MushyPeek – an experiment framework for controlled investigation of human-human interaction control behaviour

Jens Edlund, Jonas Beskow & Mattias Heldner
KTH Centre for Speech Technology

Abstract

This paper describes MushyPeek*, a experiment framework that allows us to manipulate* interaction control *behaviour – including* turn-taking *– in a setting quite similar to face-to-face human-human dialogue. The setup connects two subjects to each other over a VoIP telephone connection and simultaneuously provides each of them with an avatar representing the other. The framework is exemplified with the first experiment we tried in it – a test of the effectiveness interaction control gestures in an animated lip-synchronised talking head.*

Introduction

People take a great number of things into consideration in order to manage the flow of the interaction when conversing face-to-face. We call this *interaction control* – the term is wider than *turn-taking* and does not presuppose the existence of "turns". Examples of features that play a part in interaction control include auditory cues such as pitch, intensity, pause and disfluency, hyperarticulation, etc.; visual cues such as gaze, facial expressions, gestures, and mouth movements; and cues like pragmatic, semantic and syntactic completeness.

People commonly use these cues in combination and seem to mix them or shift between them seamlessly. In order fully understand human interaction control, we need to know how these features work in combination. In order to reach that goal, however, it seems fair to first get a handle on how the cues are used and perceived on their own.

We have previously tested the perception and the production of a number of such cues in various user experiments, often involving users talking to different configurations of a spoken dialogue system where the interaction behaviour can be varied in a controlled manner (e.g. Edlund & Nordstrand, 2002, Bell et al., 2001), but also by analysing human-machine as well as human-human dialogues (Edlund & Heldner, 2005, Edlund et al., 2005).

This paper describes *MushyPeek*, a different experimental design that allows us to investigate interaction control behaviour by manipulating certain parameters in somehing quite similar to face-to-face human-human dialogue. Similar methods have been used by others; we were especially inspired by Gratch et al. (2006).

In this experimental setup, two subjects are connected to each other over a VoIP telephone connection. Furthermore, each speaker sees an avatar, a visual representation of the other speaker, at all times. (We use *avatar* to denote a virtual representation of a *human*, whereas a virtual representation of a *system* as a creature would be an ECA – an *embodied conversational agent* – in our terminology. For the avatars, we use SynFace lip synchronised talking heads (Beskow et al., 2004). Finally, the setup consists of a simple model of multi-party interaction which functions as the control mechanism for the manipulation of the interaction, and of a number of logging mechanisms. In the following, we will present this experimental setup and include a brief report of the first experiment we've undertaken in it as an illustration.

The *MushyPeek* framework

In order to better be able to investigate people's turn-taking behaviour, we have designed an experiment framework in which two interlocutors speaks freely. The participants are placed in separate rooms, and each participant is equipped with a head-set connected to a Voice-over-IP call. Currently, we use Skype (http://www.skype.com/) for this. On both sides, the call is enhanced with SynFace – a lip synchronised animated talking head functioning as an avatar, representing each participant. These

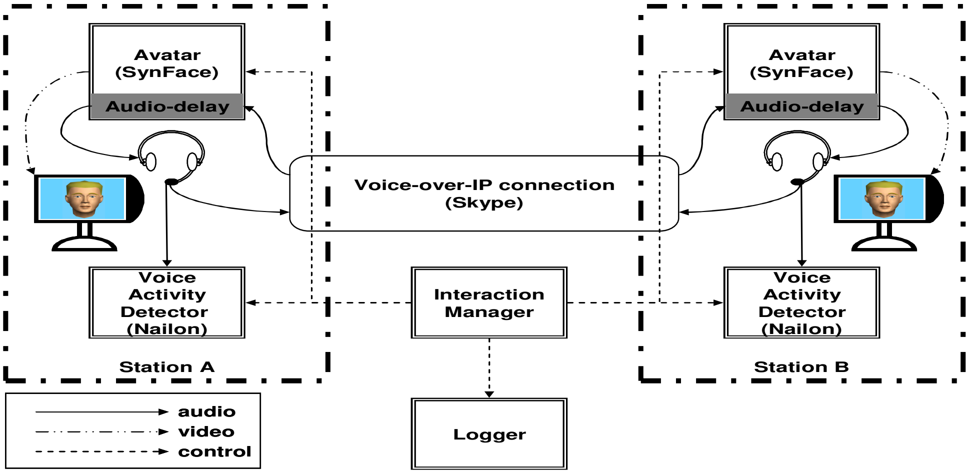

Figure 1: The basic experiment setup

components constitute the communicative backbone of the system. On top of this, the framework contains a component for voice activity detection (VAD) on each side and a interaction model that takes the results of the VADs as its input, and a logger. In addition, the framework allows for experiment-specific components for tasks such as gesture realisation. All components communicate over TCP/IP. The framework is symmetrical in that both participants have the same setup. The general layout is shown in Figure 1. The components are described in more detail in the following.

Voice activity detector

The audio signal from each participant was processed locally by the voice activity detection (VAD) included in the **/nailon/** software package (Edlund & Heldner, 2006). This VAD is quite fast, and although the algorithms used are quite rudimentary, they produce good results in the quiet environments of the experimental setup. The VAD reports to the interaction model. It reports a change to the SPEECH state each time it detected a certain number of consecutive speech frames whilst in the SILENCE state, and vice versa. The number of speech frames used is configurable. The equivalent of 100-300 ms produces good results.

Avatar

The SynFace talking head is an application originally developed to provide real-time lip-reading support to hard-of-hearing persons during telephone conversation. It uses phoneme recognition and facial animation to re-create important features of the articulation of the speaker at the other end of a telephone connection. SynFace introduces a small delay (200 ms) on the audio channel to obtain full audio-visual synchrony between the animated face and the audio stream. In MushyPeek, SynFace can be supplemented with gestures related to interaction control.

Interaction model

The interaction model in MushyPeek is computationally simple yet powerful. It bears similarities to other computational models of interaction, such as the AVTA system (Jaffe & Feldstein, 1970), but differs in that it can be used to model multilogue (dialogue with more than two interlocutors) and in that it models the interaction from the perspective of each participant separately. The model consists of three parts: a state derived directly from each participant's speech activity, a state derived from the speech activity of all participants, and events representing changes in these states.

The first state (SPEECH/SILENCE) continuously models speech/non-speech as a binary state on a per-participant level. At any given point in time, each participant may be either speaking or not speaking. The only input the model takes is speech/non-speech decisions from each participant's VAD.

The second part of the model is a four-way decision of the communicative state (SELF/OTHER/NONE/BOTH), again repeated for each participant. These states are derived from the SPEECH/SILENCE state of each participant. From participant P's point of view, the state is NONE if none of the participants are

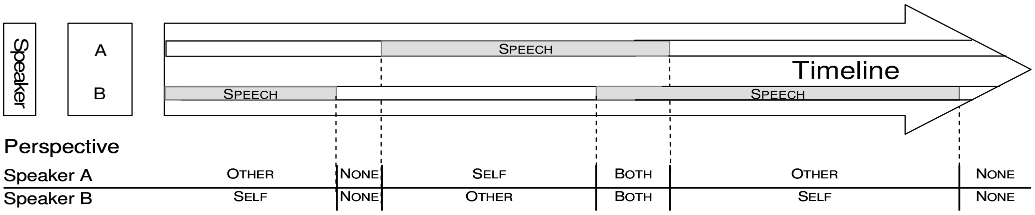

Figure 2: illustration of the *MushyPeek* interaction model

speaking. It is SELF if P is speaking but no one else. If one or more other participants are speaking and P is silent, it is OTHER, and finally, if both P and some other participant is speaking, it is BOTH. Note that whereas the SPEECH/SILENCE state models each participant without considering relations to other participants, the communicative states NONE, SELF, OTHER and BOTH takes all participants into consideration.

Finally, the model includes transitions between one communicative state to another for each participant. If P is in state NONE and someone else starts speaking, P goes from NONE to OTHER (and the participant who started speaking goes from NONE to SELF). Figure 2 illustrates the model.

Logger

A prerequisite for the experiment framework was that it should not require human post-processing. Once an experiment is done, it should be possible to calculate the results directly from the log files. We achieved this by deriving turn-taking statistics from the SPEECH/SILENCE decisions provided by **`/nailon/`** and from the interaction model. The measure used in the experiment we are about to describe in brief is the quotient between the number of spoken contributions or inter-pausal units (IPUs) by a speaker that are followed by a speaker change and the number that are followed by the same speaker. In order to fully operationalise this measure and access it without human intervention, the following definitions are adopted from [2]:

- An IPU is speech of a given minimum length (300ms) in one speaker's channel delimited in both ends by a given amount (300ms) of non-speech in the same channel.
- A speaker CHANGE occurs when the end of an IPU from one speaker is followed by a speech by another speaker.
- A speaker KEEP occurs when the end of an IPU from one speaker is followed by a speech by the same speaker.

Using these simple definitions, the framework can test whether a participant takes the floor more often under one condition than under another.

Note that speaker CHANGE and KEEP events are not symmetric, but reflect the perspective of one speaker only. They are logged separately from each participant's perspective. At the end of a session, the quotient of speaker KEEPs and speaker CHANGEs for each user under each condition is calculated. A larger proportion of KEEPs indicates a speaker exhibiting a more pushy turn-taking behaviour, whilst a meeker behaviour results in a larger proportion of CHANGESs, as the speaker looses the floow more regulalrly after pausing.

Experiment example

MushyPeek was inaugurated with an experiment investigating whether know turn-taking gestures, such as looking away and looking back up at the speaker, could be used in a talking head to influence interlocutants behaviour, and given that they could, whether such influence would be noticed by the interlocutors. Subjects were equipped with headsets and placed in front of monitors in separate rooms. They consisted of pairs who knew each other from before, and their task was to speak about any topic freely for around ten minutes. 6 different pairs were used, resulting in 12 participants all in all. None of the participants had any previous knowledge of the experiment setup.

MushyPeek is intended for within-group designs where the conditions are varied systematically and a large number of times, and in this experiment there were two sets of gestures – one that was hypothesised to produce a more meek turn-taking behaviour in the person subjected to the gestures, and one that was hypothesised to do the opposite. The sets were always applied one to the one user and the other to the other, so that the users never were subjected to the same gesture set. The sets were swapped at random intervals quite often – about every 10-20 second – to prevent the users from getting used to them.

Given that the gestures had the intended effect, the relation between CHANGE and KEEP should vary depending on the direction of the gesture sets: when participants face ACTIVE (and their partner PASSIVE), they should have a higher degree of KEEP, and when they face PASSIVE, they should have a higher degree of CHANGE.

Results

The results clearly confirmed the hypothesis, making the experiment a success and validating the experimental framework. The percentage of all contributions followed by CHANGE was larger under the PASSIVE condition than under the ACTIVE condition for each participant without exception. The difference is significant at the 0.01 level with a paired two-sample t-test for means: df=11, t=4,66, P<0,01 (two-tailed).

Interestingly, the data shows that different users exhibited very different behaviour in all manners. For example, the total number of utterances spoken varies from around 30 to well above 100; one user let the turn pass on in almost 90% of the cases overall whereas another kept the floor in 75% of the cases; and some pairs spoke an almost equal number of utterances whereas others had a ratio of 2 to 1 or more. There are many reasons for this: the subjects' personalities and the way they felt about the experiment is an influence, of course. Another reason may be that they chose to speak about quite varying topics. The results have been submitted for publication (Edlund & Beskow, submitted).

Conclusion and future work

We have presented a framework for conducting controlled interaction control studies of people talking to each other in a near-face-to-face condition. The results from the first experiment we have conducted are very encouraging, supporting the hypothesis tested as well as validating the experimental framework.

While experiments in this framework are intended to be carried out in an avatar-mediated human-human communication setting, we believe that the framework can easily be extended to make human-computer studies as well. Furthermore, experiments in the framework ought to be useful not only for testing hypothesis about human-human interaction, but for human-computer interaction scenarios in the design of human like spoken dialogue systems as well.

The experiment described here tested artificial manipulation of head pose and gaze, rather than auditory cues. While manipulating the interlocuters' voices on-line, without raising too much supsision, may be more difficult than adding gestures in a talking head, we believe it is certainly possible.

Finally, the interaction model presented, as well as the framework as a whole, is not confined to two-party dialogue but can be easily extended to handle multilogues, which will an be interesting aspect to explore in the future.

Acknowledgements

This work was funded by the Swedish research council projects #2006-2172 and #621-2005-3488 and the European Commission's FP6 projects IP-506909 and IP-035147.

References

Bell, L., Boye, J., & Gustafson, J. (2001). Real-time handling of fragmented utterances. In *Proc. NAACL 2001 Workshop: Adaptation in Dialogue Systems*.

Beskow, J., Karlsson, I., Kewley, J., & Salvi, G. (2004). SYNFACE - A talking head telephone for the hearing-impaired. In Miesenberger, K., Klaus, J., Zagler, W., & Burger, D. (Eds.), *Computers Helping People with Special Needs* (pp. 1178-1186). Springer-Verlag.

Edlund, J., & Beskow, J. (submitted). Pushy versus meek ▯ using avatars to influence turn-taking behaviour. Submitted to *Proceedings of Interspeech 2007 ICSLP*. Atwerp, Belgium.

Edlund, J., & Heldner, Mattias (2005). Exploring Prosody in Interaction Control. *Phonetica, 62*(2-4), 215-226.

Edlund, J., & Heldner, Mattias (2006). /nailon/ - software for online analysis of prosody. In *Proc of Interspeech 2006 ICSLP*. Pittsburgh PA, USA.

Edlund, J., & Nordstrand, M. (2002). Turn-taking gestures and hour-glasses in a multi-modal dialogue system. In *Proc of ISCA Workshop Multi-Modal Dialogue in Mobile Environments*. Kloster Irsee, Germany.

Edlund, J., Heldner, Mattias, & Gustafson, J. (2005). Utterance segmentation and turn-taking in spoken dialogue systems. In Fisseni, B., Schmitz, H-C., Schröder, B., & Wagner, P. (Eds.), *Computer Studies in Language and Speech* (pp. 576-587). Frankfurt am Main, Germany: Peter Lang.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., & Morency, L-P. (2006). Virtual rapport. In *Proceedings of 6th International Conference on Intelligent Virtual Agents*. Marina del Rey, CA, US.

Jaffe, Joseph, & Feldstein, Stanley (1970). *Rythms of dialogue*. New York: Academic Press.