

Automatic Prediction of Consistency among Team Members’ Understanding of Group Decisions in Meetings

Joseph Kim and Julie A. Shah
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{joseph_kim, julie_a_shah}@csail.mit.edu

Abstract—Occasionally, participants in a meeting can leave with different understandings of what had been discussed. For meetings that require immediate response (such as disaster response planning), the participants must share a common understanding of the decisions reached by the group to ensure successful execution of their mission. In such domains, inconsistency among individuals’ understanding of the meeting results would be detrimental, as this can potentially degrade group performance. Thus, detecting the occurrence of inconsistencies in understanding among meeting participants is a desired capability for an intelligent system that would monitor meetings and provide feedback to spur stronger group understanding.

In this paper, we seek to predict the consistency among team members’ understanding of group decisions. We use self-reported summaries as a representative measure for team members’ understanding following meetings, and present a computational model that uses a set of verbal and nonverbal features from natural dialogue. This model focuses on the conversational dynamics between the participants, rather than on what is being discussed. We apply our model to a real-world conversational dataset and show that its features can predict group consistency with greater accuracy than conventional dialogue features. We also show that the combination of verbal and nonverbal features in multimodal fusion improves several performance metrics, and that our results are consistent across different meeting phases.

I. INTRODUCTION

Meetings are an integral component in many collaborative and organized work environments [1]. Each day, over 11 million meetings take place in the United States, and over 2.6 billion occur each year [2]. However, meetings are often not as efficient as they could be: Every year, an estimated \$54 million to \$3.4 billion is lost as a result of inefficient meetings [3]. Consequently, there is great interest in improving meeting productivity and efficiency.

One possible source of inefficiency is inconsistency between team members in their understanding of the outcome of a meeting [3]. This occurs when each team member leaves the meeting with a different understanding of what has been decided by the group (i.e. group decisions). This can potentially lead to miscommunication and confusion after a meeting. In certain applications, these mishaps can have severe consequences; such as disaster response planning, where

degradation in team performance can lead to high public safety costs [4].

Regardless of the purpose behind a meeting, it is important that all team members are ‘on the same page.’ We are interested in developing an intelligent system that would monitor meetings with the goal of helping team members to have consistent understandings of their group decisions. We envision a system capable of detecting inconsistencies among members so that it can provide cues to the group to revisit relevant discussion points. A system with this capability would help to reduce the number of misunderstandings among meeting participants and result in more efficient meetings.

Monitoring dialogue during group decision-making is a challenging task. Human dialogue is complex; discussions unfold in cycles, agreements are fluid, and proposals of ideas are often communicated and accepted implicitly [5]. Despite these challenges, researchers have developed useful qualitative models for group decision-making processes based on philosophical, linguistic and psychological foundations [6], [7], [8]. In the field of computational linguistics, several models have been proposed to capture levels of joint consensus [9], commitment [10] and agreement [5] within a group. Overall, prior art provides theoretical foundations for mapping natural human dialogue to a set of *dialogue features* that concretely capture information during group decision-making [5], [11], [12].

In this paper, we utilize a set of dialogue features to develop a computational model for predicting consistency among team members’ understanding of group decisions (referred to as **group consistency**). We use a particular set, called *Eugenio’s features*, which qualitative studies indicate better captures levels of joint commitment compared to conventional features [5]. We incorporate these features into a quantitative framework to perform prediction. In addition to verbal features, we also utilize a set of nonverbal features (head gestures) to capture a parallel layer of team members’ communicative intent. To the best of our knowledge, our model is the first to automatically predict levels of group consistency, and also the first to do so through sequences of verbal and nonverbal features taken from natural dialogue.

We conduct our analysis on a corpus of meeting data focused on group decision-making to solve a design problem [13]. First, we validate the use of Eugenio’s features for predicting group consistency and show improvement relative to the use of conventional dialogue features. Next, we incorporate a layer of head gestures into our model and show that this further improves several performance metrics. Finally, we show that the above results are consistent across different meeting phases.

II. RELATED WORK

In this section, we provide a brief summary of prior qualitative and quantitative studies that analyzed group decision-making.

Qualitative study of group decision-making is an active area of multidisciplinary research, involving a number of studies with philosophical, linguistic and psychological foundations [6], [7], [8]. These works focus on the development of theoretical models for conversational dynamics involved in group decision-making. Researchers have proposed models for consensus [9], commitment [10] and agreement processes [5], while other works have proposed models for various social decision schemes and considered human socio-emotional and behavior models [14], [15]. To the best of our knowledge, none of the theoretical models mentioned above have yet to be generalized to an automatic, predictive framework in natural human dialogue. One reason for this is the difficulty of mapping complicated human dialogue to a concrete set of dialogue features.

To address this challenge, several sets of dialogue features have been proposed for the capture of relevant information during group decision-making. Several studies incorporate features such as dialogue acts (DAs) and adjacency pairs [16] [17] to capture a level of *agreement*. Other works use word-based features (e.g., the number of positive and negative keywords spoken during a conversation) and prosodic cues to perform prediction tasks [18]. Although various sets of dialogue features have been used, these studies only analyze decision-making process from the perspective of a single participant; consequently, they do not capture the level of joint agreement among team members as a group.

More recently, Eugenio et al. [5] conducted a study of the modeling of collaborative dialogues, with a focus on design problems in which information is equally distributed and must be shared among team members to successfully result in a decision. Eugenio et al. found that the notion of commitment is more useful than that of simple acceptance or rejection (conventional DAs) for monitoring group decision-making. A new set of features, referred to as *Eugenio’s features*, were introduced to help monitor the evolving attitude of participants’ commitment toward options¹, and how a joint commitment is achieved by the group. Eugenio’s features have also been shown to facilitate the recognition of implicit

and/or passive acceptance of options by team members. These characteristics make Eugenio’s features useful for predicting group consistency, since joint commitment toward an option would naturally lead to joint understanding of group decisions.

The key differences between previous works and ours are: (1) Our approach generalizes dialogue features derived from studies of group decision-making into an automatic, predictive framework. (2) Specifically, we develop a computational model for predicting group consistency using Eugenio’s features and quantitatively verify their utility. (3) Our approach also integrates a layer of nonverbal features that provide additional information toward the prediction of group consistency.

III. DEFINITIONS AND HYPOTHESES

In this section, we provide a formal definition of ‘group consistency’, which is the primary object of our investigation; as well as our definition of a ‘discussion point.’ Then we present our hypotheses and describe our approaches for evaluating them.

A. Definition: Group consistency

Group consistency is defined as “the consistency among team members’ understanding of group decisions,” and essentially captures the alignment of team members’ understanding. It is important to highlight that our focus is on the level of *consistency*, and not on what the actual group decisions are.

In our study, we define two levels of group consistency: strong or weak. **Strong consistency** occurs when all team members have the same understanding of the group’s decisions; **weak consistency** occurs when one or more team members’ understandings differ. (Weak consistency would occur in situations where participants may have misunderstood an idea or missed an important detail during the meeting.) Figure 1 provides a visual illustration of weak consistency, where the team member on the rightmost side carries an understanding that is conceptually different from that of the other members.

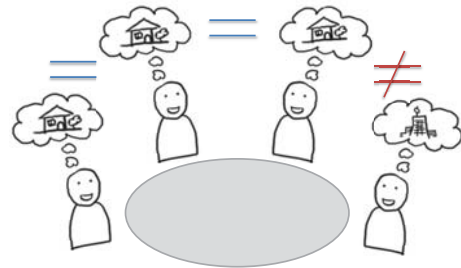


Fig. 1: Visual illustration of weak consistency

In order to assess the level of group consistency, we use self-reported summaries as a representative measure. These summaries contain information regarding group decisions from the perspective of each participant. If the contents of all the summaries are aligned, there is strong consistency among team members. If one or more summaries indicate a different conclusion, there is weak consistency. Prior work has used

¹‘Options’ here refers to proposed ideas or choices to be decided on by the group. [5]

a similar approach of using self-reports from team members to establish ground truth on agreements [19]; similarly, we assume that self-reported summaries can provide an effective ground truth to assess group consistency.

B. Definition: Discussion point

We define a **discussion point** as a conversation segment wherein deliberations are focused toward decision-making about a single topic. Discussion points closely resemble a list of items within a meeting agenda. Since meetings frequently involve the discussion of several different topics, we assess group consistency per discussion point.

C. Hypotheses and Approaches

The aim of our study is to develop a computational model capable of predicting group consistency using features from natural dialogue. We incorporated a set of verbal and nonverbal features important for capturing the dynamics of the group decision-making process. In doing so, we evaluated the following hypotheses:

H1 (Validation of Eugenio’s features)

Eugenio’s features can predict group consistency with improved performance over conventional DAs. We use Eugenio’s features as input for our computational model, and compare its prediction performance to the use of conventional DAs.

H2 (Incorporating head gestures)

Multimodal fusion of Eugenio’s features and head gestures improves overall prediction performance compared with using Eugenio’s features alone. Literature suggests the importance of nonverbal features for modeling human communication [20]. Head gestures have been used in prior computational models to infer a state of agreement, disagreement, concentration, interest or confusion [21]. In our study, we test whether the combination of head gestures with Eugenio’s features improves the prediction performance of our model.

H3 (Robustness across different topics)

Our model is robust where its prediction results are consistent across meetings of different topics. Since Eugenio’s features and head gestures focus on conversational dynamics rather than what is being talked about, our model should be robust to specific keywords and topics. We seek to evaluate whether the results from **H1** and **H2** are consistent across different meeting phases (described in Table I), where each one is fundamentally unique in its agenda and discussion topics. We train and test our model separately within each meeting phase, and compare performances.

IV. DATASET AND FEATURES

The dataset we used for our study is from the AMI meeting corpus [13]. It is one of the largest corpora of meeting data,

containing over 100 hours of recordings. In each of these meetings, a team of four people collaborated in order to design a remote controller. The meetings are divided into four distinct phases of the design process (descriptions provided in Table I) and are *scenario-driven*, with each person playing one of four specific roles: project manager, industrial designer, marketing expert or user interface designer. Although each participant was playing a role, the conversations that occurred during the meetings reflect natural, human-human interaction.

TABLE I: Four distinct meeting phases in the dataset

Meeting Phase	Discussion
Project kick-off	Getting acquainted with each other and discussing project goals
Functional design	Setting user requirements, technical functionality, and working design
Conceptual design	Determining conceptual specifications for components, properties, and materials
Detailed design	Finalizing user interface and evaluating the final product

The AMI meeting corpus is well-suited for our study, because the conversations that occurred during these meetings were tailored toward a group decision-making process. The use of Eugenio’s features is also appropriate, due to the meetings’ collaborative environment, wherein all decision points were consensual. The dataset provides a rich collection of annotations². In our study, we used annotations of participant summaries, topic segmentations, dialogue acts and head gestures. Here, we describe how each annotation layer was used to construct the components necessary to build our computational model.

A. From topic segmentations to discussion points

Topic segmentations partition each meeting according to related topics. They naturally represent our definition of discussion points by providing conversation segments that focus on the decision-making process for a single topic³.

B. From participant summaries to group consistency

Self-reported participant summaries were used to establish ground truth for group consistency. At the end of each meeting phase, participants were asked to provide written summaries of all decisions made during the meeting. As mentioned in Section III-A, we compared their contents and established *strong consistency* for the given meeting whether all summaries were aligned, or *weak consistency* if one or more differed. This comparison was made for each discussion point during the meeting. Two annotators performed the comparison (inter-rater agreement, $\kappa = 0.73$), resulting in group consistency labels for a total of 140 discussion points. (There was an imbalance in the distribution: Out of 140 discussion points, 93 had strong consistency and 47 had weak consistency.)

²For a full list of available annotations, we refer readers to [13].

³Some examples of topics from the AMI dataset include: handle design, battery options, target audience, etc.

Line	Speaker ID	Topic Discussion = "Remote locator"	Eugenio's features	Head gestures
			
1	B	Do we incorporate the idea of trying to locate the remote control again via a beeping noise?	PDO	
2	D	Yeah, think so.		D: Concord
3	C	Um, I think so, because it's so small		
4	C	I mean if we only have like two, three buttons it might be essential to have to have that [pause]		B: Concord
5	B	The ability to locate it again.		
6	C	Yeah.		B: Concord
7				A: Concord
8	B	So that would require a transmitter maybe attached to the TV and a basically small microphone on the actual unit, ...	UO	
			
9	B	If you could look into what we've suggested so far, the feasibility of small transmitter, and ...	Proposal	
10	C	Okay. Sure.	Commit	C: Concord

Fig. 2: A sample conversation segment taken from the AMI corpus. The participants here are discussing on a topic of a remote locator. Corresponding layers of Eugenio's features and head gestures are shown on the right columns.

C. From dialogue acts to Eugenio's features

The AMI dataset provides full annotations of dialogue acts (DAs), but not Eugenio's features. However, DAs can be used to form Eugenio's features, given knowledge of 'solution sizes.' A solution size is defined as 'determinate' when sufficient relevant information has been exchanged between meeting participants to form options. 'Indeterminate' refers to instances when further balancing of information is required.

We applied the heuristic of marking a portion of conversation as 'indeterminate' until the last DA label of 'inform' is displayed, after which the conversation segment is marked as 'determinate.' With DAs and solution sizes, we applied the coding scheme described in [5] to form Eugenio's features. Table II provides an overview of Eugenio's features, including their descriptions and coding schemes. Note that 'action-directives (AD)' correspond to suggestions and all elicit forms of DAs, which require actions from partners. For illustration, Figure 2 shows a sample conversation segment with the layer of corresponding Eugenio's features.

TABLE II: Eugenio's features, descriptions, and coding schemes

Feature	Description	Coding
Partner decidable option (PDO)	Occurs when a speaker offers an option that partners can use in decision-making. Corresponds to options that require further deliberation and balancing of information within the group.	AD, offer + indeterminate
Proposal	Occurs when a speaker offers an option following its full deliberation by the group.	AD, offer + determinate
Commit	Occurs when a speaker shows commitment to an option after its full deliberation.	Offer, assessment (positive) + determinate
Unendorsed option (UO)	Occurs when an option is simply presented during deliberation, without the speaker expecting any corresponding action from the other group members.	Open-options + determinate

D. Head gestures

The AMI corpus provides annotations of head gestures that reflect one's intentionality rather than simple form. A head nod

is further evaluated in order to distinguish between signals of comprehension, emphasis, etc. We incorporated gestures intended to communicate understanding and comprehension between participants. Table III highlights the description of head gestures used in our study. Figure 2 also shows a layer of head gestures in the conversation segment. Overall, sequences of Eugenio's features and head gestures are used in our computational model for prediction of group consistency.

TABLE III: Description of head gestures used in our study

Head gesture	Description
Concord	Signals comprehension, agreement or positive response; often characterized by a head nod.
Discord	Signals comprehension failure, uncertainty or disagreement; often characterized by a head shake or tilt.
Negative	Signals negative response to a yes-no question; usually characterized by a head shake.

V. COMPUTATIONAL MODEL

In this section, we describe our computation model, which incorporates feature sets described in Section IV. We aimed to study the effect of using these sets, so rather than exploring and comparing the accuracy of various learning algorithms, we focused on hidden Markov models (HMMs) as the primary tool for prediction. Given a sequence of features per discussion point, HMMs are used to predict either strong or weak group consistency (a form of binary classification). We incorporated HMMs because of their applicability to modeling systems with temporal sequences, as well as for their prior success in modeling human communication and social interactions [22], [23], [24].

An HMM can be described as a tuple $\{S, O, A, B, \pi\}$, where S is the set of hidden states, O is the set of observations, A is the state transition matrix, B is the observation probability matrix and π is the initial state distribution. In our formulation, O represents a finite set of Eugenio's features and head gestures, listed in Tables II and III, respectively. S represents hidden states underlying the temporal process of group decision-making. Unlike O , the concrete representation of S is unknown; however, only its cardinality, ($|S| = m$),

is necessary in order to learn the remaining distributions of A , B , and π . These distributions are learned iteratively through an expectation-maximization (EM) algorithm known as the Baum-Welch algorithm [25]. With HMMs learned separately for both strong and weak cases of group consistency, test sequences are classified according to maximum likelihood estimation. We leave further details and properties of standard HMMs to [26].

A. HMM: Eugenio’s features

Figure 3 depicts a graphical model of our HMM with a sequence of Eugenio’s features. A Eugenio’s feature coupled with the speaker ID forms a unique observation in the HMM sequence. To train and test the model, we performed leave-one-out cross-validation (LOOCV) in order to maximize the size of the training data. The only meta-parameter for the learning algorithm is the number of hidden states (m), which we varied from 1-5. In addition to HMM with Eugenio’s features, we trained and tested another HMM using conventional DAs⁴ relevant to group decision-making. This served as a baseline case for comparison.

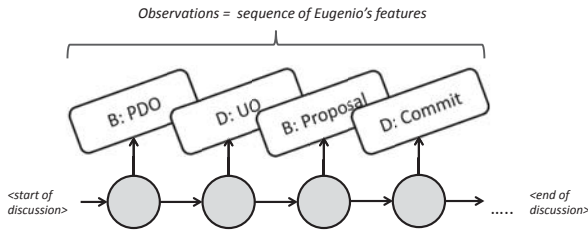


Fig. 3: An HMM with Eugenio’s features as observations (follows the order shown in the sample conversation segment in Fig. 2)

B. HMM: with head gestures

In order to incorporate head gestures into our model, we used an early fusion scheme wherein two feature sets are concatenated into one larger set. The two modality streams were ordered chronologically and combined to form a single stream of observations. Figure 4 shows an illustration of an HMM incorporating both Eugenio’s features and head gestures. This combined HMM captures occurrences of both feature sets and learns information regarding their transitions. We compared the performance of the combined HMM to HMMs trained only with Eugenio’s features.

VI. RESULTS AND DISCUSSION

In this section, we present the prediction performance of the computational model and evaluate our hypotheses.

A. Prediction performance: Eugenio’s features

Standard performance measures are presented in Figure 5. This figure depicts the average results from five different iterations of varying m . $HMM_{Eugenio}$ had a mean accuracy of 62.1% – an increase of 11% compared to the baseline of

⁴Four conventional DAs were used: assessment, elicit-assessment, comment-about-understanding (CAU) and elicit-CAU

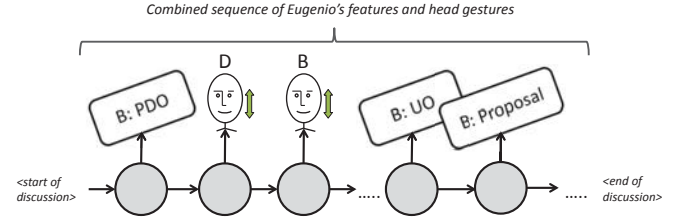


Fig. 4: An HMM with a combined set of Eugenio’s features and head gestures (follows the order shown in the sample conversation segment in Fig. 2)

conventional DAs. Other measures, such as recall, precision and F1 score, all showed improvement, each with an increase of approximately 10%. There was also a 12% reduction to the false positive rate (FPR).

In order to statistically test our hypothesis, we ran paired-sample t-tests on each performance metric. The assumption of normality on paired differences was not rejected by one-sample Kolmogorov-Smirnov tests [27]. P-values from paired-sample t-tests ($df=4$) are listed in Figure 5; all show a statistically significant difference at α of 0.05. These results support our first hypothesis **H1**, i.e., that using Eugenio’s features improves overall prediction performance compared with the use of conventional DAs.

	Acc. [%]	Rec. [%]	Prec. [%]	F1 [%]	FPR [%]
HMM_{DAs}	51.4	36.5	31.0	33.5	41.1
$HMM_{Eugenio}$	62.1	44.7	43.8	44.2	29.5
P-value	0.01	0.02	0.001	0.007	<0.001

Fig. 5: Performance measures between HMM with DAs and HMM with Eugenio’s features

B. Prediction performance: combined features

We compared the results of combined HMM to those of $HMM_{Eugenio}$, but also to an $HMM_{Eugenio+DAs}$ wherein four conventional DAs are incorporated to $HMM_{Eugenio}$. This served as a baseline to level out any potential improvement to prediction simply due to the addition of more features. Figure 6 highlights the comparison of results. We first noted that the baseline case performed worse than $HMM_{Eugenio}$, and that the addition of four conventional DAs actually reduced overall performance. Consequently, we ruled out the baseline case and directly compared $HMM_{Eugenio+Head}$ to $HMM_{Eugenio}$.

With $HMM_{Eugenio+Head}$, there was an overall increase in mean accuracy, recall, precision and F1 score. However, the increases to accuracy and precision were small: approximately 2-4%. Paired-sample t-tests indicated statistical significance for the improvements to recall and F1 score, but not the other metrics. We observed a small increase to FPR, but this change was not significant. Although the inclusion of head gestures in the model yielded positive signs of improved overall prediction, more statistical evidence is required to fully support **H2**.

	Acc. [%]	Rec. [%]	Prec. [%]	F1 [%]	FPR [%]
HMM _{Eugenio}	62.1	44.7	43.8	44.2	29.5
HMM _{Eugenio+DAs}	45.1	39.1	39.1	39.1	50.0
HMM _{Eugenio+Head}	64.2	55.3	47.3	51.0	31.1
P-value	0.28	0.02	0.18	0.03	0.49

Fig. 6: Performance measures across HMMs using different feature sets

C. Robustness across different meeting phases

We performed four-fold cross validation and compared prediction performances across the four distinct meeting phases. As described earlier in Table I, each meeting phase is fundamentally unique in its agenda and discussion topics. Consistency of prediction performance would indicate robustness of our model to specific keywords and topics. Figure 7 shows the comparison, highlighting the accuracies of HMM_{Eugenio+Head}, HMM_{Eugenio} and HMM_{DAs}.

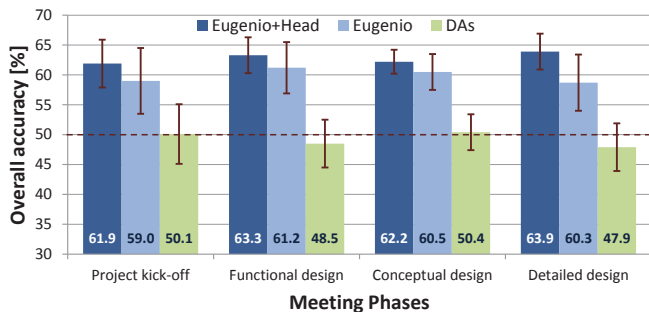


Fig. 7: Comparison of accuracies across different meeting phases

The mean accuracies for all three HMMs remained similar across the different meeting phases, though their values were slightly lower than global results presented in previous sections. This is to be expected, since four-fold CV has less available training data per fold than LOOCV. We observed an increasing trend in accuracy per meeting phase from HMM_{DAs} → HMM_{Eugenio} → HMM_{Eugenio+Head}, which reflects results presented in Sections VI-A and VI-B. This trend was repeated across different meeting phases as well.

Paired-sample t-test indicates that the difference between HMM_{Eugenio} and HMM_{DAs} is statistically significant for all meeting phases (df=4, p=[0.001-0.03]). However, more statistical evidence is required to support difference between HMM_{Eugenio+Head} and HMM_{Eugenio} (p=[0.28-0.58]). An identical approach was taken with the four other performance metrics, with analogous results. Overall, we observed that results from **H1** and **H2** are consistent across different meeting phases, supporting **H3**.

D. Discussion

Overall, our ‘best’ computational model of HMM_{Eugenio+Head} is able to predict group consistency with 64.2% accuracy. There is statistical evidence suggesting the model’s robustness to any specific keywords and topics different meeting phases.

We also highlight that our computational model has the flexibility to receive as input any set of dialogue features involved in the group decision-making process. Other features, such as vocal intonation, sentiments and/or facial expressions, could have easily been used as an additional observation sequence for our HMM.

When integrating our computational model for an on-line system, high recall and low FPR are especially important. High recall would have a high hit rate of capturing discussion points with weak consistency; the system can then provide feedback and spur stronger group consistency. Low FPR is also important to reduce the rate of false alarms within the system. Incorrect prediction of weak consistency and false feedback would be disruptive, and may cause humans to lose trust in the system. With HMM_{Eugenio+Head}, recall and FPR are 55.3% and 31.1%, respectively. There is definitely room for improvement, especially for reducing the FPR as much as possible while maintaining a moderate level of recall. However, we believe that a system capturing only 55.3% of weak consistencies can still be helpful to human teams, as long as FPR is low (i.e. that the system predicts ‘weak’ cases only when it is highly confident). In future work, we aim to test our computational model during live group meetings and investigate its utility using both objective (e.g. prediction performance) and subjective (e.g. human response, social impacts) measures.

Finally, our computational approach relies on a text transcription with segmented topics. This naturally prompts working with automatic speech recognition and a topic segmentation tool, with the end goal of developing a real-time system. Outputs from those tools may induce early noise for our input.

VII. CONCLUSION

In this paper, we developed a computational model capable of predicting consistency among team members’ individual understandings of group decisions, referred to as ‘group consistency.’ Our model focuses on conversational dynamics involved in the group decision-making process, and uses a set of dialogue features, *Eugenio’s features*, that has been qualitatively validated to capture levels of joint commitment during group decision-making.

We demonstrated the utility of *Eugenio’s* features for the prediction of group consistency, and showed an improvement in prediction performance over conventional dialogue acts. We also investigated a multimodal approach of incorporating head gestures into our model and observed a statistically significant increase in recall and F1 score.

Overall, our full model predicts group consistency with 64.2% accuracy and shows signs of robustness for meetings of different topics. Our work combines the strength of human communications research and machine learning with a vision for developing an intelligent system that would help teams to achieve stronger group understanding.

ACKNOWLEDGMENT

The funding for this project was provided in part by the National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP) under Grant No. 2012150705.

REFERENCES

- [1] P. Monge, C. McSween, and J. Wyer, "A profile of meetings in corporate America: results of the 3M meeting effectiveness study." Annenberg School of Communications, University of Southern California, Los Angeles, CA, 1989.
- [2] "Meetings in America: A study of trends, costs and attitudes toward business travel, teleconferencing, and their impact on productivity," in *A network MCI Conferencing White Paper*, vol. 3. Greenwich, CT: INFOCOMM, 1998.
- [3] N. Romano and J. Nunamaker Jr., "Meeting analysis: findings from research and practice," in *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*. IEEE, 2001.
- [4] J. F. Desforges and J. F. Waeckerle, "Disaster planning and response," *New England Journal of Medicine*, vol. 324, no. 12, pp. 815–821, 1991.
- [5] B. Di Eugenio, P. W. Jordan, R. H. Thomason, and J. D. Moore, "The agreement process: an empirical investigation of human-human computer-mediated collaborative dialogs," *International Journal of Human-Computer Studies*, vol. 53, no. 6, pp. 1017–1076, 2000.
- [6] M. E. Bratman, "Shared cooperative activity," *The Philosophical Review*, vol. 101, no. 2, pp. 327–341, 1992.
- [7] H. H. Clark, *Arenas of language use*. University of Chicago Press, 1992.
- [8] —, *Using language*. Cambridge University Press, 1996.
- [9] F. Herrera and E. Herrera-Viedma, "A model of consensus in group decision making under linguistic assessments," *Fuzzy Sets and Systems*, vol. 78, no. 1, pp. 73–87, 1996.
- [10] B. J. Grosz and S. Kraus, "Collaborative plans for complex group action," *Artificial Intelligence*, vol. 86, no. 2, pp. 269–357, 1996.
- [11] B. A. Goodman, F. N. Linton, R. D. Gaimari, J. M. Hitzeman, H. J. Ross, and G. Zarrella, "Using dialogue features to predict trouble during collaborative learning," *User Modeling and User-Adapted Interaction*, vol. 15, no. 1-2, pp. 85–134, 2005.
- [12] D. R. Traum and E. A. Hinkelman, "Conversation acts in task-oriented spoken dialogue," *Computational Intelligence*, vol. 8, no. 3, pp. 575–599, 1992.
- [13] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, and M. Kronenthal, "The AMI meeting corpus: a pre-announcement," in *Machine Learning for Multimodal Interaction*, 2006, pp. 28–39.
- [14] J. H. Davis, "Group decision and social interaction: A theory of social decision schemes." *Psychological Review*, vol. 80, no. 2, pp. 97–125, 1973.
- [15] S. G. Green and T. D. Taber, "The effects of three social decision schemes on decision group process," *Organizational Behavior and Human Performance*, vol. 25, no. 1, pp. 97–106, 1980.
- [16] S. Germesin and T. Wilson, "Agreement detection in multiparty conversation," in *Proceedings of the 2009 International Conference on Multimodal Interfaces*, 2009.
- [17] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004.
- [18] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: training with unlabeled data," in *Proceedings of HLT-NAACL Conference*, 2003.
- [19] S. R. Hiltz, K. Johnson, and M. Turoff, "Experiments in group decision making communication process and outcome in face-to-face versus computerized conferences," *Human Communication Research*, vol. 13, no. 2, pp. 225–252, 1986.
- [20] N. Ambady and M. Weisbuch, "Nonverbal behavior," *Handbook of Social Psychology*, pp. 464–497, 2010.
- [21] R. Kaliouby and P. Robinson, "Real-time inference of complex mental states from facial expressions and head gestures," *Real-time Vision for Human-Computer Interaction*, pp. 181–200, 2005.
- [22] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interaction in meetings," in *In Proc. IEEE ICASSP, Hong Kong*, 2003.
- [23] J. J. Lee, W. Knox, J. Wormwood, C. Breazeal, and D. DeSteno, "Computationally modeling interpersonal trust," *Frontiers in Psychology*, vol. 4, no. 893, 2013.
- [24] L. Yang, Y. Xu, and C. S. Chen, "Human action learning via hidden Markov model," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 27, no. 1, pp. 34–44, 1997.
- [25] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.
- [26] Z. Ghahramani, "An introduction to hidden Markov models and Bayesian networks," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, no. 1, pp. 9–42, 2001.
- [27] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.