Clustering and Multimodal Analysis of Participants in Task-Based Discussions

David Johnson
University of British Columbia, Canada

Gabriel Murray
University of the Fraser Valley, Canada

ABSTRACT

Participants in task-based conversational interactions are clustered using outcomes of interest that include task performance, satisfaction ratings, and demographic traits. Each cluster is described in terms of the member participants' common characteristics, and we perform participant outlier detection as well. We extract multimodal features of the conversational interaction and analyze how the participant groups differ in terms of these features.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; • **Applied computing** → *Sociology*.

KEYWORDS

clustering, multimodal interaction, meetings, group affect, survival task

ACM Reference Format:
David Johnson and Gabriel Murray. 2021. Clustering and Multimodal Analysis of Participants in Task-Based Discussions. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3461615.3485416

1 INTRODUCTION

Some recent work on group interaction has used multimodal features of a conversation which may be sometimes empirically chosen or hypothesized to be informative in order to automatically predict conversation outcomes or participant characteristics. This includes automatic prediction of group performance [3, 14] and prediction of group satisfaction or affect [15, 19], as well as work on predicting social, psychological, or personality traits from conversational data [1, 4, 5, 10, 12, 17]. The goal that is common to much of this research is to find a fusion of multimodal features that are effective for predicting a particular conversational outcome or participant trait.

In another vein of research by Avci and Aran [2, 4], groups in a survival-task are partitioned into lower-scoring and high-scoring cohorts, based on their group score on the task, and subsequently the cohorts are compared in terms of multimodal features and verbal content. In contrast to all of these related works, we are automatically clustering the groups based on a large space of outcome variables, including performance scores, satisfaction ratings, and demographic traits. We then compare and contrast the resulting clusters in terms of a large set of multimodal features. Although other works allow some understanding of which multimodal features may be helpful for predicting performance or predicting satisfaction, our work goes beyond to also allow an understanding of how the participants see themselves and see various dimensions of their group, and how this relates to their scored performance and their own stated satisfaction.

Part of our motivation for person-level clustering is based on previous research on personality types (e.g. [17]), with our intention being to explore how different types of people respond to group tasks, and to try to identify groups of people that are similar in their traits. In addition to these results being interesting in their own right from a social psychological perspective, understanding the relationships between a user's sense of self, their satisfaction, their performance, and their conversational behaviours may inform AI systems within the domain of group interaction; for example, it could be useful for an intelligent meeting assistant to be able to infer personality traits when monitoring a group interaction to help it determine what types of interventions to make, a strategy that has been used in the domain of intelligent tutoring systems [8]. Our analysis of multimodal features may also be informative for researchers working on predicting and explaining group performance and participant affect, e.g. by suggesting efficacious features for these predictive tasks. And as our work proceeds by analyzing a rich set of outcome variables jointly rather than individually, it may provide motivation for deploying joint task modeling when attempting to automatically predict these outcomes, e.g. through the use of joint task deep neural explainable models [11].

2 DATASET

Our data were extracted from the Group Affect and Performance (GAP) corpus [6]. The GAP corpus uses a winter survival task scenario in which 28 groups of participants (84 participants total, all of which are undergraduate students) are required to rank 15 items (eg. compass, cigarette lighter, chocolate bar, etc.) on a scale of usefulness for survival after a hypothetical plane crash. The task has the participants begin by individually ranking the items, before having participants work with others in their group to cooperatively rank the 15 items. After completing their group task, the participants each fill out a post-task questionnaire using a Likert scale intended to measure participant's satisfaction with the group task in which 1 is strongly disagree and 5 is strongly agree. The questionnaire measures the following:

- **Time Expectation (TE):** "This task took longer than expected to complete"

ICMI '21 Companion, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8471-1/21/10…$15.00
https://doi.org/10.1145/3461615.3485416

- **Worked Well Together (WWT):** "Our group worked well together"
- **Time Management (TM):** "Our group used its time wisely"
- **Efficiency:** "Our group struggled to work together efficiently on this task"
- **Quality of Work (QoW):** "Overall, our group did a good job on this task"
- **Leadership:** "I helped lead the group during this task"

Though some questions are negations (ie. TE, Efficiency) those questions are reverse scored so that in all of the resulting outcomes a score of 5 reflects the strongest positive view of the group task and 1 reflects the strongest negative view.

Additionally, the GAP corpus computes derived scores from the participant and group ranking performance as follows:

- **Satisfaction:** A combined average of TE, WWT, TM, Efficiency, and QoW
- **Absolute Individual Score (AIS):** A sum of the difference between the participant's item ranking and the item ranking of a survival expert
- **Absolute Individual Influence (AII):** A sum of the difference between the participant's item ranking and the group's cooperative item ranking

In both AIS and AII the lower the score, the better the performance (ie. A lower AIS indicates the participant scored closer to the expert's item ranking, and a lower AII indicates a participant's individual item ranking was more similar to the group's item ranking, suggesting the participant influenced the group).

The demographic characteristics available in the GAP corpus consist of age, gender, whether English is their first language, and level of education. All of the group conversations are in English.

The subjective ratings, performance scores, influence scores, and demographic traits are collectively referred to in this paper as the outcome variables. These outcome variables serve as the features of interest for the unsupervised clustering methods used in this work.

We also extract multimodal features from the corpus which we use to compare between clusters as described in section 3.2.

This corpus was used because there are only a small number of public survival task datasets containing both audio and video. These survival tasks have been used for decades in social psychology literature and more recently in the multimodal interaction community (e.g. [20, 21]), and are interesting tasks because they are complex in that they require much real world knowledge which is aggregated over participants, requiring them to work together, and objective performance scores are obtained at the end of the task.

3 METHODS

Here we describe the methods we used to cluster participants based on their outcome scores, as well as features and demographics we use to compare clusters. The results of these methods are presented in section 4.

3.1 Participant Clustering Methods

We use k-means clustering on the participant outcome variables to discover latent clusters of participants. We use the k-means implementation from scikit-learn [18] with N=5 for the number of clusters as determined by the "elbow method" [22]. The elbow method is a heuristic which involves fitting the model with a range for K, plotting the model's explained variance, and selecting the value of K in which there is an "elbow" in the plot indicating further increases in K do little to increase explained variance and may lead to overfitting.

3.2 Multimodal Features

After clustering participants using the set of outcome variables, we extract a set of multimodal features and compare features between clusters to identify conversational behaviours that differ across the previously identified groups. The multi-modal features include linguistic features, movement features, and speech features.

The linguistic features we extract are:

- **Pronoun usage:** First person (eg. "I", "my"), first person plural (eg. "we", "us"), second person (eg. "you", "your"), third person (eg. "he", "she"), third person plural (eg. "they", "them"). For each pronoun type, we calculate its use proportional to the total pronoun use for each participant.
- **Coordination:** Coordination was extracted using Convokit [7] and is a measure which reflects the tendency for those in lower power positions to change their speaking patterns to become more linguistically similar to those in higher power positions within groups [9]. While there are no assigned hierarchical roles in the GAP corpus, previous research using similar survival tasks has explored how some participants emerge as leaders of the group [20, 21].

In addition to pronoun use and coordination, the GAP corpus also has gold-standard labels for speech acts which we use as part of our cluster comparisons. The speech acts we are use are: proposals (eg. "So, I would say cigarette lighter is two"), agreements (eg. "Yeah, okay"), confirmations (eg. "Okay, so four is ball of steel wool."), and disagreements (eg. "I would put the chocolate bar above the newspaper.")

As speech features, we extract the average utterance duration among all participants over each cluster, and average utterances per participant over each cluster.

We chose these features because they have been previously shown to be indicative of interesting group task characteristics (e.g. linguistic coordination has been associated with group task performance and cohesion, average utterance duration has been associated with dominance, speech acts such as proposals have been associated with influence, etc.)[5, 6]

Additionally, we include four psycholinguistic features: concreteness (the degree to which a word refers to a tangible perceived entity), imageability (the degree to which a word draws a clear image in the mind), age of acquisition, and familiarity as they have been shown to be predictive of group task performance [16].

To facilitate a multimodal analysis, we also use movement features extracted from video recording of the meetings, which show the average individual movement for each participant in a cluster. Individual participant videos were recorded, converted to greyscale, and smoothed using Gaussian blurring. Individual images were analyzed and for each image the pixel intensity is compared with the intensity of a previous background image, which is regularly updated. We used a threshold to detect only significant amounts of

movement, and aggregated these over the whole video to get the total amount of movement by the participant.

4 RESULTS

In this section we look at the resulting clusters and how they differ from one another, as well as interesting multimodal feature results of both participants and clusters. In section 5 we interpret and analyze the resulting clusters.

4.1 Findings of Participant Clustering

Clustering participants with k-means produced five clusters total: four clusters with distinguishing outcome differences, and one cluster with only a single substantially outlying participant. Table 1 shows the differences in outcome values averaged over each cluster. Table 2 shows the differences in speech, linguistic, and movement features averaged over clusters.

Cluster 1 (N=26) had the best average scores in AII (36.88), TE (3.69), WWT (4.85), TM (5), and QoW (4.77) as well as being only slightly behind Cluster 4 for the best Efficiency score (4.85 Efficiency for Cluster versus 4.87 Efficiency for Cluster 4), all of which lead to the cluster also having the best average satisfaction score of 4.63.

Cluster 2 (N=22) had the worst AII score of 50.41, and had the second worst WWT (3.98), TM (3.61), QoW (3.7), and Satisfaction (3.56) scores (all of which are ahead of only the outlying Cluster 5 participant).

Cluster 3 (N=12) had the worst AIS (83.25), and second worst TE (1.75). This cluster had the best leadership score of 4.5.

Cluster 4 (N=23) had the best Efficiency (4.87), and the second best AIS (74.2), TE (3.39), WWT (4.48), and Satisfaction (4.32). Interestingly, this cluster had the worst leadership score of 2.61.

Cluster 5 (N=1) had a single participant which was a substantial outlier on most measures. The participant had both the best AIS of 56 and the worst AII of 67. The participant also had the worst scores on all of: TE (1), WWT (1), TM (2), Efficiency (2), QoW (1), and Satisfaction (1.4).

4.2 Multimodal Feature Results

In this section we analyze the relationships between outcome variables and the multimodal features.

We used the Kruskal-Wallis test to determine whether a feature's mean rank differences between groups (i.e. across clusters) were significantly different. Given that Cluster 5 had only a single outlying participant, we excluded Cluster 5 from the Kruskal-Wallis test as a Cluster with size 1 is not supported by the method. First using a significance level of 0.05 we were able to identify the mean rank differences between groups in the use of proposals ($H(3) = 8.68$, $P = 0.03$) and disagreements ($H(3) = 12.35$, $P = 0.006$) as being significant. If loosening the significance level to 0.1, concreteness ($H(3) = 6.36$, $P = 0.08$), familiarity ($H(3) = 6.72$, $P = 0.08$), first-person plural pronoun usage ($H(3) = 6.85$, $P = 0.077$), second-person pronoun usage ($H(3) = 6.75$, $P = 0.08$), and confirmations ($H(3) = 7.51$, $P = 0.057$) would also be considered significant. Features which did not show significant difference between groups were: age of acquisition ($H(3) = 1.28$, $P = 0.73$), imageability ($H(3) = 5.59$, $P = 0.13$), first person singular pronoun usage ($H(3) = 3.41$, $P = 0.33$), third person pronoun usage ($H(3) = 2.58$, $P = 0.46$), third person plural pronoun usage ($H(3) = 2.71$, $P = 0.44$), movement ($H(3) = 0.44$, $P = 0.93$), and agreement usage ($H(3) = 1.909$, $P = 0.59$).

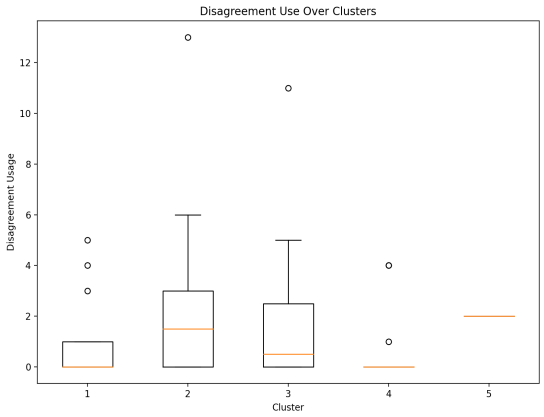

Figure 1: Distribution of Disagreement speech act usage among clusters.

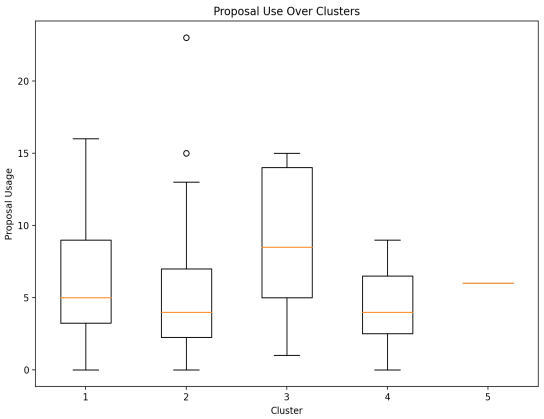

Figure 2: Distribution of Proposal speech act usage among clusters.

Figures 1 and 2 show the distributions of disagreements and proposals, respectively, across the identified clusters. We will further analyze these patterns as we characterize the identified clusters in the following section.

5 CLUSTER DISCUSSION

In this section we describe and characterize the groups that were identified through the clustering process.

Cluster 1 had participants who were on average most satisfied with the group task, while also being the cluster on average most effective at group influence. Given that this cluster used the highest proportion of confirmations and third person plural pronouns, the second highest proportion of proposals and agreements, and the second lowest proportion of disagreements, we can suggest viewing this cluster as people who often proposed their ideas, focused on the task, were agreeable with others, and who used confirmations

Cluster	AIS	AII	TE	WWT	TM	Efficiency	QoW	Satisfaction	Leadership
C1 (N=26)	78.27	36.88	3.69	4.85	5	4.85	4.77	4.63	3.85
C2 (N=22)	74.59	50.41	2.77	3.98	3.61	3.75	3.7	3.56	3.32
C3 (N=12)	83.25	37.58	1.75	4.33	4.67	3.58	4.75	3.8	4.5
C4 (N=23)	74.2	47.78	3.39	4.48	4.52	4.87	4.3	4.32	2.61
C5 (N=1)	56	67	1	1	2	2	1	1.4	3

Table 1: Frequency of Absolute Individual Score (AIS), Absolute Individual Influence (AII), Time Expectation (TE), Worked Well Together (WWT), Time Management (TM), Efficiency, Quality of Work (QoW), Satisfaction, and Leadership (described in Section 2).

Cluster	Utterances per person	(%) Proposals	(%) Disagreements	(%) Agreements	(%) Confirmations	Movement	Coordination
C1 (N=26)	91.27	7	0.72	8.09	3.7	2.23	-0.01
C2 (N=22)	109.36	5.53	1.95	7.94	2.62	2.24	-0.03
C3 (N=12)	125.58	7.3	1.66	7.3	1.59	2.78	-0.02
C4 (N=23)	70.91	6.19	0.55	11.04	0.61	2.41	-0.03
C5 (N=1)	92	6.52	2.77	5.43	0	1.23	0.02

Table 2: This table shows multimodal features used in our analysis as described in Section 3.2.

to reiterate what the group had accomplished, emphasizing their agreements. Though the coordination value was slightly negative, this cluster did have a higher coordination than Cluster 2, 3, or 4, suggesting their influence.

Cluster 2 stands in clear contrast to Cluster 1, since this cluster was simultaneously the least satisfied and least influential in their groups. The contrast is seen again in their speech acts, with Cluster 2 using the smallest proportion of proposals, and second highest proportion of disagreements (behind only the outlying cluster 5 participant). We can gain a sense of this cluster as being much more likely to disagree when others propose ideas, while also being less likely to propose their own ideas, leading unsurprisingly to the participants in this cluster having the worst influence score.

We can intuit Cluster 3 as being a cluster of participants who saw themselves as leaders, but who did not succeed in influencing their group and were ultimately not as satisfied as Cluster 1 or Cluster 4. As may be expected from those who see themselves as leaders, these participants spoke more often on average than all other clustered participants. Additionally, as seen in Table 2, these participants also moved the most on average. As shown by Kacewicz et al. [12] those in leadership roles consistently use fewer first person singular pronouns and more first person plural and second person pronouns. We see this effect clearly with Cluster 3 which had the lowest proportion of first person singular pronouns, highest proportion of first person plural pronouns and second highest proportion of second person pronouns.

It seems reasonable to categorize Cluster 4 as less experienced participants who are quite agreeable, and happy to go along with whatever it is that the group decides. This is seen clearly in their having the lowest proportion of disagreements, and highest proportion of agreements, as well as having the lowest number of average utterances per person. Interestingly, although we intuit Cluster 4 as a cluster of participants agreeable with accepting the group decision, they still did better overall at influencing the group than Cluster 2 did.

Cluster 5 seems to be a substantial outlier. In determining explanations for the participant's extremely low outcome scores, we inspected transcripts from the participant's group task. We were not able to ascertain anything notable from the group task discussion which might have contributed to the participant's scores. The participant had a very strong individual score, an objectively poor influence score, and scored their task outcomes extremely low. Taken together, we believe there is a clear picture that this outlying participant was frustrated that they could not influence the group better, and were therefore dissatisfied with the group task, and scored their experience of the group task low on nearly all measures. In addition to our clustering work, we used isolation forests for outlier detection in which this participant was also identified as the single greatest outlier in the dataset.

6 CONCLUSION

Using a task-based conversational interaction dataset, participants were clustered using outcomes of interest that include their individual performance and influence, their satisfaction ratings, and their demographic traits. We described the distinguishing characteristics of each identified group, giving a sense of what types of participants took part in the study, as well as identifying and describing outlier participants. We then extracted multimodal features from the conversational interaction and examined whether the identified groups also differed from each other in terms of their conversational behaviours. We highlighted key findings relating to pronoun usage and satisfaction across groups, as well as movement and task performance across groups. In contrast with recent works that have used a set of multimodal features to predict a particular outcome of interest, in our work herein we have focused on exploring the space of outcomes, grouping participants according to those outcomes, and examining how multimodal features differ across the identified groups. In future work, we will perform a similar analysis on an aggregation of group interaction datasets that include participant data on the big five personality types [13, 21].

REFERENCES

[1] Firoj Alam and Giuseppe Riccardi. 2014. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM multi media on workshop on computational personality recognition*. 15–18.

[2] Umut Avci and Oya Aran. 2014. Effect of nonverbal behavioral patterns on the performance of small groups. In *Proceedings of the 2014 workshop on Understanding and Modeling Multiparty, Multimodal Interactions*. 9–14.

[3] U. Avci and O. Aran. 2016. Predicting the performance in decision-making tasks: From individual cues to group interaction. *IEEE Transactions on Multimedia* 18, 4 (2016), 643–658.

[4] Umut Avci and Oya Aran. 2019. Analyzing group performance in small group interaction: Linking personality traits and group performance through the verbal content. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–7.

[5] Ligia Batrinca, Bruno Lepri, Nadia Mana, and Fabio Pianesi. 2012. Multimodal recognition of personality traits in human-computer collaborative tasks. In *Proceedings of the 14th ACM international conference on multimodal interaction*. 39–46.

[6] McKenzie Braley and Gabriel Murray. 2018. The Group Affect and Performance (GAP) Corpus. In *Proceedings of the Group Interaction Frontiers in Technology*. ACM, 2.

[7] Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246* (2020).

[8] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (2021), 103503.

[9] Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*. 699–708.

[10] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37, 1 (2010), 3–19.

[11] David Johnson, Giuseppe Carenini, and Gabriel Murray. 2020. NJM-Vis: interpreting neural joint models in NLP. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 286–296.

[12] Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology* 33, 2 (2014), 125–143.

[13] Maria Koutsombogera and Carl Vogel. 2018. Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[14] Uliyana Kubasova, Gabriel Murray, and McKenzie Braley. 2019. Analyzing Verbal and Nonverbal Features for Predicting Group Performance. In *Proceedings of Interspeech 2019, Graz, Austria*.

[15] Catherine Lai and Gabriel Murray. 2018. Predicting group satisfaction in meeting discussions. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*. ACM, 1.

[16] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of ICMI 2018, Boulder, USA*. 14–20.

[17] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 15–22.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[19] Navin Raj Prabhu, Chirag Raman, and Hayley Hung. 2020. Defining and Quantifying Conversation Quality in Spontaneous Interactions. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 196–205.

[20] Dairazalia Sanchez-Cortes, Oya Aran, and Daniel Gatica-Perez. 2011. An audio visual corpus for emergent leader analysis. In *Workshop on multimodal corpora for machine learning: taking stock and road mapping the future, ICMI-MLMI*.

[21] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia* 14, 3 (2012), 816–832.

[22] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*. IEEE, 166–171.