Look at Me When I Talk to You: A Video Dataset to Enable Voice Assistants to Recognize Errors

ANDREA CUADRA, Cornell Tech, USA
HANSOL LEE, Cornell University, USA
JASON CHO, Cornell University, USA
WENDY JU, Cornell Tech, USA

People interacting with voice assistants are often frustrated by voice assistants' frequent errors and inability to respond to backchannel cues. We introduce an open-source video dataset of 21 participants' interactions with a voice assistant, and explore the possibility of using this dataset to enable automatic error recognition to inform self-repair. The dataset includes clipped and labeled videos of participants' faces during free-form interactions with the voice assistant from the smart speaker's perspective. To validate our dataset, we emulated a machine learning classifier by asking crowdsourced workers to recognize voice assistant errors from watching soundless video clips of participants' reactions. We found trends suggesting it is possible to determine the voice assistants's performance from a participant's facial reaction alone. This work posits elicited datasets of interactive responses as a key step towards improving error recognition for repair for voice assistants in a wide variety of applications.

CCS Concepts: • **Human-centered computing**; • **Computing methodologies → Intelligent agents**;

Additional Key Words and Phrases: error-recognition, self-repair, voice assistants, conversational design

1 INTRODUCTION

"Alexa, spell 'Seven.' "
"Salmon is spelled S-A-L-M-O-N."

Today's voice agents are blind to the giggling and the eyerolls of the people they are interacting with. On one hand, this may spare the digital feelings of machine assistants; on the other, it keeps the assistants from recognizing the errors they make and from learning from them. People often get frustrated with voice agents making frequent mistakes and not responding to any feedback other than explicit voice commands.

In human face-to-face interaction, people monitor each other to see if their meaning is being understood by others, and they stop and self-correct if they recognize that they have made an error. Computers and machines could more easily perform error recovery if they used cameras to see how people react to their statements. For example, a machine would know it had just made an error if it could observe the user shaking his/her head in response to its actions. For machines to be able to make sense of these cues, however, they need some model of how people respond when they hear a correct versus an incorrect response.

In this work, we introduce an open-source video dataset of people reacting to a voice agent's right or wrong answers and explore how this dataset could be used to train machines to recognize error based on user's facial expressions by emulating a video classifier. This is a first step in a broader initiative to build a machine-learning model for voice assistants to recognize error based on people's reactions. For this project, we first collect video samples of people reacting to Amazon Alexa's right or wrong answers, then we remove the sound and ask crowdsourced workers to guess whether or not a mistake was made based on the participant's facial expressions. This work contributes on a novel approach in human machine interaction, using datasets of interactive responses to inform the development artificial intelligence to improve human machine interaction, and a dataset on which machine learning can be performed.

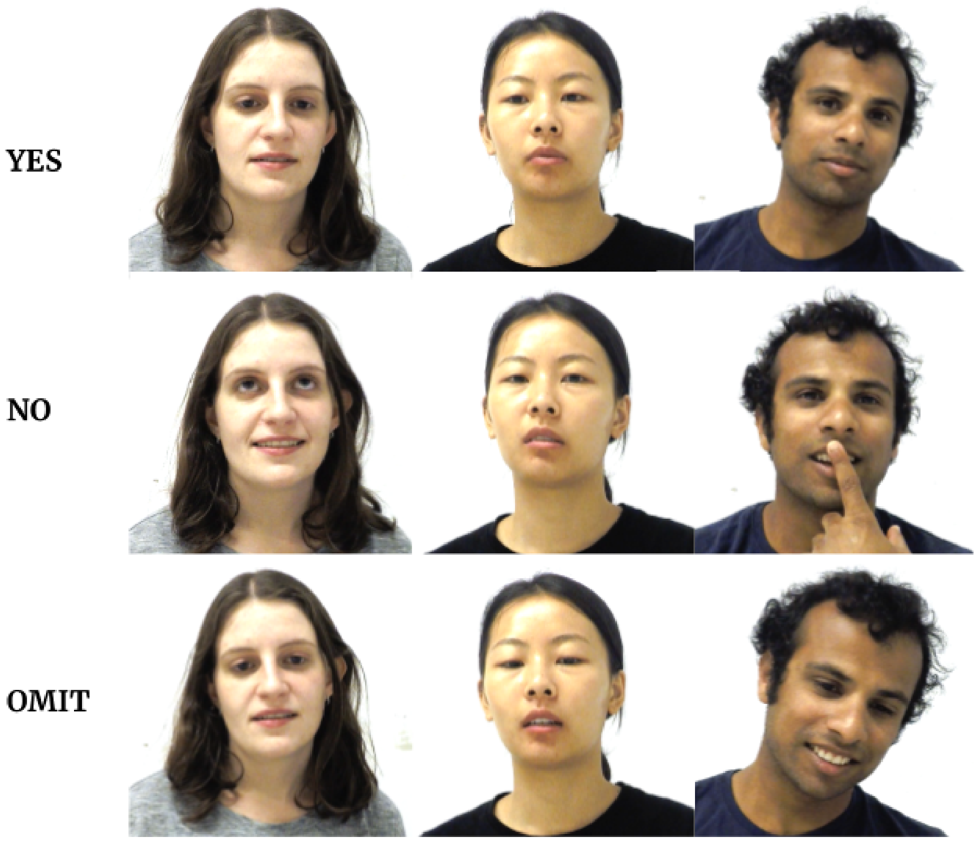

Fig. 1. Our dataset includes people reacting to correct responses (YES), incorrect responses (NO), or non-responses (OMIT).

2 BACKGROUND

Our dataset is constructed from people interacting with current-day voice agents, but the intent of the research is to inform the response people to have to voice assistants more generally. The advantage of working with current-day voice agents is that people are familiar with them, and their lack of embodiment presents a more generic class of responses than people might have in response to a speech-enabled robot or a screen-based avatar. Here, we survey a variety of related research from a number of arenas to make the case for the need for a more general model of how users respond to errors in speech recognition.

2.1 Speech-enabled devices

The commercial viability of voice assistants is presaged by the plethora of new voice-based devices and user interfaces now on the market. Many are voice-enabled assistants, such as Microsoft's Cortana [46], Amazon's Alexa [1], Apple's Siri [5], or Google Assistant [30]. In some of the first-generation voice assistant research published in the larger HCI community, such as [13, 51, 65], the computer-generated voice that people were speaking with was often on the other side of a phone line. In other cases, like the embodied conversational agents of Cassell, Sullivan, Churchill and Prevost [14], the voice was front-ended by on-screen virtual agents. Unlike in these initial examples, today's voice agents are embodied in standalone devices such as Amazon's Echo [2] or Dot, Apple's Homepod [6], or Google's Home or Home mini [31, 32], and are usually situated proximate to people in their living or working quarters.

The widespread explosion in speech-enabled interfaces is driven by advances in natural language processing, text-to-speech and dialog generation systems driven by big data, as well as hardware breakthroughs in far-field microphone arrays. While improvements to the hardware and software of these speech-enabled devices might improve the recognition of individual words people say, common-sense intelligence is not yet in grasp [23]. The limited capabilities of today's speech systems would be well complimented by interactive savvy that would help them recognize and recover from conversational errors.

2.2 Repairing Error

Voice agents may make mistakes, but human dialog is far from error-free itself. A key difference is that people perform repair in communication [60], monitoring listeners to see if they have been heard and understood before moving forward in the conversation. This process, in which speakers seek and provide evidence of understanding, is called grounding. [20] In collaborative conversations [19], addressees must therefore also indicate their understanding, or lack of understanding, to help the speaker understand the state of the communication.

Linguists Schegloff, Jefferson and Sacks define repair to be the practices that interactants use to handle troubles in hearing, speaking and understanding that occur regularly in social interaction. [57–59] They found from analyzing naturalistic conversation that people had a preference for self-correction over being corrected by others: in moments when repair was necessary or possible, the distribution of repairs was strongly skewed towards self-repair [60]. Very often self-repair occurs when the speaker notices a mistake, in the transition space between speaking turns, before the listener even has a chance to respond.

Gieselmann ran a small experiment to look at what error recovery strategies people use when talking to robots compared to when they talk to other people. Gieselmann found that very different repair tactics were used, largely due to the limited interaction capabilities of robot, and that the most common indicator that an error was made is a sudden change in the dialogue topic. In this research, the focus of the error detection lay in analysis of the discourse [29]. More recently, Salazar-Gomez et al. experimented with using EEG-based feedback methods to correct robot mistakes in real time; because the EEG signals were analyzed in real-time in closed-loop fashion, the robot was able to respond to possible signs of error by hyper-articulating actions to elicit stronger response to help it determine if it was making a mistake [56].

2.3 Facial Displays for Conversational Facilitation

Between human interactors, the visual feedback channel can play an important role in the coordination of joint understanding. Clark and Krych performed an experiment where one participant directed another on assembling Lego models; pairs which could not visually see each other performed much slower if the director could not see the builder's workspace, and much worse when the instructions were audio-taped. [18] The face in particular is critical to providing feedback in discourse. Birdwhistell noted in 1970 that facial displays perform linguistic functions, particularly as listener commentaries. [8] Ekman and Friesen characterized the category of nonverbal acts which maintain and regulate the back-and-forth nature of speaking and listening as *regulators. Regulator* actions occur in the attentional periphery; people perform them without thought, but can recall and repeat them if asked. Addressees are sensitive to the lack of these cues, but are rarely aware of them when they are present. [25] In 1991, Chovil performed an experiment with people listening to a story in a face-to-face, partition, and telephone and answering machine condition, and found that listeners primarily react facially when they would be seen by the storyteller [16].

The importance of identifying and incorporating responses to such conversational signals was recognized early in the human-computer interaction community by Nagao and Takeuchi [50]. While linguists and behavioral psychologists have recognized and analyzed the *regulatory* use of facial displays, the machine learning and computer vision community has largely focused on emotion recognition in their analysis of faces [42, 52]. This is in part due to the widespread availability of emotional expression image databases such as Ekman's Pictures of Facial Affect [24], the Belfast database [21], the Extended Cohn-Kanade Dataset [43], or the Affectiva-MIT Facial Expression Dataset [45]. These datasets have been used for a wide variety of applications, such as to evaluate effectiveness of advertisements [66] or political branding [44].

While emotion is certainly an important factor in interaction, strong shows of emotion often take place only after the user is angry, and the interaction is beyond repair. [7] Analysis of facial displays for conversational grounding and efficacy should be just as important, since linguists have found, at least anecdotally, that at least two-thirds of facial actions in dialogue are communicative rather than emotional expressions [27]. Bousmalis et al. for instance, have surveyed the conversation analysis literature for nonverbal audiovisual cues that indicate agreement and disagreement between human speakers, with the goal of developing machine recognition of these cues. [11]

2.4 Embodiment in Conversational Facilitation

The human-robot interaction community has also examined affect recognition [12, 55], but, in that community, there is greater recognition of the use of embodied signals for conversational *regulation*. Fujie et al. made a robot that recognized head motions, like nodding, for paralinguistic information that clarifies speaker intent [28]. Sidner et al. found that participants that knew their robots recognized conversational head nods would nod more. [62] Huang and Mutlu have proposed developing a Robot Behavior Toolkit that uses the social cues that people use to achieve interaction goals to make robots that are able to adapt their behaviors to people. [36] Mutlu et al. have focused predominantly on gaze cues to signal attention and intent, largely for humanoid robots [4, 35, 47–49, 54].

We know of no research in the DIS, HCI, or HRI communities to date that focuses on facial displays for error recognition. Part of the challenge for these communities has been the lack of datasets focusing on facial displays for conversational facilitation to train machine learning models. Our project seeks to address this shortcoming.

3 DATASET GENERATION

To generate the dataset of people reacting to voice agent answers, we recruited $N = 21$ participants to come interact with an Amazon Tap with Alexa in a lab setting for approximately 30 minutes each. The participants were prompted to ask simple queries with right or wrong answers. They were given a set of example questions, but were instructed to ask anything they would like. Participants were compensated with $5 gift cards each for their participation, and signed a consent form that informed them that "Your participation will be audio- and video-recorded, will be used for academic & public presentations, and made publicly available as part of an open video dataset."

3.1 Recording setup

Fig. 2 shows the setup of the room. Participants were video-recorded throughout the course of the the interaction, with the camera situated so that the video reflected the vantage point of the speech-enabled device. Aside from the

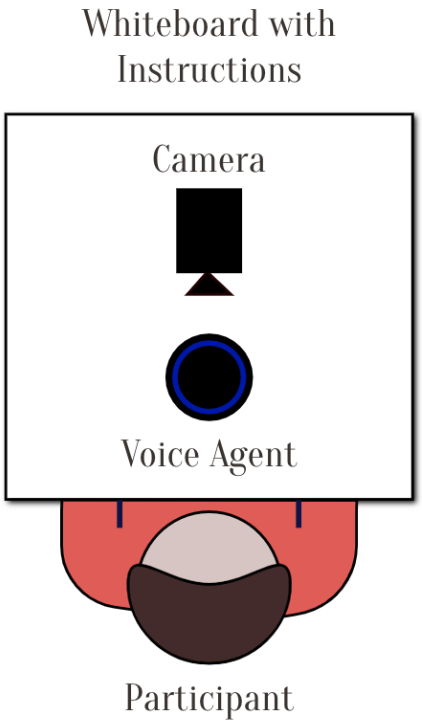

Fig. 2. Illustration of the setup: participants sat in front of the voice agent facing a camera and a whiteboard with handwritten instructions.

instruction and training segment of the sessions, the actual recorded portion of the sessions was generally about 15 minutes long.

3.2 Clip formation

Participant recordings were manually trimmed and classified by a researcher into separate clips. The clips start when the voice agent starts talking, and end after the participant's main reaction to that response occurs, in most cases when the participant says "Alexa" to start a new request.

3.3 Yes/No/Omit

We originally planned to compare only Yes and No responses, but found that Amazon Alexa missed participant commands often enough that an Omit category was necessary to address how people reacted when the voice agent failed to respond to their request at all.

From the participant's perspective, the Omit and No conditions are similar, but from the machine perspective, they are completely different. In the case of *Yes vs. No*, the machine knows it has just said something, and is evaluating whether it made an error. In the case of Omit, the machine does not know that any request has been made, and would need to distinguish the Omit reaction from the reaction of participants who are just idle.

3.4 Resulting dataset

The resulting dataset contains 1238 clips from 21 participants. Alexa answered correctly (YES) in 597, incorrectly (NO) in 458, did not respond (OMIT) in 92 clips, and were idle (IDLE) in 91 clips. Figure 4 illustrates the distribution of reactions in the dataset, Figure 3 illustrates the distribution of reactions per participant, and Figure 1 shows some examples of the types of reactions recorded. Note that in about half of all reactions, the voice agent is not telling the participant what the participant wants to hear.

To enable a point of comparison for the Omit condition, we created a set of I(Idle) clips in which participants were not waiting for Alexa to respond. We created approximately as many I(Idle) clips as O(Omit) clips. Clips vary in lengths from a couple of seconds to minutes.

Each clip was labeled following this set of guidelines:

Y — the voice agent successfully understood and responded to the participant;

N — the voice agent misheard, misunderstood, was not able to fulfill, or misanswered the request; or

O — the voice agent did not hear or ignored the request.

I — the voice agent was idle, because no request was made.

Each clip file contains the reaction label (N, Y, O, or I), the location where the clip was recorded, the participant number, and a clip number. The enclosing directories also contain the dates in which the data was collected. Aside from the inclusion of participant faces, which are necessary given the nature of the study, the clips were anonymized. We had 21 participants, fourteen which identified as female, and the rest who identified as male. Eighteen were college students in their early twenties, and 3 were older adults. 11 participants were from the United States, and 10 participants were from outside the United States.

3.5 Dataset

The dataset contains 1238 video clips of naturalistic reactions (or non-reactions) to voice agent interaction from 21 participants. The clips vary in length from a couple seconds to minutes. The video clips are in color, include sound, and are at least 1352 x 760 DPI in resolution.

We are hosting the dataset on our institution's digital repository. (Following publication), access is freely available to people who secure IRB approval from their own institution for secondary data analysis, indicating that they will use the data only for the research purposes intended. This provision is intended to protect human participants, who gave consent for their videos to be used for scientific research and machine learning, but not for other purposes.

4 DATASET INITIAL ANALYSIS

To validate the dataset, we performed data classification using human crowd workers, online. We implemented two sets of analyses: *Yes vs. No*, and *Idle vs. Omit*. The goal of this data validation is to show that the dataset has signal that is detectable by an human classifier, and thus can more likely be used by machine learning scientists to computationally classify people's reactions.

4.1 Method

For the validation, video clips from our dataset were shown to human classifiers viewing the videos online, to test whether facial expressions can aid in helping humans or machines recognize error. We intentionally took worsened the quality and quantity of the data to match what we anticipated a commercial voice assistant would be able to process in real time, which at the moment does not usually include the context awareness from previous interactions or audio-prosodic features such as rhythm, tone, or intonation. To do so, we removed the sound from the clips in question, trimmed the video to leave out the participants' questions, and reduced the resolution to make the videos easily streamable. We envision machine learning scientists doing similar modifications to our dataset to test their individual hypotheses. The clips were reduced in size to 640x360 pixels, and clipped to be at most 12 seconds long.

This validation helps to establish baseline feasibility of our method; the addition of sound, higher resolution or longer reactions should improve the recognition rates, as would the addition of additional participants to the dataset.

4.1.1 Human Classification. In the machine learning community, human classification is sometimes used as a proof-of-concept and benchmark for machine classification [3, 53, 61]. While there are debates about the quality of Amazon Mechanical Turk worker [1] annotations or classifications [26], nevertheless, seeing that people are able to make a classification using only the data that would be provided to a machine classifier is often used as proof that there is enough information in the data for classification to occur.

To ensure quality and consistency in our classification task, we recruited Mechanical Turk workers with a HIT (Human Intelligence Task) approval rate higher than 95% who were located in the United States. We used online Qualtrics [2] surveys with embedded videos which Mechanical Turk workers were asked to classify. The overall assignment lasted about 12 minutes, and we paid workers $1.80 to complete the assignment.

4.1.2 Classifying Yes vs. No. Before starting the prediction tasks, each worker completed an introductory tutorial with video demonstrations for each label category. Each video demonstration had three parts. The first part showed a clip showing a full interaction of a participant with Alexa. The clip included the sound, the participant's question, and the participant's reaction. Then a black screen with white text indicated, "In this example, the person got what they wanted. In the real tasks, you will only see the reaction without sound." This was followed by a clip like the one they would be asked to rate. Then the worker was presented with a survey question, "Did the person get what they wanted?," and had the option to say "yes" or "no." This process was repeated with a video demonstration of a *No* video. We used the workers responses on these initial trial questions to identify which workers successfully understood the prompt of the assignment.

After the introductory tutorial, each assignment contained 23 videos: 13 videos randomly selected from the pool of Y-labeled clips, and 10 videos randomly selected from the pool of N-labeled clips. The proportions were determined based on label prevalence in the dataset. After watching each video, humans were asked, "Did the person get what they wanted?" and had the options to say "yes" or "no."

170 Mechanical Turk workers completed the assignment, out of which 125 accurately responded to the two initial training videos. We removed 45 entries from workers who did not answer the training correctly. In total, we analyzed 2875 *Yes* or *No* predictions. Table 1 summarizes our findings.

Table 1. Human classification of *Yes vs. No*

N = 2875	Predicted: YES	Predicted: NO
Actual: YES	843	782
Actual: NO	583	667

[1] https://www.mturk.com
[2] https://www.qualtrics.com

4.1.3 Classifying Idle vs. Omit. The protocol for the *Idle vs. Omit* classification followed the one we used for classifying *Yes vs. No*. We recruited workers using the same guidelines, and had them complete an introductory tutorial with video demonstrations of participants in the Idle or Omit scenarios.

The Idle or Omit assignment lasted about 6 minutes, and we paid participants $0.90 to complete the assignment. After the introductory tutorial, each assignment contained 12 videos evenly selected at random from the pools of Idle- and Omit-labeled clips. After watching each video, workers were asked "Was this person initially waiting for Alexa to respond?" and had the options to say "yes", or "no."

57 Mechanical Turk workers completed the assignment, out of which 38 accurately responded to the two initial training videos. We removed entries from 19 workers who did not answer the training correctly. Like for the Yes/No Human Classifier, even though we removed responses in which people failed the training, we estimate that about 25% of the workers that did not watch or understand the training videos might have accurately responded to the tutorial questions by chance as there were only two 2-option questions. We analyzed a total of 456 Idle or Omit predictions. Table 2 summarizes our findings.

Table 2. Human classification of IDLE vs. OMIT

N = 456	Predicted: IDLE	Predicted: OMIT
Actual: IDLE	129	99
Actual: OMIT	111	117

5 RESULTS

The human classifiers performed slightly better than chance at predicting the labels of videos, 2.5% better at distinguishing Y- and N-labeled clips, and 3.9% better at distinguishing I- and O-labeled clips. The paired t-tests comparisons with chance *Yes vs. No* and *Idle vs. Omit* were, .0642 and .1128, respectively. Only the *Idle vs. Omit* finding was statistically significant at a 90% confident level. The *Yes vs. No* comparisons suggest a trend that needs further testing.

6 DISCUSSION

The process of classifying and validating the videos in our dataset turned out to have a number of complexities that were non-obvious, and which we feel would be good for anyone researchers following in our footsteps to know.

6.1 Complexity in Error Classification

One surprising aspect of this process was that the separation of YES/NO/OMIT clips was not straightforward, and involved a greater degree of subjectivity than the authors initially expected.

Coding the ground truth for each clip of whether an answer was right or wrong was challenging. For example, there are many ways in which Alexa can be wrong, and many ways in which a person might be displeased by Alexa's response. Only some of these cases should require Alexa to perform self-repair, but how might we distinguish a frown due to a bad odor in the room from a frown due to Alexa hearing the name of the wrong song? Additionally, some responses are layered, having correct and incorrect components — for example, in one case, a German participant asks about the weather in Dortmund, Germany. When Alexa gets the location right, the person is satisfied; however, fractions of a second later, Alexa says the weather in degrees Fahrenheit, so the person immediately becomes dissatisfied as he was expecting the weather in degrees Celsius.

Retrospectively, as we dug deeper into these issues, we discovered several works suggesting error taxonomies for chat and speech systems. Higashinaka et al. distinguished utterance-level, response-level, context-level and environment level errors, and further proposes multiple subcategories to each [34]. Bohus and Rudnicky's work suggests the following non-understanding errors: out-of-application (conversational level), out-of-grammar (intent level), ASR Error (signal level) and End-pointer error (channel level) [10]. These taxonomies were helpful to us when classifying errors, and will be useful again when deciding what type of repair to perform. However, the distinction between errors did not seem to cause differences that we could detect in the human reaction.

6.2 Can People Detect Error?

This validation of our reaction dataset with human classifiers was performed to test whether we could recognize interaction errors using facial reactions alone. The results from using humans to recognize error based only on visual cues suggest that error recognition is possible, but that this is a very difficult task, even for people. Our human classifiers performed only slightly better than chance. This is an important finding on its own, and opens up ways in which our dataset can be used. For example, instead of making the clips low-resolution and silencing them, dataset users can keep the clips' original resolution and audio content. By doing so, they can do a multimodal study similar to what D'Mello and Graesser did over ten years ago [22]. In a multimodal semi-automated affect detection study, D'Mello and Graesser found that the accuracy of a multichannel model (face, dialogue, and posture) was statistically higher than the best single-channel model for fixed (but not spontaneous) affect expressions [22]. This suggests that combining the audio and visual elements from our dataset could yield surprisingly better results, and a dataset like ours opens up possibilities for many to do so without having to do they data collection step themselves. In retrospect, we should have done our initial validation analysis with more than just one modality, but doing another round of testing was beyond the scope of this work, especially given our findings were sufficient to demonstrate the dataset's usefulness.

On the whole, this validation indicates that facial reactions *can* be used by machine learning scientists to further explore the possibilities for error-recognition to enable recovery. This dataset brings us a first step closer at enabling voice agents to successfully recognize error in order to perform self-repair to reliably improve one-on-one interactions.

6.3 The cost of repair

One key question that our finding that people are only slightly better than chance at recognizing error from only facial displays uncovers is: how good does error recognition need to be for successful interaction? Is it possible that repair and recovery works as a process in human dialogue *despite* our weak recognition of error because attempts at conversational repair are not costly?

After improving error-recognition accuracy, the most immediate next step is to focus on self-repair. In the current work, the modeling of likely error occurred post-facto, and off-line. By developing models for error recognition that can occur on-line, we would be able to have machines experiment with different ways that a voice agent could perform self-repair. Bohus, et al. performed online supervised repair of error telephone-based spoken dialog system that provides bus route and schedule information in the greater Pittsburgh area [9]. Similar experiments for in-situ voice interaction could employ a similar approach to on-line learning; we expect the right repair policy probably differs in different contexts, and it would be interesting to see how visual approaches to error recognition and subsequent recovery differ from speech-based approaches.

6.4 Hedging

Based on our qualitative observation of the interactions, the main point of confusion was often that Alexa states her answers very definitively, as if she were sure of her answers. When Alexa is making a mistake, the mismatch between that tone and the actuality of the error causes people to take longer to understand that a breakdown has occurred. This type of error, an error in belief or conviction of response, is one that Alexa makes with regularity that seems not to be named at all in aforementioned taxonomies [10, 34]. If a respondent is giving an answer she is unsure about, she hedges by expressing some level of uncertainty in her voice and expressions. This provides "feedforward" information to the requester to provide more marked signals about the success or failure of the answer, and is usually accompanied by greater attention by the respondent to the requester's reactions of approval or further confusion. This type of closed loop interaction could be made possible if the computer's responses were modulated by its confidence that it had the right response to the user's query.

6.5 Scaling the method

To collect data at a larger scale, we believe it is important to come up with a method and setup that motivates longer streams of interaction between the user and the voice agent. There are many ways to achieve this goal; for example, by using scenarios where there are many faces simultaneously providing input about the same event, like the one in Kateva's Robot Comedy Lab in which a robotic comedian collects facial data from a large number of people at once [38].

Also, it is possible that defining a task for participants to perform with the help of the voice agent might have yielded a larger dataset of interactions. In this study, although we made suggestions for topics for people to discuss with the voice agent, we left the actual queries up to users. We felt people's facial reactions were more likely to be natural if they were motivated by real interest in the answers. In retrospect, however, this setup also had the effect of limiting the number of interactions, because people randomly recruited for our study did not intrinsically have a lot of things they really wanted to ask the voice agent.

It could also be that the best way to deploy this experiment is not in a laboratory setting, but in situ, employing code in "beta" chatbots that collects reactions "in the wild," and then asking users how they felt in order label the reactions. Learning individual models would also help improve error recognition accuracy as we noticed more consistency in reactions within participants, rather than between participants.

6.6 Priming

The final reflection we had on our method is that we could have primed participants to believe that the machine was looking for visual cues in order to recognize its own errors. Previously cited research from Fujie et al. and Sidner et al. had found that participants who knew that robots were responding to their head nods would nod their heads more [28, 62]; by priming the participants, we might help them to become more expressive and responsive when interacting with the voice agent. More pronounced visual signals could help increase prediction accuracy. Another experiment to measure the value of visual cues could compare error recognition in clips with only sound or only video against clips with sound and video. The combination of audio and visual cues might be a lot more useful than either set of cues alone.

Finally, it is also possible that people interacting with a machine that interactively performs self-repair would be more emotive, and more pointedly make facial displays that reflect positive or negative feedback, making the tasks of predicting easier.

6.7 Ethical Considerations

We recognize that both the collection of people's faces and the analyzing of their expressions, can introduce a wide set of ethical concerns, for example, about privacy infringements [15, 17], about giving personal information to companies [37], and even about exposing people to risk of deep fakes [39] built with their likenesses. Here, we mention the ethical considerations and implications that we encountered in this work.

6.7.1 Privacy. One of the challenges of our dataset and other computer-vision datasets is that they are based on videos of actual people. Our videos were of participants who had gone through consent review, were offered incentives that were in line with fair wage compensation, and had given permission for their images and likenesses to be used for a machine learning dataset.

Because our dataset includes the voices and faces of our participants, obviously personal identifiable information [40], we are restricting our distribution of the dataset to researchers who receive an internal review board (IRB) clearance in advance, acknowledging limitations on the use of the videos for machine learning and interaction research purposes. We acknowledge the privacy, security and trust issues that arise from cloud computing in general, and facial recognition in specific [41, 64, 67]. We do not advocate (nor do we find practical) the large-scale transmission and collection of face image data from people's personal environments to companies with voice-enabled products. It is our intent that our research eventually be built into algorithms and methods that enable real-time error recognition and response on the local devices in people's pockets, cars and homes.

6.7.2 Fairness and Transparency for Crowdwork. As mentioned previously, we recruited Mechanical Turk workers from the United States, and paid crowdworkers who performed our original classification task (*YES*/*NO*) 1.80 USD, and our second classification task (*IDLE*/*OMIT*) 0.90 USD. Based on our pilot testing of the task, we estimated that the first task should take 12 minutes, and the second 6 minutes. The equivalent hourly wage for the work is 9.00 USD, above the US Federal minimum wage [33, 63]. The actual mean time on tasks was 10.8 minutes for the first task and 5.2 minutes for the second task. We approved payment for all completed tasks, regardless of quality of response.

7 CONCLUSION

This work marks the beginning of a much larger body of research to be performed in error-recognition and self-repair for human-machine interaction. We gathered videos of humans interacting with a voice agent, created clips to include the participant's reaction, and labeled them by type of reaction. We have verified this method as a way of moving towards developing classifiers and predictors of errors, and offer our dataset for others would be interested in this foundational interaction problem.

Moving forward, based on this experience, we have identified complexities in error classification, hedging, repair, priming and scaling of method as key new directions for research in this domain.

Finally, we aim to apply this research towards applications of people interacting with speech devices in naturalistic settings. Although some voice agent applications can easily have a fixed camera position, like voice agents in cars, future speech-enabled agents might be appliances or mobile robots in people's living or working quarters without a consistent field of view of their user(s). It is important to explore how error-recognition and self-repair plays out in these contexts.

Today, we interact with incredibly smart machines with no common sense about interaction. The ability to negotiate conversation–not only to model the words that people are saying correctly, but to know if the conversation is going well or not–is a skill that people do exceedingly well, and one that has long been out of reach for machines. The ability to use the multiple channels and the interactive strategies that people use to negotiate meaning can make interactions with voice agents less stilted and frustrating. This research moves us in the direction of being able to create more natural human-machine interactions.

ACKNOWLEDGMENTS

This research was deemed to be IRB exempt by Cornell University. We obtained consent for collection and dissemination of audio-visual data for the purposes of machine learning from each participant. We would like to thank David Goedicke, Jofish Kaye, and Chinmayi Dixit for their considerable help and advice, and the Mozilla Foundation for their financial support of this project.

REFERENCES

[1] Amazon.com. 2014. Alexa. Virtual Assistant.

[2] Amazon.com. 2015. Amazon Echo. Smart Speaker.

[3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[4] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 25–32.

[5] Apple Inc. 2011. Siri. Virtual Assistant.

[6] Apple Inc. 2017. Homepod. Smart Speaker.

[7] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker, and Elmar Nöth. 2003. How to find trouble in communication. *Speech communication* 40, 1-2 (2003), 117–143.

[8] Ray Birdwhistell. 1970. Kinesics and context. *Essays on body-motion communication. Philadelphia* (1970).

[9] Dan Bohus, Brian Langner, Antoine Raux, Alan W Black, Maxine Eskenazi, and Alex Rudnicky. 2006. Online supervised learning of non-understanding recovery policies. In *2006 IEEE Spoken Language Technology Workshop*. IEEE, 170–173.

[10] Dan Bohus and Alexander I Rudnicky. 2005. Sorry, I didn't catch that!-An investigation of non-understanding errors and recovery strategies. In *6th SIGdial workshop on discourse and dialogue*.

[11] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. 2013. Towards the automatic detection of spontaneous agreement and disagreement based on non-verbal behaviour: A Survey of related cues, databases, and tools. *Image and vision computing* 31, 2 (2 2013), 203–221. https://doi.org/10.1016/j.imavis.2012.07.003 eemcs-eprint-24491.

[12] Cynthia Breazeal and Brian Scassellati. 1999. How to build robots that make friends and influence people. In *Intelligent Robots and Systems, 1999. IROS'99. Proceedings. 1999 IEEE/RSJ International Conference on*, Vol. 2. IEEE, 858–863.

[13] Ivan Bretan, Anna-Lena Ereback, Catriona MacDermid, and Annika Waern. 1995. Simulation-based dialogue design for speech-controlled telephone services. In *Conference Companion on Human Factors in Computing Systems*. ACM, 145–146.

[14] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. *Embodied conversational agents*. MIT press.

[15] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, and Julie A. Kientz. 2011. Living in a Glass House: A Survey of Private Moments in the Home. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. ACM, New York, NY, USA, 41–44. https://doi.org/10.1145/2030112.2030118

[16] Nicole Chovil. 1991. Social determinants of facial displays. *Journal of Nonverbal Behavior* 15, 3 (1991), 141–154.

[17] H. Chung, M. Iorga, J. Voas, and S. Lee. 2017. "Alexa, Can I Trust You?". *Computer* 50, 09 (sep 2017), 100–104. https://doi.org/10.1109/MC.2017.3571053

[18] Herbert H Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of memory and language* 50, 1 (2004), 62–81.

[19] Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science* 13, 2 (1989), 259–294.

[20] Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.

[21] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. 2000. A new emotion database: considerations, sources and scope. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

[22] Sidney K D'mello and Arthur Graesser. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction* 20, 2 (2010), 147–187.

[23] The Economist. 2017. Terry Winograd: Where Humans still Beat Computers. *The Economist* (Jan 2017).

[24] Paul Ekman. 1976. Pictures of facial affect. *Consulting Psychologists Press* (1976).

[25] Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica* 1, 1 (1969), 49–98.

[26] Karën Fort, Gilles Adda, and K Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics* 37, 2 (2011), 413–420.

[27] Alan J Fridlund, Paul Ekman, and Harriet Oster. 1987. *Facial expressions of emotion.* Lawrence Erlbaum Associates, Inc.

[28] Shinya Fujie, Yasuhi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi. 2004. A conversation robot using head gesture recognition as para-linguistic information. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on.* IEEE, 159–164.

[29] Petra Gieselmann. 2006. Comparing error-handling strategies in human-human and human-robot dialogues. In *Proc. 8th Conf. Nat. Language Process.(KONVENS). Konstanz, Germany.* 24–31.

[30] Google. 2016. Google Assistant. Virtual Assistant.

[31] Google. 2016. Google Home. Smart Speaker.

[32] Google. 2017. Google Home Mini. Smart Speaker.

[33] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 449.

[34] Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* 87–95.

[35] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology* 6 (2015), 1049.

[36] Chien-Ming Huang and Bilge Mutlu. 2013. The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior. *Journal of Human-Robot Interaction* 2, 2 (2013), 80–102.

[37] James Jacoby. 2018. The Facebook Dilemma. https://www.pbs.org/wgbh/frontline/film/facebook-dilemma/

[38] Kleomenis Katevas, Patrick G. T. Healey, and Matthew Tobias Harris. 2015. Robot Comedy Lab: experimenting with the social dynamics of live performance. *Frontiers in Psychology* 6, August (2015), 1–9. https://doi.org/10.3389/fpsyg.2015.01253

[39] Will Knight. 2017. AI algorithms are creating a frighteningly realistic fake future. https://www.technologyreview.com/s/604270/real-or-fake-ai-is-making-it-very-hard-to-know/

[40] Gina Kolata. 2019. Your Data Were 'Anonymized'? These Scientists Can Still Identify You. https://www.nytimes.com/2019/07/23/health/data-privacy-protection.html

[41] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 102.

[42] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Emotional Triggers and Responses in Spontaneous Affective Interaction: Recognition, Prediction, and Analysis. *Transactions of the Japanese Society for Artificial Intelligence* 33, 1 (2018), DSH–D_1–10. https://doi.org/10.1527/tjsai.DSH-D

[43] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on.* IEEE, 94–101.

[44] Daniel McDuff, Rana El Kaliouby, Evan Kodra, and Rosalind Picard. 2013. Measuring voter's candidate preference based on affective responses to election debates. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on.* IEEE, 369–374.

[45] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. 2013. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 881–888.

[46] Microsoft Corporation. 2013. Microsoft Cortana. Virtual Assistant.

[47] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid robots, 2006 6th IEEE-RAS international conference on.* Citeseer, 518–523.

[48] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 2 (2012), 12.

[49] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction.* ACM, 61–68.

[50] Katashi Nagao and Akikazu Takeuchi. 1994. Speech dialogue with facial displays: Multimodal human-computer conversation. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 102–109.

[51] Jakob Nielsen. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 373–380.

[52] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in Virtual Agents and Robots. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (2017), 1–40. https://doi.org/10.1145/2912150

[53] Devi Parikh and C Lawrence Zitnick. 2010. The role of features, algorithms and data in visual recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* IEEE, 2328–2335.

[54] Tomislav Pejsa, Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2015. Gaze and attention management for embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 1 (2015), 3.

[55] Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. 2006. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications* 9, 1 (2006), 58–69.

[56] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus. 2017. Correcting robot mistakes in real time using EEG signals. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. 6570–6577. https://doi.org/10.1109/ICRA.2017.7989777

[57] Emanuel A Schegloff. 1997. Practices and actions: Boundary cases of other-initiated repair. *Discourse processes* 23, 3 (1997), 499–545.

[58] Emanuel A Schegloff. 1997. Third turn repair. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4* (1997), 31–40.

[59] Emanuel A Schegloff. 2000. When'others' initiate repair. *Applied linguistics* 21, 2 (2000), 205–243.

[60] Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53, 2 (1977), 361–382.

[61] Pradeep Shenoy and Desney S Tan. 2008. Human-aided computing: utilizing implicit human processing to classify images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 845–854.

[62] Candace L Sidner, Christopher Lee, Louis-Philippe Morency, and Clifton Forlines. 2006. The effect of head-nod recognition in human-robot conversation. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 290–296.

[63] M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 3 (2018), 39–41.

[64] Olivia Solon. 2019. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

[65] Bernhard Suhm, Josh Bers, Dan McCarthy, Barbara Freeman, David Getty, Katherine Godfrey, and Pat Peterson. 2002. A comparative study of speech in the call center: natural language call routing vs. touch-tone menus. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 283–290.

[66] Thales Teixeira, Rosalind Picard, and Rana El Kaliouby. 2014. Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study. *Marketing Science* 33, 6 (2014), 809–827.

[67] Jennifer Valentino-devries, Natasha Singer, Michael H. Keller, and Aaron Krolik. 2018. Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret. https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html

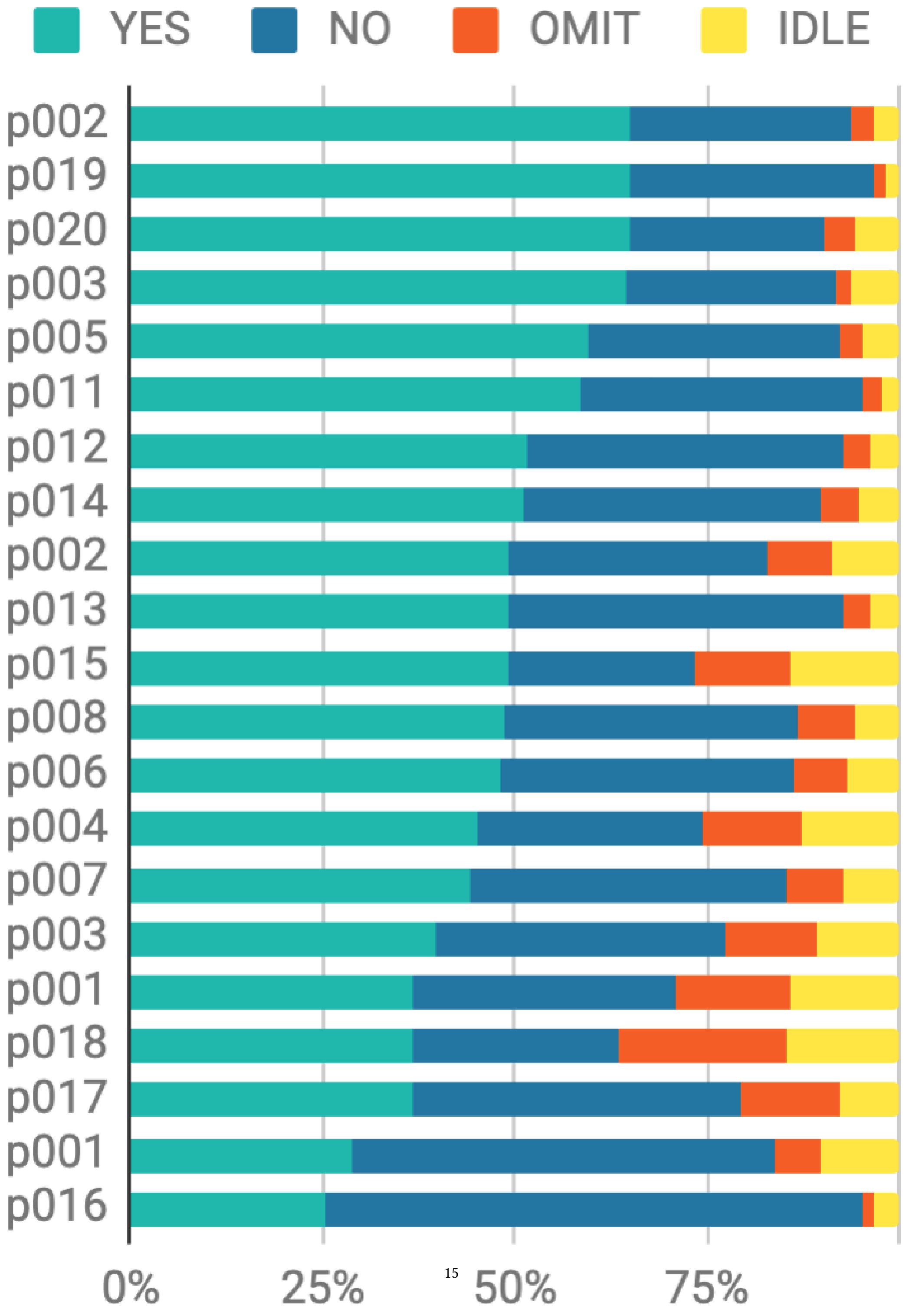

Fig. 3. Prevalence of labels by participant.

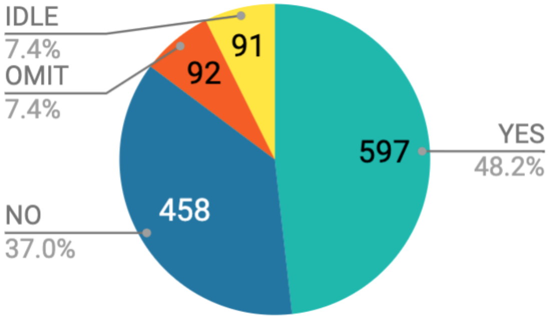

Fig. 4. Out of 1147 reaction clips, Alexa answered correctly (YES) in 597, incorrectly (NO) in 458, and did not respond (OMIT) in 92. We additionally clipped 91 non-reaction (IDLE) clips.