Pushy versus meek – using avatars to influence turn-taking behaviour

Jens Edlund and Jonas Beskow

KTH Centre for Speech Technology
Stockholm, Sweden

edlund@speech.kth.se, beskow@kth.se

10.21437/Interspeech.2007-289

Abstract

The flow of spoken interaction between human interlocutors is a widely studied topic. Amongst other things, studies have shown that we use a number of facial gestures to improve this flow – for example to control the taking of turns. This type of gestures ought to be useful in systems where an animated talking head is used, be they systems for computer mediated human-human dialogue or spoken dialogue systems, where the computer itself uses speech to interact with users. In this article, we show that a small set of simple interaction control gestures and a simple model of interaction can be used to influence users' behaviour in an unobtrusive manner. The results imply that such a model may improve the flow of computer mediated interaction between humans under adverse circumstances, such as network latency, or to create more human-like spoken human-computer interaction.

1. Introduction

The mechanisms regulating the flow of conversation are complex and rely on many cues, both auditory and visual. In the present study, we show that it is possible to unobtrusively influence the turn-taking behaviour of two interlocutors in a given direction – that is to make a person take the turn more or less often – by way of facial gestures in an animated talking head. The gestures are smoothly injected in the conversation, and are governed by voice activity detection and a simple set of interaction rules in an automated experiment framework.

2. Background

Spoken face-to-face communication is one of the most intuitive, robust and effective forms of communication between humans that we can imagine. The use of *embodied conversational agents* (ECAs) - human-like representations of a system, for example animated talking heads that are able to interact with a user in a natural way using speech, gesture and facial expression - is one way of leveraging the inherent abilities that we all possess in terms of decoding information in speech, visible articulation, intonation, voice quality, facial displays, gestures and gaze, and holds the potential of improving both effectiveness, robustness and naturalness of human-computer interaction. Similarly, a talking head acting as an *avatar* – a representation of a human that is not present – can also prove useful as a mediator in human-human communication. For example, [1] provides lip-reading support during ordinary telephone conversations to hearing impaired individuals by re-creating the visible articulation of the speaker at the other end.

An area where talking heads can potentially make a difference is turn-taking. In current spoken dialogue systems, turn-taking is typically handled in an un-humanlike manner. Methods for improving the systems' decisions about when to take turn have been devised (e.g. [2]), but systems are still not very good at showing when they want the floor. Gestures have, however, been shown to be more appreciated by users than some other signals for turn-taking [3].

[4,5,6,7] have all described ECA-like systems capable of generating turn-taking and back-channelling signals. It is however difficult to verify that these signals have the desired effect. Global evaluation metrics (e.g. task completion, disfluency rate, user satisfaction) give an indication, but metrics targeted specifically at turn-taking in ECA systems are lacking. Hence, one of the goals of the present study is to provide an experimental paradigm that can be used to evaluate cues and models for turn-taking in systems involving animated talking characters.

3. Experiment framework

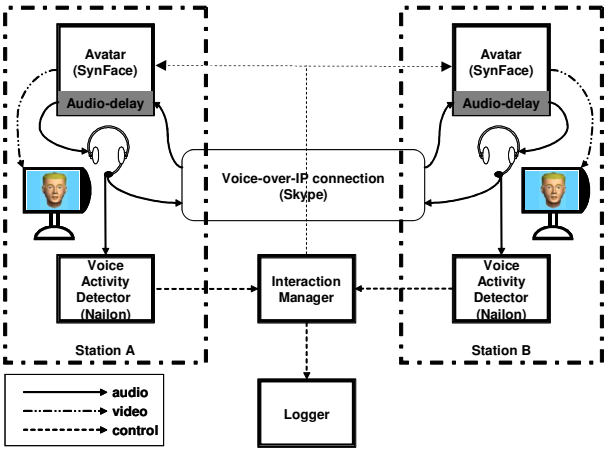

Figure 1: The experiment system

In order to test whether the people's turn-taking behaviour could be affected by system controlled gestures, we designed an experiment framework in which two interlocutors speak freely. The framework is inspired by [8], where listeners can hear the speakers' real voices while watching what they are told to be graphic representations of the speakers and their gestures on monitors. Our participants are placed in separate rooms, and each participant is equipped with a head-set connected to a Voice-over-IP call [9]. On both sides, the call is enhanced with SynFace [1] - a lip synchronised animated talking head representing each participant. As both talking heads represent real persons (the participants), we refer to them as *avatars* in the following. This basic setup constitutes the communicative back-bone of the framework. In addition, the framework contains experiment-specific components for voice activity detection (VAD), interaction modelling and control, gesture realisation, and logging. All components communicate over TCP/IP connections. The framework is symmetrical in that both participants have the same setup. The general layout is shown in Figure 1. The experiment specific components are described in detail in the following.

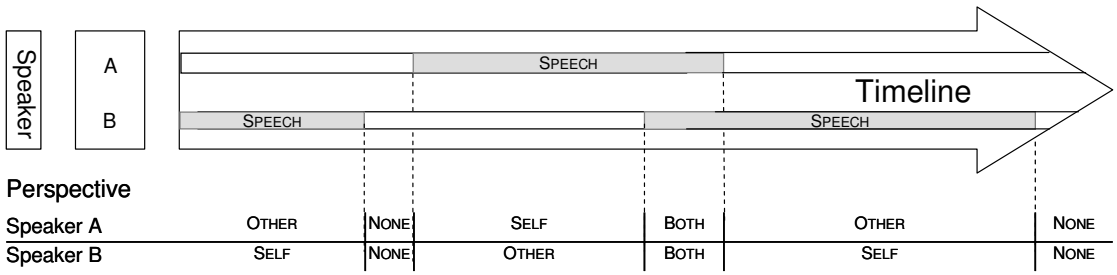

Figure 2:elations between speakers, states and transitions in the interaction model

3.1. Voice activity detector

The audio signal from each participant was processed locally by the voice activity detection (VAD) included in the **`/nailon/`** software package [10]. 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 manager. It reports a change to the SPEECH state each time it detected more than 300ms worth of consecutive speech frames whilst in the SILENCE state, and vice versa.

3.2. 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 the 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 this experiment, SynFace was supplemented with a small set of gestures related to turn-taking and interaction control.

Turn-taking in human-human communication has been widely studied. [11] suggested a large set of acoustic and visual cues that are in play in the turn taking system, related to prosody, lexical content, hand/body gestures, head pose and gaze. Several other studies have focused specifically on gaze behaviour and turn-taking [12,13]. In vastly simplified summary, these studies all generally agree that the speaker tends to turn away or shift the gaze away at the beginning of a turn, and look towards the listener towards the end of the turn. The listener, on the other hand, tends to look at the speaker to a larger extent.

Based on these findings, and our previous experiences, we decided to implement a bare minimum of interaction control gestures:

- A turn-taking/keeping gesture, where the avatar makes a slight turn of the head to the side in combination with shifting the gaze away a little.
- A turn yielding/listening gesture, where the avatar looks straight forward, at the subject, with slightly raised eyebrows.
- A feedback/agreement gesture, consisting of a small nod. In the experiment described here, this gesture is not used alone, but is added at the end of the listening gesture to add to its responsiveness. In the following, simply assume it is present in the turn yielding/listening gesture.

3.3. Interaction manager

The model used is computationally simple yet powerful. It 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 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.

Finally, the model includes transitions from 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.

The transition events in the interaction model control the gestures in the avatar. Three sets of mappings between gestures and events were used, one of which (NEUTRAL) left the face looking straight ahead for all states, resulting in a relatively immobile avatar. The label NEUTRAL refers to the fact that there was no hypothesis as to how the gesture-less avatar would affect the subjects' behaviour. The state was used as padding between the two other states. These, on the other hand, were designed to nudge the turn-taking behaviour of the subjects. ACTIVE, in which the avatar looks straight at the subject under all states except SELF, was intended to encourage a more pushy or active behaviour by presenting a listening appearance. PASSIVE, in which the avatar always looks away from the subject except in OTHER, was intended to obtain a meeker or more passive behaviour. The ACTIVE and PASSIVE gesture sets were used in pairs, so that whenever one user was confronted with the ACTIVE gestures, the other one had the PASSIVE set (Figure 3).

For every 10 SPEECH/SILENCE transitions, the gesture sets were shifted cycling through speaker A-speaker B mappings of ACTIVE-PASSIVE, NEUTRAL-NEUTRAL, PASSIVE-ACTIVE, and again NEUTRAL-NEUTRAL. SPEECH/SILENCE transitions is a measure that gives some variation in terms of time as well as number of utterances – a timer or exact utterance counter was avoided in order to make it less obvious that the gesture behaviour was varied systematically. 10 SPEECH/SILENCE

transitions is roughly equivalent to a shift every fifth utterance. The "neutral" gesture sets were active roughly 50% of the time in an attempt to avoid tiring the participants, whereas the ACTIVE-PASSIVE configuration was active roughly 25% of the time in each direction.

3.4. Logger

A prerequisite for the experiment framework was that it should not require human post-processing. Once an experiment is done, the results should be calculated directly from the log files. This is achieved by deriving turn-taking statistics from the SPEECH/SILENCE decisions provided by **`/nailon/`**. The measure used in the current experiment is the quotient between the number of spoken contributions 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]:

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

The framework tests whether a participant takes the floor more often under one condition than under another. In order to measure this, the logger inspects events following a pause of more than 300ms in either of the speakers' channels. This is done using the same system that controls the gestures. Anytime speaker A goes silent, the following may follow:

1) The other participant is already speaking. This counts as a CHANGE from speaker A's perspective.
2) The other participant is silent. The next registered change is that the other participant starts speaking. This counts as a CHANGE from speaker A's perspective.
3) The other participant is silent. The next registered change is that the first speaker resumes speaking. This counts as KEEP from speaker A's perspective.

Speaker CHANGE and KEEP decisions are not symmetric, but reflect the perspective of one speaker only. They were logged separately from each participant's perspective.

At the end of each session, the quotient of speaker KEEPs and speaker CHANGEs for each user under each condition is calculated. A speaker with a pushy turn-taking behaviour will register a larger proportion of KEEPs, whilst meek behaviour results in a larger proportion of CHANGESs.

4. Experiment

In the present experiment, subjects were equipped with headsets and placed in front of monitors in separate rooms. The type of everyday face-to-face dialogue that we strive to model is common between friends, and in an attempt to get dialogues that were as relaxed and natural as possible, we chose subjects consisting of pairs who knew each other previously. 6 different pairs were used, resulting in 12 participants all in all. None of the participants had any previous knowledge of the experiment setup. Subjects ages varied from the 20s to the 60s, and there were eight men and four women.

The subjects were asked to talk freely to each other for around ten minutes. In two cases, the participants preferred to make up a few topics in advance. The other four pairs conversed without any pre-selected topics.

The experiment is a within-group design where the conditions are varied systematically and a large number of times.

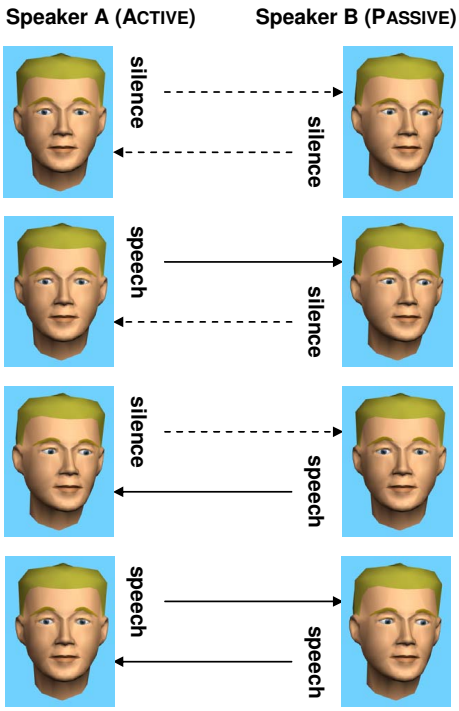

Figure 3 Resulting head poses after transition gestures for the ACTIVE-PASSIVE condition

5. Hypothesis

Given that the gestures have 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. The effect should be discernable for individual participants.

6. Results

Figure 4 shows that the percentage of all contributions followed by CHANGE is 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).

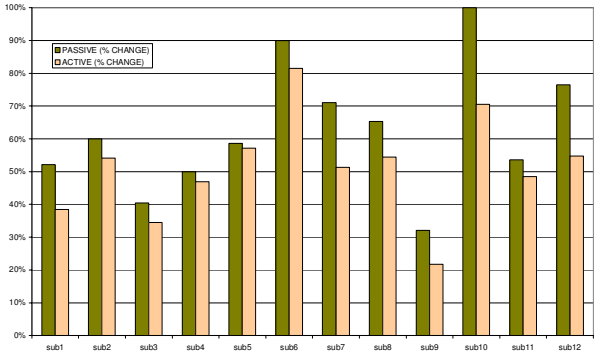

Figure 4 CHANGE quotient per user and condition

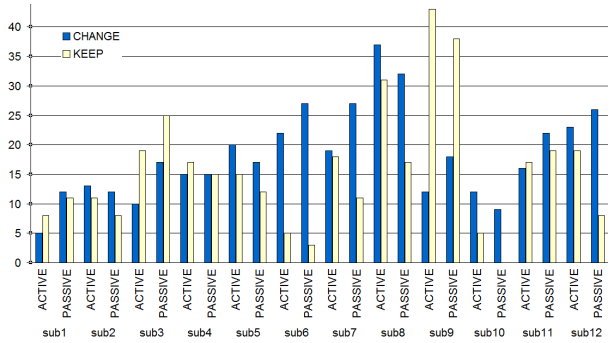

Figure 5 CHANGE and KEEP per user and condition

Figure 5 shows, for each user, the absolute number of times an utterance was followed by an utterance by the other user (CHANGE) and an utterance by the same user (KEEP) for both conditions (PASSIVE and ACTIVE). It is evident that users exhibited a great variety of behaviours. For example, the total number of utterances spoken varies from around 30 (subjects 1 and 10) to well above 100 (subjects 8 and 9); one user let the turn pass on in almost 90% of the cases overall (subject 6) whereas another kept the floor in 75% of the cases (subject 9); and some pairs spoke an almost equal number of utterances (subjects 1+2, 3+4, 5+6, and 11+12) whereas others had a ratio of 2 to 1 or more (subjects 7+8 and 9-10). 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. Two pairs preferred to make up a few topics in advance and go through them, whereas three couples (who knew each other well) chose to treat it as unprepared small-talk. Finally the VAD is not flawless and some noise is to be expected. This is distributed evenly over the conditions for each speaker, however.

In summary, the hypothesised results were achieved for each user in the study, regardless of individual speaking style, topic of conversation, attitude towards the task (although all subjects were cooperative).

7. Conclusion

A few notes to conclude:

While the experiment described in this paper is carried out in an avatar-mediated human-human communication setting, we hold that the results are equally useful for human-computer interaction scenarios, in the design of human like spoken dialogue systems.

Evaluation of ECA systems have to a large extent been focused on macro evaluation, but there is also a need for more fine-grained micro-evaluation tools to gain insight into individual aspects of the ECA design. One example of this is the audio-visual intelligibility studies that have been conducted as part of the development of [14], providing detailed feedback on the performance of a specific function of the ECA, in this case the visible articulation. The present experiment framework could be similarly used to do micro-evaluation of turn-taking behaviour.

The simple interaction model presented is not confined to two-party dialogue but can be easily extended to handle multilogues. It is also possible to use it to provide more fine grained interaction control in spoken dialogue systems, by unobtrusively guiding the user's behaviour. This is a step towards more human-like spoken human-computer interaction, but could also be useful to improve interaction flow under adverse circumstances, such as network latency.

8. Acknowledgements

The work presented here was in part funded by the Swedish research council projects #2006-2172 (Vad gör tal till samtal/What makes speech special) and #621-2005-3488 (Modelling multi-modal communicative signal and expressive speech for embodied conversational agents) and the European Commission's Sixth Framework Program projects IP-506909 (CHIL) and IP-035147 (MonAMI).

9. References

[1] 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.

[2] Edlund, J., & Heldner, M. (2005): Exploring Prosody in Interaction Control. Phonetica, 62(2-4) , 215-226.

[3] 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.

[4] Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H. & Yan, H (2001). Human Conversation as a System Framework: Designing Embodied Conversational Agents. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors, *Embodied Conversational Agents*, pages 29–63. MIT Press, Cambridge, MA.

[5] Thórisson, K. R. (1999). A Mind Model for Multimodal Communicative Creatures and Humanoids. *International Journal of Applied Artificial Intelligence*, 13(4-5):449–486.

[6] Pelachaud, C., Badler, N. & Steedman. (1996) Generating Facial Expressions for Speech. Cognitive Science, 20(1).

[7] Beskow, J., Edlund, J., & Nordstrand, M. (2005). A model for multi-modal dialogue system output applied to an animated talking head. In Minker, W., Bühler, D., & Dybkjaer, L. (Eds.), Spoken Multimodal Human-Computer Dialogue in Mobile Environments, Text, Speech and Language Technology (pp. 93-113). Dordrecht, The Netherlands: Kluwer Academic Publishers.

[8] 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.

[9] http://www.skype.com

[10] Edlund, J., & Heldner, M. (2006): /nailon/ - software for online analysis of prosody. In Proc of Interspeech 2006 ICSLP . Pittsburgh PA, USA.

[11] Duncan, S., Fiske, D. (1977), Face-to-face interaction: Research, methods, and theory. Hillsdale: Erlbaum.

[12] Cassell, J., Torres, O., and Prevost, S. Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation, in Machine Conversations, Y. Wilks, Ed. The Hague: Kluwer, pp. 143-154, 1999.

[13] Hugot, V. (2007) Eye gaze analysis in human-human commuication, M.Sc. thesis, KTH, Stockholm

[14] Siciliano, C., Williams, G., Beskow, J., & Faulkner, A. (2003). Evaluation of a Multilingual Synthetic Talking Face as a communication Aid for the Hearing Impaired. In Proc of ICPhS, XV Intl Conference of Phonetic Sciences (pp. 131-134). Barcelona, Spain.