Online Behavior Evaluation with the Switching Wizard of Oz

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,m.termaat,d.k.j.heylen}@utwente.nl

1 Introduction

Advances in animation and sensor technology allow us to engage in face-to-face conversations with virtual agents [1]. One major challenge is to generate the virtual agent's appropriate, human-like behavior contingent with that of the human conversational partner. Models of (nonverbal) behavior are pre-dominantly learned from corpora of dialogs between human subjects [2], or based on simple observations from literature (e.g. [3–6]).

Humans are particularly sensitive to flaws in the displayed behavior, both in form and timing [7, 8]. This effect also occurs when certain behaviors are not animated, which is common in experimental settings where the behavior of the virtual agent is varied systematically only one or a few modalities [9, 10]. This leads to biased perceptual ratings, which hampers progress in the design and implementation of behavior synthesis algorithms.

To this end, we propose a methodology and implementation that combines ideas behind the human Turing test with those of a Wizard of Oz setup. At the heart of our approach is a distributed (video-conferencing) setting with two human conversational partners. Each of the subjects is observed with camera and microphone and algorithms are employed to analyze the verbal and nonverbal behavior in real-time (similar to e.g., [11–13]). These observations are used as input to a behavior synthesis model. Both subjects are shown a virtual representation of the other (see Fig. 2), animated based on one of two sources: (1) directly on the observed behavior of the other, or (2) on the output of the synthesis model. Both sources share the same behavior animation capabilities and limitations. We can therefore analyze the effect of the quantity, type and timing of the nonverbal behaviors on the perception thereof. During a conversation, the source of animation of the representation of each subject switches occasionally.

The idea behind the framework is that, when the displayed behavior deviates from what is typically regarded as human-like, the observer should notice. In this case, he or she is instructed to press a button (the *yuck* button [10]). The ratings can be used to evaluate and improve the behavior synthesis models (e.g. [14]). As observations of the subjects are continuously recorded, the framework doubles as a tool for study into nonverbal behavior.

[⋆] This publication was supported by the Dutch national program COMMIT.

2 Switching Wizard of Oz

In the Switching Wizard of Oz (SWOZ) setting, two human subjects A and B, seated at distributed locations, are shown virtual representations of each other. The representation of B displays either the behavior performed by B, or behavior synthesized by an algorithm, based on audio or video observations of A. The behaviors displayed by the virtual representations can be discrete (e.g. nods) or continuous (e.g. head movement). During a conversation, the source of a virtual representation is switched at random time intervals. To evaluate the quality of behavior synthesis models, both subjects are presented with a yuck button which they press whenever they believe the displayed behavior does not originate from the other subject.

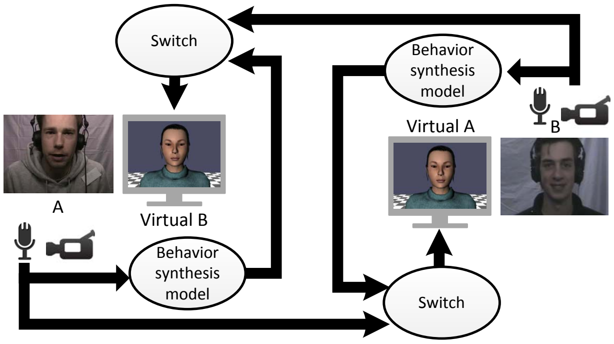

Fig. 1. Schematic representation of the Switching Wizard of Oz framework.

Subject observation The conversational partners are observed via sensors such as cameras, microphones, Kinects and gaze trackers. The observations are encoded into features in real-time. It should also be possible to regenerate the observed behavior on the virtual representation of the subject.

Behavior synthesis These extracted features are subsequently used in a behavior synthesis algorithm, to determine whether or not certain behaviors should be animated. These algorithms can be manually engineered sets of classification rules (e.g. [3, 5]) or machine learning classifiers trained on previously recorded corpus data (e.g. [6]). Based on the outcome of the algorithm or the observations of the actual conversational partner, the behavior is animated on a virtual agent. Behaviors can be verbal and nonverbal, discrete and continuous.

Behavior switching The framework switches between the two sources at random time intervals. The displayed behavior should be continuous. For discrete events, this implies that the currently animated behavior should be finished and a new behavior should not be directly animated. For continuous behaviors, it should also be ensured that the displayed behavior is continuous so the switching moment will not be perceived as such to the observer. As the switching component of the framework is presented with the behavior of both the conversational partner and the algorithm, the switching time can be selected when the two sources are more or less similar, to allow for interpolation between the two.

References

1. Heylen, D., Bevacqua, E., Pelachaud, C., Poggi, I., Gratch, J., Schröder, M.: Generating Listening Behaviour. In: Emotion-Oriented Systems Cognitive Technologies - Part 4. Springer (2011) 321–347
2. Martin, J.C., Paggio, P., Kuehnlein, P., Stiefelhagen, R., Pianesi, F.: Introduction to the special issue on multimodal corpora for modeling human multimodal behavior. Language Resources and Evaluation **42**(2) (2008) 253–264
3. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics **32**(8) (2000) 1177–1207
4. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. In: Proceedings of the Conference of the European chapter of the Association for Computational Linguistics - Volume 1, Budapest, Hungary (2003) 51–58
5. Truong, K.P., Poppe, R., Heylen, D.: A rule-based backchannel prediction model using pitch and pause information. In: Proceedings of Interspeech, Makuhari, Japan (2010) 490–493
6. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multi-Agent Systems **20**(1) (2010) 80–84
7. McDonnell, R., Ennis, C., Dobbyn, S., O'Sullivan, C.: Talking bodies: Sensitivity to desynchronization of conversations. ACM Transactions on Applied Perception **6**(4) (2009) A22
8. Hodgins, J., Jörg, S., O'Sullivan, C., Park, S.I., Mahler, M.: The saliency of anomalies in animated human characters. ACM Transactions on Applied Perception **7**(4) (2010) A22
9. Poppe, R., Truong, K.P., Reidsma, D., Heylen, D.: Backchannel strategies for artificial listeners. In: Proceedings of the International Conference on Interactive Virtual Agents (IVA), Philadelphia, PA (2010) 146–158
10. Poppe, R., Truong, K.P., Heylen, D.: Backchannels: Quantity, type and timing matters. In: Proceedings of the International Conference on Interactive Virtual Agents (IVA), Reykjavik, Iceland (2011) 228–239
11. Bailenson, J.N., Yee, N., Patel, K., Beall, A.C.: Detecting digital chameleons. Computers in Human Behavior **24**(1) (2008) 66–87
12. Edlund, J., Beskow, J.: Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. Language and Speech **52**(2–3) (2009) 351–367
13. Huang, L., Morency, L.P., Gratch, J.: Virtual rapport 2.0. In: Proceedings of the International Conference on Interactive Virtual Agents (IVA), Reykjavik, Iceland (2011) 68–79
14. de Kok, I., Poppe, R., Heylen, D.: Iterative perceptual learning for social behavior synthesis. Technical Report TR-CTIT-12-01, University of Twente (2012)