Incremental Coordination: Attention-Centric Speech Production in a Physically Situated Conversational Agent

Zhou Yu[1]
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
15213
zhouyu@cs.cmu.edu

Dan Bohus
Microsoft Research
One Microsoft Way
Redmond, WA
98052
dbohus@microsoft.com

Eric Horvitz
Microsoft Research
One Microsoft Way
Redmond, WA
98052
horvitz@microsoft.com

Abstract

Inspired by studies of human-human conversations, we present methods for incrementally coordinating speech production with listeners' visual foci of attention. We introduce a model that considers the demands and availability of listeners' attention at the onset and throughout the production of system utterances, and that incrementally coordinates speech synthesis with the listener's gaze. We present an implementation and deployment of the model in a physically situated dialog system and discuss lessons learned.

1 Introduction

Participants in a conversation coordinate with one another on producing turns, and often co-produce language by using verbal and non-verbal signals, including gaze, gestures, prosody and grammatical structures. Among these signals, patterns of attention play an important role.

Goodwin (1981) highlights a variety of coordination mechanisms that speakers use to achieve *mutual orientation* at the beginning and throughout turns, such as pausing, adding phrasal breaks, lengthening spoken units, and even changing the structure of the sentence on the fly to secure the listener's attention. His work suggests that, beyond a simple errors-in-production view, "disfluencies" help to coordinate on turns, and generally facilitate co-production among speakers and listeners. Goodwin (1981) presents sample snippets of conversations recorded in the wild, annotated to show when the gaze of a listener turns to meet the gaze of the speaker (marked with *) and when mutual gaze is maintained (marked with an underline). In the examples reproduced below from Goodwin's work, pauses and repeats are used to align grammatical sentences with a listener's gaze:

Anyway, Uh:, We went *t- I went ta bed

Restarts can be used as a means of aligning the timing of a full grammatical utterance with the start of the process by which gaze is moving towards the speaker (process indicated by the broken underline), as in the following:

She- she's reaching the p- she's at the *point I'm

While most work to date in spoken dialog systems has focused on the acoustic channel in physically situated multimodal systems, an opportunity arises to use vision to take the participants' attention into account when coordinating on the production of system utterances. We investigate this direction and introduce a model that incrementally coordinates language production and speech synthesis with the listeners' foci of attention. The model centers on computing whether the listener's attention matches a set of attentional demands for the utterance at hand. When attentional demands are not met, the model triggers a sequence of linguistic devices in an attempt to recover the listener's attention and to coordinate the system's speech with it. We introduce and demonstrate the promise of incremental coordination of language production with attention in situated systems.

Following a brief review of related work, we describe the proposed approach in more detail in Section 3. In Section 4, we discuss lessons learned

[1] Research conducted during an internship at Microsoft Research

from an in-the-wild deployment of this approach in a directions-giving robot.

2 Related work

The critical role of gaze in coordinating turns in dialog is well known and has been previously studied (*i.a.*, Duncan, 1972; Goodwin, 1981). Kendon (1967) found that speakers signal their wish to release the turn by gazing to the interlocutor. Vertegaal et al. (2003) found evidence that lack of eye contact decreases the efficiency of turn-taking in video conferencing.

Most previous work on incremental processing in dialog has focused on the acoustic channel, including efforts on recognizing, generating, and synthesizing language incrementally. For instance, Skantze and Hjalmarsson (2010) showed that an incremental generator using filled pauses and self-corrections achieved (in a wizard of Oz experiment) shorter response times and was perceived as more efficient than a non-incremental generator. Guhe and Schilder (2002) have also used incremental generation for self-corrections.

Situated and multiparty systems often incorporate attention and gaze in their models for turn taking and interaction planning (Traum and Rickel, 2002; Bohus and Horvitz, 2011). Sciutti et al. (2015) used gaze as an implicit signal for turn taking in a robotic teaching context. In an in-car navigation setting, incremental speech synthesis that accommodates user's cognitive load was shown to improve user experience but not users' performance on tasks (Kousidis, et al., 2015).

3 Model

Motivated by observations from human-human communication dynamics, we propose a model to coordinate speech production with the listeners' focus of attention in a physically situated dialog system. We believe that close coordination between language production and listeners' attention is important in creating more effective and natural interactions.

The proposed model subsumes three subcomponents. The first component defines *attentional demands* on each system output. For successful collaboration, certain utterances require the listener's focus of attention to be on the system or on task-relevant locations (*e.g.*, the direction the robot is pointing towards), while other utterances do not carry high attentional demands. The second component is an inference model that tracks the listener's focus of attention, *i.e.*, the *attentional supply*. The third component alters the system's speech production in an incremental manner to coordinate in stream with the listeners' attention. The component regulates production based on identifying when the attention supply does not match the demands.

In the following subsections, we discuss the model's components in more detail, and their implementation in the context of *Directions Robot*, a physically situated humanoid (Nao) robotic system that interacts with people and provides directions inside our building (Bohus, Saw and Horvitz, 2014). Figure 1 shows a sample dialog with the robot. The proposed coordination model can be adapted to other multimodal dialog systems with adjustments based on the task and the situational context.

3.1 Attentional demands

We consider two types of attentional demand. The first one, which we refer to as *onset demand*, encapsulates Goodwin's observation (1981) that participants in a conversation generally aim to achieve mutual orientation at the *beginnings* of turns. The model specifies that, at each system phrase onset, the listeners' attention must be on the system. In our implementation, we require that at least one of the addressees of the current utterance is attending. The system infers attention under uncertainty from visual scene analysis, and we express the attentional demand by means of a probability threshold. In the current implementation, this threshold was set to 0.6: the onset attentional demand is satisfied if the probability that at least one of the addressees is attending to the robot is greater than 0.6 when the system is launching a phrase.

In addition, a second type of attentional demand, denoted *production demand*, is defined at the level of the dialog act by the system developer. During certain system acts, for instance ones that

1 S:	Hi there!
2 S:	Do you need help finding something?
3 U:	Yes
4 S:	Where are you trying to get to?
5 U:	Room 4505
6 S:	To get to room 4505, ● walk along that hallway, ● turn left and keep on walking down the hallway. ● Room 4505 will be the 1st room on your right.
7 S:	By the way, ● would you guys mind swiping your badge on the reader below so I know who I've been interacting with?

Figure 1. Sample interaction with the Directions Robot.

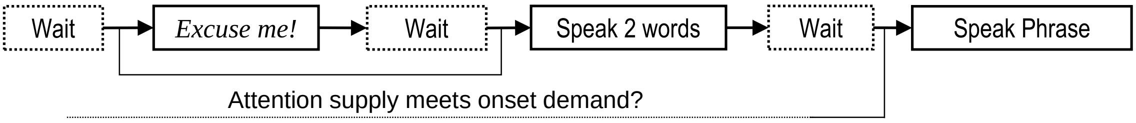


Figure 2. Actions taken to coordinate with attentional demands at phrase onset.

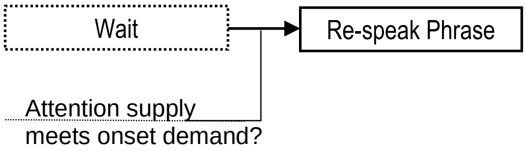


Figure 3. Actions taken to coordinate production with attentional demands.

carry important content or that are deemed as unexpected for the listeners, it is important for addressees to attend to the system or to certain task-relevant objects. The production demand defines where the listeners' attention is expected during the production of the system's utterances, *i.e.*, it defines a set of permitted targets. For instance, when the robot is giving directions in turn 6 from Figure 1, the production demand is set to *Robot* or *PointingDirections*—the locations that the robot points to via its gestures as it renders directions. Similarly, when the robot asks users to swipe their badge in turn 7, the production demand is set to *Robot* and *Badge* indicating that these are the appropriate targets of attention throughout that particular utterance. In contrast, other dialog acts, such as the robot asking *"Where are you trying to get to?"* in turn 4, are naturally expected at that point in the conversation, do not impose high cognitive demands, and can be conveyed without requiring attention on the robot throughout the utterance.

3.2 Attention supply

The Directions Robot is deployed in front of a bank of elevators. In this environment, the attention of engaged participants can shift between a variety of targets including the robot, other task-related attractors (*e.g.*, the direction that the robot is pointing, the sign next to the robot, the user's badge, and the badge reader), personal devices such as smartphones and notepads, and other people in the environment. To simplify, in the implementation we describe here, we model attention supply only over the three targets already mentioned above: *Robot*, *PointingDirection*, *Badge*, and we cluster all other attentional foci as *Elsewhere*.

The robot tracks the (geometric) direction of visual attention for each participant in the scene via a model constructed using supervised machine learning methods. The model leverages features from visual subsystems (*e.g.*, face detection and tracking, head-pose detection, etc.) and infers the probability that a participant's visual attention is directed to the robot, or to the left, right, up, down, or back of the scene. These probabilities are then combined via a heuristic rule that takes into account the dialog state and the robot's pointing to infer whether the participant's attention is on *Robot*, *PointingDirection*, *Badge*, or *Elsewhere*.

3.3 Coordinative policy

The third component in the proposed model, the *coordinative policy* controls the speech synthesis engine and deploys various mechanisms, such as pauses, restarts, interjections, to coordinate the system's speech with the listeners' attention.

Figure 2 shows a diagram of the currently implemented coordinative policy for onset attentional demand. If the listeners' attention does not meet the attentional demand at the beginning of a phrase, the system will perform a sequence of actions, starting with a wait (pause), followed by an attention drawing interjection such as "*Excuse me!*", followed by another wait action, followed by launching the phrase. If the onset attentional demand is still not satisfied the phrase is interrupted after 2 words, then another wait action is taken, followed finally by launching the entire phrase. The wait actions are chosen with a random duration between 1.5 and 2.5 seconds. The interjection is skipped if it was already produced once in this utterance, or if the preceding phrase or the remainder of the utterance contains only one word. As soon as the attention supply matches the onset demand, the system launches the phrase. If the demand is met during the interjection, the interjection will still be completed. In addition, the policy will not switch from a wait action to a verbal action if the system detects that the user is likely speaking.

We set both onset and production attentional demands on a per dialog act basis. The surface realization of a single dialog act can however involve multiple *phrases*, defined here as continuous speech units separated by a pause longer than 250 ms, as signaled by runtime events generated by the speech synthesis engine (● is used to demark phrases in the example from Figure 1.) The coordinative policy uses the attentional demand 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 current 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.

Appendix A: Description of demonstrative sample trace (video at http://1drv.ms/1GQ1ori): At time t_1 the participant's (P_{11}) attention is on the robot and the robot begins giving directions. At the end of the first phrase (t_2), P_{11}'s attention has switched to the other participant as they discuss whether 4800 is really the room they're looking for. Overall the production attention supply (mean of instantaneous attention level over the duration of the phrase, shown in plot A) has exceeded production demand on the initial phrase, so the system deems that no repetition of the phrase is necessary. At the same time, instead of launching the next phrase, the system waits because onset attentional demand is not met. At t_3, onset demand is still not met. Thus, the system launches an interjection followed by launching the first two words at t_4. At t_5, P_{11}'s attention is still not on the robot (according to the inference model, displayed in plot B), and the robot pauses. At t_6, the onset attentional demand is met and the robot re-launches the phrase "go along that hallway". At the end of the phrase (t_7), both the production demand for this phrase and the onset demand for the next phrase are met. However the system has detected that P_{11} is speaking and, instead of launching the next phrase, it waits, allowing P_{11} to finish his contribution. Next, at t_8, the robot provides directions to the new room while P_{11} is attending.

References

Bohus, D., and Horvitz, E., 2011. Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions, *In Proc. of SIGDial 2011*, Portland, OR.

Bohus, D., Saw, C.W., and Horvitz, E. 2014. Directions Robot: In-the-Wild Experiences and Lessons Learned. *In Proc. of AAMAS 2014*, Paris, France.

Duncan, S. 1972. Some Signals and Rules for Taking Speaking Turns in Conversation, *Journal of Personality and Social Psychology*, 23, 283-292.

Goodwin, C., 1981. *Conversational Organization: Interaction Between Speakers and Hearers*, New York: Academic Press.

Guhe, M., & Schilder, F. 2002. Incremental Generation of Self-corrections Using Underspecification. *Language and Computers*, 45(1), 118-132.

Kendon, A. 1967. Some Functions of Gaze Direction in Social Interaction. *Acta Psychologica* 26, 22–63

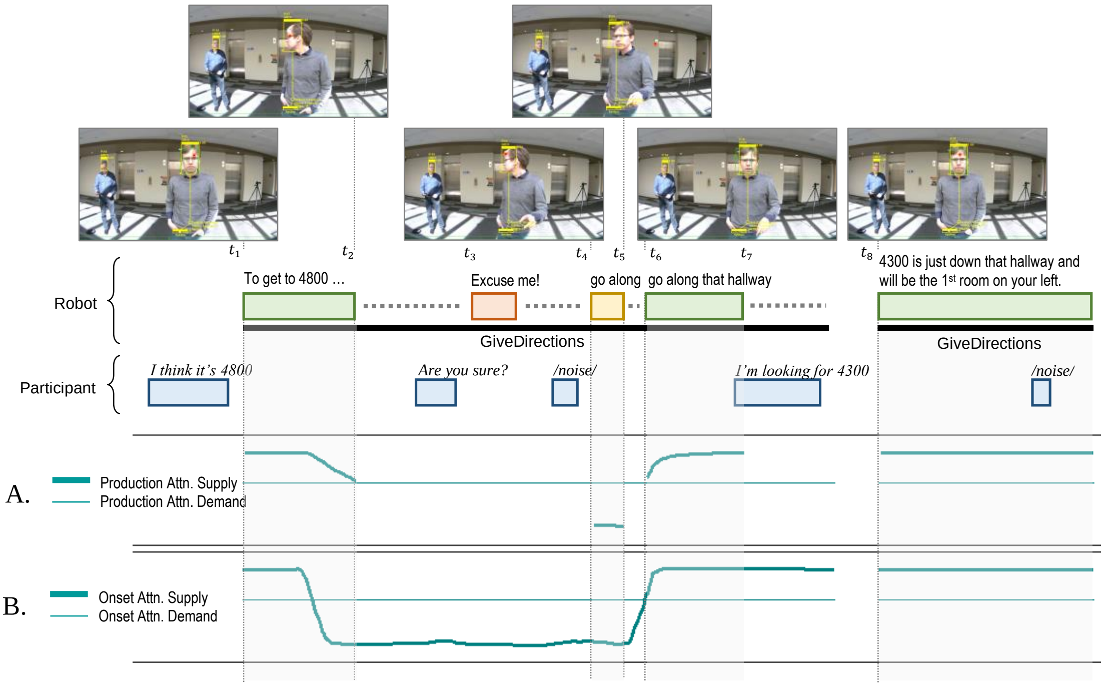
Kousidis, S., Kennington, C, Baumann,T., Buschmeer, H., Kopp, S., and Schlangen, D. 2014. A multimodal In-car Dialogue System that Tracks the Driver's Attention. *In Proc. of ICMI 2015*, Istanbual,Turkey.

Traum, D., and Rickel, J., 2002. Embodied Agents for Multi-party Dialogue in Immersive Virtual World, *In Proc. of AAMAS 2002*,Bologna, Italy.

Sciutti, A., Schillingmann, L., Palinko, O., Nagai, Y., and Sandini, G., 2015. A Gaze-contingent Dictating Robot to Study Turn-taking. *In Proceedings of HRI 2015*, Portland, OR, USA.

Skantze, G., and Hjalmarsson, A., 2010. Towards Incremental Speech Generation in Dialogue Systems, *In Proc. of SIGDial 2010*, Tokyo, Japan.

Vertegaal, R., Weevers, I., Sohn, C. and Cheung, C. 2003. GAZE-2: Conveying Eye Contact in Group Videoconferencing Using Eye-controlled Camera Direction. *In Proc. of CHI 2003*, Fort Lauderdale, FL.


t_1
t_2
t_3
t_4
t_5
t_6
t_7
t_8
Robot
To get to 4800 …
Excuse me!
go along
go along that hallway
4300 is just down that hallway and will be the 1st room on your left.
GiveDirections
GiveDirections
Participant
I think it's 4800
Are you sure?
/noise/
I'm looking for 4300
/noise/
A.
Production Attn. Supply
Production Attn. Demand
B.
Onset Attn. Supply
Onset Attn. Demand