significance level by only a small margin, we hypothesized that the effect might reach significance if more trials were recorded. As we have at our disposal four recordings with the full hesitation condition (cf. section 3.2), we re-did the analyses with the same 17 trials for the baseline condition and with all 18 hesitation trials combined as the hesitation condition. The effect on finding rate still 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 becomes significant, when all trials are considered ($W = 83.5, p = 0.02$), (cf. Figure 4).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2017 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at Multimodal Technologies Interact. 2018, 2, 9; doi:10.3390/mti2010009

This suggests that there is indeed an impact of hesitations that needs to be considered. We assume that these effects will be confirmed in a follow-up study with a bug-fixed version of the system and with more participants.

4.2.6. Summary

The results gathered here point in 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 suggests 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 user's perception of synthesis quality.

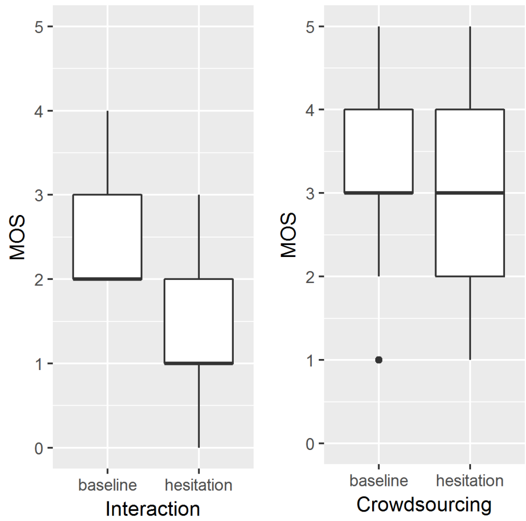

Figure 6. 5-point MOS scale user feedback on synthesis quality.

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 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 than we 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 [43], 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2017 doi:10.20944/preprints201712.0058.v1

Peer-reviewed version available at *Multimodal Technologies Interact.* **2018**, *2*, 9; doi:10.3390/mti2010009

For maximal comparison with the interaction study, we again chose a between-subjects design with a 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 is 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 is 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 [37] 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 2000ms, in all other stimuli, silences are set to 1000ms. 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 dependent variable, *hesitation condition* as 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, 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 s/he 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 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 a reasonable number of participants for both conditions (> 20), as 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 rated the stimuli containing hesitations as less good 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

This study yields several insights that demand discussion. We improve the conversational capabilities of a dialogue system by integrating a strategy for dynamic insertion of synthesized hesitations. 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 observe that: (1) In dialogue system evaluation, the speech synthesis quality is often not assessed. (2) 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 uttered in the speech synthesis community lately: Non-interactive evaluation of speech synthesis does not work, or at least, it assesses aspects of 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 totally unclear what to do with this information. Speech synthesis is not used in the void, there is always some application or interaction associated with it.

Out 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, 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 he 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.)

An important question 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: use the dialogue system evaluation to infer something for speech synthesis quality, or 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 [16].

Peer-reviewed version available at *Multimodal Technologies Interact.* **2018**, *2*, 9; doi:10.3390/mti2010009

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 off-line 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.

Turning to other objectives of this study, it is to be asked 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 to the fact that the technical realization of our hesitation model yielded 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 a previous study that used silence only as an attention-driven hesitation strategy [2], an increase the visual attention and hesitations in terms of silence increase the task performance was noticed at well [28], but the hesitating system was perceived as comparably 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 people regard 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 plainly being silent is credited by users. In a follow-up study we will explore further the applicability of our model with some extensions regarding the realization of hesitations in order to minimize the irritating effects reported for this first prototype.

To conclude, given some necessary improvements on the technical side, we expect the hesitation model to have future application and we will explore that in follow-up studies. The evaluation itself also needs improvements; 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 to go beyond pure MOS scales.

Acknowledgments: This research was carried out as part of the CITEC Large Scale Project "Computational Service Robotics Apartment" (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; Christoph Draxler, who invested much of his time and helped setting up experiment 2 in virtually no time; 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 analyzed the data jointly. Birte Carlmeyer and Simon Betz conducted experiment 1. Simon Betz constructed the stimuli for experiment 2. Petra Wagner conducted experiment 2. Simon Betz, Birte Carlmeyer and Petra Wagner wrote the paper.

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

Appendix 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 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."

Conclusion

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

References

1. Carlmeyer, B.; Schlangen, D.; Wrede, B. Exploring self-interruptions as a strategy for regaining the attention of distracted users. Proceedings of the 1st Workshop on Embodied Interaction with Smart Environments - EISE ’16. Association for Computing Machinery (ACM), 2016.
2. Carlmeyer, B.; Schlangen, D.; Wrede, B. "Look at Me!": Self-Interruptions as Attention Booster? Proceedings of the Fourth International Conference on Human Agent Interaction - HAI ’16. Association for Computing Machinery (ACM), 2016.
3. Skantze, G.; Hjalmarsson, A. Towards incremental speech generation in conversational systems. *Computer Speech and Language 27* **2013**.
4. King, S. What speech synthesis can do for you (and what you can do for speech synthesis). Proceedings of the 18th International Congress of thePhonetic Sciences (ICPhS 2015).
5. Mendelson, J.; Aylett, M. Beyond the Listening Test: An Interactive Approach to TTS Evaluation. Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017, Stockholm), 2017, pp. 249–253.
6. Rosenberg, A.; Ramabhadran, B. Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores. Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017, Stockholm), 2017, pp. 3976–3980.
7. Wester, M.; Braude, D.A.; Potard, B.; Aylett, M.; Shaw, F. Real-Time Reactive Speech Synthesis: Incorporating Interruptions. Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017, Stockholm), 2017, pp. 3996–4000.
8. Wagner, P.; Betz, S. Speech Synthesis Evaluation – Realizing a Social Turn. Tagungsband Elektronische Sprachsignalverarbeitung (ESSV), 2017, p. 167–172.
9. Eklund, R. Disfluency in Swedish human–human and human–machine travel booking dialogues. PhD thesis, Linköping University Electronic Press, 2004.
10. Shriberg, E. Preliminaries to a Theory of Speech Disfluencies. *Ph D. thesis University of California* **1994**.
11. Clark, H.H.; Tree, J.E.F. Using uh and um in spontaneous speaking. *Cognition* **2002**, *84*, 73–111.
12. Goodwin, C. Conversational organization. *Interaction between speakers and hearers* **1981**.
13. Tree, J.E.F. Listeners’ uses ofum anduh in speech comprehension. *Memory & cognition* **2001**, *29*, 320–326.

14. Collard, P. Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech. PhD thesis, University of Edinburgh, 2009.
15. Corley, M.; Stewart, O.W. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass* **2008**, *2*, 589–602.
16. Betz, S.; Zarrieß, S.; Wagner, P. Synthesized lengthening of function words - The fuzzy boundary between fluency and disfluency. Proceedings of the International Conference Fluency and Disfluency, 2017.
17. Kempen, G.; Hoenkamp, E. Incremental sentence generation: Implications for the structure of a syntactic processor. Proceedings of the 9th conference on Computational linguistics-Volume 1. Academia Praha, 1982, pp. 151–156.
18. Levelt, W.J.M. *Speaking: From Intention to Articulation*; MIT Press, 1989.
19. Shriberg, E. To 'errrr'is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* **2001**, *31*, 153–169.
20. Clark, H. Speaking in Time. *Speech Communication 36* **2002**.
21. Shriberg, E. Disfluencies in switchboard. Proceedings of International Conference on Spoken Language Processing, 1996, Vol. 96, pp. 11–14.
22. Shriberg, E. Toerrrr'is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* **2001**, *31*, 153–164.
23. Li, J.; Tilsen, S. Phonetic evidence for two types of disfluency. Proceedings of ICPhS 2015, 2015.
24. Skantze, G.; Schlangen, D. Incremental Dialogue Processing in a Micro-Domain. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), 2009, pp. 745–753.
25. Schlangen, D.; Skantze, G. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue and Discourse* **2011**, *2*, 83–111.
26. 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. Proceedings of the EACL 2014 Workshop on Dialogue in Motion, 2014, pp. 68–72.
27. Bohus, D.; Horvitz, E. Managing Human-Robot Engagement with Forecasts and... Um... Hesitations. Proc. of the 16th International Conference on Multimodal Interaction; ACM: New York, USA, 2014; pp. 2–9.
28. Chromik, M.; Carlmeyer, B.; Wrede, B. Ready for the Next Step?: Investigating the Effect of Incremental Information Presentation in an Object Fetching Task. Proc. of the Companion of the HRI 2017 ACM/IEEE. ACM, 2017.
29. Betz, S.; Wagner, P.; Schlangen, D. Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis. Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015, Dresden), 2015, pp. 2222–2226.
30. Betz, S.; Voße, J.; Zarrieß, S.; Wagner, P. Increasing Recall of Lengthening Detection via Semi-Automatic Classification. Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech 2017, Stockholm), 2017, pp. 1084–1088.
31. Betz, S.; Wagner, P.; Vosse, J. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. Phonetik und Phonologie 12, 2016.
32. Betz, S.; Voße, J.; Wagner, P. Phone Elasticity in Disfluent Contexts. Fortschritte der Akustik - DAGA 2017, 2017.
33. 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.; 1989.
34. Lundholm Fors, K. Production and Perception of Pauses in Speech. PhD thesis, 2015.
35. Carlmeyer, B.; Schlangen, D.; Wrede, B. Towards Closed Feedback Loops in HRI: Integrating InproTK and PaMini. Proceedings of the 2014 Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction. ACM, 2014, ICMI-MMRWHRI '14, pp. 1–6.
36. Baumann, T.; Schlangen, D. The InproTK 2012 Release. NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; SDCTD '12, pp. 29–32.
37. Schroeder, M.; Trouvain, J. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology, 6:365-377.* **2003**.

38. Wrede, S.; Leichsenring, C.; Holthaus, P.; Hermann, T.; Wachsmuth, S. The Cognitive Service Robotics Apartment: A Versatile Environment for Human-Machine Interaction Research. *KI - Kuenstliche Intelligenz (Special Issue Smart Environments)* **2017**.
39. Lütkebohle, I.; Hegel, F.; Schulz, S.; Hackel, M.; Wrede, B.; Wachsmuth, S.; Sagerer, G. The Bielefeld Anthropomorphic Robot Head "Flobi". 2010 IEEE International Conference on Robotics and Automation. IEEE, 2010, pp. 3384–3391.
40. Schillingmann, L.; Nagai, Y. Yet another gaze detector: An embodied calibration free system for the iCub robot. 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), 2015, pp. 8–13.
41. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* **2009**, *1*, 71–81.
42. 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.; Wrede, S. How to Address Smart Homes with a Social Robot? A Multi-modal Corpus of User Interactions with an Intelligent Environment. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); European Language Resources Association: Paris, France, 2016.
43. Draxler, C. Online Experiments with the Percy Software Framework - Experiences and some Early Results. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 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.