Investigating perception of prosody in conversation using a spoken dialogue system
Margaret Zellers & Anna Hjalmarsson
Dept. of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
zellers@kth.se; annah@speech.kth.se

In order to promote smooth turn-taking during conversation (cf. Sacks et al., 1974), speakers must have ways of providing information to one another about whether they have finished speaking or intend to continue. Multiple characteristics of interlocutors’ production contribute to turn-taking simultaneously, and recent studies involving large corpora of data (e.g. Koiso et al., 1998; Gravano & Hirschberg, 2009, 2011; Meena et al., 2013) report a hierarchy of various linguistic and non-linguistic features correlated with turn transition or turn hold. Most studies investigating turn-taking cues have focused on production data, and those that have studied listeners’ perception of turn-taking cues (Hjalmarsson, 2011; Zellers, 2013, 2014) have often relied upon relatively off-line tasks, in which listeners make judgments about speech that they are “listening in” on, rather than participating in. One exception to this generalization is a study by Edlund & Beskow (2009), which replaces visual contact between conversation participants with on-screen avatars. Edlund & Beskow report that by modifying the avatars’ eye gaze towards or away from the participants, they were able to influence how often those speakers initiated conversational turns.

We will present a methodology for on-line investigation of the influence of phonetic features, specifically pausing and intonation, on turn-taking in conversation, and preliminary results from a study applying this methodology. Participants see an image on-screen while a spoken dialogue system (SDS) offers descriptions of a similar, but not identical, image. The task is to “spot the differences” – that is, to interact with the SDS in order to identify as many differences between the two images in a short amount of time. The SDS’s production is modified to create situations in which the pausing and intonation imply turn-change or turn-hold, and the human participants’ response times and task responses are then analyzed in the context of the input intonational and pausing patterns. For example, we expect that low-falling intonation, which is typically associated with turn-change in English conversation, will lead to more and earlier turn initiation by the human participants than mid-rising intonation, which is typically associated with turn-hold. The use of an SDS allows for close control of the input provided to the human participant, thus creating a context for controlled experiments in prosody perception while at the same time maintaining a more natural conversational setting. Since listeners have been shown to respond to synthesized speech in a similar manner to natural speech (Hjalmarsson, 2011; Huang et al., 2010), our spot-the-difference system represents a powerful tool in the attempt to extend controlled linguistic research to more natural conversational settings.

References


