Crowdsourcing Ratings of Caller Engagement in Thin-Slice Videos of Human-Machine Dialog: Benefits and Pitfalls

Vikram Ramanarayanan
Educational Testing Service R&D
San Francisco, CA
vramanarayanan@ets.org

Chee Wee Leong
Educational Testing Service R&D
Princeton, NJ
cleong@ets.org

David Suendermann-Oeft
Educational Testing Service R&D
San Francisco, California
suendermann-oeft@ets.org

Keelan Evanini
Educational Testing Service R&D
Princeton, NJ
kevanini@ets.org

ABSTRACT

We analyze the efficacy of different crowds of naïve human raters in rating engagement during human–machine dialog interactions. Each rater viewed multiple 10 second, thin-slice videos of native and 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 crowdsourced paradigms in designing, developing and analyzing engagement-aware dialog systems.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**;

KEYWORDS

engagement,multimodal dialog, thin-slicing, crowdsourcing

ACM Reference Format:
Vikram Ramanarayanan, Chee Wee Leong, David Suendermann-Oeft, and Keelan Evanini. 2017. Crowdsourcing Ratings of Caller Engagement in Thin-Slice Videos of Human-Machine Dialog: Benefits and Pitfalls. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3136755.3136767

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. Toward this end, multiple studies in the literature have attempted to define and develop annotation schemes for user engagement and user involvement in interactions to aid subsequent manual analysis as well as automated processing [5, 12, 22].

Psychologists have long been studying how well laypeople rate different aspects of human behavior by only viewing short durations, or *thin slices*, of video. 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 [2]. 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 [8], analysis of medical dialog [28], evaluation of sales effectiveness [1], assessment of socioeconomic status [16], assessment of personality traits [6, 24], detecting conflict in team interactions [14] and even detection of psychopathy [11], 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 [21].

Recent research in the literature has extensively analyzed the role of engagement in multimodal dialog systems. For example, Yu *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 [31]. 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 [3, 19, 23]). 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 [9, 10]. Closest to our study is the work of Salam

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

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136767

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 [29]. 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. The present study extends a recent smaller-scale study we conducted [25] to a much larger scale.

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 caller 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 five CALL applications designed to expose English language learners to commonplace workplace scenarios where they can practice their conversational English skills. 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 implications 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. For more details on the HALEF architecture and components, see [26]. We used the Amazon Mechanical Turk platform 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 [15, 20, 27]. Researchers have also developed tools to rapidly annotate videos of interactions that exploit the power of the crowd (see for e.g. [18]). 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, 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. In the data sample considered in this study, approximately 59% of videos were from people who self-reported their native language to be 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 investigated three crowdsourced rating paradigms – a small-scale internal study, as well as two larger-scale studies on Amazon Mechanical Turk, the first of which targeted breadth (rating a large number of video segments), while the second targeted depth (obtaining a large number of ratings for each video segment). See Table 2 for a breakdown of number of participants and video segments in each study paradigm.

For the internal study, we requested 31 participants[3] from within our 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[4]. 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.

We deployed an extension version of this study on Amazon Mechanical Turk with more raters and more video segments. The first one, had a total of 102 crowdsourced participants rate 1020 video segments, with each rater rating 30 videos to allow sufficient ratings to compute rater reliability. However, this study does not source a large number of ratings per unique video segment (only 3). The second study aims to address this gap, sourcing 30 ratings per video segment from a larger rater pool of 300 raters, for a smaller number of video segments (100). In this manner, the first study allows us to

[2] http://halef.org.

[3] 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.

[4] While we automatically discarded files that had both no audio and no 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.

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

Item	Brief Task Description	# of Calls	Call Duration (sec) Mean	Call Duration (sec) Std. Dev.
Job Placement Interview	Interact with an interviewer at a job placement agency	206	345.2	114.1
Coffee Shop Order	Order food and drink from a coffee shop	359	135.3	66.8
Billing Dispute	Dispute charges on a customer phone bill	139	154.0	79.4
Conference Ad	Answer a caller's questions about a conference ad posting	61	112.7	86.9
Meeting Request	Request your boss for a meeting and to review slides	178	80.2	35.9

Table 2: *Experimental design.*

	Crowdsourcing Paradigm		
	Internal Study	AMT Study 1	AMT Study 2
# of video segments	300	1020	100
# of ratings/video segment	3	3	30
# of raters	30	102	300
# of ratings/rater	30	30	10
Types of sampling	Both	Both	Uniform only

Table 3: *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 (1) to highly engaged (5).	✓	✓
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 user's 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.		✓

obtain ratings for a large number of video segments which can subsequently be used to train engagement classifiers, while the second allows us to study the accuracy of ratings assigned by the crowd for a smaller set of video segments.

3.2 Experimental Design

We processed the videos using the following steps for use in all study paradigms (refer to Table 2 for statistics):

(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 (i.e., 30 samples from each of the class labels from 1 to 5). 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 required each video segment to be rated by at least 3 unique raters (and in the case of AMT Study 2, 30 raters). Note that we took care to ensure that no rater rated the same segment twice.

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

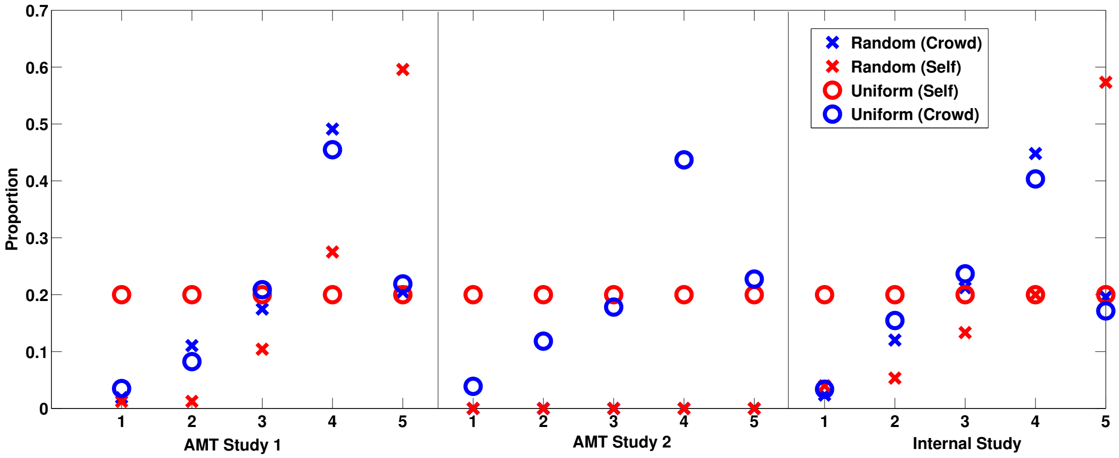

Figure 1: Engagement distributions across the two sampling conditions.

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

4.1 How did engagement ratings vary as a function of how the data was sampled?

Figure 1 shows the distribution of non-zero engagement ratings, across all three study paradigms, assigned by our 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 originally 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 (ranging from highly disengaged to highly engaged) 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 (i.e., ratings of 4 or 5) than disengaged (i.e., ratings of 1 or 2). This observation, which is consistent across different study paradigms, might be because the level of engagement observed in 10 second thin slices is not indicative of the overall engagement of the caller over the entire interaction, which stands to reason considering that a person's engagement level evolves over time depending on multiple factors [4, 5]. We will revisit this hypothesis in Section 4.4.

4.2 How did engagement ratings vary as a function of who was speaking?

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 trend was, again, consistent across study paradigms. 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.

4.3 How many ratings are sufficient?

In order to understand how consistently crowd engagement ratings were between our three study paradigms, we computed correlations between average ratings obtained for the 100 video segments that all three studies have in common. We observed the correlation between the two large-scale MTurk studies to be as high as 0.74 ($p \approx 0$), while correlations between the internal study and Studies 1 and 2 were also high and statistically significant: 0.58 and 0.73, respectively. This clearly suggests that (i) the crowdsourced ratings are generally consistent with each other, and (ii) the correlations between ratings increase as a function of the number of ratings per video considered.

4.4 How well did the crowd ratings correlate with caller self-ratings?

We found the Pearson correlation between the average crowd engagement rating and the corresponding counterpart self-rated by the original caller to be low and close to zero for all study paradigms ($\rho_{AMTStudy1} = 0.03 (p = 0.73); \rho_{AMTStudy2} = -0.05 (p = 0.15);$ $\rho_{InternalStudy} = 0.15 (p = 0.009)$). This could be due to the lack of rater training and/or calibration; however, as we have mentioned earlier, 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 likely varies over the course

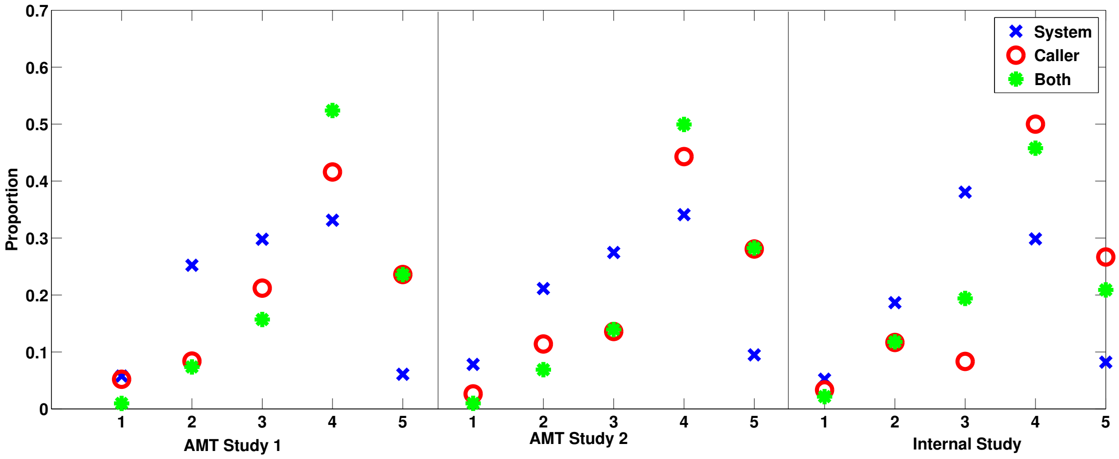

Figure 2: Engagement distributions as a function of interlocutor identity, i.e., who was speaking in the 10 second segments – the automated agent, the caller, or both.

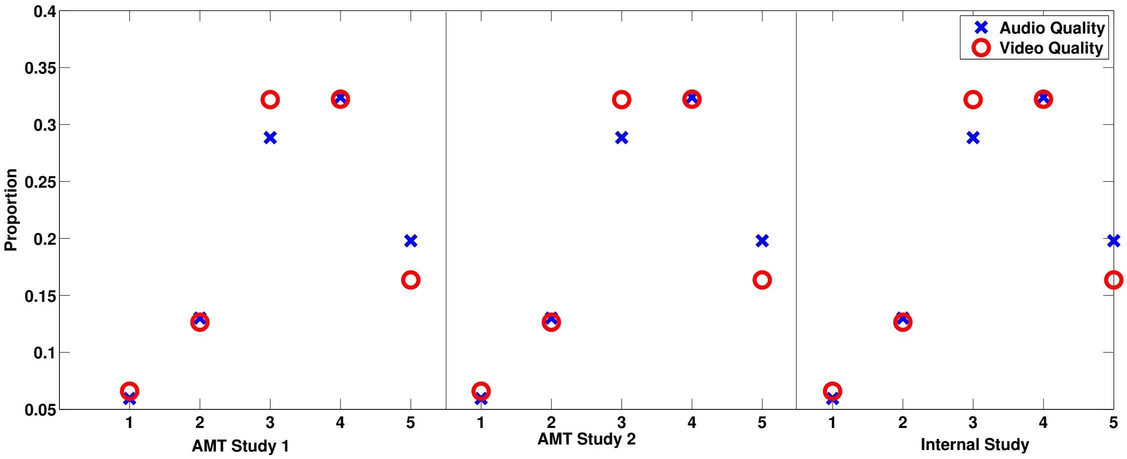

Figure 3: Histogram distributions of audio and video quality as rated by the crowd.

of the interaction (see for example [4, 5]), 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 factor influencing the crowd rating could be the quality of data from either the audio or video channel. However, 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 a large number of video segments were rated as being of satisfactory quality (mean ≈ 3.5), so while there might have been some audio or video files which were of poor quality, this is unlikely to have greatly impacted the observed correlation trends.

Table 4: *Inter-rater agreement statistics computed for our three different experimental paradigms.*

Statistical Metric	Value		
	AMT Study 1	AMT Study 2	Internal Study
Krippendorff's α [17]	0.273	0.273	0.401
Conger's κ [7]	0.272	0.272	0.399
Scott's π [30]	0.272	0.272	0.400
Gwet's γ [13]	0.7	0.7	0.699

4.5 How consistently did raters rate video segments?

In order to understand how consistently raters rated each of the videos in the different study paradigms, we computed various statistical measures of inter-rater agreement on our dataset. See Table

4. We see 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: 0.27 for the large-scale AMT paradigms and 0.4 for the internal study. This suggests a low to moderate agreement between raters, which is understandable given that these are naïve raters who were not given too much instruction or rater training. However, notice that the Gwet's γ is high (0.7). Nonetheless, keep in mind that our case is different from the traditional use cases for these statistical metrics (where the number of raters is much lower than the numbers examined in this study), since the canonical matrix in our case is very sparse.

The low inter-rater agreement suggests that one perhaps needs more video context for accurate rating (assuming that the metrics are correct and make sense in our use case, which they may not). Therefore, a potential avenue for future work may involve examining thin-slices of longer length.

5 DISCUSSION AND OUTLOOK

This paper has presented an experimental design and statistical analysis paradigm to understand how well crowds of human annotators rate engagement in 10 second thin-slice videos of a caller interacting with a spoken dialog system. We explored different crowdsourcing study paradigms designed to either obtain a large number of ratings to train automatic classifiers or to analyze the accuracy and reliability of the crowd, and found that the observations and rating trends from these three studies were largely consistent with each other. We further 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 it is unlikely that presenting both sets of videos together could have hypothetically influenced the rating distribution in the latter case to mirror that of the former. Rather a more likely explanation is our finding that caller self-ratings over an entire video dialog are uncorrelated with, and not predictive of the engagement values in thin slices of that interaction. This could be because engagement ratings vary considerably over the entire interaction, a proposition also supported by multiple studies in the literature. This calls into question the usefulness of an aggregated rating over the entire call as well as the reliability of user self-ratings[5]. Furthermore, while independent trained experts' ratings of our 10 second clips might be one of the best indicators of the accuracy of our crowd ratings, the relatively high correlations between our three crowdsourcing study paradigms already suggests that these ratings might be useful for the purpose of training automated engagement detectors and classifiers. Having said that, future work will investigate in more detail the moderate values obtained for the inter-rater reliability metrics presented here; specifically, how applicable these metrics are to our case of a sparse canonical matrix and how to modify or extend them appropriately in case the specific mathematical assumptions involved in their calculation are violated by our use-case.

The study also presented other 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 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 internal R&D project team for taking their valuable time to participate in our rating experiment.

REFERENCES

[1] Nalini Ambady, Mary Anne Krabbenhoft, and Daniel Hogan. 2006. The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology* 16, 1 (2006), 4–13.

[2] Nalini Ambady and Robert Rosenthal. 1993. Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology* 64, 3 (1993), 431.

[3] Roman Bednarik, Shahram Eivazi, and Michal Hradis. 2012. 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, 10.

[4] Ronald Böck. 2013. *Multimodal automatic user disposition recognition in human-machine interaction*. Ph.D. Dissertation. Magdeburg, Universität, Diss., 2013.

[5] Francesca Bonin, Ronald Bock, and Nick Campbell. 2012. How do we react to context? annotation of individual and group engagement in a video corpus. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 899–903.

[6] Dana R Carney, C Randall Colvin, and Judith A Hall. 2007. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality* 41, 5 (2007), 1054–1072.

[7] Anthony J Conger. 1980. Integration and generalization of kappas for multiple raters. *Psychological Bulletin* 88, 2 (1980), 322.

[8] Jared R Curhan and Alex Pentland. 2007. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology* 92, 3 (2007), 802.

[9] Sidney S D'Mello, Patrick Chipman, and Art Graesser. 2007. Posture as a predictor of learner's affective engagement. In *Proceedings of the Cognitive Science Society*, Vol. 29.

[10] Kate Forbes-Riley and Diane Litman. 2012. 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, 217–226.

[11] Katherine A Fowler, Scott O Lilienfeld, and Christopher J Patrick. 2009. Detecting psychopathy from thin slices of behavior. *Psychological assessment* 21, 1 (2009), 68.

[12] Nadine Glas and Catherine Pelachaud. 2015. Definitions of engagement in human-agent interaction. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 944–949.

[13] Kilem L Gwet. 2008. Intrarater reliability. *Wiley encyclopedia of clinical trials* (2008).

[14] Malte F Jung. 2016. Coupling interactions and performance: Predicting team performance from thin slices of conflict. *ACM Transactions on Computer-Human Interaction (TOCHI)* 23, 3 (2016), 18.

[15] Filip Jurcıcek, Simon Keizer, Milica Gašic, Francois Mairesse, Blaise Thomson, Kai Yu, and Steve Young. 2011. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In *Proceedings of INTERSPEECH*, Vol. 11.

[5] Note that this means we cannot guarantee that the distribution of segments obtained for the "uniform" sampling condition was truly representative of the five different engagement levels.

[16] Michael W Kraus and Dacher Keltner. 2009. Signs of socioeconomic status a thin-slicing approach. *Psychological Science* 20, 1 (2009), 99–106.

[17] Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. *Departmental papers (ASC)* (2007), 43.

[18] Walter S Lasecki, Mitchell Gordon, Danai Koutra, Malte F Jung, Steven P Dow, and Jeffrey P Bigham. 2014. Glance: Rapidly coding behavioral video with the crowd. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 551–562.

[19] Andres Levitski, Jenni Radun, and Kristiina Jokinen. 2012. Visual interaction and conversational activity. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*. ACM, 11.

[20] Ian McGraw, Chia-ying Lee, I Lee Hetherington, Stephanie Seneff, and Jim Glass. 2010. Collecting Voices from the Cloud.. In *LREC*.

[21] Laurent Son Nguyen and Daniel Gatica-Perez. 2015. 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, 51–58.

[22] Catharine Oertel, Céline De Looze, Stefan Scherer, Andreas Windmann, Petra Wagner, and Nick Campbell. 2011. Towards the automatic detection of involvement in conversation. *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues* (2011), 163–170.

[23] Catharine Oertel and Giampiero Salvi. 2013. 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, 99–106.

[24] Thomas F Oltmanns, Jacqueline NW Friedman, Edna R Fiedler, and Eric Turkheimer. 2004. Perceptions of people with personality disorders based on thin slices of behavior. *Journal of Research in Personality* 38, 3 (2004), 216–229.

[25] Vikram Ramanarayanan, Chee Wee Leong, and David Suendermann-Oeft. 2017. Rushing to Judgement: How Do Laypeople Rate Caller Engagement in Thin-Slice Videos of Human–Machine Dialog? *Proc. Interspeech 2017* (2017), 2526–2530.

[26] Vikram Ramanarayanan, David Suendermann-Oeft, Patrick Lange, Robert Mundkowsky, Alexei V Ivanov, Zhou Yu, Yao Qian, and Keelan Evanini. 2017. Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System. In *Multimodal Interaction with W3C Standards*. Springer, 295–310.

[27] Emmanuel Rayner, Ian Frank, Cathy Chua, Nikolaos Tsourakis, and Pierrette Bouillon. 2011. For a fistful of dollars: Using crowd-sourcing to evaluate a spoken language CALL application. (2011).

[28] Debra L Roter, Judith A Hall, Danielle Blanch-Hartigan, Susan Larson, and Richard M Frankel. 2011. Slicing it thin: new methods for brief sampling analysis using RIAS-coded medical dialogue. *Patient education and counseling* 82, 3 (2011), 410–419.

[29] Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2016. Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions. *IEEE Access* (2016).

[30] William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly* (1955), 321–325.

[31] Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alex I Rudnicky. 2016. 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*. 55.