Multimodal Analysis of Vocal Collaborative Search: A Public Corpus and Results

Daniel McDuff
Microsoft
Redmond, WA, USA
damcduff@microsoft.com

Paul Thomas
Microsoft
Canberra, Australia
pathom@microsoft.com

Mary Czerwinski
Microsoft
Redmond, WA, USA
marycz@microsoft.com

Nick Craswell
Microsoft
Bellevue, WA
nickcr@microsoft.com

ABSTRACT

Intelligent agents have the potential to help with many tasks. Information seeking and voice-enabled search assistants are becoming very common. However, there remain questions as to the extent by which these agents should sense and respond to emotional signals. We designed a set of information seeking tasks and recruited participants to complete them using a human intermediary. In total we collected data from 22 pairs of individuals, each completing five search tasks. The participants could communicate only using voice, over a VoIP service. Using automated methods we extracted facial action, voice prosody and linguistic features from the audio-visual recordings. We analyzed the characteristics of these interactions that correlated with successful communication and understanding between the pairs. We found that those who were expressive in channels that were missing from the communication channel (e.g., facial actions and gaze) were rated as communicating poorly, being less helpful and understanding. Having a way of reinstating non-verbal cues into these interactions would improve the experience, even when the tasks are purely information seeking exercises. The dataset used for this analysis contains over 15 hours of video, audio and transcripts and reported ratings. It is publicly available for researchers at: http://aka.ms/MISCv1.

CCS CONCEPTS

• **Computing methodologies** → *Intelligent agents*;

KEYWORDS

Multimodal; search; dataset; agent

ACM Reference Format:
Daniel McDuff, Paul Thomas, Mary Czerwinski, and Nick Craswell. 2017. Multimodal Analysis of Vocal Collaborative Search: A Public Corpus and Results. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3136755.3136813

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

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3136813

1 INTRODUCTION

Intelligent agents have great potential for making us more efficient. Voice-enabled search is becoming increasingly common and involves an intelligent agent performing information seeking tasks based on voice commands. However, voice-enabled search presents new challenges for human-computer interaction. For example, unlike in desktop search, the user typically has less visual feedback about the results of the task and relies on the agent to select the information that is appropriate. Research is needed to understand how the personality and emotional cues between the searcher and the agent influence how people view the agent and the value of the assistance it provides.

Prior work has analyzed how non-verbal signals impact bonding [14], encourage more engaging conversations [6] and increase rapport [11]. However, in the specific context of voice-enabled assistants (such as Cortana, Siri and Alexa) there remain questions about how to design an agent that interprets and expresses emotion. Especially, what value should be placed on visual cues in this context and how damaging is it if an agent cannot "see" and express these cues?

We created an experiment to mimic complex voice search tasks in order to understand how affective signals can be used to improve information retrieval. We designed a set of information seeking tasks and asked participants to complete these using a human intermediary. The participants were only able to communicate via an audio channel. This setup was designed to mimic the interactions one might have with a voice-enabled software agent. A total of 22 pairs completed five information seeking tasks resulting in over 15 hours of data. We extracted numerous audio-visual features from the recordings using a set of automated methods (see Figure 1 for an overview of the features extracted from the audio-visual data).

We present analyses of this rich multimodal non-verbal and verbal information. Specifically, we examine how each set of features relates to measures of how helpful, understanding and communicative a partner was during each task. The Microsoft Information-Seeking Conversation (MISC) dataset contains all the data collected in this study. The videos and audio recordings, reported rating measures and the subsequent facial coding data, audio features and transcripts are publicly available as part of this corpus. The

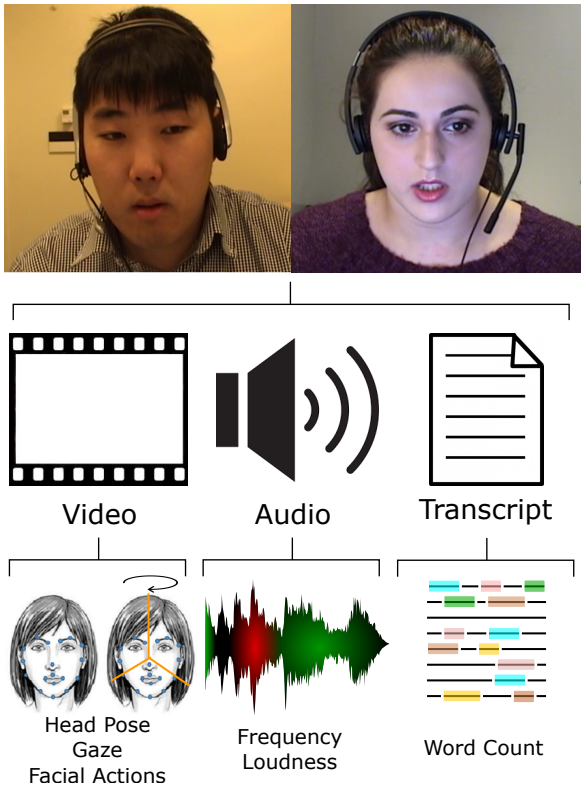

Figure 1: We analyzed verbal and non-verbal features of interactions with an intelligent agent in order to identify what characteristics are correlated with success.

download instructions are available at: http://aka.ms/MISCv1. We hope that access to this data will help researchers further improve the state-of-the-art in information-seeking agents.

2 AFFECTIVE SIGNALS

The face is one of richest sources of affective information for humans. Although the participants were communicating via an audio link in our study, their facial responses were very rich. We used automated facial coding to extract moment-to-moment measures of their facial activity. The Facial Action Coding System (FACS) [8] provides an objective taxonomy for coding the face and is extremely useful for quantitative measurement of emotional and social signals [9].

Non-verbal speech also contains significant affective information [15]. Automated analysis can be used to code characteristics such as loudness and pitch of the voice and extract estimates of valence and arousal [12]. There is not a standardized taxonomy for coding speech, as is the case with facial coding, but public software tools [10] support transparent and repeatable analysis using well defined features.

Linguistic patterns and word choice can also be very informative of a person's affective state and intentions. Linguistic style typically synchronizes between individuals during interactions [19]. Speech-to-text (STT) tools enable automatic language transcription from audio signals. The Linguistic Inquiry and Word Count program (LIWC [20]) is a tool for text analysis that captures and measures many characteristics of speech, including positive and negative sentiment and functional word usage. Combining STT and LIWC provides an efficient way to code linguistic patterns from audio inputs.

All the analyses described above can be extracted from audio-visual recordings collected via an off-the-shelf camera and microphone. The result being that we can create natural experimental protocols that do not require the participants to wear uncomfortable and obtrusive contact sensors. Furthermore, the collection of audio-visual data can be scaled using Internet frameworks [18].

3 BACKGROUND

3.1 Multimodal Analysis in Interactions

Affective responses are multimodal and researchers consistently find improvement in the automated understanding of nonverbal behavior by combining signals from numerous modalities (such as speech, gestures and language). A meta-analysis of 30 studies showed that multimodal classification led to better results than the best unimodal alternative [7]. Dyadic interactions have been studied in a number of contexts including, mental health [23], remote collaboration [22], in-car systems [1] and learning [16]. Multimodal datasets can contain many different types of features visual, audio, physiological, self-reported, language.

For a current review of multimodal interaction research see [25]. A helpful survey of machine learning methods for multimodal interaction was published by Baltrusaitis et al. [4].

The RECOLA project [22] used a similar paradigm to ours, having participants complete collaborative tasks over a web link. However, while theirs was a video conference, we intentionally required the participants to interact only via an audio channel, thus simulating the interaction with a voice agent. We believe our MISC dataset will act as a complement to the valuable RECOLA dataset for researcher wishing to compare the two scenarios. By comparison the RECOLA dataset features 47 participants (compared to 44 in MISC) and 3.5 hours of recordings (compared to 15.3 hours in MISC).

3.2 Affect and Search

Picard [21] identified that there were applications of affective computing in information retrieval. Given that emotions influence memory and decision-making, it is natural to hypothesize that agents performing information retrieval tasks will be more effective if they can respond to the affective state of the user.

Arapakis et al. [2, 3] found that affective signals (i.e., facial expression and peripheral physiological measurements) were useful in determining topic relevance in information search tasks. However, the physiological signals were captured through obtrusive contact devices. Thus this approach does not scale well or naturally extend to real-world applications. While we use search-based tasks in this study, we believe that the results are likely to be generally applicable to many cooperative tasks.

Table 1: Audio-visual features extracted from the recordings of the seeker and searcher pairs during each task.

Channel	Modality	Features	# Features
Video	Head Pose and Gaze	Head rotation and displacement and eye gaze means and standard deviations.	24
	Facial Actions	Facial action output means and standard deviations.	36
Audio	Nonverbal Speech	Pitch (f0) and loudness means and standard deviations.	4
	Language	Linguistic inquiry word counts.	65
All			129

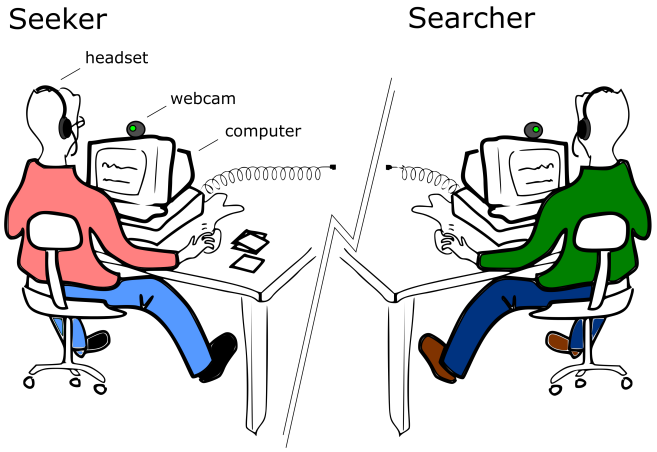

Figure 2: Participants were recruited to complete information seeking tasks via a human intermediary. The participants could only communicated by voice. Audio-visual recordings were made of both participants while they completed the tasks.

4 DATASET

4.1 Participants

Twenty-two pairs of participants (N=44, 24 females; 24 to 65 years) were recruited to complete a set of information seeking tasks. The participants did not know each other and were randomly assigned to their partner. The participants were all fluent English speakers with self-reported experience using Internet search engines.

4.2 Apparatus

Videos of the participants were recorded using a Sony EVI camera at 30 frames-per-second (FPS) and a resolution of 700×900 pixels. The audio from each participant was recorded separately as a single-channel 48 kHz signal. The videos are stored in WMV format and the audio as WAV files. The audio and video data is available for research. More information about the distribution of the dataset is given at the end of the paper.

4.3 Methods

For each pair, one of the participants was randomly assigned as the information "seeker" and the other as the information "searcher". The participants were given instructions that they would complete five tasks and needed to work together but would only have an audio link to communicate. They completed the experiment in different rooms on desktop computers (see Figure 2). The participants completed survey questions at the start of the experiment, following each search task and at the end of the experiment. The "seekers" did not have access to any on-line resources to complete the tasks, they were only allowed to use the computer to record their answers and respond to the survey questions. The list of survey questions can be found in the documentation of the MISC dataset [24]. The participants were compensated with $150 in the form of a gift-card for taking part in the study.

Pre-survey Questions: Prior to completing the search task both the "seeker" and "searcher" completed a survey featuring the Positive and Negative Affect Schedule (PANAS) [26] and "Big Five" personality traits questions. These provide valuable context within which to interpret the affective responses observed during the tasks.

Post Task Questions: Following each search task the participants completed questions from the NASA TLX [13] and reported the emotions they experienced during the task. Participants also completed questions from the User Engagement Scale, and the three questions used as dependent variables described in Section 6.

Post-survey Questions: Following all the search tasks and post task questions the participants ranked the tasks in order of difficulty and noted things they liked and did not like about searching with another human.

Further details of methods are available in the description by Thomas et al. [24].

4.4 Search Tasks

A set of five search tasks were designed in order to reflect a range of difficulties (how hard the information is to find) and complexities (how many steps are required to find the information). The task difficulty and complexity was verified by a pilot test. The difficulty was further verified by the responses of the participants, who ranked the difficulty of the tasks at the end of the experiment. Below are the exact descriptions of the task provided to the participants:

HPV Vaccine: Mary has been hearing a lot about the HPV vaccine, a vaccine that protects against several types of the human papillomavirus, a common sexually transmitted infection (STI). Mary is considering getting the vaccine. Using the Internet, find out who can get the HPV vaccine.

Heroic Acts: Recently you had dinner with your cousin. She is very cynical and kept telling you that nobody ever helps others unless there's something in it for them. You'd love to prove her wrong,

so you want to find accounts of selfless heroic acts by individuals or small groups for the benefit of others or for a cause.

Treating Migraines: Imagine that you recently began suffering from migraines. You heard about two possible treatments for migraine headaches, beta-blockers and/or calcium channel blockers, and you decided to do some research about them. At the same time, you want to explore whether there are other options for treating migraines without taking medicines, such as diet and exercise.

Olympic Venues: For a work project you're looking at international sport in the developing world. You're making a list of Olympic host cities to see how well different areas are represented. Find the venues of the 2024 Olympic Games and the 2016 Winter Olympic Games.

Summer Transportation: This summer, during your vacation, you are planning to go on a touring trip of North America. You want information to help you plan your journey and there are many tourist attractions you would be interested in visiting. You have set aside 3 months for the trip and hope to see as much of the continent as you can. As you cannot drive, you will have to use public transport, but are unsure which type to take. Task: Bearing in mind this context, your task is to decide on the best form of transportation between cities in North America that would be suitable for you.

If the participants had not completed or submitted their answer after 10 minutes working on a task they were asked to move on. The average duration of the tasks was 8 minutes 20 seconds (standard deviation: 2 minutes 29 seconds). On 42% of occasions they reached the 10 minute limit, showing that the tasks were not trivial and required several steps to find the necessary information. Note: The Olympic Venues question was intentionally designed to be difficult in that there was no 2016 Winter Olympics, and the 2024 Olympic venue had not been decided.

5 AUTOMATED CODING

Table 1 provides a summary of the features extracted from the audio-visual data. The following section describes how these features were extracted, normalized and fused. The automatically coded facial features, prosodic features and linguistic word count features are also provided to researchers in our public dataset.

5.1 Facial Features

Automated facial coding was performed using OpenFace [5]. Three axis displacement and three axis rotation of the head pose was extracted for each frame. Three axis translation of gaze was also calculated for each frame, for each eye. Outputs for 18 facial action units were calculated for each frame, each a continuous value from 0 to 1. From each of these features we calculated the mean and standard deviation resulting in 60 features per participant per task (6×2 pose, 6×2 gaze, and 18×2 facial actions).

5.2 Voice Features

The audio channels from the video recordings were striped and processed to extract non-verbal speech features. We used the audio feature extractor from openSMILE [10]. We extracted fundamental frequency (F0) and PCM loudness features sampled at 100Hz from the audio signals. We calculated the mean and standard deviation of the F0 and loudness values resulting in four (2×2) features.

5.3 Language Features

Transcriptions were made of the interactions to allow us to capture linguistic features. This was first performed using an automated speech-to-text (STT)[1]. Following this we used the LIWC software [20] to extract linguistic and word count features from the transcripts for each task. The linguistic features are output as scores and represent things such as the frequency of positive words or how social the vocabulary is. This resulted in 65 features. Automated STT transcription may result in some errors; however, we found that overall the transcripts reflected the conversation well. Manual annotation would have been much more laborious and problematic for end-to-end automation.

5.4 Feature Fusion

The facial, voice and language features have different ranges and units of measurement. Therefore, they need to be normalized in order to be fused effectively and used in the same analysis.

For each set of features we take the absolute mean across the five tasks to form what we term "base rates" for each participant for our analysis. We then take the mean of the base rates across the two participants in a given pair to provide a measure of the magnitude of each feature for that set of interactions.

Following this we normalize the values for each feature across all the pairs such that they fall within the range 0 to 1. This gives each of the features equal importance.

Finally, we take a mean of all the features for the pair that results in one value per pair that broadly reflects their overall expressiveness during the tasks. We performed separate analyses for each modality (visual, audio and language) and all the features combined, in each case we summarize and normalize the features as above.

6 CLASSIFICATION OF SUCCESSFUL INTERACTIONS

As there are a number of ways one might characterize successful pairings, we performed multiple tests. The target labels in these analyses were based on the reported experience of the seeker and searcher. Following each task the participants were asked to rate the other participant using the following Likert scale questions:

- The other participant **helped me work on this task**.
- The other participant **understood what I needed**.
- The other participant **communicated clearly**.

For each they responded on a scale of 1 (not at all) to 7 (a lot). We use the average of each score across the five tasks and the two participants as the dependent variables in our subsequent analysis. This yields three scores per pair which we call the: *help* score, *understanding* score and *communication* score respectively from now on.

We do not analyze the responses of the other survey questions in this paper; however, the wordings of all the other questions asked during the study are provided in the dataset.

7 RESULTS

We perform a correlation analysis between the expressiveness measures and the three reported rating scores. We report Spearman's

[1] https://www.microsoft.com/cognitive-services/en-us/speech-api

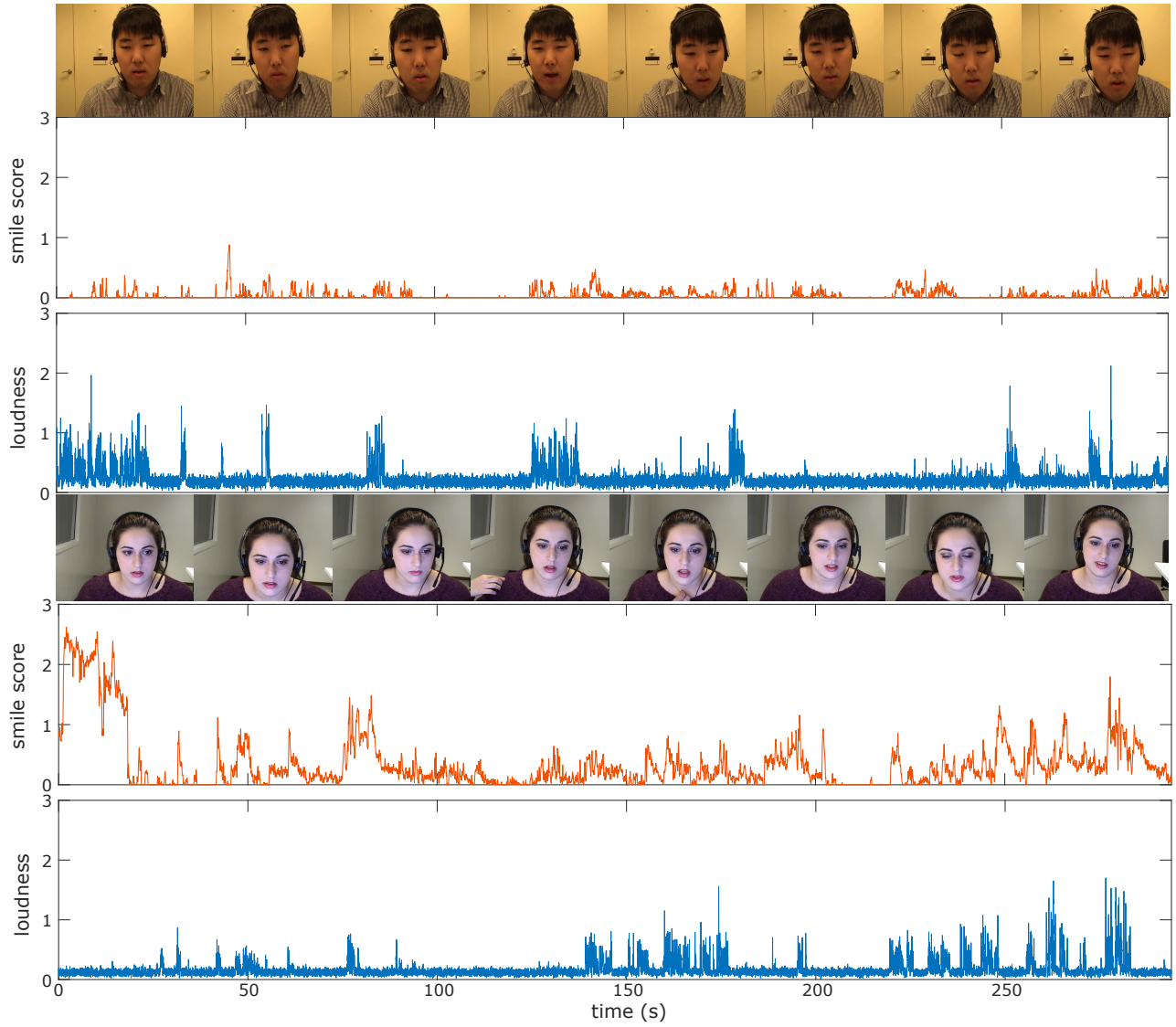

Figure 3: Examples of recordings for one of the tasks. The searcher's (top) and seeker's (bottom) smile responses (orange) and loudness (blue) are shown. Thumbnail example frames from the video recordings illustrate the set-up.

rank correlation coefficients. Spearman's correlation is suitable for both continuous and discrete ordinal variables. Table 2 shows the correlations between expressiveness and the *help*, *understanding* and *communication* scores. Figure 4 shows the data graphically. All three scores were negatively correlated with our measure of expressiveness (help: -0.66, p = 0.001; understanding: -0.52, p = 0.016; communication: -0.39, p = 0.082). To help us understand this relationship, we performed a similar analysis of each modality separately and thus reveal the contribution of each to this result.

Table 3 shows the correlations between each of the expressiveness measures for the three modalities and the three scores. Figure 5 shows the data graphically for the help score. Analysis with each modality separately revealed that the visual channels (facial actions and gaze) were significantly correlated with success (facial actions: -0.58, p = 0.006; head pose and gaze: -0.69, $p < 0.001$). The nonverbal speech and language features were not significantly correlated with the score (p = 0.23 and 0.44 respectively). This result was consistent across all three of the scores.

To shed more light on whether it was expressions of positive or negative affect that were primarily driving the negative relationship between facial actions and the reported rating measures, we divided the action units into those with primarily negative valence (AUs 1, 4, 5, 7, 9, 10, 15, 17, 20) and those with primarily positive valence (AUs 6, 12). This was based on a large prior study of facial actions and their relationship with emotional valence [17].

The negative valence facial actions were the main contributors to the relationship (negative valence facial actions correlation with

Table 2: Correlations between verbal and nonverbal measures of expressiveness and reported help, understanding and communication scores.

Score	r	*p*
Understanding	-0.52	0.016
Help	-0.66	0.001
Communication	-0.39	0.082

Table 3: Correlations between verbal and nonverbal measures of expressiveness separated by modality and reported help, understanding and communication scores.

Channel	Features	Help r (*p*)	Understanding r (*p*)	Comm. r (*p*)
Video	Pose & Gaze	-0.69***	-0.63***	-0.45*
	AUs	-0.58***	-0.67***	-0.62***
Audio	Prosody	NS	NS	NS
Linguistic	Word Count	NS	NS	NS

NS = Not significant, ^<0.1, *<0.05, **<0.01, ***≪0.01.

helped: -0.70, $p \ll 0.01$). The relationship between the positive valence actions and the responses for the three questions was weaker and only mildly significant for one of the questions (*understood*: p=0.05), (*helped*: p=0.33, *communicated*: p=0.14).

Finally, one might question whether the features from each mode correlated with those from other modes, i.e., are people with lots of facial actions also those who speak more. Only the gaze and AU features were correlated (0.45, p=0.042).

8 DISCUSSION

The results show a negative correlation between expressiveness and successful communication between the pairs that is driven primarily by features in the visual channels (facial actions and gaze). The participants in our experiment were relying on the audio channel alone to communicate and thus these expressions were not seen by the other person.

The results could be interpreted in several ways. First, it is possible that those that rely more on facial expressions and head gestures for communication are rated as poorer communicators when the visual channel is not present. As the relationships between language and voice were not significantly correlated with success it does not seem that other participants are relying on the audio channel more, they are just not missing important visual cues.

Second, when there is no visual channel, some people may express more negative behaviors when the interaction is going poorly, perhaps because they do not feel the need to mask these and desire to be polite. Overall, the participants were more likely to modulate their language and tone of voice as they knew the other participant could hear them at all times during the experiment.

Whatever the reason for the correlations, the results reinforce the importance of the need for agents that can *see* as well as *hear* someone to enable the most effective communication. If these signals are absent the agent cannot try to moderate its behavior when the interactions are going poorly. The results suggest that visual feedback from the agent could also be important, though this remains as future work.

It is logical that nonverbal communication and language features would be related to ratings of the clarity of communication. However, there were significant correlations with all three scores, adding confidence that these reported measures are consistent with one another and captured how helpful the other participant was, in addition to how well they communicated.

We should note that as the participants were speaking this may have led to more false positives from the facial action classifiers, and in turn contributed to the correlation that we observed (i.e., people who talked more were rated as poorer communicators, helping less and being less understanding). However, the fact that average loudness (that would be higher with more talking) did not correlate strongly with the measures suggests this was not the case.

Finally correlation between non-verbal measures and the reported ratings does not inform us of the causation between these variables. Further analyses would be necessary to determine causation.

9 DISTRIBUTION OF DATA

Participants provided informed consent for use of their audio-visual recordings for scientific research purposes. Distribution of the dataset is governed by the terms of their informed consent. The data may be used for research purposes. Approval to use the data does not allow recipients to redistribute it and they must adhere to the terms and confidentiality restrictions. The details can be found at: http://aka.ms/MISCv1. This data is available for distribution to researchers online.

10 CONCLUSION

Intelligent agents have the potential to help with many tasks. We designed a set of information seeking tasks and recruited participants to complete these tasks using a human intermediary. In total, we collected data from 22 pairs of individuals each completing five search tasks, resulting in over 15 hours of rich multimodal data. This was supplemented with self-reported survey questions. Using automated methods we extracted facial action, voice prosody and linguistic features from the audio-visual recordings. The data is publicly available for researchers as part of the MISC dataset.

We studied the relationship between the nonverbal and verbal features and three measures of successful interaction. Each participant rated their partner on how much they *helped* them, *understood* them and how clearly they *communicated*. We found that expressiveness was negatively correlated with success across all three scores and that this was driven by the visual cues. We interpret this as evidence that those that rely on facial expressions and gestures more for communication are rated as poorer communicators and less helpful when these channels are absent. In addition, participants in less successful interactions may have expressed negative behaviors visually as they knew the other participant could not see them.

Designing successful agents that interpret and express emotion probably requires visual cues, in addition to audio cues, even if the interactions are by voice alone. Our results show that for people who tend to express these emotional cues more, the impact of

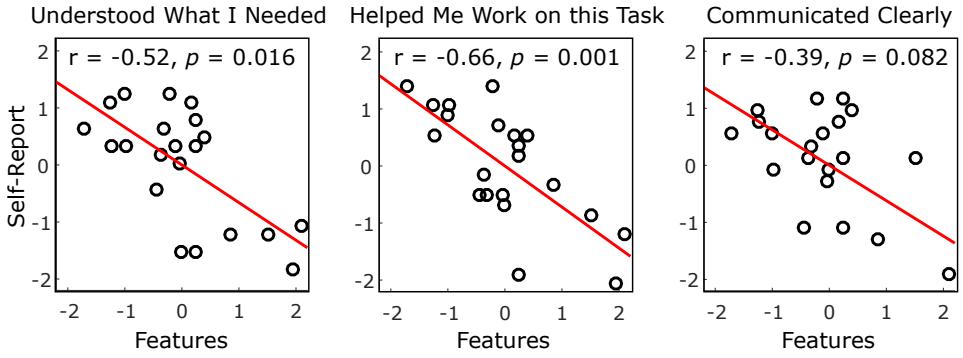

Figure 4: Relationships between the language, voice and facial features and the help, understanding and communication scores respectively. All were negatively correlated with the features suggesting that those that were more expressive were rated lower on all the measures. The features have been normalized to zero mean and unit standard deviation (z-transform).

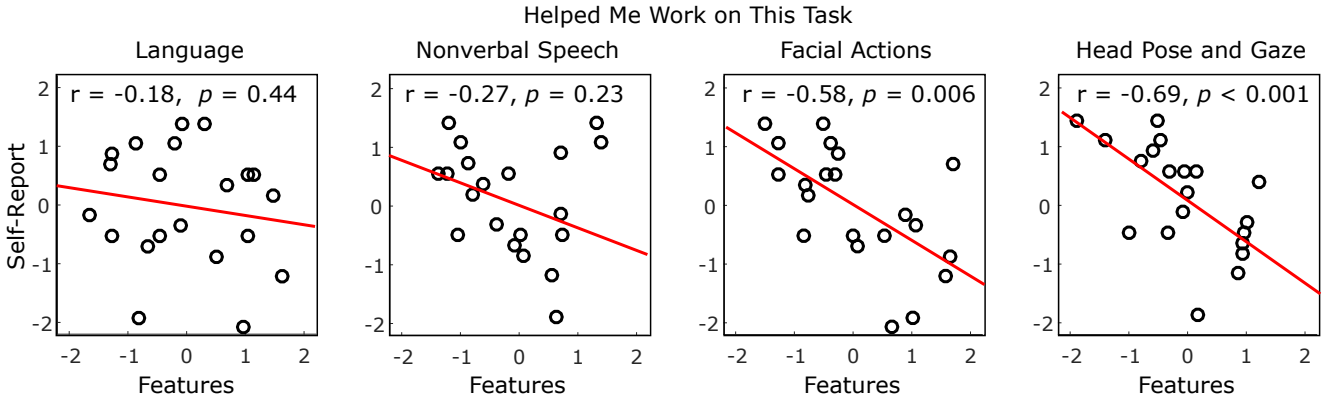

Figure 5: Relationships between the features from each modality and the help score. Language and voice features were not correlated with the help score. Facial actions, head pose and gaze were strongly negatively correlated with the help score.

not having a visual channel for communication might be greater. Agents that lack these abilities are likely to be rated as less helpful in addition to being poorer communicators. However, testing this with an artificial agent remains as future work.

11 FUTURE WORK

There are many aspects of analysis that could be explored using the public MISC dataset. First, it could be that seekers and searchers reacted differently from one another. A study of synchrony between the individuals would be of great interest. The Big Five questionnaire data would allow one to analyze if synchrony is related to specific personality types, or personality pairings. Second, we looked for correlation between successful collaborations but predicting participants' responses would allow design of an agent that could make inferences about how helpful it appeared. Third, the self-report data has many measures beyond the three that we used (*helped*, *understood* and *communicated*). These could be used to test further hypotheses about how multimodal cues can be used effectively in voice-based interactions.

REFERENCES

[1] Irman Abdic, Lex Fridman, Daniel McDuff, Erik Marchi, Bryan Reimer, and Björn Schuller. 2016. Driver Frustration Detection from Audio and Video in the Wild. *KI 2016: Advances in Artificial Intelligence* (2016), 237.

[2] Ioannis Arapakis, Joemon M Jose, and Philip D Gray. 2008. Affective feedback: an investigation into the role of emotions in the information seeking process. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 395–402.

[3] Ioannis Arapakis, Ioannis Konstas, and Joemon M Jose. 2009. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 461–470.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal Machine Learning: A Survey and Taxonomy. *arXiv preprint arXiv:1705.09406* (2017).

[5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[6] Justine Cassell and Kristinn R Thorisson. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence* 13, 4-5 (1999), 519–538.

[7] Sidney D'Mello and Jacqueline Kory. 2012. Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 31–38.

[8] Paul Ekman, Wallace V Friesen, and John Hager. 2002. *Facial action coding system: A technique for the measurement of facial movement*. Research Nexus, Salt Lake City, UT.

[9] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, USA.
[10] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.
[11] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 125–138.
[12] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. 2017. Prediction-based learning for continuous emotion recognition in speech. In *42nd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2017*.
[13] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.
[14] Natasha Jaques, Daniel McDuff, Yoo Lim Kim, and Rosalind Picard. 2016. Understanding and Predicting Bonding in Conversations Using Thin Slices of Facial Expressions and Body Language. In *International Conference on Intelligent Virtual Agents*. Springer, 64–74.
[15] Tom Johnstone and Klaus R Scherer. 2000. Vocal communication of emotion. *Handbook of emotions* 2 (2000), 220–235.
[16] Ashish Kapoor, Winslow Burleson, and Rosalind W Picard. 2007. Automatic prediction of frustration. *International journal of human-computer studies* 65, 8 (2007), 724–736.
[17] Karim Sadik Kassam. 2010. *Assessment of emotional experience through facial expression*. Harvard University.
[18] Daniel Jonathan McDuff. 2014. *Crowdsourcing affective responses for predicting media effectiveness*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[19] Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21, 4 (2002), 337–360.
[20] James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: LIWC [Computer software]. *Austin, TX: liwc. net* (2007).
[21] Rosalind W Picard. 1997. *Affective computing*. Vol. 252. MIT press Cambridge.
[22] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–8.
[23] Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. 2013. Automatic behavior descriptors for psychological disorder analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–8.
[24] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2017. MISC: A data set of information-seeking conversations. In *Proc. Int. W'shop on Conversational Approaches to Information Retrieval*.
[25] Matthew Turk. 2014. Multimodal interaction: A review. *Pattern Recognition Letters* 36 (2014), 189–195.
[26] David Watson, Lee A Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.