The Effect of Multiple Modalities on the Perception of a Listening Agent

Ronald Poppe, Mark ter Maat, and Dirk Heylen

Human Media Interaction Group, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{r.w.poppe,d.k.j.heylen}@utwente.nl

Listening agents are IVAs that display attentive listening behavior to a human speaker. The research into listening agents has mainly focused on (1) automatically timing listener responses; and (2) investigating the perceptual quality of listening behavior. Both issues have predominantly been addressed in an offline fashion, e.g. based on controlled animations that were rated by human observers. This allows for the systematic investigation of variables such as the quantity, type and timing of listening behaviors. However, there is a trade-off between the control and the realism of the stimuli. The display of head movement and facial expressions makes the animated listening behavior more realistic but hinders the investigation of specific behavior such as the timing of a backchannel.

To migitate these problems, the Switching Wizard of Oz (SWOZ) framework was introduced in [1]. In online speaker-listener dialogs, a human listener and a behavior synthesis algorithm simultaneously generate backchannel timings. The listening agent is animated based on one of the two sources, which is switched at random time intervals. Speakers are asked to press a button whenever they think the behavior is not human-like. As both human and algorithm have the same limited means of expression, these judgements can solely be based on aspects of the behavior such as the quantity and timing of backchannels. In [1], the listening agent only showed head nods. In the current experiment, we investigate the effect of adding facial expressions. Facial expressions such as smiles and frowns are known to function as backchannels as they can be regarded as a signal of understanding and attention.

Experiment Setup

We use an asymmetric version of the SWOZ setting [1]. A human speaker and listener are at different locations and their communication is mediated via the framework. They engage in a conversation where the speaker does the talking and the listener displays backchannel feedback at appropriate times. To this end, the listener is shown the video and audio of the speaker, whereas the listener is represented as a virtual listener to the human speaker. The source of the virtual listener is switched at random times. We ask the human speaker to press a button (the yuck button) everytime the behavior is perceived as unhuman-like.

We use a 2 (role) $\times$ 2 (condition) within-subjects design. In the *static* condition, the listening agent shows only nods, indicated by the human listener by pressing the space bar. Simultaneously, the algorithm from [1] generates

backchannel timings. In the face condition, facial expressions are also shown. These are always animated based on those of the human listener using [2]. From the detected activations, we animated those of the mouth and the brows.

The SWOZ framework switches after a random amount of time (between 10 and 50 seconds). The interaction starts with the human listener as the source, and when the speaker presses the yuck button, the source is always set to the human listener. Speakers were informed of this, but were unaware of the duration of the switching interval. Speakers were free to choose any topic, and were provided with some suggestions. Interactions were stopped by the experimenter after approximately five minutes.

Results and Discussion

Ten subjects (five pairs) took part in the experiment. We recorded 111.61 minutes of dialog in the 20 interactions. In 61.56% of the time, the listener agent's head nods originated from the actual listener. The yuck button was pressed 86 times, 57 times (66.28%) while the listening agent's head nods were generated by the algorithm. For the algorithm and human respectively, this amounts to 1.33 and 0.42 yucks per minute. The difference between static and face condition is small: 0.82 and 0.72 yucks per minute, respectively. A repeated measures ANOVA shows a significant effect for source ($F(1) = 5.964, p < .05$), but not for condition ($F(1) = 1.418, p = .26$) or the interaction between the two ($F(1) = 2.303, p = .16$). Closer analysis, however, reveals that the diffence between the two conditions is much larger when the listener is animated by the algoritm: 1.11 and 1.58 yucks per minute for the static and face conditions, respectively. We expect this difference is because observers are more forgiving when the timing of the backchannels is less human-like. We compare the timings of both and consider a displayed backchannel matching if it is produced within a margin of one second of a backchannel produced in the other source. Of all backchannels shown to the speaker, 19.5% of those generated by the human listener, and 10.2% of those of the algorithm match the other source. Indeed the timings of the backchannels produced by the algorithm are less appropriate.

The percentage of matching backchannels for the algorithm in the static and face condition is similar (10.2% and 10.0%). However, the number of yucks is much lower in the latter. Apparently, the additional display of facial expressions causes the speakers to reduce their yuck presses only when the timing is less accurate. This finding is important as adding more modalities might similarly bias the results when performing experiments to analyze the human perception of listening behavior. A;sp, when developing listening agents, this finding might be used to improve the human-likeliness by adding more modalities of expression.

References

1. Poppe, R., ter Maat, M., Heylen, D.: Online backchannel synthesis evaluation with the Switching Wizard of Oz. In: Joint Proceedings IVA Workshops, pp. 75–82 (2012)
2. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: Proceedings ICCV, pp. 1034–1041 (2009)