Crowdsourced Multimodal Corpora Collection Tool

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

KTH Royal Institute of Technology, Sweden
Lindstedsvägen 24, 114 28 Stockholm, Sweden
{pjjonell, catha, diko, beskow}@kth.se; jocke@speech.kth.se

Abstract

In recent years, more and more multimodal corpora have been created. To our knowledge there is no publicly available tool which allows for acquiring controlled multimodal data of people in a rapid and scalable fashion. We therefore are proposing (1) a novel tool which will enable researchers to rapidly gather large amounts of multimodal data spanning a wide demographic range, and (2) an example of how we used this tool for corpus collection of our "Attentive listener" multimodal corpus. The code is released under an Apache License 2.0 and available as an open-source repository, which can be found at `https://github.com/kth-social-robotics/multimodal-crowdsourcing-tool`. This tool will allow researchers to set-up their own multimodal data collection system quickly and create their own multimodal corpora. Finally, this paper provides a discussion about the advantages and disadvantages with a crowd-sourced data collection tool, especially in comparison to a lab recorded corpora.

Keywords: crowdsourcing, human-computer interaction, multimodal corpus

1. Introduction

In the last decade more and more efforts have been carried out to create multimodal corpora. Efforts have come from diverse fields such as psychology, linguistics and computer science. They shared the common goal of enabling researchers to gain a better understanding of how humans converse with one another.

The traditional approach has been to record corpora in a lab environment. This approach has several disadvantages, however. One such disadvantage is that due to the constraint of having to attend the recording on campus, most resulting corpora are quite homogeneous with regards to participants' demographics (e.g. education level, socio-economic background, native language). Participants might also exhibit some degree of the Hawthorne effect, while being observed in sterile lab environments, leading to biased behaviour (Parsons, 1974). A further disadvantage is the monetary cost of staff, equipment and the time spent when doing in-lab recordings, for example setting up equipment.

Due to the constraints mentioned above more and more researchers have been turning to crowdsourcing as a means of gathering large quantities of data. In particular, crowdsourcing has been extensively used for annotations as well as transcriptions (Gruenstein et al., 2009; Hipp et al., 2013; Rashtchian et al., 2010). While it has, to some degree, been used for creation of conversational corpora as well, these corpora have mainly been restricted to speech and text (Filatova, 2012; Orkin and Roy, 2009; Breazeal et al., 2013). One of the reasons for the lack of these kinds of corpora might be that there are few tools which facilitate the acquisition of such data. While there are tools for collecting conversational data such as Voxforge[1], there are currently, to our knowledge, no publicly available open-source tools that allow for crowdsourcing multimodal corpora while simultaneously enabling researchers to control for experimental factors. We therefore propose and release an open-source tool that facilitates large-scale collection of multimodal corpora. Figure 1 depicts how the tool is used to collect a multimodal corpus using crowdworkers from different crowdsourcing platforms. In this paper, we will first provide an overview of the capabilities the tool provides and will then provide a summary of an example corpus which was collected using this data collection tool. Finally, we discuss the advantages and disadvantages when using a crowd-sourced approach and the limitations of the system.

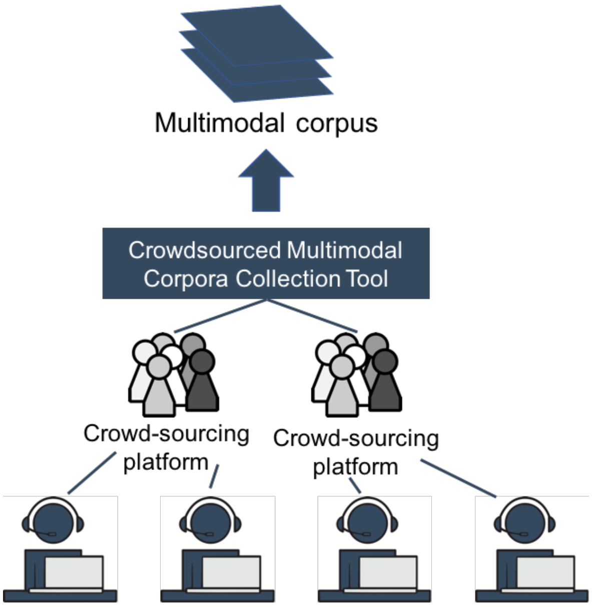

Figure 1: Corpus creation using multiple crowdsourcing services where participants record themselves.

[1] `http://www.voxforge.org/`

2. Background

2.1. Lab Recorded Multimodal Corpora

More and more corpora have been collected using a multitude of sensors. Both multi-party as well as dyadic corpora have been collected. Examples of dyadic corpora are: (van Son et al., 2008; Edlund et al., 2010) and examples of multi-party corpora are: (Carletta, 2007; Mostefa et al., 2007; Hung and Chittaranjan, 2010; Oertel et al., 2014).

An example of a corpus which tries to escape the lab setting is the D64 corpus (Oertel et al., 2013). It spans approximately two days of interactions. Four to five participants met in an apartment in Dublin which was equipped with a variety of sensors. They discussed any topic which sprang to mind during the two days.

In addition to human-human interaction corpora, there are also corpora which focus on human-robot interaction. An example of such a corpus is, for instance, the Tutorbot corpus (Koutsombogera et al., 2014). The aim of this corpus was to gather data, in order to teach a robot how to take on the role of a tutor in a multi-party tutoring scenario. In a similar fashion the data described in (Vollmer et al., 2014) was also gathered in order to understand tutoring scenarios. In this instance however, the emphasis was on understanding how a robot could communicate to a human how it wants to be taught. This approach is in some ways similar to the approach described in the "Attentive listener" scenario where we also provide a situational context in order for the robot to optimally learn certain skills.

Similarly to some of the recordings in (Carletta, 2007), we chose for the "Attentive listener" example corpus (Oertel et al., 2017) to define a scenario and let the participant play different roles when interacting with a fictive job applicant. As opposed to most previous corpus collection, e.g. (van Son et al., 2008; Oertel et al., 2013; Oertel et al., 2014), where long interactions with few participants were recorded, we gathered short interactions with a large amount of participants enabling to better capture the variety within human behaviour.

2.2. Crowdsourced Corpora

Crowdsourcing has been used in various ways for corpora creation. It has, for example, previously been used in the domain of automatic generation of narratives (Li et al., 2013; Leite et al., 2016). Leite et al. used crowdsourcing as a dialogue creation tool for (repeated) human-robot interaction. This was done in order to increase variation in a robot's dialogue responses.

There has been work on collection of text-based corpora, for example Filatova investigated irony and sarcasm by creating a corpus based on Amazon[2] reviews (Filatova, 2012). Another example is "sketchy" [3] where the authors had crowd workers to produce sketches using their computer.

Examples of speech corpora which have been created using crowdsourcing are (Lane et al., 2010), where the authors created a corpus for the purpose of improving speech recognition or (Gruenstein et al., 2009) where a speech corpus of orthographical transcriptions was created through the means of an online educational game.

Crowdsourcing for dialogue generation has been pioneered by (Orkin and Roy, 2009). Similarly, (Breazeal et al., 2013) proposed a data-driven approach to dialogue generation for a social robot by crowdsourcing dialogue and action data from an online multi-player game. These studies, however, have taken place in virtual environments and primarily focus on immediate interactions.

Like Filatova, we focused on the modelling of a paralinguistic phenomenon but in contrast to their work we focused on the modelling of scepticism vs. support instead of irony and sarcasm (Filatova, 2012). Moreover, our corpus includes additional modalities such as speech and video. While there have been other speech corpora created such as (Lane et al., 2010), or (Gruenstein et al., 2009), we are not aware of any other crowdsourced corpora which include both audio and video modalities.

2.2.1. Quality Control

While the implementation of quality control mechanisms varies from study to study, the following section provides an overview of approaches which have been used in order to ensure sufficiently high quality.

One possibility is the use of a second batch of crowdworkers to rate the work of the initial batch of crowdworkers. Mechanisms such as majority voting as well as inter-rater reliability can be used in order to assess which data to keep and which data to discard (Filatova, 2012).

Leite et al. also used a second batch of crowdworkers to rate the first batch of crowdworkers' performance. In addition, they included a test-item. If crowdworkers failed to correctly tag their data, it would be excluded and a new crowdworker would be recruited instead. They also set thresholds for the amount of agreement they expect crowdworkers to have when labelling a specific item (Leite et al., 2016).

It might, due to privacy concerns, not be possible to let a second batch of crowdworkers rate videos of the first batch, as has been done with text (Filatova, 2012; Leite et al., 2016). It might however be possible to map the participants' faces to a virtual agent, similar to (Edlund and Beskow, 2009; Jonell et al., 2017) and then let the crowdworkers rate the other participants indirectly using majority voting.

3. The tool

The tool was designed for rapid collection of rich multimodal data. It also provides researchers with 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 special technical skills are required from the participants as no installation on the participant's computer is required. As can be seen in Figure 2, 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. The following sections describe the key components of the tool in more detail.

[2] http://www.amazon.com/

[3] http://sketchy.eye.gatech.edu/

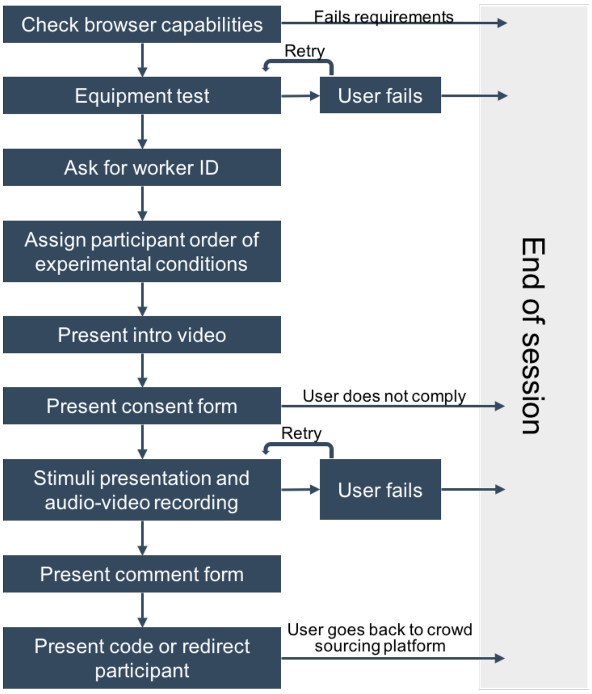
Check browser capabilities
Fails requirements
Retry
Equipment test
User fails
Ask for worker ID
Assign participant order of experimental conditions
Present intro video
Present consent form
User does not comply
Retry
Stimuli presentation and audio-video recording
User fails
Present comment form
Present code or redirect participant
User goes back to crowd sourcing platform
End of session

Figure 2: Flow chart depicting the user flow for a participant.

3.1. Data and Synchronisation

The data being collected using the tool is synchronised audio and video data. It is retrieved from the participants' microphone and webcamera while they are being exposed to stimuli. The participants are presented with stimuli which are streamed over the internet in the form of video or audio clips presented on HTML pages. The recorded data is also synchronised with the playback of the stimuli material so that it is possible to analyse the collected data in relation to the stimuli. The data is stored using common file formats for video files such as mp4 or webm. The tool does not handle retrieval of the data, as this can easily be done through other tools, depending on which storage solution was implemented (see section 4.2.). For analysis of the data it is essential that synchronisation between the crowdworker's video stream and the stimulis's video-stream is guaranteed. We used insertion of black frames into the video at given moments in order to synchronize the stimuli and the recordings. If for technical reasons the crowdworker experiences a lag in the video streaming, the same method is applied. Black frames are inserted up until the point in time of when the video is started again.

3.2. Assignment of experimental condition

To guarantee that an equal amount of videos are recorded for each experimental condition, an automatic balancer has been implemented. This balancer assigns a trial order to each new participant at the start of the session. If a participant timed out, i.e. took longer than 30 minutes, their data was removed so that another user could be assigned that experiment condition order.

3.3. Quality control

In order to ensure that the resulting corpus is of high recording quality and subsequent multimodal analysis is possible, an initial quality control test is performed to ensure that the following prerequisites are fulfilled.

First, the user needs to be using a modern web browser in order for the tool to work. If the tool detects that the user is using an outdated browser, they are notified about this and the session is ended. Second, it is important for the prosodic analysis to ensure that data is recorded in a quiet environment. To ensure this, participants are asked to speak at predefined moments and to be quiet at others. The noise-to-speech ratio is then calculated. If participants do not meet the requirements, i.e. the noise-to-speech ratio is too high, the users are informed about the problem and given the possibility to correct their set-up and redo the test. In order to proceed they have to pass the quality control test.

Finally, it is important that no cross-talk occurs. By cross-talk we refer to audio from the stimuli being picked up by the microphone of the participant. In order to ensure that this does not happen, participants are asked to use a headset with a close-talking microphone. A similar procedure as described above for detecting a suitable recording environment is used to ensure that no cross-talk occurs.

In order to achieve an automatic quality control for speech recordings after each recording, 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.

4. Technical overview

For detailed instructions on how to set up and configure the tool, refer to the projects README file which can be found in the git repository. In order to set up a data collection, stimuli material must be prepared and HTML pages created. The tool does not provide any user interface to set this up, but there are many tools available to create such HTML pages. There are some examples of HTML pages used for the “Attentive listener” scenario in the code repository.

4.1. Web Application

The web application's backend was written in Python, and was responsible for serving the HTML code to the participants. It was also designed to handle video uploads, data storage and database operations.

Its frontend was written with modern standardized web technologies. Using HTML5 the tool is able to access the participant's peripherals such as their webcamera or microphone. These features are enabled in most modern web browsers.

4.2. Storage

As the file sizes of video material can become large, the application has the capability to use an external storage service. In the example of the “Attentive listener” corpus,

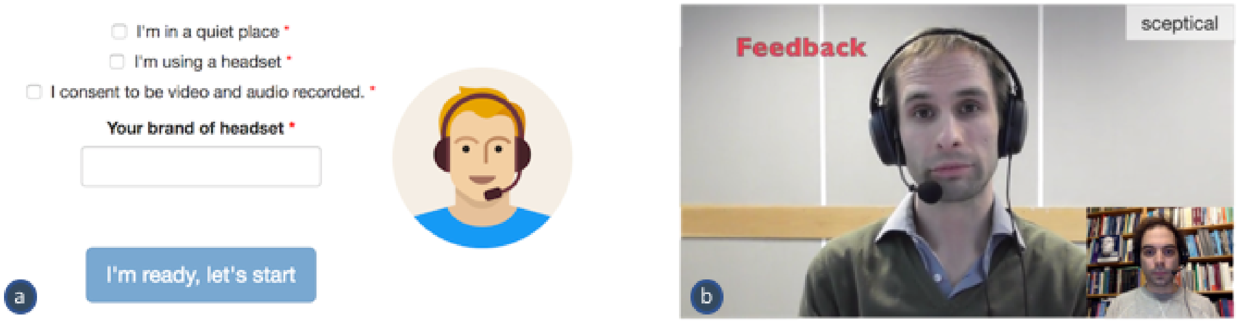

Figure 3: Interface elements from the tool; a) consent form that the user is presented with and b) stimuli presentation and participant recording from the “Attentive listener” scenario.

Amazon S3 was used, as it provides a scalable storage solution for large sized data. However, the tool can be integrated with any storage service or data can directly be stored on the local file system on the server. The files can be stored as either webm or mp4 files. The tool only facilitates saving the data at the given storage location, but does not retrieve it, as this can easily be done using other tools depending on the chosen storage solution.

4.3. Database

For the database a PostgreSQL server was used. The database keeps records of each participant's progress and meta data concerning their trials. The database stores the participants crowdsourcing ID, which is the ID given from the crowdsourcing platform. This ID is used in order to link the participant back to the crowdsourcing platform for reimbursement purposes, but also to ensure avoiding duplicate participant recordings. Each time the participant viewed or uploaded a new video to the server, the database was updated accordingly. The important information that the database stores about the user is the participant's ID from the crowdsourcing platform, the order of the trial conditions, participant's final comment, current progress of the participant and finally subjective quality measure of each participant's videos.

5. Attentive listener use-case scenario

In this section we would like to briefly discuss a corpus creation using the data collection tool described in this paper.

Feedback generation is an important component of human-human communication. Humans can choose to signal support, understanding, agreement or also scepticism by means of feedback tokens. In order to also make human-agent and human-robot robot interactions smoother, many projects and studies have focused on modelling feedback generation, e.g. (Douglas-Cowie et al., 2008; Schroder et al., 2012; Park et al., 2017). Many studies in particular have focused on the timing of feedback behaviours (Morency et al., 2010; Ward and Tsukahara, 2000; Cathcart et al., 2003). To our knowledge, very few studies have kept the timing constant (Oertel et al., 2016a; Oertel et al., 2016b) and instead focused on the variation of lexical form and prosody of feedback tokens within identical contexts.

For our “Attentive listener” corpus, crowdsourced participant's feedback behaviour is captured in identical interactional contexts in order to model a virtual agent that is able to provide feedback as an attentive/supportive as well as attentive/sceptical listener. We chose a scenario where a confederate was asked to pretend to apply for several jobs. We instructed the crowdworkers to interview our confederate by means of a video conference call Figure 3b. We asked them specifically to give short feedback utterances at predefined moments. These moments were indicated through a countdown visualised in the right hand side of the video. The resulting models were realised in an artificial agent which was evaluated by third-party observers.

We recorded 92 participants, with 3 recordings per participants, given the conditions supportive, sceptical and neutral in a random order. This resulted in a total of 276 recordings. The recordings took place during 3 consecutive days and we chose to only recruit participants with English as their native language living in the United States or Canada. This corpus is not publicly available but is described in more detail in (Oertel et al., 2017).

6. Discussion

As previously mentioned the proposed tool can help to collect large data amounts in a short amount of time. This can be used in many applications such as for example the above mentioned corpus collection. It was also successfully used to automatically generate virtual agents (Jonell et al., 2017) from the recorded data. Further it can be used to analyse human behaviour or to automatically learn human behaviour in given situations by applying machine learning algorithms on the data.

In addition to obvious factors of time and scalability, capturing high degrees of variation in behaviour was also an important factor when deciding to design the audio-visual crowd-sourced data collection tool.

Because crowdsourcing platforms were used in order to recruit participants, it's possible to reach a wide variety of participants. This enables taking a wider range of people into account when building new multimodal systems. It is also possible to recruit participant's with very specific requirements, e.g., a certain language as native language.

6.1. Quality of task performance

In addition to the quality of the signal the quality of task performance is also essential for determining the overall quality of the audio-visual corpus. For instance, did the crowd-workers complete the task as intended or was there confusion? Were there instances where the crowd-workers lacked motivation for completing the task? We encountered instance where some crowd-workers understood providing feedback not as saying "yeah", "okay"- as illustrated in the instructions but as providing more of a general kind of feedback directed towards the applicant such as: “You should be more confident about yourself”. In other cases the crowd-worker remained still during the whole duration of the task and did not provide any feedback. These examples were discarded from the general corpus.

6.2. Limitations

Despite many advantages of using a crowdsourced multimodal corpora creation over an in-lab data collection there are also limitations and shortcomings. One limitation is that the approach is restricted to dyadic corpus creation. This implies that the study of multi-party conversational phenomena is not possible within the current setup. However, while we are currently using crowdsourcing platforms that are mostly used by individual workers, it would be possible to, for example, email known participants in pairs and ask them to sign in to the same session or to sit together in front of one computer.

Another disadvantage of our approach is that it does not provide the researcher with full control of the environment nor participants’ equipment. One way to handle differences in participants’ equipment is to let the participants list the equipment they are using and for example only use participants with the same or similar equipment. The downside with this approach is that it will heavily limit the potential participants. Furthermore, when recordings were bad due to faulty or low-quality equipment, we discarded those participants as it was both easy and cheap to recruit new participants instead.

Also the authenticity of our approach is debatable with regards to three points. First of all, the authenticity of the stimuli provided. If the stimuli which are presented to crowd-workers are based on acted scenarios, the reactions of crowd-workers might also not be as they would be in a completely spontaneous co-located interactions. However, in a completely spontaneous co-located interaction, participants might also be influenced by other events or actions going on at the same time. Therefore, the resulting actions might also be not optimal for learning. Second of all, the authenticity of participants reactions. We are asking participants to take on a stance and provide feedback in a specific way e.g. “supportive", “sceptical" etc. While it is true that these reactions might not be completely genuine, it is also true that as humans we often masking our true emotions and attitudes. However, it cannot be denied that some people might be better skilled in acting in such a scenario than other people. Yet, in the case of backchannels we found that people could generally convey the conversational function very well.

And finally, being co-located during an interaction or reacting to a video might influence crowd-workers behaviours. While being co-located of course makes a difference with regards to conversational behaviours, it is also true that we are more and more used to video-mediated interactions. Therefore, we believe that our approach while maybe not as ecological valid as other co-located, spontaneous conversation approaches provides a good comprise between, scalability, control and authenticity.

The last limitation discussed in this paper, is that it is not trivial to capture participants’ gaze direction. This is a similar limitation to any conference call. This is partly due to the software currently available for gaze estimation, but also for other unknown factors such as screen size, distance to screen and so on. This could, partly be helped by an initial gaze calibration process.

6.3. Ethical and legal concerns

As the participants are being recorded on both video and audio this gives rise to both ethical, privacy and legal concerns. Therefore, it is important to make the participants aware of the fact that their data is being recorded and how this data will be handled. The participants should actively agree to being recorded, and any local laws regarding data collection and storage has to be obeyed. This paper presents a tool for data collection, but it is the responsibility of the person using the tool to make sure both the ethical and legal concerns are being properly addressed.

7. Conclusions

In this paper, we have proposed a novel tool that allows for fast, scalable, demographically varied and quality controlled multimodal corpora creation. The tool is available for download as an open source project at: `https://github.com/kth-social-robotics/multimodal-crowdsourcing-tool` under Apache License 2.0. We also gave an example of how we successfully employed it to collect our “Attentive listener” corpus. We are hoping that by releasing this tool, users who find it useful will contribute to the code base.

7.1. Future work

Improvements can still be made regarding the tool, such as in the area of quality control and video streaming features. While we tested the usefulness of the tool in the use-case scenario described in this paper, we did not carry out a formal usability study. The conduction of such a study will be part of future work. Additional modalities being captured in form of, for example, capturing keyboard and mouse activity or eye gaze in combination with the audio-visual data could also be highly interesting extensions.

8. Acknowledgements

The authors feel particularly thankful to Joseph Mendelson and Todd Shore for making the data recordings possible and Sarah Ramsay for reviewing the paper. The authors would also like to acknowledge the support from the Swedish Foundation for Strategic Research project EACare (RIT15-0107), the Swedish Research Council Project InkSynt (2013-4935), the EU Horizon 2020 project BabyRobot (687831) and the Swedish Foundation for Strategic Research project FACT (GMT14-0082).

9. References

Breazeal, C., DePalma, N., Orkin, J., Chernova, S., and Jung, M. (2013). Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1):82–111.

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Cathcart, N., Carletta, J., and Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 51–58. Association for Computational Linguistics.

Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., and Heylen, D. (2008). The sensitive artificial listner: an induction technique for generating emotionally coloured conversation.

Edlund, J. and Beskow, J. (2009). Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. *Language and Speech*, 52(2-3):351–367.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., and House, D. (2010). Spontal: A swedish spontaneous dialogue corpus of audio, video and motion capture. In *LREC*.

Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. pages 392–398. Citeseer.

Gruenstein, A., McGraw, I., and Sutherland, A. M. (2009). A self-transcribing speech corpus: collecting continuous speech with an online educational game. pages 109–112.

Hipp, J. A., Adlakha, D., Gernes, R., Kargol, A., and Pless, R. (2013). Do you see what i see: Crowdsource annotation of captured scenes. In *Proceedings of the 4th International SenseCam &#38; Pervasive Imaging Conference*, SenseCam '13, pages 24–25, New York, NY, USA. ACM.

Hung, H. and Chittaranjan, G. (2010). The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 879–882. ACM.

Jonell, P., Oertel, C., Kontogiorgos, D., Beskow, J., and Gustafson, J. (2017). Crowd-powered design of virtual attentive listeners. In *International Conference on Intelligent Virtual Agents*, pages 188–191. Springer, Cham.

Koutsombogera, M., Al Moubayed, S., Bollepalli, B., Abdelaziz, A. H., Johansson, M., Lopes, J. D. A., Novikova, J., Oertel, C., Stefanov, K., and Varol, G. (2014). The tutorbot corpusâĂŢa corpus for studying tutoring behaviour in multiparty face-to-face spoken dialogue. In *LREC*, pages 4196–4201.

Lane, I., Waibel, A., Eck, M., and Rottmann, K. (2010). Tools for collecting speech corpora via mechanical-turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 184–187, Stroudsburg, PA, USA. Association for Computational Linguistics.

Leite, I., Pereira, A., Funkhouser, A., Li, B., and Lehman, J. F. (2016). Semi-situated learning of verbal and nonverbal content for repeated human-robot interaction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 13–20. ACM.

Li, B., Lee-Urban, S., Johnston, G., and Riedl, M. O. (2013). Story generation with crowdsourced plot graphs. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pages 598–604. AAAI Press.

Morency, L.-P., de Kok, I., and Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84.

Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S. M., Tyagi, A., Casas, J. R., Turmo, J., Cristoforetti, L., Tobia, F., et al. (2007). The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3-4):389–407.

Oertel, C., Cummins, F., Edlund, J., Wagner, P., and Campbell, N. (2013). D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1-2):19–28.

Oertel, C., Funes Mora, K. A., Sheikhi, S., Odobez, J.-M., and Gustafson, J. (2014). Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 27–32. ACM.

Oertel, C., Gustafson, J., and Black, A. W. (2016a). Towards building an attentive artificial listener: On the perception of attentiveness in feedback utterances.

Oertel, C., Lopes, J., Yu, Y., Mora, K. A. F., Gustafson, J., Black, A. W., and Odobez, J.-M. (2016b). 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*, pages 21–28. ACM.

Oertel, C., Jonell, P., Kontogiorgos, D., Mendelson, J., Beskow, J., and Gustafson, J. (2017). Crowdsourced design of artificial attentive listeners. In *INTERSPEECH: Situated Interaction, Augusti 20-24 Augusti, 2017*.

Orkin, J. and Roy, D. (2009). 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*, pages 385–392. International Foundation for Autonomous Agents and Multiagent Systems.

Park, H. W., Gelsomini, M., Lee, J. J., and Breazeal, C. (2017). Telling stories to robots: The effect of backchanneling on a childâĂŹs storytelling. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 100–108. ACM.

Parsons, H. M. (1974). What happened at hawthorne? *Science*, 183(4128):922–932.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier,

J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Ter Maat, M., McKeown, G., Pammi, S., Pantic, M., et al. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183.

van Son, R., Wesseling, W., Sanders, E., van den Heuvel, H., et al. (2008). The ifadv corpus: a free dialog video corpus. In *LREC*, pages 501–508.

Vollmer, A.-L., Mühlig, M., Steil, J. J., Pitsch, K., Fritsch, J., Rohlfing, K. J., and Wrede, B. (2014). Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning. *PloS one*, 9(3):e91349.

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.