

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

Zhou Yu¹

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