

information specified on the dialog act, but operates at the phrase level. In other words, the onset demand is checked at the beginning of every phrase in the dialog act.

In addition to reasoning about onset attention, the proposed model also assesses if production demand is met at the end of phrases, *i.e.* if the accumulated attention throughout the phrase matched the production demand specified for the dialog act. If this is not the case, a wait is triggered (to re-acquire onset attention), and then the phrase is repeated. If the onset demand is met at any point during the wait, the system immediately repeats the phrase. The variability of the wait durations, coupled with variability in the attention estimates and the times when the specified onset or production attentional demand is met, leads to a variety of production behaviors in the robot.

4 Deployment and lessons learned

We implemented the model described above in the Directions Robot system and deployed it on three robots situated in front of the bank of elevators on floors 2, 3, and 4 of a four-story building. Appendix A contains an annotated demonstrative trace of the system’s behaviors. Additional videos and snippets of interactions are available at: <http://1drv.ms/1GQ1ori>. While a comprehensive evaluation of the model is pending further improvements, we discuss below several lessons learned from observing natural interactions with the robots running the current implementation.

A first observation is that the usefulness and naturalness of the behaviors triggered by the robot hinges critically on the accuracy of the inferences about attention. When the model incorrectly concludes that the participants’ attention is not on the robot (false-negative errors), the coordinative policy triggers unnecessary pauses, interjections and phrase repeats that can be disruptive and unnatural. The attention inference challenge includes the need to recognize both the participants’ *visual* focus of attention (which in itself is a difficult task in the wild) and *cognitive* attention as being on task. Cognitive attention does not overlap with visual attention all the time. For example, at times participants would shift their visual attention away from the robot as they leaned in and cocked their ear to listen closely. Problems in inferring attention are compounded by lower-level vision and tracking problems.

Second, we believe that there is a need for better integration of the coordinative policy with cur-

rent existing models for language generation, gesture production, multiparty turn-taking and engagement. Beyond the number of words in a phrase, the current policy does not leverage information about the contents of phrases that are about to be generated. This sometimes leads to unnatural sequences, such as “*Excuse me! By the way, would you mind [...]*” Another important question is how to automatically coordinate the robot’s physical pointing gestures when repeating phrases or when phrases are interrupted. With respect to turn taking, problems detected in early experimentation led to an adjustment of the coordinative policy that we described earlier: the system does not move from a wait to a verbal action if it detects that the user is likely speaking. Beyond this simple rule, we believe that the floor dynamics in the turn-taking model need to take into account the system’s discontinuous production, *e.g.*, take into account the fact that the pauses injected within utterances might be perceived by the participants as floor releases. Further tuning of the timings of the pauses, contingent on the dialog state and expectations about when the attention might return, as well as a tighter integration with the engagement model might be required. For instance, we observed cases where the robot’s decision to pause to wait for a participant’s attention to return from the direction that the robot was pointing (before continuing to the next phrase) was interpreted as the end of the utterance and the participant walked away before session completion.

Third, we find that the definition of attentional demands (both onset and production) need to be further refined (in some cases on a per-dialog state basis) and modeled at a finer level of granularity, down to the phrase level. In an utterance like “By the way, would you mind swiping your badge?”, the “By the way” phrase is in fact an attention attractor, and itself does not require attentional demands and thus should be modeled separately.

5 Conclusion

We presented a model for incrementally coordinating language production with listeners’ foci of attention in a multimodal dialog systems. An initial implementation and in-the-wild deployment of the proposed model has highlighted a number of areas for improvement. While further investigation and refinements are needed, the interactions collected highlight the potential and promise of the proposed approach for creating more natural and more effective interactions in physically situated settings.