Spot the Difference
- A dialog system to explore turn-taking in an interactive setting

Anna Hjalmarsson
Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
annahj@kth.se

Margaret Zellers
Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
zellers@kth.se

Abstract

Much research has been devoted to understanding the principles that control the flow of dialog contributions between speakers in dialog. This demonstration paper describes a dialog system that was developed as test-bed to experiment with turn-taking aspects in an interactive setting.

1 Introduction

Over the years, much effort has been devoted to understanding the principles that control the flow of dialog contributions between speakers (see for example Sacks et al., 1974). The motivation for this research includes both the desire to understand the underlying mechanisms of human communication as well as to build dialog systems with more sophisticated turn-taking capabilities. A first step towards increased understanding of these phenomena has been to identify behaviors that correlate with speaker changes in human-human dialog (see for example Duncan, 1972). One approach to further understand how these behaviors influence listeners' expectations of a speaker change is to study listeners' expectations of who will speak next in an off-line setting where subjects listen to pre-recorded dialog excerpts (Hjalmarsson, 2011 and Zellers, 2013). However, in order to understand to what extent the target behaviors actually influence listeners' turn-taking decisions, these behaviors needs to be explored in an interactive setting. The aim of the dialog system presented in this demonstration paper is to serve as a test-bed for such experimentation. An advantage of using a dialog system to do this is that a system's behavior, as opposed to a human's behavior, can easily be controlled. Furthermore, a dialog system is also suitable for studies that aim to identify human behaviors that can be used to regulate turn-taking in human-machine interaction.

The paper is structured as follows. In section 2, we will present the motivation and theoretical background of an initial planned study and in section 3, we will present the domain and implementation of the dialog system that we will use in this research.

2 Timing in utterance generation

Most of today's dialog systems have no strategies to adjust the timing of speech to the local dialog context. Utterances are produced as whole units as soon as they become available to the speech generator, and the timing of individual speech segments is typically based on a shallow syntactic analysis of the isolated utterance. However, dialog systems that use incremental models for processing (Schlangen & Skantze, 2011) process utterances in smaller sub-segments in a way that is more similar to human speech processing. Such incremental speech processing opens up for more fine-grained generation of utterances where small variations in the system's output can be used to accommodate the semantic and pragmatic dialog context. Analyses of human-human dialog data suggest that the temporal flow of speech has several important structural functions (cf. Goldman-Eisler, 1972). The timing of different speech events – a phoneme, a prolonged syllable or a pause – in conversation affects listeners' perception of an utterance and is influenced by the dialog context (Zellner, 1994).

In a recent series of articles (Skantze & Hjalmarsson, 2013 and Skantze et al., 2014), we have explored how the preceding context affects users' reactions to temporary silences in the system's speech. The aim of the system

presented here is to serve as a testbed for pursuing this research in the setting of fully functional dialog system. In an initial experimental study, we will explore how various non-lexical behaviors, such as variation in pitch and duration as well as inhalations and fillers (e.g. "eh" and "ehm") affect users' turn-taking decisions when these behaviors are followed by silence.

3 The Spot the Difference system

The domain that was chosen for the dialog system is similar to the frequently used map-task domain (Anderson et al., 1991). However, instead of identifying differences between maps, the players' task is to identify differences between two versions of a picture (see Figure 1).

Figure 1: Two versions of a scene in the system.

In this domain, nominal phrases of various complexity are used to refer to objects, and whether it is appropriate to take the turn or not is often ambiguous when relying on lexical context alone (see the dialog example in Figure 2). This makes the domain suitable for experimenting with non-lexical turn-taking cues.

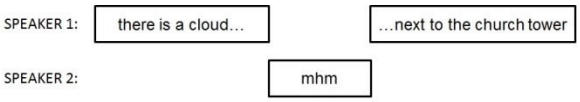

Figure 2: Human-human dialog excerpt.

3.1 System implementation and setup

The dialog system was implemented using IrisTK (Skantze & Al Moubayed, 2012), a framework for building multimodal conversational systems, and the GUI was implemented in Java. For automatic speech recognition and end-of-speech-detection, we use an off-the-shelf speech recognizer, and for speech synthesis, we use the CereVoice system developed by CereProc[1]..In order to explore the effect of mid-utterance pauses, the system's utterances are realized in utterance segments with short silences in-between. As the aim of the experiment is to explore effect of non-lexical behaviors, all utterance segments are semantically complete.

[1] http://www.cereproc.com

Acknowledgments

This work was supported by the Swedish Research Council (VR) project Classifying and deploying pauses for flow control in conversational systems (2011-6152).

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351-366.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15*, 103-113.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1), 23-35.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Schlangen, D., & Skantze, G. (2011). A General, Abstract Model of Incremental Dialogue Processing. *Dialogue & Discourse, 2*(1), 83-111.

Skantze, G., & Al Moubayed, S. (2012). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI*. Santa Monica, CA.

Skantze, G., & Hjalmarsson, A. (2013). Towards Incremental Speech Generation in Conversational Systems. *Computer Speech & Language, 27*(1), 243-262.

Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, Feedback and Joint Attention in Situated Human-Robot Interaction. *Speech Communication, 65*, 50-66.

Zellers, M. (2013) Pitch and lengthening as cues to turn transition in Swedish. Proceedings of 14th Interspeech, Lyon, France, 248-252.

Zellner, B. (1994). Pauses and the Temporal Structure of Speech. In Keller, E. (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41-62). Chichester: Wiley.