Automatically Analyzing Brainstorming Language Behavior with Meeter

BERND HUBER, Harvard University, USA
STUART SHIEBER, Harvard University, USA
KRZYSZTOF Z. GAJOS, Harvard University, USA

Language both influences and indicates group behavior, and we need tools that let us study the content of what is communicated. While one could annotate these spoken dialogue acts by hand, this is a tedious, not scalable process. We present Meeter, a tool for automatically detecting information sharing, shared understanding, word counts, and group activation in spoken interactions. The contribution of our work is two-fold: (1) We validated the tool by showing that the measures computed by Meeter align with human-generated labels, and (2) we demonstrated the value of Meeter as a research tool by quantifying aspects of group behavior using those measures and deriving novel findings from that. Our tool is valuable for researchers conducting group science, as well as those designing groupware systems.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; *User studies.*

Additional Key Words and Phrases: groups, brainstorming, machine learning, natural language processing, speech processing

ACM Reference Format:
Bernd Huber, Stuart Shieber, and Krzysztof Z. Gajos. 2019. Automatically Analyzing Brainstorming Language Behavior with Meeter. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 30 (November 2019), 17 pages. https://doi.org/10.1145/3359132

1 INTRODUCTION

There is extensive research in CSCW on designing groupware systems that engage participants in a joint task [15, 23, 31, 42]. However, despite the promise of computing capabilities to make studying groups more scalable, relatively little work exists on using automated computational techniques to *study* collocated groups [17], and research that investigates computational ways of studying spoken group brainstorming is limited to just a few studies [22, 23].

With the goal of enabling automatic group brainstorming behavior research, we introduce Meeter, a system to automatically quantify several group processes from spoken interactions. Meeter measures four group behaviors that have been shown important for successful groups: (1) the act of telling group members a new kind of information, which we refer to as *information sharing*, (2) the act of elaborating on one's own or another person's understanding, which we refer to as *shared understanding*, (3) the communication quantity that occurred, independent of content, as measured by word count, and (4) the engagement and energy that is expressed in the

Authors' addresses: Bernd Huber, bhb@seas.harvard.edu, Harvard University, Cambridge, MA, USA; Stuart Shieber, shieber@harvard.edu, Harvard University, Cambridge, MA, USA; Krzysztof Z. Gajos, kgajos@eecs.harvard.edu, Harvard University, Cambridge, MA, USA.

2573-0142/2019/11-ART30 $15.00
https://doi.org/10.1145/3359132

Type	Dialogue Act	Example
Information Sharing	Share Information	This is a coin.
	Elicit Information	How high is Mount Everest?
Shared Understanding	Assess	That's great.
	Elicit Assessment	What do you think?
	Show Understanding	Okay yeah.
	Elicit Understanding	Do you understand?

Table 1. Example dialogue acts for information sharing and shared understanding.

interaction, which we refer to as *group activation*. Meeter is the first system to automatically quantify information sharing and shared understanding from spoken interaction by analyzing the content of the interaction. While there exists previous work that automatically analyzed communication quantity and group activation [23], Meeter includes two novel measures to provide a single rich analytical tool. Furthermore, there are only limited previous studies that provide empirical evidence on the links between those measures and successful group behaviors, with most of them relating word counts to group outcomes [29, 33, 41]. With Meeter, we introduce a tool that allows a more effective study of groups, and we provide empirical evidence for the links between group measures detected by Meeter and group outcomes.

We evaluated Meeter on the existing *AMI meeting corpus* [3] consisting of 135 brainstorming meetings, as well as on a newly collected corpus of 15 brainstorming meetings. In a technical evaluation, we used the AMI meeting corpus to evaluate the accuracy of the dialogue act classifier in Meeter when classifying information sharing and shared understanding from transcribed brainstorming sessions. The results of this analysis show that Meeter performs at an average of 73% precision and 74% recall.

We further validated Meeter in a laboratory study with 15 group brainstorming sessions, collecting group outcome data (performance on the brainstorming task and group satisfaction). The audio recordings of those meetings were analyzed in two ways: automatically by Meeter and manually by human annotators. The results show that the output of Meeter (word count, fraction of information sharing dialogue acts, fraction of dialogue acts indicating shared understanding) was highly correlated with the output produced by human annotators ($r > 0.87$), showing the feasibility of using Meeter for conducting group research. Furthermore, we studied the relationship between Meeter measures and group outcomes. The same conclusions were reached using both Meeter-generated and human-generated measures: (1) Groups that showed higher levels of information sharing and shared understanding were more satisfied; (2) Groups that used fewer words were more satisfied; (3) More activated groups as measured by Meeter through speech prosody, showed higher levels of group performance.

In summary, the contributions of this paper are as follows:

- **Meeter, a tool for automatic language analysis of spoken group interactions.** The system detects and combines multiple measures: (1) a set of dialogue acts in spoken group brainstorming meetings, (2) word counts, (3) speech prosody. Our validation study demonstrates that quantities automatically reported by Meeter are highly correlated with those produced by human annotators.
- **We study the relationship between the Meeter measures and group outcomes.** Our findings provide novel insights into how group processes relate to group outcomes in brainstorming settings, with results by Meeter aligning with results from human-generated measures. These findings highlight the value of Meeter as a research tool to derive novel findings.

2 BACKGROUND

2.1 Automatic Group Language Analysis

Previous research has shown that group language can be analyzed automatically, with most of the tools being built for text-based interaction [25, 30, 43]. *GroupMeter* used agreement word counts and overall word counts to predict group outcomes. This system used the linguistic inquiry and word count (LIWC) dictionary to count agreement words [37], and it used total word counts to estimate the number of factual statements. More recently, Tausczik et al. [43] used word counts to indirectly detect information sharing, positivity, balance, and engagement from chat. Our work extends this line of research in two ways: by introducing a more direct way to identify interactions related to information sharing and shared understanding, and by supporting spoken interactions.

In group brainstorming studies with spoken interaction, one frequently measured group process is speaker participation rate [2, 9, 23]. The *ConversationClock*, for example, used sounds per person as a measure of speaker balance [2]. *Meeting Mediator* detected enthusiasm, interest, and persuasiveness using prosody features as well as body movement and proximity among people [23]. A more recent study presents the TalkTraces system, which provides visualizations of topic clusters in meetings [4]. Our work extends this research by introducing linguistic, data-driven measures for information sharing and shared understanding.

2.2 Language Behavior in Successful Teams

2.2.1 Information Sharing and Shared Understanding. The concept of groups as information processors posits that groups with better outcomes often share more information, and show more shared understanding within the group [6–8, 18, 32, 34]. Communicating information is an important mechanism for effective group brainstorming because all the necessary information might be in the room, but if the group members do not communicate the information well, then the discussed problem may remain unsolved [7]. Developing a shared understanding is the process in a group to develop a representation on a topic shared between members of the group. The process allows group members to make full use of the information that is being shared [1, 34].

Previous research demonstrated that information sharing and shared understanding are expressed in the language used by the group [20, 39]. This suggests that linguistic analysis could be used to determine information sharing and shared understanding levels in groups. We present a set of six specific dialogue acts that indicate information sharing and shared understanding (see Table 1) [10]. Information sharing dialogue acts can be either a group member sharing information or a group member eliciting information from other group members. Shared understanding dialogue acts can be comments on what has been said or done so far, or statements about group members' understanding (such as "oh okay", "this is not clear to me"). This framework allows to better quantify information sharing and shared understanding in group brainstorming.

2.2.2 Word Count. Previous work on group interaction shows that the overall amount of communication in a group is an important factor determining the effectiveness of groups [29, 33, 41]. Group performance and word counts were, for example, studied in flight simulation tasks, where people who spoke more performed better on the task [41]. Yet previous works seem inconclusive with the effect of communication amount on outcomes. In brainstorming sessions, the number of ideas that come out of a group brainstorming session was found to be positively related with the word count during the conversation [29], which however was not replicated in other settings [30]. Group satisfaction has been studied in chat conversations, where fewer total word counts led to higher group satisfaction outcomes [33]. This work explained the phenomenon with the fact that one dominant participant that talks a lot—and thus inflates the word count—may suppress the overall perceived group experience.

2.2.3 Speech Prosody. Non-verbal speech signals play an active role in human conversations. Emotional vocalizations cover a broad range, from short bursts of laughter [35] or nonverbal *affect bursts* such as "ah," or "eww" [40] to more complex speech utterances with elaborate suprasegmental and prosody features such as pausing, rhythm, and intonation. These features are often used in speech research to detect non-verbal characteristics such as emotion in voice. Collectively, these non-verbal speech signals are referred to as *speech prosody*.

In brainstorming, speech prosody can be indicative of a few social regulation group processes. Positivity, the degree to which group members are encouraging one another, can enhance interpersonal relationships and motivate individuals to work harder [27]. However, positivity can also detract from task effort when it leads to off-topic conversations [29]. Groups that are positive use more activated voice and raise their voice more frequently.

Another important social process is how activated people are in a group [33]. Group activation, the degree to which group members are paying attention and connecting with each other, can enhance group satisfaction. If group members are activated, they are more likely to stay motivated and enjoy the task. Speech prosody has been previously used to automatically capture group activation in conversations [38]. In many situations, non-linguistic social signals (body language, facial expression, prosody) are as important as linguistic content in predicting behavioral outcome [6, 8], but prosody is among the most powerful of these social signals [8]. In summary, speech prosody has been linked to a number of group processes, each of which—in turn—can influence group satisfaction, performance, or both. We extend this work to the brainstorming domain.

2.3 Studying Group Discussions

The current practice of studying spoken group discussions typically requires such burdensome steps as transcription of discussions, manual data annotation, or multiple humans to annotate data. Hence, researchers often study small numbers of groups, small samples of the total interactions, or avoid studying groups at such a level of detail altogether [21]. Computational developments provide novel ways to gain insights on human social interactions, leading to the emergence of the field of computational social science [26]. Computational social science comes with the promise to make research more efficient by relieving the burden from human coders, making data analysis potentially more private, and overall more scalable.

All of these benefits allow researchers to gain new insights into human social interactions by enabling new, improved ways to study groups, for example in such complex interactions as software development groups [19], and to uncover complex interactions among different agents in group discussions [36]. However, there are still many challenges involved in using computational models in social science successfully, often coming from the large gap of expectations between the social science field and computer science field [19]. For example, computational studies of social behavior might be of descriptive nature (e.g. analysis of sentiment in GitHub commit comments [16]), while social science might look for theory-driven approaches to understand the underlying processes. Meeter helps to bridge this gap by bringing computational capabilities into settings with spoken interaction and focusing on group process measures, allowing researchers to gain new insights into what makes groups effective.

3 MEETER SYSTEM

In the previous section, we identified several language-related group processes that have been previously linked to group performance and group satisfaction. Since we aim at analyzing these processes automatically, this section will go over the design and implementation of the Meeter system. Figure 1 illustrates the overall information flow in the Meeter system.

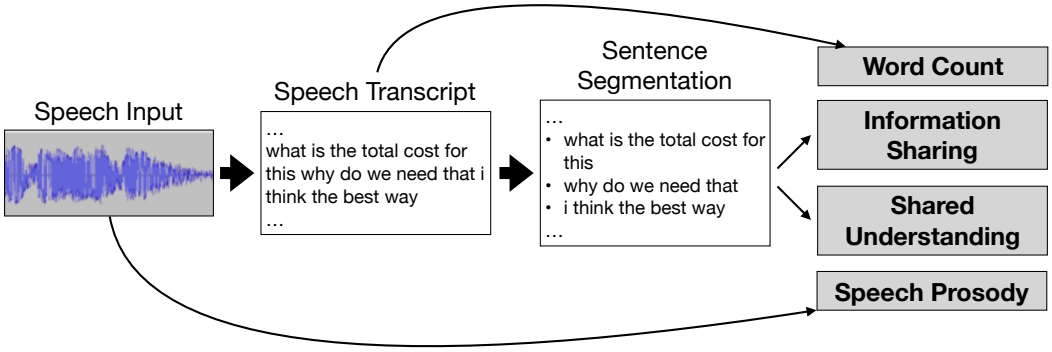

Fig. 1. An overview of the information flow with the different computational units incorporated in the system architecture of Meeter. Speech first gets transcribed and segmented between pauses. The transcript is then further segmented into separate sentences using a neural network segmentation algorithm. Every sentence is then categorized into dialogue act labels with our classifier and its prosody features are extracted. This then gives the automatically tagged dialogue act label and prosody features for every dialogue act.

3.1 Dialogue Acts

Given an audio stream of a brainstorming session, Meeter outputs a segmented transcript of the conversation, with each dialogue act classified as Information Sharing, Shared Understanding or Other.

To classify spoken brainstorming interaction, Meeter first uses the *Google Cloud Speech Recognition Service*[1] to transcribe the content of the conversation. The transcript is then segmented into sentences using a neural network–based sentence segmentation algorithm [44].

As a second step, Meeter classifies every sentence into one of the three dialogue act classes. To train our classifier, we used dialogue act annotated sentences from the AMI meeting corpus [3]. This corpus consists of a collection of transcribed and segmented design meeting recordings, where each sentence in the transcript is time stamped and annotated with a dialogue act class. The corpus consists of 135 meetings in which groups of four people collaborated on a product design task for 20 minutes. The corpus consists of a total of N=110,000 samples (dialogue acts), which we divided into a fixed set of 90,000 training samples, 10,000 validation samples used for hyperparameter tuning (such as feature selection), and 10,000 test samples used for the final evaluation of classifier accuracy.

We constructed and evaluated three types of models for classifying dialogue acts:

SVM The support vector machine (SVM) classifier was used with a linear kernel and a complexity parameter value of 0.1. For this classifier, each sentence is represented as a feature vector using a text frequency-inverse document frequency (TFIDF) representation. This model has proven successful for a broad range of text classification tasks, while still being relatively simple. In our analysis, the TFIDF representation is learned on the training set and represents each sentence as a vector, giving each word in the sentence a score and a fixed position in the vector. We also automatically tag the text with part-of-speech (POS) tags and use both words and POS to construct the TFIDF vectors. We train the TFIDF vectors with a combined uni- and bigram feature representation. Bigram models use two consecutive words as a feature, and therefore the word order is taken into account. This led to a total number of 16,005 features. From this full feature set, a feature subset was selected with a frequency filter of a

[1] https://cloud.google.com/speech/

Model	Actual \ Predicted	Information Sharing	Shared Understanding	Other
SVM	**Information Sharing**	2768 (80%)	305 (9%)	401 (11%)
	Shared Understanding	531 (15%)	2732 (77%)	271 (8%)
	Other	729 (24%)	413 (14%)	1850 (62%)
	Precision/Recall/F1	**80%**/69%/**0.74**	77%/**79%**/**0.78**	62%/**73%**/0.67
Overall Precision/Recall/F1/Accuracy: 73%/74%/0.73/73%				
CNN	**Information Sharing**	2471 (80%)	455 (9%)	548 (11%)
	Shared Understanding	409 (15%)	2621 (77%)	504 (8%)
	Other	433 (24%)	423 (14%)	2136 (62%)
	Precision/Recall/F1	74%/71%/0.72	75%/74%/0.74	67%/71%/**0.69**
Overall Precision/Recall/F1/Accuracy: 72%/72%/0.72/72%				
Sequential	**Information Sharing**	2475 (80%)	351 (9%)	548 (11%)
	Shared Understanding	359 (15%)	2703 (77%)	472 (8%)
	Other	537 (24%)	423 (14%)	2032 (62%)
	Precision/Recall/F1	73%/**73%**/0.73	**78%**/76%/0.77	**67%**/68%/0.67
Overall Precision/Recall/F1/Accuracy: 73%/72%/0.72/73%				

Table 2. Confusion matrices of the SVM, CNN, and sequential dialogue act classifier on the three classes *Shared Understanding*, *Information Sharing* and *Other*. Classifiers were evaluated on the AMI meeting corpus (N=110,000). The corpus was divided into a fixed training set with 90,000 samples, validation set with 10,000 samples for tuning the model hyperparameters, and test set with 10,000 samples for evaluating the models. Our results show an overall F1 score of 0.73 for SVM, which was the highest performing classifier on the test set.

minimum of 10 occurrences, as determined via our validation set. This filtering step led to a subset of 6,735 features which were ultimately taken to train our SVM classifier.

CNN The convolutional neural network (CNN) was used as presented by Kim et al [24], and Collobert et al [5]. For this classifier, each sentence is represented as a sequence of word embeddings, which are vector representations of words. The sequence of word embeddings is then fed through a CNN to classify the sentence. Previous work shows that this model can be effective for various sentence classification tasks [24].

Sequence classifier The sequential CNN short-text classification model was used as presented by Lee et al [28]. For this classifier, each sentence is represented as a sequence of word embeddings, and each conversation is represented as a sequence of sentences. To classify the sequence of sentences, each sentence is first fed through a CNN, and the sequence of sentences is then fed through a sequence classifier to classify each sentence. Previous work shows that this model can moderately, in some cases, improve accuracy over non-sequential models in settings with sequential dependencies between sentences [28].

We evaluated the dialogue act classifiers using cross-validation and data from the AMI meeting corpus. Table 2 shows the resulting performance of the sentence classifier, with the best performing classifier being the SVM at an F1 score of 0.73 (vs. 0.72 for the CNN model and 0.72 for the sequence classifier), as compared to a majority based score of 0.35. This accuracy level has been shown acceptable for analyzing group interactions, e.g., [22]. In the remainder of the paper, we report the analysis results of the SVM classifier, since it showed the highest accuracy. Of the three models tested, the linear SVM is also the simplest and the most easily interpretable.

Measure	Explanation
Information Sharing	The absolute number of sentences classified as information sharing. At the end of a session, this measure is divided by the absolute number of sentences.
Shared Understanding	The absolute number of sentences classified as shared understanding. At the end of a session, this measure is divided by the absolute number of sentences.
Word Count	Absolute number of words in the brainstorming session transcript.
Harmonics-to-noise-ratio (HNR)	The amount of noise in a voice signal measured in decibels. The lower the HNR, the more noise in the voice.
Jitter	The variation of pitch in a voice signal measured in percent of the pitch. Jitter values in normal voices range from 0.2 to 1 percent.
Loudness peaks per second	The number of loudness peaks per second in a voice signal.
Voiced segments per second	The number of segments with voice signal in an audio signal. One segment is a fixed size frame of 40 milliseconds, a typical segment length in speech analysis tasks coming from average phoneme and pause duration.

Table 3. Overview of the measures that Meeter provides.

3.2 Word Counts

Word counts were computed as the absolute number of words in the brainstorming session transcript.

3.3 Speech Prosody

Meeter also extracts sentence-level speech prosody in the group brainstorming session. The system leverages the *openSMILE* feature extraction toolkit to extract a set of prosody features [11, 12]. In particular, Meeter extracts the following four features that have been shown to be important voice characteristics to capture voice activation and engagement [11]: (1) Harmonics-to-noise-ratio (HNR), the amount of noise in a voice signal; (2) Jitter, the variation of pitch in a voice signal; (3) loudness peaks per second, and (4) voiced segments per second. Every feature provides a numeric value and the final prosody representation is a 4-dimensional feature vector representing the prosody characteristics of the sentence. Table 3 shows an overview of the measures that Meeter returns.

3.4 Implementation

In our experiment, we used an Android frontend showing timing information to participants. In our setup, this Android app sends the Meeter classification results per dialogue act to an online database. Every row of this database contains a classification of the dialogue act, the word count, the four prosody features, and a session ID unique to the currently analyzed session. These data can be accessed for further analysis by the researcher. A screenshot of the experiment interface can be seen in Figure 2. The Meeter backend has been implemented with an audio processing pipeline that allows for real-time computation of the measures. The code for our Meeter setup is available at github.com/BerndHuber/meeter_code.

4 METHODS

In our evaluation study, we wanted to know (1) how Meeter performs on a newly collected data set of 15 brainstorming meetings compared to measures produced by human annotators, and (2)

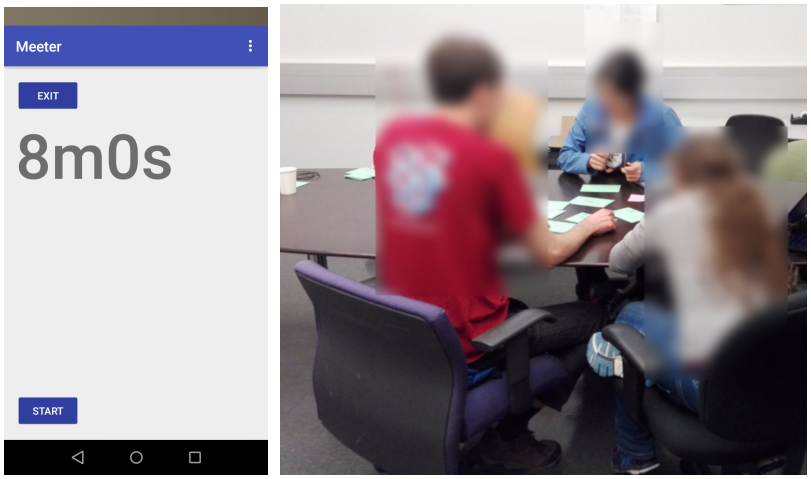

Fig. 2. The study setup for evaluating Meeting: (Left) The graphical user interface of Meeter with the timer. (Right) The study setup, with three participants brainstorming about one given problem. The smartphone with the running app was positioned on the same table in the middle of the discussion. No further microphone setup was used.

how the group process measures detected by Meeter relate to group outcomes. To answer these research questions, this study tests the following hypotheses on Meeter performance:

H1: Meeter group process measures are significantly correlated with group process measures determined by humans.

H2: Meeter measures are related to group outcome measures in the same way as human-annotated measures.

Furthermore, we provide new evidence for the relationship between the group processes and group outcomes, and we hypothesize the following:

H3: Groups with more information sharing and shared understanding dialogue acts perform better and are more satisfied.

H4: Groups using fewer words perform better and are more satisfied.

H5: A statistical model combining a number of prosody features that measure group activation predicts group performance and group satisfaction.

The remainder of this section describes the methods of this study.

4.1 Tasks and Procedures

To collect the brainstorming interaction data, we put participants in groups of three and we asked them to solve open-ended problems and come up with a wide range of ideas. Participants were given ten minutes per problem and were instructed to come up with as many solutions to the given problem as possible. Participants were instructed to write down ideas on cards that were provided to them at the beginning and were instructed to write down one idea per card. To ensure a certain level of quality for the ideas, participants were instructed to rank the ideas for two minutes after the session and filter out ideas that they thought were not feasible.

Participants were given, in counterbalanced order, the following problems:

- Develop strategies to prevent bullying in schools.
- Develop strategies to reduce cell phone disturbances on public transportation.

- Develop strategies to encourage people to read diverse news.

We collected a total of 15 sessions of ten minutes by five groups, where each group solved a total of three problems. We report the number of cards given to us at the end of every brainstorming session as the number of ideas generated in this session. Group performance was operationalized as the total number of ideas that were handed in at the end of the session, a common way to measure group performance in brainstorming settings [45].

We recorded the conversations using an off-the-shelf smartphone. We placed the smartphone on the table where it was visible to the participants. The only information displayed on the smartphone screen was a timer so that participants could see how much time was left in the session (Figure 2).

After each round of brainstorming, participants were asked about their subjective impressions about the brainstorming session. We adopted measures from the interaction rating questionnaire (IRQ), which is frequently used in group interaction research [33]. Each group member rated their enjoyment of the interaction, how well they clicked with their group, and whether they would work with their group again. Questions were scored on 5-point Likert scales from strongly disagree to strongly agree [33]. Using these question responses, the overall group satisfaction was computed as the average of the group members' rated satisfaction.

4.2 Participants

We recruited 15 participants through an on-site University-managed recruitment pool. None of the participants knew each other before participation in our study. The average age was 32, with 7 male and 8 female participants, with two of them being current students. All of the participants spoke English as their first language and were college educated.

4.3 Human Data Annotation

To validate the performance of Meeter with respect to the current practice of manually transcribing and annotating interactions, we collected data annotations for the recorded group brainstorming sessions. The annotation required two steps, first we collected human transcriptions, and second, we collected human annotations of the sentences with dialogue act labels.

The speech transcripts were collected using the *rev.com* transcription service, which reports an accuracy level of > 99%. We used this transcript as ground truth for the spoken words in the brainstorming sessions.

To annotate the transcribed sentences with dialogue act labels, we created a crowdsourcing task on Amazon Mechanical Turk. In the task, workers were provided an educational module explaining the different kinds of dialogue acts and providing the context of conversations. Each worker was then given 30 sentences of a conversation and was asked to annotate each sentence as instructed. To ensure the quality of the annotations, we had four workers annotate every sentence. We also seeded the task with gold standard sentences for which we generated labels ourselves. Workers that had more than 20% of the gold standard sentences wrong were filtered out in our analyses. Demographics of the crowdworkers were restricted to English language speakers in the UK and USA. Workers were paid $1.5 (USD) per task (expected hourly wage of $9).

We recruited 60 raters for the 15 sessions, which provided us four labels per sentence. The agreement rates were sufficiently high, with a kappa of 0.75. We, therefore, computed the final labels using the label most agreed upon (three or more). In 21 utterances, no agreement was reached, in which case we manually annotated the final label.

	μ (σ) Meeter	μ (σ) Human	Pearson Correlation
Word Count	926 (191)	1041 (215)	$r = 0.91$
Information Sharing	0.49 (0.068)	0.44 (0.069)	$r = 0.87$
Shared Understanding	0.14 (0.041)	0.25 (0.076)	$r = 0.93$
Harmonics to noise ratio	-0.370 (0.566)	-	-
Jitter	0.338 (0.981)	-	-
Loudness peaks/second	-0.456 (0.530)	-	-
Voiced segments/second	0.094 (0.716)	-	-

Table 4. Overall group statistics for the 15 meetings that we recorded and analyzed. All measures were computed automatically from the output of Meeter, as well as from the ground truth human annotations. Pearson correlations were computed per session.

4.4 Measures

Meeter detects dialogue acts, word counts, and speech prosody features, which were the measures in our analyses. Table 4 shows the statistics of the measures on our dataset. For our analyses, measures were preprocessed as follows:

- Dialogue act ratios were computed as the total number of information sharing/shared understanding dialogue acts, divided by the total number of spoken utterances.
- Word counts were computed as the absolute number of words spoken in the brainstorming session.
- Prosody feature values were aggregated per brainstorming session by calculating the mean over the whole session. Furthermore, prosody features were standardized with the z - transformation (with mean and standard deviation taken from the AMI meeting corpus), which is a common preprocessing step for speech prosody features to account for speaker variation [11].

4.5 Design and Analysis

To validate Meeter with respect to the current practice of manually transcribing and annotating dialogue acts, we performed two analyses. First, we conducted a correlation analysis between human annotations and Meeter annotations on word counts and ratios of information sharing and shared understanding dialogue acts. Second, we conducted an analysis of the Meeter classifier accuracy using human transcripts as input (to allow for a direct comparison to human-generated labels) and human labels as ground truth.

To analyze the relationship between group process and group outcome, we looked at information sharing, shared understanding, word counts, and speech prosody features. There were two dependent variables: (1) average group satisfaction; and (2) group performance operationalized as the number of ideas that were generated by the group during the session.

To investigate the relationship between group process measures and group satisfaction and group performance, an ordinal logistic regression was performed with the following between-session factors: word count, ratios of information sharing and shared understanding dialogue acts, as well as the four prosody features. Information sharing and shared understanding ratios were analyzed in two separate models. We performed these analyses twice: once using solely Meeter-generated measures and once using solely the human annotations for word counts, information sharing and shared understanding (but including Meeter-generated prosody features). The analyses of the link between prosody features and group outcomes used all four prosody features simultaneously.

Actual \ Predicted	Information Sharing	Shared Understanding	Other
Information Sharing	451 (82%)	53 (10%)	47 (9%)
Shared Understanding	69 (18%)	283 (74%)	31 (8%)
Other	51 (29%)	39 (22%)	86 (49%)
Precision/Recall/F1	82%/79%/0.80	74%/75%/0.74	49%/52%/0.50

Overall Precision/Recall/F1/Accuracy: 68%/68%/0.68/68%

Table 5. Confusion matrix of the dialogue act classifier of Meeter on the three classes *Shared Understanding*, *Information Sharing* and *Other*. The overall F1 score is 0.68. The data for evaluation comes from our newly collected brainstorming data and ground truth is based on human annotation. Both human annotations and Meeter annotation were collected from human transcript for comparison.

5 RESULTS

5.1 Validation Results

Table 4 shows the summary statistics from the 15 brainstorming sessions. The three statistics that were computed by both Meeter and human annotators (word counts, ratio of information sharing dialogue acts, ratio of dialogue acts indicating shared understanding) were highly correlated ($r = 0.87$ or higher) between Meeter and the human annotators, which supports our H1.

The average number of ideas for the bully problem was 8.4 (2.0), for the cell phone disturbance 9.2 (3.7), and for the news diversity 11.2 (2.3). None of these pairwise differences were statistically significant.

Furthermore, we analyzed the performance of Meeter to classify information sharing and shared understanding dialogue acts with our new dataset. To do this, we classified all dialogue acts in the human transcript using Meeter (as originally trained on the AMI corpus) and used the human annotations as ground truth. The confusion matrix resulting from this test can be seen in Table 5, with an overall F1 score of 0.68. This result also demonstrates Meeter's ability to generalize to an entirely different corpus from the one on which it was originally trained. Table 6 shows excerpts from three different brainstorming sessions and the corresponding dialogue act annotations. The examples show where Meeter classifies sentences correctly, but it also reveals where Meeter classifies sentences incorrectly.

5.2 More satisfied groups show more information sharing and shared understanding

Our analysis did show a statistically significant main effect of information sharing dialogue acts on group satisfaction, both for human-annotated data ($\chi^2_{1,N=15} = 10.74, p = 0.0010$), as well as the data from Meeter ($\chi^2_{1,N=15} = 11.85, p = 0.0006$) — groups that shared more information were more satisfied.

We also observed a significant positive main effect of the ratio of shared understanding dialogue acts on group satisfaction, both for human-annotated data ($\chi^2_{1,N=15} = 4.06, p = 0.0438$), as well as the data from Meeter ($\chi^2_{1,N=15} = 7.53, p = 0.0061$) — groups that showed more shared understanding were more satisfied.

No significant main effect of information sharing or shared understanding on group performance was found in either human-annotated or Meeter data.

Table 7 summarizes the group performance analysis results. The odds ratios give an idea of the effect sizes of the measures. For example, the odds ratio of 2 with 100 words unit-step size means that if 100 more words occur, the probability of group satisfaction being below a certain value doubles. Figure 3 shows the relationship between the measures of linguistic behavior and group satisfaction in our data.

Human/Meeter: **Bully problem**	Human/Meeter: **Cell phone disturbance**	Human/Meeter: **News diversity**
IS / IS : ✓ the students like student classes or students	IS / IS : ✓ so ban is one of them	OT/IS: you can use that pen again
SU/SU: ✓ yeah	SU/ IS: ban would be the quickest way or one	SU/SU: ✓ okay
IS / IS : ✓ maybe requiring young students to go through some sort of girl scout boy scout type training on not necessarily bullying topics but other leadership development	OT/SU: ban	SU/SU: ✓ the aggregator was a really good idea
SU/IS: like character development	SU/SU: ✓ okay	IS/OT: how about on facebook the newsfeed that you get is based on your preferences so maybe you could have based on the opposite of your preferences
SU/SU: ✓ yeah	IS /SU: there could just be more signage	SU/SU: ✓ sure
SU/SU: ✓ character development	SU/SU: ✓ yeah like be courteous	SU/SU: ✓ if that makes sense
SU/SU: ✓ exactly	SU/SU: ✓ yeah add signage	SU/IS: so you're getting if you're a liberal getting fox news or vice versa or something like that
IS / IS : ✓ i would say strong or stronger administrative responses to like alleged incidents	IS / IS : ✓ the acoustics of the vehicle could be constructed so that it minimizes sound	SU/SU: ✓ okay
IS / IS : ✓ stronger punishment as you get older	SU/OT: it could	SU/SU: ✓ somehow i don't think that would be very popular
OT/IS: i guess as a kid it's different but like in high school	SU/SU: ✓ if that's possible okay	SU/SU: ✓ i know it wouldn't be popular but
SU/SU: ✓ yeah	SU/IS: sure i don't know a lot about vehicle construction but maybe	OT/IS: just a thought

Table 6. Example brainstorming discussion excerpts for each of the topics. Every column is an excerpt of a dialogue on one of the three problems, and every row represents one dialogue act. Dialogue acts are annotated with *Human/Meeter* annotation, with the labels Information Sharing (IS), Shared understanding (SU), and Other (OT). Correct Meeter classifications are highlighted with checkmarks (✓).

	Measure \Predictor	**Odds Ratio (unit-step size)**	χ^2	**p**
Meeter	Word Count	2 (100)	6.12	0.0134*
	Information Sharing	0.65 (0.01)	11.85	0.0006*
	Shared Understanding	0.66 (0.01)	7.53	0.0061*
Human	Word Count	1.8 (100)	5.2	0.0225*
	Information Sharing	0.69 (0.01)	10.74	0.0010*
	Shared Understanding	0.86 (0.01)	4.06	0.0438*

Table 7. Summary of statistical analyses of information sharing, shared understanding, word counts, and group satisfaction. The odds ratios give an idea of the effect sizes of the measures — for example, the odds ratio of 2 with 100 words unit-step size means that if 100 more words occur, the probability of group satisfaction being below a certain value doubles.

In summary, previous literature suggests that both group performance and group satisfaction increase when groups share more information and develop a better shared understanding (H3). Our analysis of the dialogue acts and group outcomes aligns partially with these results, and we observed these results both for human-annotated data, as well as data reported by Meeter (H2).

5.3 More satisfied groups use fewer words

Our analysis did show a statistically significant main effect of word count as measured by Meeter ($\chi^2_{1,N=15} = 6.12, p = 0.0134$), as well as measured by human generated transcripts ($\chi^2_{1,N=15} = 5.2, p = 0.0225$), on group satisfaction. Groups that used fewer words reported higher overall satisfaction. No significant main effect of word count on group performance was found for either Meeter words counts ($\chi^2_{1,N=15} = 0.04, p = 0.84$) or human word counts ($\chi^2_{1,N=15} = 0.28, p = 0.59$). Previous literature suggests that both group performance and group satisfaction are affected by the word count in a group meeting (H4). Our analyses partially align with those prior results: the fewer words groups used, the more satisfied group members were with the brainstorming session.

Measure \Predictor	Odds Ratio (unit-step size)	χ^2	p
Harmonics-noise-ratio (HNR)	3.6 (0.1)	8.13	<.0001
Jitter	3.5 (0.1)	6.69	0.0097
Loudness peaks/second	0.31 (0.1)	4.7	0.0301
Voiced segments/second	1.1 (0.1)	0.22	0.64

Table 8. Summary of statistical analyses for speech prosody and group performance. Performance is measured as the number of ideas submitted. Since the data has been z-transformed, unit-step sizes can be interpreted in fractions of standard deviations — for example, , the odds ratio of 3.6 with 0.1 unit-step size means that if increasing the HNR by 0.1 its standard deviation, the probability of group performance being below a certain value increases to 3.6 its original value.

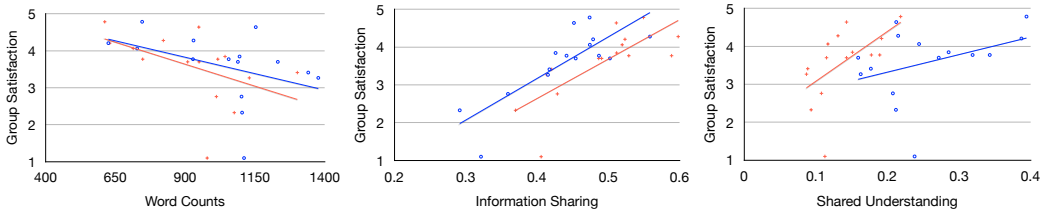

Fig. 3. Overview of how group satisfaction is affected by (A) word counts, (B) information sharing, and (C) shared understanding, as evaluated by both Meeter (red) and humans (blue). It can be seen that human and Meeter measures are correlated for all three measures.

5.4 Higher performing groups show higher levels of speech activation

Our data analysis with the speech prosody measures shows that three of the four prosody features have a significant main effect on group performance — the overall model shows that higher performing groups show higher levels of speech activation as measured by the speech prosody features ($\chi^2_{4,\,N=15} = 12.5, p = 0.0138$). No significant effect of any of the features on group satisfaction was found, overall leading to a partial support of H5. Table 8 shows the results of the group performance analysis for the four different prosody features.

6 DISCUSSION

With our dialogue act measures of information sharing and shared understanding, we introduce novel automatic quantitative language measures to analyze spoken conversations. In an analysis of a novel corpus of brainstorming meeting recordings, we found large and significant positive correlations between human and Meeter dialogue act labels. Further, both human annotations and automatic Meeter annotations lead to the same conclusions on the relationships between linguistic behaviors and group outcomes. Our findings suggest the promise of this approach for conducting group research more effectively through automatic language behavior detection.

With an overall accuracy result of 73%, our technical analysis shows the potential but also challenges involved in automatically analyzing group language behavior. The analysis revealed that erroneous classification mostly comes from ambiguous sentences or contextual references to previously spoken sentences. However, the comparison of different existing text classification models revealed no large improvement through modeling conversation context and other long term dependencies. One research question that we did not investigate is whether *pragmatics* could be modeled to capture more subtle signals in language. Further research in this area might be necessary to improve sequence classifiers.

A common concern with machine learning classification models is that of overfitting, which may lead to poor generalizability on novel examples. Our classifier performed well on the previously unseen examples from the AMI data set and on a completely new data set that we collected. Hence, our results suggest that the threat of overfitting has been largely averted.

Even if not all of the individual sentences were correctly classified, our study of group outcomes shows that the system accuracy was sufficient to capture the overall group signal characteristics — we provide empirical evidence that in group brainstorming settings, higher ratios of information sharing and shared understanding dialogue acts correlate with higher group satisfaction. While we could not provide further evidence for all of the measures (e.g., group performance seemed not to be affected by the language measures in our settings), our findings suggest that computer-based detection of language group processes is feasible.

We also found that more satisfied groups used fewer words, which might be explained by analyzing the relationship between information sharing/shared understanding, and word count — compared to the groups that used a lot of words, groups that used fewer words had higher proportions of dialogue acts related to information sharing and shared understanding. The transcript content suggests that this may have to do with a shift in conversation content of more talkative groups, leading to less relevant conversations and lower group satisfaction.

Furthermore, with the speech prosody measures that Meeter detects, we quantify not only *what* people talk about in a meeting but also *how* people talk. Our work provides further evidence for speech prosody being an important predictor of a group's performance — higher performing groups show an increase in group activation as measured by speech prosody. This aligns with findings from other settings such as negotiation, where social signals are important determiners for the effectiveness in a negotiation task [38]. These findings suggest that both prosody and language-based group processes are important elements of effective group interactions and that computational approaches are feasible for group research.

Meeter can be a useful tool for the CSCW community and others that study group conversations. The tool enables studying group language behaviors in a less tedious, more scalable way compared to manual analysis of group discussions. Meeter is available for download at https://github.com/BerndHuber/meeter_code and can be used as an off-the-shelf tool for group studies.

One potential further application of Meeter for groupware is in coaching: Educational research suggests that the most effective feedback is specific and immediate [14]. However, because coaching on group processes occurs in the context of actual work tasks, prior systems chose intervention strategies that provided aggregate and delayed (rather than specific and immediate) feedback so as to cause fewer interruptions [29, 43]. Whether such feedback is delivered via an ambient display (e.g., [29]) or via natural language (e.g., [43]), Meeter, which can be used to analyze conversation in real-time, is an important step towards supporting such specific and immediate feedback strategies— although its classification accuracy is not perfect, it captures well the relative proportions of different dialogue acts. We are particularly excited about the possibility of building on Meeter to develop a system that can be deployed outside the laboratory setting to offer tracking and feedback in the context of meetings with participants genuinely motivated to improve their group's outcomes.

7 FUTURE WORK

While we have not studied speech prosody and dialogue act related models on other group characteristics and processes such as identity development and group background, we believe this will be an important future research direction that may be studied using the Meeter system. We also believe that longitudinal studies with Meeter will be informative, e.g., through analyzing long-term trends of information sharing and shared understanding, or the long-term development of prosody, as, for example, those suggested by Gersick et al [13].

8 CONCLUSION

In this paper, we presented the Meeter tool for automatic detection of language group processes from speech using linguistic and prosody analysis. We showed that with Meeter, a number of important group behavior measures, can be detected automatically and that these measures are highly correlated with measurements produced by human annotators. We further validated the tool by studying relationships between group processes detected by Meeter and group outcomes: (1) Groups that showed higher levels of information sharing and shared understanding were more satisfied; (2) Groups that used fewer words were more satisfied; (3) More activated groups as measured through speech prosody, showed higher levels of group performance. Our study shows that tools that automatically and accurately quantify group behaviors could enable larger-scale studies and make it possible to conduct group behavior measurements in the wild, and such tools could enable novel automated interventions.

9 ONLINE APPENDIX

The code related to this paper can be found at https://github.com/BerndHuber/meeter_code.

10 ACKNOWLEDGEMENTS

The authors would like to thank Ofra Amir, Kenneth Arnold, Pao Siangliulue, Joseph Kim, and Karen Brennan for helpful comments and suggestions.

REFERENCES

[1] Michael Baker, Tia Hansen, Richard Joiner, and David Traum. 1999. The role of grounding in collaborative learning tasks. *Collaborative learning: Cognitive and computational approaches* 31 (1999), 63.

[2] Tony Bergstrom and Karrie Karahalios. 2007. Conversation Clock: Visualizing audio patterns in co-located groups. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*. IEEE, 78–78.

[3] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 28–39.

[4] Senthil Chandrasegaran, Chris Bryan, Hidekazu Shidara, Tung-Yen Chuang, and Kwan-Liu Ma. 2019. TalkTraces: Real-Time Capture and Visualization of Verbal Content in Meetings. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI 2019)*.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research* 12, Aug (2011), 2493–2537.

[6] Gregorio Convertino, Helena M Mentis, Mary Beth Rosson, John M Carroll, Aleksandra Slavkovic, and Craig H Ganoe. 2008. Articulating common ground in cooperative work: content and process. In *proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1637–1646.

[7] Carsten KW De Dreu, Bernard A Nijstad, and Daan van Knippenberg. 2008. Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review* 12, 1 (2008), 22–49.

[8] Barbara Di Eugenio, Pamela W Jordan, Richmond H Thomason, and Johanna D Moore. 2000. The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies* 53, 6 (2000), 1017–1076.

[9] Joan Morris DiMicco, Katherine J Hollenbach, Anna Pandolfo, and Walter Bender. 2007. The impact of increased awareness while face-to-face. *Human-Computer Interaction* 22, 1 (2007), 47–96.

[10] Gijsbert Erkens and Jeroen Janssen. 2008. Automatic coding of dialogue acts in collaboration protocols. *International journal of computer-supported collaborative learning* 3, 4 (2008), 447–470.

[11] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, and Shrikanth S Narayanan. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202.

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[13] Connie JG Gersick. 1988. Time and transition in work teams: Toward a new model of group development. *Academy of Management journal* 31, 1 (1988), 9–41.

[14] Arthur C Graesser, Mark W Conley, and Andrew Olney. 2012. Intelligent tutoring systems. (2012).

[15] Saul Greenberg and Michael Rounding. 2001. The notification collage: posting information to public and personal displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 514–521.

[16] Emitza Guzman, David Azócar, and Yang Li. 2014. Sentiment analysis of commit comments in GitHub: an empirical study. In *Proceedings of the 11th Working Conference on Mining Software Repositories*. ACM, 352–355.

[17] Gahgene Gweon, Soojin Jun, Joonhwan Lee, Susan Finger, and Carolyn Penstein Rosé. 2011. A framework for assessment of student project groups on-line and off-line. In *Analyzing interactions in CSCL*. Springer, 293–317.

[18] Verlin B Hinsz, R Scott Tindale, and David A Vollrath. 1997. The emerging conceptualization of groups as information processors. *Psychological bulletin* 121, 1 (1997), 43.

[19] Jesse Hoey, Tobias Schröder, Jonathan Morgan, Kimberly B Rogers, Deepak Rishi, and Meiyappan Nagappan. 2018. Artificial Intelligence and Social Simulation: Studying Group Dynamics on a Massive Scale. *Small Group Research* 49, 6 (2018), 647–683.

[20] David H Jonassen and Hyug Kwon. 2001. Communication patterns in computer mediated versus face-to-face group problem solving. *Educational technology research and development* 49, 1 (2001), 35.

[21] Norbert L Kerr and Scott Tindale. 2014. Methods of small group research. (2014).

[22] Joseph Kim and Julie A Shah. 2016. Improving Team's Consistency of Understanding in Meetings. *IEEE Transactions on Human-Machine Systems* 46, 5 (2016), 625–637.

[23] Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. 2008. Meeting mediator: enhancing group collaborationusing sociometric feedback. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 457–466.

[24] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[25] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 2017. What Makes a Strong Team?: Using Collective Intelligence to Predict Team Performance in League of Legends.. In *CSCW*. 2316–2329.

[26] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Computational social science. *Science* 323, 5915 (2009), 721–723.

[27] Linda Lebie, Jonathan A Rhoades, and Joseph E McGrath. 1995. Interaction process in computer-mediated and face-to-face groups. *Computer Supported Cooperative Work (CSCW)* 4, 2-3 (1995), 127–152.

[28] Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827* (2016).

[29] Gilly Leshed, Jeffrey T Hancock, Dan Cosley, Poppy L McLeod, and Geri Gay. 2007. Feedback for guiding reflection on teamwork practices. In *Proceedings of the 2007 international ACM conference on Supporting group work*. ACM, 217–220.

[30] Gilly Leshed, Diego Perez, Jeffrey T Hancock, Dan Cosley, Jeremy Birnholtz, Soyoung Lee, Poppy L McLeod, and Geri Gay. 2009. Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 537–546.

[31] Joseph F McCarthy, Ben Congleton, and F Maxwell Harper. 2008. The context, content & community collage: sharing personal digital media in the physical workplace. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 97–106.

[32] Joseph E McGrath. 1997. Small group research, that once and future field: An interpretation of the past with an eye to the future. *Group Dynamics: Theory, Research, and Practice* 1, 1 (1997), 7.

[33] 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.

[34] Eyal Ofek, Shamsi T Iqbal, and Karin Strauss. 2013. Reducing disruption from subtle information delivery during a conversation: mode and bandwidth investigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3111–3120.

[35] Michael J Owren and Jo-Anne Bachorowski. 2003. Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior* 27, 3 (2003), 183–200.

[36] Wei Pan, Wen Dong, Manuel Cebrian, Taemie Kim, James H Fowler, and Alex Sandy Pentland. 2012. Modeling dynamical influence in human interaction: Using data to make better inferences about influence within social systems. *IEEE Signal Processing Magazine* 29, 2 (2012), 77–86.

[37] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[38] Alex Pentland. 2004. Social dynamics: Signals and behavior. In *Proceedings of the third international conference on developmental learning (ICDL'04). Salk Institute, San Diego. UCSD Institute for Neural Computation*. 263–267.

[39] Yan Qu and Derek L Hansen. 2008. Building shared understanding in collaborative sensemaking. In *Proceedings of CHI 2008 Sensemaking Workshop*.

[40] Marc Schröder. 2003. Experimental study of affect bursts. *Speech communication* 40, 1-2 (2003), 99–116.

[41] J Bryan Sexton and Robert L Helmreich. 2000. Analyzing cockpit communications: the links between language, performance, error, and workload. *Journal of Human Performance in Extreme Environments* 5, 1 (2000), 6.
[42] Yang Shi, Yang Wang, Ye Qi, John Chen, Xiaoyao Xu, and Kwan-Liu Ma. 2017. IdeaWall: Improving creative collaboration through combinatorial visual stimuli. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 594–603.
[43] Yla R Tausczik and James W Pennebaker. 2013. Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 459–468.
[44] Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In *Interspeech 2016*.
[45] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science* 330, 6004 (2010), 686–688.

Received April 2019; revised June 2019; accepted August 2019