Crowd-Powered Design of Virtual Attentive Listeners

Patrik Jonell ✉, Catharine Oertel, Dimosthenis Kontogiorgos, Jonas Beskow, Joakim Gustafson

Department of Speech, Music and Hearing,
KTH Royal Institute of Technology, Stockholm
[pjjonell; catha; diko; beskow]@kth.se, jocke@speech.kth.se

Abstract. This demo presents a web-based system that generates attentive listening behaviours in a virtual agent acquired from audio-visual recordings of attitudinal feedback behaviour of crowdworkers.

1 Introduction

In the last decade, there have been increasing efforts on making robots more human-like. Most of the studies which investigated the design of virtual agents and social robots have used actors or rule-driven approaches in order to design specific social behaviour. These approaches come with certain disadvantages, however. Behaviours that are gathered in an artificial environment will remain, at least to a certain degree, artificial. Furthermore, it is hard to get a natural variation of behaviour using this approach. Various studies have therefore explored crowdsourcing techniques as an alternative or supplement to the more traditional approaches [4, 7, 2, 9, 6].

In this paper we demonstrate our approach of using crowdsourcing techniques in order to collect a wide array of attitudinal backchannel responses in identical conversational context and how we translate them in a virtual agent. Applications where this could be useful are for example virtual agents used in education, counselling and elderly care, where it is important to give the impression of the interlocutor being listened to.

2 The Crowd-Powered Design Tool

2.1 Data collection

The Crowd-Powered Design Tool was designed for rapid collection of rich multimodal data. It allows for collecting demographically varied data as the crowdsourcing platforms reach a wide audience. Additionally these platforms often provide very detailed demographic data about each participant. It also gives researchers the capabilities to control for experimental factors. The tool is a web-based application which utilises modern web technologies in order to access the participant's webcamera and microphone. No particular technical skills are required from the participants and no installation of software is required. The main features of the tool are presented below.

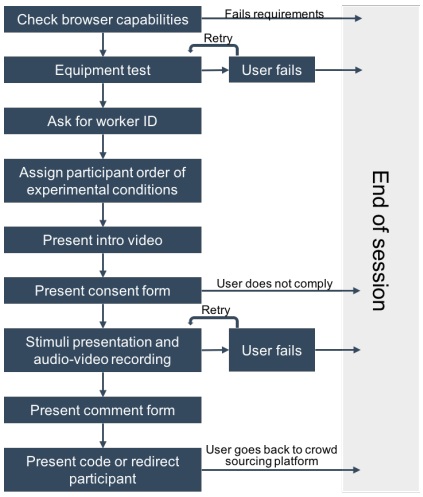

Fig. 1. Flow diagram depicting the data collection process.

User flow As can be seen in Fig. 1, the participants go through several steps during the session. If a participant does not successfully complete a step, the session ends and the data recording is dismissed.

Initial quality control Participants are asked to perform the recordings in a quiet environment, and an automatic process makes sure that the user's equipment is working, and that no background sounds are collected when the participant is silent. The system also makes sure that no cross-talk occurs by playing audio while asking the participant to be silent.

Stimuli presentation through video stream Participants are presented with stimuli which are streamed over the internet. Both pre-recorded and live audio-video stimuli are supported and can be assigned to experimental conditions.

Automatic quality control after stimuli presentation As with any crowdsourcing application, quality control is essential. In addition to the general recording environment requirements detailed above, it is necessary to ensure that the crowdworkers actually do the task as intended.

In order to be able to make an automatic quality control, we implemented the following simple but efficient procedure; (1) Record a test recording where a participant responds to each stimuli and (2) aggregate the duration of the speech. Then (3) set a generous lower and upper duration limit for each stimuli. If participants are too far outside the time-span, the participant is notified about it and given the option to repeat the recording.

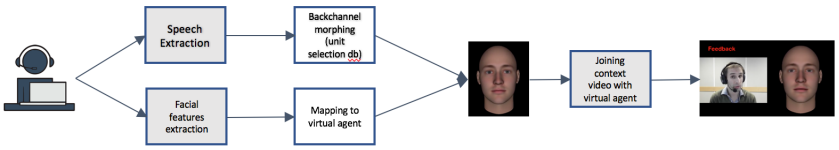

Fig. 2. Crowdworker recording being processed and transformed into a virtual agent.

2.2 Audio-Visual Processing

Audio Processing We processed the audio files in the following way. We used the ProMo (Prosody Morph) library [8] in order to morph a backchannel from our unit-selection database towards the duration and pitch of a crowdworker's backchannel. In order to do so, we first selected a backchannel token from our database by filtering on backchannels that shared the same lexical form. We then picked the one with the smallest difference in duration with the crowdworker's backchannel. Finally, we morphed the duration and then the pitch.

Video Processing In order to transform the video of the crowdworker into the face of the animated agent we did the following: we used OpenFace [1] in order to extract the visual features from the video. We extracted head pose, facial landmarks, and FACS action units [3]. We mapped these features onto a 3D-model (generated by FaceGen[1]) using morph targets for the corresponding action units. In addition, we smoothed the signal and normalised the pose of the participant; they were all normalised to face forwards towards the camera. The face was then generated through Open Scene Graph[2].

Finally, the generated face of the agent and the morphed audio were merged and either used interactively or saved to a video file. This audio-visual processing pipeline is detailed in Fig. 2

3 Studies

In an initial study [5] we wanted to investigate whether it was possible to learn lexical and prosodical backchannel generation models for different attitudes when using the approach described above. We found significant differences in the distribution of both lexical token and prosodic features in backchannels across attitudinal data. For an initial evaluation we presented crowdworkers with a dialogue in which a robot took on the role of a supportive or sceptical listener and could show that crowdworkers were able to perceive the attitudinal state of the robot with an accuracy of 63%.

[1] https://facegen.com/
[2] http://www.openscenegraph.org/

4 Conclusions and future work

Despite many advantages of using our proposed approach there are some limitations, such as not providing the researcher with full control of the environment nor the equipment being used. But as each recording is relatively cheap to perform, participants who do not meet a desired quality criteria can easily be discarded. The crowdsourcing platforms often provide good pre-screening.

Future research should investigate how the data collected through the presented tool compares to data collected in an in-lab high quality environment.

Acknowledgements

The authors feel particularly thankful to Joseph Mendelson and Todd Shore for making the data recordings possible. The authors would also like to acknowledge the support from the Swedish Research Council Project InkSynt (2013-4935), the EU Horizon 2020 project BabyRobot (687831) and the Swedish Foundation for Strategic Research project EACare (RIT15-0107).

References

1. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. pp. 1–10. IEEE (2016)
2. Breazeal, C., DePalma, N., Orkin, J., Chernova, S., Jung, M.: Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. Journal of Human-Robot Interaction 2(1), 82–111 (2013)
3. Ekman, P., Friesen, W.V.: Facial action coding system (1977)
4. Leite, I., Pereira, A., Funkhouser, A., Li, B., Lehman, J.F.: Semi-situated learning of verbal and nonverbal content for repeated human-robot interaction. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 13–20. ACM (2016)
5. Oertel, C., Jonell, P., Kontogiorgos, D., Mendelson, J., Beskow, J., Gustafson, J.: Crowd-sourced design of artificial attentive listeners. In: accepted at Interspeech 2017 (2017)
6. Oertel, C., Lopes, J., Yu, Y., Mora, K.A.F., Gustafson, J., Black, A.W., Odobez, J.M.: Towards building an attentive artificial listener: on the perception of attentiveness in audio-visual feedback tokens. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 21–28. ACM (2016)
7. Orkin, J., Roy, D.: Automatic learning and generation of social behavior from collective human gameplay. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1. pp. 385–392. International Foundation for Autonomous Agents and Multiagent Systems (2009)
8. Tim Mahrt: ProMo: The Prosody-Morphing Library. https://github.com/timmahrt/ProMo (2016), online; accessed 15 May 2017
9. Yu, Z., Xu, Z., Black, A., Rudnicky, A.: Chatbot evaluation and database expansion via crowdsourcing. In: Proc. of the chatbot workshop of LREC. 2016 (2016)