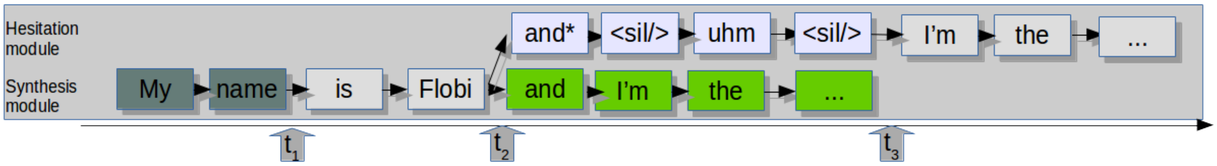

Figure 2: Production of the *wordIU's* of the *hesitation module*: (t_1) module receives *hesitation start* event; (t_2) best entry point for the hesitation; (t_3) module receives *hesitation end* signal. Color scheme of the different incremental units: (dark gray) already synthesized; (blue) hesitation insertions; (green) revoked *wordIU's*.

Incremental processing capabilities. In order to realize the proposed hesitation insertion strategy - in this case the answer of the question *how to react?* - we need the possibility to change the ongoing speech plan. Therefore the capability of incremental processing, especially on the synthesis level, is mandatory. We realize this by using the incremental processing framework *InproTK* [2].

Figure 1 shows an overview of the main parts of the current system. The agent can perceive the human interaction partner via several sensors, in this case via a microphone and a small webcam. The attention management receives information from the gaze detector [9] - the VFoA - and context information from the dialogue management - the current FoD - and provides information about the estimated attention state of the human interaction partner (*attentive*, *not-attentive*). The dialogue management [5] can trigger the motor control of the robot, e.g., to show attention at the current FoD by looking at it and is able to send utterances to text-to-speech synthesis module (tts). Based on the attention state, the dialogue manager can also start and stop the *hesitation strategy*.

Figure 2 shows an example of the hesitation strategy, which is implemented as separate module in *InproTK*, an implementation of the general, abstract model for incremental processing [10]. The model consists of a network of processing modules, which exchange incremental data in form of incremental units (IU). The IU-modules receive information on their *left buffer*, perform some kind of processing on these IU's and provide output on their *right buffer*. At the moment t_1 the module receives the event *'start hesitation'*. The strategy takes the IUs from the *left buffer* of the synthesis module, in this case a list of *wordIUs*, each representing a single word. It searches for the best entry point and lengthens the most appropriate segments. In this example, the synthesis module already played back the first two *wordIU*s "my" and "name". The rest of the current phrase ("is Flobi and I'm the ...") is still in the playback pipeline. According to the proposed strategy, the best entry point for the hesitation strategy is the *wordIU* "and" (t_2), which is then stretched by a factor based on the findings of [3]. Then the synthesis module will be paused up to 1000ms ($< sil/>$). If this is not enough time, the module inserts a filler ("uhm"), also applied with lengthening, followed by a second pause until the dialogue management stops the hesitation strategy (t_3). In the case the dialogue management wants to stop the hesitation strategy earlier (e.g., the estimated attention state of the human interaction partner changed to *attentive*), the strategy can be interrupted at several points: (i) before the entry point t_2: without any effect in the synthesis (ii) before the filler ("uhm"): the lengthening will be produced, but the silence can be interrupted (iii) during or after the filler: the filler will be produced, but the silence is again interruptible.

3 CURRENT STATE & FUTURE WORK

We tested our system in a small pilot study (n=4) in an interaction scenario in a smart apartment to get some insights and first impressions of our system. As a platform we use the simulation of the anthropomorphic robot head Flobi. The system is mostly working as expected. The hesitation strategy starts if the human is not attentive (looking away) and stops when the user refocuses. We observed some issues, which need to be addressed before we can evaluate our system in an interaction study. The insertion of the filler sometimes leads to noise inferences, which need to be eliminated. Additionally, producing fillers such as "uhm" is not a trivial issue, because they are normally not part of the training corpus for speech synthesis voices, and at least for German, they cannot be synthesized out-off-the-box with the tts system at hand without further acoustic modification. We need to investigate if the participants correctly interpret the intention of these fillers as hesitations. After these issues are solved, the next step will be an evaluation study to investigate the effect of this hesitation strategy on the attention, task progress, and the subjective ratings of the agent in order to test if the lengthening and the insertion of fillers can counteract the in [7] perceived rudeness of self-interruptions.

4 ACKNOWLEDGMENTS

This work was funded as part of the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University.

REFERENCES

[1] Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. 6 (03 2017), 25.

[2] Timo Baumann and David Schlangen. 2012. The InproTK 2012 release *(Proc. of the NAACL-HLT Workshop in SDCTD 2012)*. ACL, 29–32.

[3] Simon Betz, Petra Wagner, and Jana VoÃ§e. 2016. Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data. In *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum.* LMU, 19–22.

[4] Dan Bohus and Eric Horvitz. 2014. Managing Human-Robot Engagement with Forecasts and... Um... Hesitations. In *Proc. of the 16th ICMI.* ACM, 2–9.

[5] Birte Carlmeyer, David Schlangen, and Britta Wrede. 2014. Towards Closed Feedback Loops in HRI. In *Proceedings of ICMI-MMRWHRI '14.* ACM, 1–6.

[6] Birte Carlmeyer, David Schlangen, and Britta Wrede. 2016. Exploring self-interruptions as a strategy for regaining the attention of distracted users. In *Proc. of - EISE '16.* ACM.

[7] Birte Carlmeyer, David Schlangen, and Britta Wrede. 2016. "Look at Me!": Self-Interruptions as Attention Booster?. In *Proc. of HAI '16.* ACM.

[8] Philip Collard. 2009. *Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech.* Ph.D. Dissertation. University of Edinburgh.

[9] L. Schillingmann and Y. Nagai. 2015. Yet another gaze detector. In *2015 IEEE-RAS 15th Humanoids.* 8–13.

[10] David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialogue and Discourse* 2, 1 (2011), 83–111.