Rushing to Judgement: How Do Laypeople Rate Caller Engagement in Thin-Slice Videos of Human–Machine Dialog?

Vikram Ramanarayanan[†]*, Chee Wee Leong*[‡] *& David Suendermann-Oeft*[†]

Educational Testing Service R&D
[†]90 New Montgomery Street, #1500, San Francisco, CA
[‡]660 Rosedale Rd., Princeton, NJ

<vramanarayanan,cleong,suendermann-oeft>@ets.org

Abstract

We analyze the efficacy of a small crowd of naïve human raters in rating engagement during human–machine dialog interactions. Each rater viewed multiple 10 second, thin-slice videos of non-native English speakers interacting with a computer-assisted language learning (CALL) system and rated how engaged and disengaged those callers were while interacting with the automated agent. We observe how the crowd's ratings compared to callers' self ratings of engagement, and further study how the distribution of these rating assignments vary as a function of whether the automated system or the caller was speaking. Finally, we discuss the potential applications and pitfalls of such a crowdsourced paradigm in designing, developing and analyzing engagement-aware dialog systems.

Index Terms: engagement, human-computer interaction, dialog systems, computer-assisted language learning, crowdsourcing.

1. Introduction

The increasing multimodality of human–computer interaction technologies affords researchers and developers more opportunities to improve the efficacy of the interaction and overall user experience. An important aspect of this process involves the measurement, tracking and maintenance of user engagement over the course of the interaction.

Psychologists have long been studying how well laypeople rate different aspects of human behavior by only viewing short durations, or *thin slices*, of video, as opposed to the entire video file, which can be tedious and time-consuming in comparison. Ambady and Rosenthal conducted a seminal study where they asked complete strangers to first view 2, 5 and 10 second silent segments of university teachers' classroom lectures, and then rate their non-verbal behavior [1]. They found that these naïve ratings predicted expert-rated[1] gold-standard ones of the same behaviors with surprising accuracy. Multiple research studies have since replicated the efficacy and sufficiency of such a thin-slice approach in a variety of application domains, including the judgement of conversational dynamics during negotiations [2], analysis of medical dialog [3], evaluation of sales effectiveness [4], assessment of socioeconomic status [5], assessment of personality traits [6, 7], and even detection of psychopathy [8], among others. Much progress has also been made in using the thin-slice approach for automated feature extraction and machine learning. For instance, Nyugen and Gatica-Perez showed that extracting audiovisual, dyadic and non-verbal feature cues from thin slices of real job interviews were predictive of hirability impressions of those employment applicants [9].

Recent research in the literature has extensively analyzed the role of engagement in multimodal dialog systems. For example, Zhou *et al.* presented a non-task oriented engagement-aware dialog system which was trained by having 2 expert annotators rate how engaging different strategies were [10]. Multiple research studies have examined the annotation and prediction of user engagement in videos of multi-party dialog, and have typically relied on gold-standard annotations rated by a few annotators (see for instance [11, 12, 13]). Such analysis and prediction of engagement and other learner states are also critical to the design and development of intelligent tutors and computer-assisted language learning (CALL) systems in the education domain [14, 15]. Closest to our study is the work of Salam *et al.*, who analyzed engagement in the human-robot interaction domain, where they had a large number of crowdsourced participants view 20-120 second video clips of people interacting with a robot and rate them on multiple aspects of engagement and personality [16]. They found a good inter-rater agreement for engagement annotations, and succesfully used these crowdsourced ratings for further automated analysis and to train engagement classifiers. However, this study aims to analyze even thinner slices of video of 10s in duration. Also, with the exception of the Salam *et al.* study, there has not been much exploration into the use of a large number of crowdsourced raters for engagement annotation.

While many studies have leveraged the use of thin slices of audio and video for automatic processing and prediction of variables of interest, there are none that have explicitly looked at this in the case of human–machine dialog interactions, to our knowledge. That being said, we want to specifically answer the following broad research questions in this particular domain: (1) how do caller engagement ratings of a small crowd of individuals compare to callers' self-assessment of their own engagement levels; (2) how do engagement ratings vary depending on whether the person is responding or listening to the automated agent; (3) how consistent are assigned ratings across a broad sample of video data and different raters; and finally, (4) can we understand how different naïve raters grossly performed on the rating task. In order to answer these questions, we will analyze audio and video data collected from interactions between a human and a dialog system in the context of a CALL application. The rest of the paper is organized as follows: Section 2 presents an overview of how we collected the videos of human–machine dialog used in this study. Section 3 describes the experimental design of the engagement rating task, followed by a detailed description of observations and experimental results in Section 4. We conclude with a discussion of the impli-

[1] The experts, in this case, were people who had substantial interactions with the same teachers in question.

cations for the design and development of engagement-aware dialog systems.

2. Audiovisual Dialog Dataset Generation

We used the open-source HALEF dialog system[2] to collect audio and video data of human–machine dialog interactions. HALEF is an open-source, modular, cloud-based dialog system that is compatible with multiple W3C and open industry standards. The HALEF architecture and components have been described in detail in prior publications [17, 18, 19]. We leveraged Amazon Mechanical Turk for our crowdsourcing data collection experiments. Crowdsourcing (particularly via Amazon Mechanical Turk) has been used in the past for the assessment of spoken dialog systems (SDSs) as well as for collection of interactions with SDSs [20, 21, 22]. We leveraged the aforementioned HALEF dialog system to develop conversational applications within this crowdsourcing framework and collect data over Amazon Mechanical Turk. In this iterative data collection framework, the data logged to the database during initial iterations is transcribed, annotated, rated, and finally used to update and refine the conversational task design and models (for speech recognition, spoken language understanding, and dialog management). In addition to calling into the system to complete the conversational tasks, callers were requested to fill out a 2-3 minute survey regarding different aspects of the interaction, such as their overall call experience, how well the system understood them and to what extent system latency affected the conversation. Importantly for our task, they also rated how engaged they felt while interacting with the system. Since the targeted domain of the tasks in this study is conversational practice for English language learners, the majority of our crowdsourcing user pool comprised non-native speakers of English; however, we also collected data from native speakers of English in order to test the robustness of the system and to obtain expected target responses from proficient speakers of English. For the purposes of this engagement study, we chose to extract video data collected from the conversational dialog tasks shown in Table 1. The selected tasks provide a good mix of different types of dialog interaction across domains, open-endedness of response, and length of the interaction, with an aim to allow for a good coverage of different engagement states for our video annotation experiments.

3. Method

3.1. Rating

We requested 31 participants from within our Educational Testing Service R&D project team to assign an engagement rating to 10-second video segments on a 1–5 Likert scale. Raters assigned a rating of ‘0’ or unscorable if there were issues with the audio or video, such as the lack of an audio or video channel[3]. We also asked them to rate the audio and video quality, as well as who was speaking – system, human, both, or neither. Note that we did not have raters go through any special training or calibration process.

3.2. Experimental Design

We processed the videos using the following steps:

1. First, in order to remove files with empty audio/video recordings, we validated the codecs of each video using the *ffmpeg* toolkit to ensure their integrity, and discarded any video that was found to have either corrupted video or audio codec.
2. Using *ffmpeg*, we split each video into segments of 10 seconds each. We discarded the first and last segments of each video during this process in order to (i) remove pixellated video or spurious audio that can be recorded at the beginning of calls during the establishment of the connection, and to (ii) control for the variations in user engagement states before and after performing the task.
3. From this newly-created corpus of 10-second video segments, we generated 300 unique segments: (a) 150 *randomly*-sampled segments, and (b) 150 segments based on uniform sampling from the distribution of engagement ratings assigned by the callers themselves. We did this in order to (i) control for the effect of class label imbalance (for instance, there are far fewer '1' ratings than '3'), (ii) ensure that we had a somewhat uniform distribution of video instances across the engagement spectrum for laypeople to rate, and (iii) ensure that we have adequate training instances from each class to train automated engagement classifiers in the future.
4. We had each video segment rated by 3 unique raters, resulting in a total of 900 segments in all to be rated. Note that we took care to ensure that no person rated the same segment twice.
5. Finally, we requested our pool of 31[4] naïve raters to each rate 30 video segments, resulting in a total of 900 segment annotations collected.

The above experimental design allows us to perform several insightful statistical analyses: (i) the performance of different individual raters in rating 30 videos, (ii) the consistency of assigned ratings for each of the 300 unique videos, and (iii) how well callers' self-assigned engagement ratings compare to those assigned by our small crowd of naïve raters.

4. Observations and Analyses

Figure 1 shows the distribution of engagement ratings assigned by our small crowd of raters in both the *randomly*-sampled 10s videos as well as those based on uniformly sampling from the distribution of engagement ratings assigned by the callers themselves. We also show for comparison the callers' self-ratings of engagement in both cases, though note that these were assigned at the level of the full-call. Plotting the latter allows us to visualize the inherent distribution of engagement labels in the original dataset. Callers rated themselves as mostly engaged, resulting in a skew towards the higher end of the rating spectrum as seen in the random sampling case. While the distributions of crowd ratings somewhat mirror the original self-ratings, as expected, this is surprisingly not the case with the uniform sampling condition; instead the distribution of crowd ratings mirrors the random condition, with disproportionately more segments being rated as engaged than disengaged. One potential reason for this

[2] http://halef.org.

[3] While we automatically discarded files that had no audio or video, we retained files that had either the audio or video channel recording and instructed raters to rate engagement based on all available channels. We did this in order to model real-life situations where sometimes only one channel might be available for automated engagement prediction.

[4] Two of our raters rated a set of 30 calls between them, which is why we have 31 raters in practice instead of the original 30 that was part of the experimental design.

Table 1: *The details of conversational tasks from which videos were sampled for the purposes of this experiment.*

Item	Brief Task Description	Call Duration (sec)		Number of
		Mean	Std. Dev.	Full Calls
Job Placement Interview	Interact with an interviewer at a job placement agency	345.2	114.1	83
Coffee Shop Order	Order food and drink from a coffee shop	135.3	66.8	83
Billing Dispute	Dispute charges on a customer phone bill	154.0	79.4	40
Conference Ad	Answer a caller's questions about a conference ad posting	112.7	86.9	22

Table 2: *Dimensions along which our pool of naïve raters rated video segments. Note, however, that while callers self-rated their engagement levels over the course of the full call, the crowd had to make engagement judgements solely based on 10 second samples of those calls.*

Rating	Description	Caller	Crowd
Caller Engagement	A qualitative measure of caller's engagement with the task or the system, ranging from highly disengaged to highly engaged.	✓	✓
Audio quality	This metric measures, on a scale from 1 to 5, how clear the caller audio is. A poor audio quality rating would be marked by user responses dropping in and out of the call, being muffled, garbled, echoing or inaudible.		✓
Video quality	This metric measures, on a scale from 1 to 5, the video quality of the call. A poor quality rating here would involve issues with lighting, other problems with the video (such as pixellation, blocking artifacts, non-constant background, etc.) and if the users head is not located in the center of the image as instructed in the caller guidelines.		✓
Interlocutor Identity	Who was speaking in the video – the automated system, the caller, both, or neither.		✓

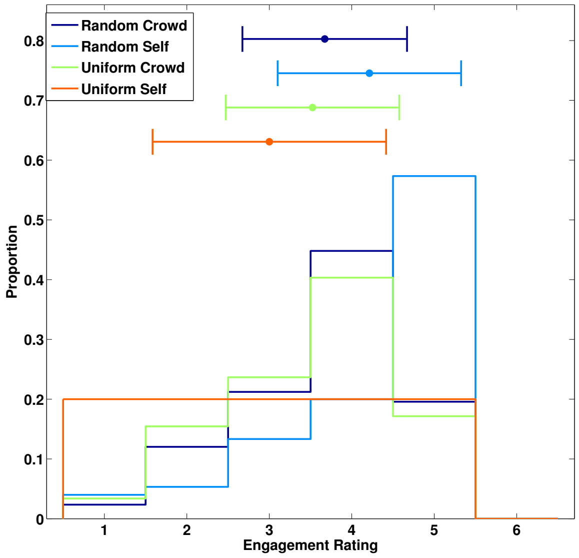

Figure 1: *Engagement distributions across the two sampling conditions.*

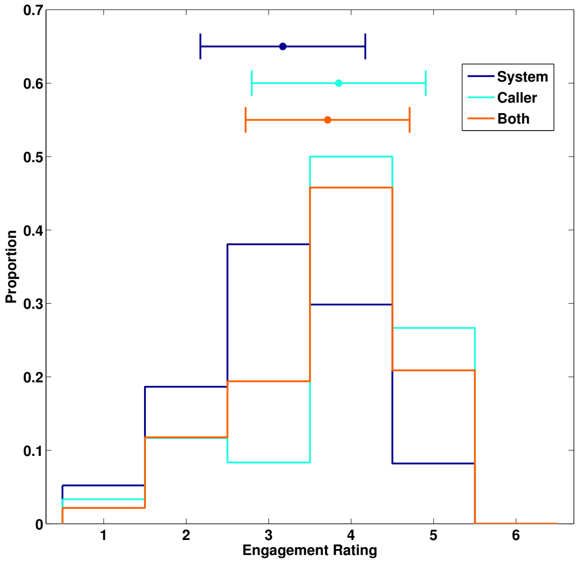

Figure 2: *Engagement distributions as a function of interlocutor identity, i.e., who was speaking in the 10 second segments – the automated agent, the caller, or both.*

could be the fact that speakers were presented with videos from both sampling conditions as part of the same experimental set of 30 videos, but a more likely reason is that the engagement level of the caller during different 10 second snippets of the call are not representative of the overall engagement level of the caller over the full call.

We next analyzed how the engagement distributions varied as a function of who was speaking in the 10s video segments – the automated system, the caller, or both (see Figure 2). We observed that most segments involved both parties speaking, and callers were rated as most engaged on average in this condition. Interestingly, crowd engagement ratings as dropped slightly on average when only the caller was speaking, and dropped further when only the system was speaking. This suggests that users were most engaged when they were listening to short system questions and getting ready to respond, but their engagement levels dropped if either (i) the system prompt was too long, or (ii) they were giving a long answer to the question posed by the system and were thinking about their response, potentially resulting in a wandering gaze.

In order to understand how consistently raters rated each of the 300 videos, we computed various statistical measures of inter-rater agreement on our dataset. See Table 3. We see

Table 3: *Inter-rater agreement statistics computed for our experimental setup of 300 videos and 31 raters.*

Statistical Metric	Value
Krippendorff's α	0.401
Conger's κ	0.399
Scott's π	0.400

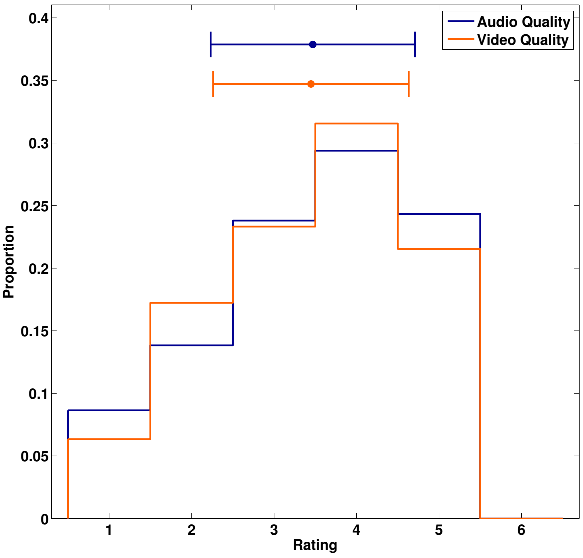

Figure 3: *Histogram distributions of audio and video quality as rated by the crowd.*

that the multiple rater versions of Krippendorff's α, Conger's κ (which is an extension of Cohen's κ to more than two raters) and Scott's π values are in close agreement: 0.4. This suggests a moderate agreement between raters, which is pretty good and encouraging given that these are naïve raters who were not given too much instruction or rater training.

We also computed the Pearson correlation between the average crowd engagement rating and the corresponding counterpart self-rated by the original caller to be a statistically significant, but small positive value ($\rho = 0.15; p = 0.0087$). While this could be due to the lack of rater training and/or calibration, a more likely contributing factor is the fact that the crowd viewed and rated only 10 second segments, while caller self-ratings were assigned for the entire video interaction. This is important since user engagement could (and most likely does) vary over the course of the interaction, and the overall caller self-rating is more like an *average* engagement value over the entire interaction. In addition, there remains the possibility of caller bias while self-rating calls, i.e., people might tend to rate themselves as more engaged than they actually were, for instance. Yet another reason could be low quality of data from either the audio or video channel, which could have increased the difficulty of judging engagement. Figure 3, which plots the distribution of audio and video quality ratings (on a 1–5 Likert scale, from least satisfactory to most satisfactory) as assigned by the crowd, suggests that while a large number of video segments were rated as being of satisfactory quality (mean $\approx$ 3.5), there were some videos which were either of poor quality or lacked an audio or video recording.

5. Discussion and Outlook

This paper has presented an experimental design and statistical analysis paradigm to understand how well a small crowd of human annotators rate engagement in 10s thin-slice videos of a caller interacting with a spoken dialog system. We explored two different sampling paradigms – one where videos were picked at random, and the other where we equally sampled videos from each rating label (based on caller self-ratings), and found, interestingly, that presenting both sets of videos together could have hypothetically influenced the rating distribution in the latter case to mirror that of the former. This has implications for the design of rating experiments – the uniform sampling paradigm is important in order to obtain sufficient ratings from each label category for the training of automatic classifiers, but one should ensure that no unwanted bias creeps into the rating process nonetheless to ensure the integrity of ratings.

The study also presented some useful findings for the design and development of engagement-aware multimodal dialog systems. Unsurprisingly, we found that caller engagement varies as a function of whether caller or system were speaking, with callers exhibiting higher engagement levels in general when they were speaking or both were speaking as compared to when the system was speaking, particularly for longer system prompts. Ensuring that caller engagement does not drop during such periods in an important consideration for dialog design. Furthermore, while we observe an influence between the crowd ratings and audio/video quality, it is important to rate and take such data into account nonetheless as this situation is representative of a real-world dialog system setting, where there could be delays and audio/video quality problems due to network bandwidth and connectivity issues.

Future work will look to extend this paper's findings to a larger number of crowdsourced raters, and leverage such ratings toward the training of more accurate dialog-context-aware engagement classification modules.

6. Acknowledgements

We would like to thank the 31 raters from within our Educational Testing Service R&D project team for taking their valuable time to participate in our rating experiment.

7. References

[1] N. Ambady and R. Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness." *Journal of personality and social psychology*, vol. 64, no. 3, p. 431, 1993.

[2] J. R. Curhan and A. Pentland, "Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes." *Journal of Applied Psychology*, vol. 92, no. 3, p. 802, 2007.

[3] D. L. Roter, J. A. Hall, D. Blanch-Hartigan, S. Larson, and R. M. Frankel, "Slicing it thin: new methods for brief sampling analysis using rias-coded medical dialogue," *Patient education and counseling*, vol. 82, no. 3, pp. 410–419, 2011.

[4] N. Ambady, M. A. Krabbenhoft, and D. Hogan, "The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness," *Journal of Consumer Psychology*, vol. 16, no. 1, pp. 4–13, 2006.

[5] M. W. Kraus and D. Keltner, "Signs of socioeconomic status a thin-slicing approach," *Psychological Science*, vol. 20, no. 1, pp. 99–106, 2009.

[6] T. F. Oltmanns, J. N. Friedman, E. R. Fiedler, and E. Turkheimer, “Perceptions of people with personality disorders based on thin slices of behavior,” *Journal of Research in Personality*, vol. 38, no. 3, pp. 216–229, 2004.

[7] D. R. Carney, C. R. Colvin, and J. A. Hall, “A thin slice perspective on the accuracy of first impressions,” *Journal of Research in Personality*, vol. 41, no. 5, pp. 1054–1072, 2007.

[8] K. A. Fowler, S. O. Lilienfeld, and C. J. Patrick, “Detecting psychopathy from thin slices of behavior.” *Psychological assessment*, vol. 21, no. 1, p. 68, 2009.

[9] L. S. Nguyen and D. Gatica-Perez, “I would hire you in a minute: Thin slices of nonverbal behavior in job interviews,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 51–58.

[10] Z. Yu, L. Nicolich-Henkin, A. W. Black, and A. I. Rudnicky, “A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement,” in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, p. 55.

[11] R. Bednarik, S. Eivazi, and M. Hradis, “Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement,” in *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*. ACM, 2012, p. 10.

[12] A. Levitski, J. Radun, and K. Jokinen, “Visual interaction and conversational activity,” in *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. ACM, 2012, p. 11.

[13] C. Oertel and G. Salvi, “A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue,” in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 99–106.

[14] S. S. D’Mello, P. Chipman, and A. Graesser, “Posture as a predictor of learner’s affective engagement,” in *Proceedings of the Cognitive Science Society*, vol. 29, no. 29, 2007.

[15] K. Forbes-Riley and D. Litman, “Adapting to multiple affective states in spoken dialogue,” in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2012, pp. 217–226.

[16] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, “Fully automatic analysis of engagement and its relationship to personality in human-robot interactions,” *IEEE Access*, 2016.

[17] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, R. Mundkowsky, A. V. Ivanov, Z. Yu, Y. Qian, and K. Evanini, “Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System,” in *Multimodal Interaction with W3C Standards*. Springer, 2017, pp. 295–310.

[18] Z. Yu, V. Ramanarayanan, R. Mundkowsky, P. Lange, A. Ivanov, A. W. Black, and D. Suendermann-Oeft, “Multimodal halef: An open-source modular web-based multimodal dialog framework,” in *International Workshop on Spoken Dialog Systems (IWSDS 2016), Saariselka, Finland*, 2016.

[19] D. Suendermann-Oeft, V. Ramanarayanan, M. Teckenbrock, F. Neutatz, and D. Schmidt, “Halef: An open-source standard-compliant telephony-based modular spoken dialog system: A review and an outlook,” in *Natural Language Dialog Systems and Intelligent Assistants*. Springer, 2015, pp. 53–61.

[20] I. McGraw, C.-y. Lee, I. L. Hetherington, S. Seneff, and J. Glass, “Collecting voices from the cloud.” in *LREC*, 2010.

[21] E. Rayner, I. Frank, C. Chua, N. Tsourakis, and P. Bouillon, “For a fistful of dollars: Using crowd-sourcing to evaluate a spoken language call application,” 2011.

[22] F. Jurcıcek, S. Keizer, M. Gašic, F. Mairesse, B. Thomson, K. Yu, and S. Young, “Real user evaluation of spoken dialogue systems using amazon mechanical turk,” in *Proceedings of INTERSPEECH*, vol. 11, 2011.