Using Crowd-Sourcing for the Design of Listening Agents: Challenges and Opportunities

Catharine Oertel
KTH Royal Institute of Technology
Stockholm, Sweden
catha@kth.se

Patrik Jonell
KTH Royal Institute of Technology
Stockholm, Sweden
pjjonell@kth.se

Kevin El Haddad
University of Mons
Mons, Belgium
kevin.elhaddad@umons.ac.be

Eva Szekely
KTH Royal Institute of Technology
Stockholm, Sweden
szekely@kth.se

Joakim Gustafson
KTH Royal Institute of Technology
Stockholm, Sweden
jocke@speech.kth.se

ABSTRACT

In this paper we are describing how audio-visual corpora recordings using crowd-sourcing techniques can be used for the audio-visual synthesis of attitudinal non-verbal feedback expressions for virtual agents. We are discussing the limitations of this approach as well as where we see the opportunities for this technology.

CCS CONCEPTS

• **Human-centered computing** → *Social tagging systems*; *Empirical studies in collaborative and social computing*;

KEYWORDS

artificial listener, listening agent, multimodal behaviour generation

ACM Reference Format:
Catharine Oertel, Patrik Jonell, Kevin El Haddad, Eva Szekely, and Joakim Gustafson. 2017. Using Crowd-Sourcing for the Design of Listening Agents: Challenges and Opportunities. In *Proceedings of 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents (ISIAA'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3139491.3139499

1 INTRODUCTION

Conversational systems are becoming more and more a part of our everyday life. Most of these systems are still restricted to short interactions, and often very specific tasks. Having a conversational system however, which is able to engage in a social interaction with a human, over an extended period of time could however open the door to many new applications. For instance, in the educational sector such an application could be the development of a study peer. In health care, this could be a virtual therapy aider. For all of these applications it is very important that the virtual agent can portray well the role of a listener.

A very important part of being a listener is to produce appropriate feedback at the right point in time.

ISIAA'17, November 13, 2017, Glasgow, UK.

ACM ISBN 978-1-4503-5558-2/17/11…$15.00
https://doi.org/10.1145/3139491.3139499

A lot of studies are concerned with the development of listening agents. However, most of the work until now has focused on the timing of feedback expression [6, 14, 15].

In the current paper however, we are going to discuss how audio-visual data collection using crowd-sourcing can inform the paralinguistic realisation/synthesis of feedback token for listening agents.

2 BACKGROUND

Most current virtual agents use a limited set of feedback expressions. This is on the one had due to a lack of control over the expression parameters, but on the other hand also due to restrictions to datasets which are recorded in the lab and therefore can only to a certain degree model the variability present in real-life conversations.

For example, most current synthesisers are optimised for speech, but not for non-verbal vocalisations such as backchannels.

Synthesisers which focused on synthesis of non-verbal are realtively rare exceptions are for example [3, 4, 7, 10, 13]. But also these synthesisers are limited to datasets which were recorded in the lab.

3 NEED FOR CROWD-SOURCED DATA

Generally, while there are many multi-modal corpora available - most of these corpora are created with very specific purposes in mind. Their content is limited by the physical constraints in which they were collected (number of participants, their background and gender, recording equipment, etc...). Collecting adequate data of a desired amount and for a specific task can therefore be difficult. Indeed the subjects' reactions are unpredictible and variable from one speaker to another. Specifically collecting conversational data which are necessary to build listening agents means dealing with expressions that are cultural dependent and even subject specific. Therefore a good compromise would be to be able to have a large amount of speakers to statistically have more chances of obtaining the required set of data for a specific task while controlling as many factors as possible. One way to do this is by using crowd-sourcing. We used this method to collect audio-visual non-verbal expressions. We provided crowd-workers with a video of a speaker and situated them in the context of a job-interviews. We asked them to provide feedback to the speaker using expressions such as e.g. "mhm", "mh" etc. We asked participants to provide the feedback whenever they saw the word "feedback" being realised on the screen. In order to

restrict the cultural background of participants, we only sampled from the US and Canada. To control for the paralinguistic stance we asked participants to respond in one of three ways ("supportive", "sceptical" or neutral). We then mapped the collected feedback expressions to a unit-selection data base and evaluated them by synthesizing them in a "Furhat" Robot and carrying out a perception experiment [8]. The multi-modal corpora collection tool which was used in this experiment is described in [12], demoed at IVA 2017 [11].

4 DISCUSSION

We are of the opinion that data collection using crowd-sourcing techniques can help us to obtain a larger and more varied database of non-verbal feedback expressions. In [8] we could show that it is possible to carry out audio- visual data collection using crowd-sourcing techniques while still controlling for the conversational context.

Moreover, the here proposed crowed-sourcing technique could help to obtain sufficient data that could at least allow to initiate investigations into feature learning for nonverbal feedback expressions.

One disadvantage of using crowd-sourcing techniques for the collection of audio-visual non-verbal feedback expressions is of course the quality of the audio data and to a certain degree also the video data.

For the creation of a synthesis voice it is very important to record in a sound studio with explicit control on the equipment. In our experience with the audio-data collected from crowd-sourcing [8] it was only possible to map to an already existing unit-selection database. It was however not possible to create a new voice This problem could be tackled by re-enacting the collected expressions in a recording studio or use signal processing and vocoding techniques to reproduce them through temporal and prosodic modification of the signals such as denoising autoencoders [5] and adaptation algorithms (CMLLR) [2, 16].

Data collection using crowed-sourcing techniques is also limited in terms of the quality of the visual data which can be obtained. Indeed the type and quality of the camera used by the crowd-workers can only be controlled in a limited way (constraining the participation to crowd-workers which are using equipment with certain technical specifications). However, the presence of robust facial detection and landmark estimation systems such as for example OpenFace [1] allows for the use of a wider range of recording cameras.

5 CONCLUSION

In future work we are planing to explore how we can enrich previously explored synthesis techniques of attitudinal head nods [9], backchannels [7] as well as smiles and laughs [3] with the here described crowd-sourced data collection techniques.

ACKNOWLEDGMENTS

The authors would 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] T. BaltruÅąaitis, P. Robinson, and L. P. Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–10. DOI: http://dx.doi.org/10.1109/WACV.2016.7477553

[2] Vassilios V Digalakis, Dimitry Rtischev, and Leonardo G Neumeyer. 1995. Speaker adaptation using constrained estimation of Gaussian mixtures. *Speech and Audio Processing, IEEE Transactions on* 3, 5 (1995), 357–366.

[3] Kevin El Haddad, Hüseyin Çakmak, Emer Gilmartin, Stéphane Dupont, and Thierry Dutoit. 2016. Towards a Listening Agent: A System Generating Audiovisual Laughs and Smiles to Show Interest. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. ACM, New York, NY, USA, 248–255. DOI: http://dx.doi.org/10.1145/2993148.2993182

[4] K. E. Haddad, S. Dupont, J. Urbain, and T. Dutoit. 2015. Speech-laughs: An HMM-based approach for amused speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4939–4943. DOI: http://dx.doi.org/10.1109/ICASSP.2015.7178910

[5] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. 2013. Speech enhancement based on deep denoising autoencoder.. In *Interspeech*. 436–440.

[6] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20, 1 (2010), 70–84.

[7] Catharine Oertel, Joakim Gustafson, and Alan W. Black. 2016. On Data Driven Parametric Backchannel Synthesis for Expressing Attentiveness in Conversational Agents. In *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI '16)*. ACM, New York, NY, USA, 43–47. DOI: http://dx.doi.org/10.1145/3011263.3011272

[8] Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, Joseph Mendelson, Jonas Beskow, and Joakim Gustafson. 2017. Crowd-Sourced Design of Artificial Attentive Listeners. In *accepted at Interspeech 2017*.

[9] Catharine Oertel, José Lopes, Yu Yu, Kenneth A Funes Mora, Joakim Gustafson, Alan W Black, and Jean-Marc Odobez. 2016. 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*. ACM, 21–28.

[10] Sathish Pammi, Marc Schröder, Marcela Charfuelan, Oytun Türk, and Ingmar Steiner. 2010. Synthesis of listener vocalisations with imposed intonation contours.. In *SSW*. 240–245.

[11] Jonell Patrik, Catharine Oertel, Dimosthenis Kontogiorgos, Jonas Beskow, and Joakim Gustafson. 2017. Crowd-Powered Design of Virtual Attentive Listeners. In *accepted at IVA 2017: The 17th International Conference on Intelligent Virtual Agents*.

[12] Jonell Patrik, Catharine Oertel, Dimosthenis Kontogiorgos, Jonas Beskow, and Joakim Gustafson. 2017. Crowdsourced Multimodal Corpora Collection Tool. In *submitted to MMC 2017: The 12th Workshop on Multimodal Corpora*.

[13] Thorsten Stocksmeier, Stefan Kopp, and Dafydd Gibbon. 2007. Synthesis of prosodic attitudinal variants in German backchannel ja.. In *INTERSPEECH*. 1290–1293.

[14] Khiet P. Truong, Ronald Poppe, Iwan De Kok, and Dirk Heylen. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. *Proc. of Interspeech* (2011), 2973–2976.

[15] Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of pragmatics* 32, 8 (2000), 1177–1207.

[16] Heiga Zen, Keiichi Tokuda, and Alan W Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51, 11 (2009), 1039–1064.