Recognizing conversational context in group interaction using privacy-sensitive mobile sensors

Dinesh Babu Jayagopi[1,2], Taemie Kim[3], Alex (Sandy) Pentland[3], and
Daniel Gatica-Perez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3] MIT Media Lab, USA
djaya@idiap.ch, taemie@media.mit.edu, sandy@media.mit.edu, gatica@idiap.ch

ABSTRACT

The availability of mobile sociometric sensors allows Computer-Supported Cooperative Work (CSCW) designers the possibility to enhance online meeting support through automatic recognition of conversational context. This paper addresses the task of discriminating one conversational context against another, specifically brainstorming from decision-making interactions using easily computable nonverbal behavioral cues. We hypothesize that the difference in the dynamics between brainstorming and decision-making discussions is significant and measurable using speech activity based nonverbal cues. We employ a set of nonverbal cues to characterize the entire group by the aggregation (both temporal and person-wise) of their nonverbal behavior. Our results on a dataset collected using privacy-sensitive sociometric badges show that the floor-occupation patterns in a brain-storming interaction are different from a decision-making interaction and we can obtain a discrimination accuracy as high as 87.5%.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Human Factors

Keywords

CSCW, Nonverbal behavior, Brainstorming, Decision-making.

1. INTRODUCTION

The automatic recognition of group interaction context in real life is a useful module for Computer-Supported Cooperative Work [4]. With the advent of ubiquitous and mobile sensing platforms, novel ways of collecting and visualizing group interaction behavior have been explored [2, 6] with the primary objective of influencing the group's behavior. Such applications would greatly benefit from the knowledge of the interaction context i.e. awareness about the interaction type, e.g. a cooperative vs competitive interaction, or a brainstorming vs decision-making phase.

Various social factors related to individual attributes (e.g. personality, social verticality, roles); relationship among individuals (close friends vs strangers, remote vs collocated); and goal at hand (cooperation vs competition) have begun to be studied in ubiquitous environments, mostly indoor environments equipped with microphones, cameras, and other sensors [3]. The availability of privacy-sensitive, mobile platforms to sense conversations [1], is opening the possibility of recording and analyzing behavioral aspects of real-life interactions without breaching the privacy of people, through online audio extraction of nonverbal cues without recording or storing raw audio.

Within this emerging domain, our work addresses the novel problem of discriminating two types of conversational context categories, namely brainstorming vs decision-making using computationally simple nonverbal cues extracted from sociometers. Laughlin and Ellis postulated that cooperative group tasks may be ordered on a continuum anchored by intellective and judgmental tasks [8]. According to them, intellective tasks are defined as tasks for which a demonstrably correct solution exists, as opposed to decision making or "judgmental" tasks where "correctness" tends to be defined by the group consensus. A different line of research asserts that group interactions have different dynamics depending on the group's objective [9]. A brainstorming session has a different objective as compared to that of a decision-making session and therefore demands a different response from the group members as well. Our work investigates whether these differences can be captured through nonverbal behavioral cues automatically extracted from sociometers; and if so whether the interaction type can be automatically inferred. With much of the work in modern workplaces becoming group-based, such interactions are indeed ubiquitous.

The specific research questions addressed in the paper are: Can brainstorming and decision-making meetings be discriminated from each other using only privacy-sensitive acoustic nonverbal cues? How good are single nonverbal cues? Does fusion of cues improve the performance?

Sections 2 discusses our approach. Section 3 introduces the experimental setup. Section 4 documents the results obtained, and Section 5 gives the conclusions of our analysis.

2. OUR APPROACH

Acoustic nonverbal cues are known to contain useful information to understand the behavior of individuals in groups [5]. Figure 1 shows our approach. We extract a number of nonverbal cues to characterize the group as a whole, and then use them to predict the group interaction context using standard machine learning techniques as described in detail in Section 3. These acoustic nonverbal cues are easily computable and privacy-sensitive [11]. The acoustic data is collected using wearable electronic badges.

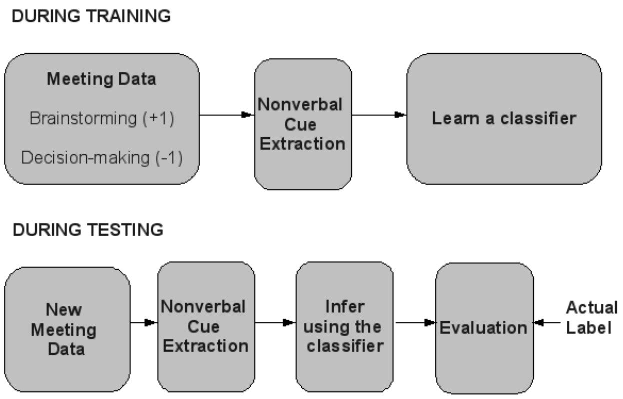

Figure 1: Our approach. Nonverbal cues are extracted to learn and infer the conversational context.

2.1 Meeting dataset

The dataset was collected from 24 groups of four members each. Each participant wore a sociometric badge - a wearable electronic badge with multiple sensors collecting interaction data. By interacting with other badges it can collect proximity data, other badges in direct line of sight, movement data, and speech features. Speech features collected by the badge include pitch, tone, volume, etc. Due to privacy concerns, we did not collect content of speech of any other features that may identify the speaker. The microphone of the sociometric badges collected speech variation data sampled at 50Hz, which is immediately processed on the badge so that only the processed data is saved on its SD card. The badges communicated with each other via 2.5GHz radio which allows synchronization error to be less than 0.003 msec. An example of the participants wearing sociometric badges can be found in Figure 2.

Figure 2: Example of an interacting group wearing sociometric badges around the neck.

The task given to subjects were based on a modification of the game "Twenty-Questions", replicating Wilson's experiments [10]. Each round consisted of two phases. In the first phase, each group was given a set of ten yes/no question-and-answer pairs. The groups were given 8 minutes to collaboratively brainstorm as many ideas that satisfy the set of question-and-answers. We label these interactions as 'brainstorming'. Then in the second phase, groups were given 10 minutes to ask the remaining ten questions of the Twenty-Question game to determine the correct solution. As this problem-solving phase mainly involved the group making decisions about the subsequent questions, we regard and label them as 'decision-making' interactions. In the second phase groups were asked to select a leader among themselves that would be the question-asker who communicates with the experimenter.

Each team began with one practice round and then participated in two rounds where their behavior was measured: one round in co-located settings and the other round separated into pairs into two rooms. When distributed, the group members were not able to see each other but were able to have verbal communication. The sequence of co-located and distribution was counter-balanced to minimize learning effect. The group leader was chosen during the practice round, and was kept consistent throughout the two measured rounds.

The dataset we used for our experiments was 9.8 hours of group conversational recordings.

2.2 Nonverbal cue extraction

We rely on robust cues that have been studied in nonverbal communication [7]. From the sociometer speech variation data, we first extract the binary segmentation (speech and non-speech for each participant) by thresholding the speaking energy values at $Fps = 10$ frames per second. A turn is a continuous period of time for which the person's speaking status is 1. A successful interruption is an event defined as follows: participant i starts talking while another participant j speaks, and i finishes his turn before j does. Conversely, an unsuccessful interruption can defined as participant i starts talking while another participant j speaks, and i finishes his turn before j does. These cues are illustrated in Figure 3. We first compute nonverbal cues at the individual level and then define cues at the group level.

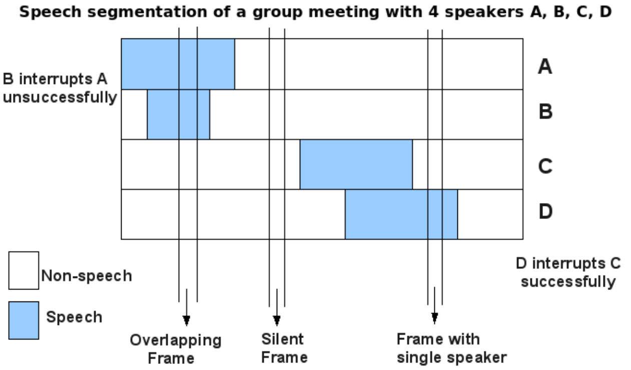
Speech segmentation of a group meeting with 4 speakers A, B, C, D
B interrupts A unsuccessfully
A
B
C
D
Non-speech
Speech
D interrupts C successfully
Overlapping Frame
Silent Frame
Frame with single speaker

Figure 3: Nonverbal cues extracted from speech segmentation.

Individual cues. From the speech segmentation, we compute Speaking Length (ISL_i) defined as the total time that participant i speaks, Speaking Turns (IST_i), Successful interruptions (ISI_i), and Unsuccessful interruptions (IUI_i) defined as the number of turns, successful interruptions, and unsuccessful interruptions accumulated over the entire meeting for every participant i, respectively.

Group cues. Three types of group cues are extracted. A first set of cues characterize the participation rates of the group by accumulating it over the participants. Let D denote the duration of the meeting. We compute the following from speaking length, turns, and interruptions of each of the participants: *Group Speaking Length(GSL)* $= \frac{\Sigma_i ISL(i)}{D}$, *Group Speaking Turns(GST)* $= \frac{\Sigma_i IST(i)}{D}$., *Group Successful Interruptions(GSI)* $= \frac{\Sigma_i ISI(i)}{D}$, *Group Unsuccessful Interruptions(GUI)* $= \frac{\Sigma_i IUI(i)}{D}$, *Group Successful Interruptions-to-Turns Ratio(GIT)* $= \frac{\Sigma_i ISI(i)}{\Sigma_i IST(i)}$, *Group Unsuccessful Interruptions-to-Turns Ratio(GUT)* $= \frac{\Sigma_i IUI(i)}{\Sigma_i IST(i)}$, resulting in 6 cues.

A second set of cues attempts to capture the overlap and silence patterns of a group as a whole. Let $T = D * Fps$ be the total number of frames in a meeting, S be the number of frames when no participant speaks, M be the number of frames when only one participant is speaking, and O be the number of frames when more than one participant talks. Then we define the following 3 cues: *Fraction of Silence(FS)* $= \frac{S}{T}$, *Fraction of Non-overlapped Speech(FN)* $= \frac{M}{T}$, and *Fraction of Overlapped Speech(FO)* $= \frac{O}{T}$.

A third set of cues characterizes which meeting is more 'egalitarian' with respect to the use of the speaking floor. Let **ISL** denote the vector composed of P elements, whose elements are $\frac{ISL(i)}{\Sigma_i ISL(i)}$ for the ith participant. Employing an analogous notation for **IST**, **ISI**, and **IUI**, these vectors are first ranked and then compared with the uniform (i.e. "egalitarian") distribution i.e. a vector of the same dimension with values equal to $\frac{1}{P}$. The comparison is done using the Bhattacharya distance (a distance measure useful to compare probability distributions and bounded between 0 and 1). For our case 0 would correspond to a egalitarian meeting and 1 corresponds to a one-man show. This results in 4 cues: *Group Speaking Length Egalitarian Measure (GLEM)*, *Group Speaking Turns Egalitarian Measure (GTEM)*, *Group Successful Interruption Egalitarian Measure (GIEM)*, and *Group Unsuccessful Interruptions Egalitarian Measure (GUEM)*.

2.3 Meeting type prediction

We used two supervised models to classify the group interaction type. The first is a Gaussian Naive-Bayes classifier, which assumes 1. the features are independent given the class and 2. the conditional densities are a unimodal Gaussian. Let A and B denote the class labels. Also, let $f_{1:N} = (f_1, f_2, ...f_N)$ denote the feature set and f_1, f_2, $...f_N$ the individual features. Then the log-likelihood ratio is given, by using Bayes' theorem and cancelling the common terms as follows:

$$log(\frac{P(A|(f_{1:N}))}{P(B|(f_{1:N}))}) = log(\frac{\prod_{k=1}^{N} P(f_k|A)P(A)}{\prod_{l=1}^{N} P(f_l|B)P(B)}) \qquad (1)$$

The probabilities $P(f_k|A)$ or $P(f_l|B)$ are estimated by fitting a Gaussian to the data from the respective class and the ratio of the priors are inferred from the data. It is to be noted that for our dataset, this ratio is 1 and the prior is uninformative. The second model is an SVM classifier, employing a linear kernel, using $(f_1, f_2, ...f_N)$ as features.

3. EXPERIMENTAL SETUP

As described in Section 2.1, we have 24 participant groups, solving two "Twenty-questions" games, one in collocated and the other in distributed settings. Each game involved a brainstorming phase followed by a decision-making phase. In order to model the difference between brainstorming and decision-making interactions, we define the following four datasets and three binary classification tasks.

Dataset A and B consists of 24 brainstorming meetings and decision-making meetings in distributed scenario respectively. Dataset C and D consists of 24 brainstorming meetings and decision-making meetings in collocated scenario respectively.

Task 1: The first task is to distinguish between brainstorming and decision-making meetings during the distributed setting. We classify Dataset A versus Dataset B. Each class has 24 datapoints.

Task 2: The second task is to distinguish between brainstorming and decision-making meetings during the collocated setting. We classify Dataset C versus Dataset D. Each class has 24 datapoints.

Task 3: The third task is to distinguish between brainstorming and decision-making meetings. We classify Dataset A+C versus Dataset B+D. Each class has 48 datapoints.

Group Adaptation Step. To account for the feature variations among the 24 groups, we perform z-normalization on the group nonverbal cues before using it for classification as follows : $\hat{f}^s = (f^s - \mu_f)/(\sigma_f), \forall s \in A, B, C, D$ where $\hat{f}$ and f are the values of the feature in a particular scenario s before and after z-normalization respectively.

In all cases, we use a leave-one-out approach.

4. RESULTS

We first analyze the performance of single cues. Figure 4 shows the performance of the group cues for Task 1 (distributed setting). Random performance for all the tasks is 50%. Though we experimented with two different classifiers, as described in Section 2.3, we report the results using the Gaussian Naive Bayes classifier only as the results are similar when a linear SVM is employed (omitted for space reasons). Fraction of Silence (FS), Fraction of Overlap (FO), and Group Speaking Length (GSL) were the top performing cues with an accuracy of 79.2%. Figure 5 shows for Task 2 (collocated setting). Fraction of Silence (FS), Group Speaking Length (GSL), and Fraction of Nonoverlapped speech (FN) were the top performing cues with a performance of 81.3%, 81.3%, and 75.0% respectively. For Task 3, a similar trend was observed. Fraction of Silence (FS), Group Speaking Length (GSL), and Fraction of Overlap (FO) gave the best classification result with an accuracy of 80.2%, 78.1%, and 74% (Figure 6). All these results are statistically significant compared to the random performance at 5% level. The results suggest that some of the investigated features indeed have discriminating power. Also, it is interesting to observe the following trend: Most groups have higher Fraction of Silence during brainstorming and higher Group Speaking

Length and Fraction of Overlap while making decisions.

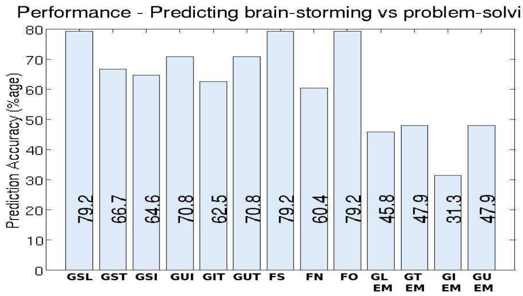

Figure 4: Performance of the group cues on classifying the brainstorming and decision-making meetings during distributed setting (Task 1).

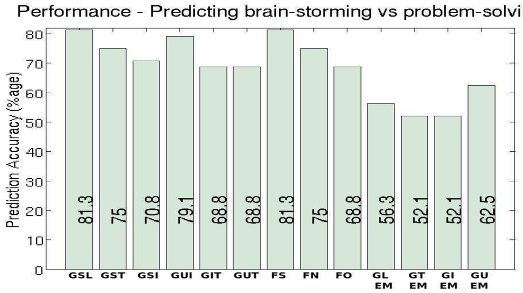

Figure 5: Performance of the group cues on classifying the brainstorming and decision-making meetings during collocated setting (Task 2).

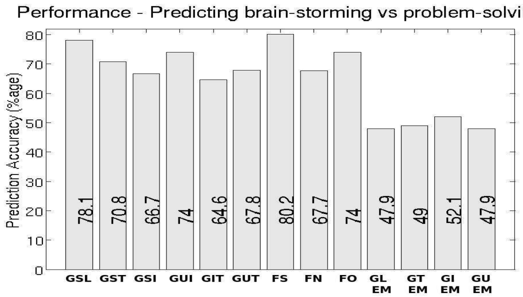

Figure 6: Performance of the group cues on classifying the brainstorming and decision-making meetings (Task 3).

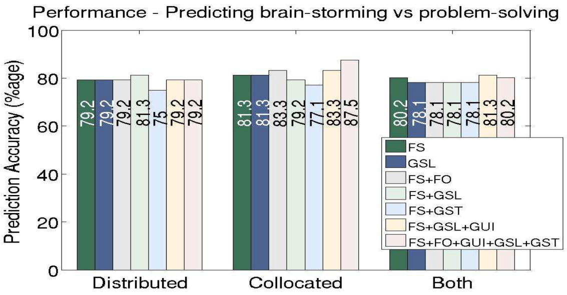

Figure 7: Performance of combination of group features on predicting the brainstorming and decision-making meetings.

Later, we also combined the cues to investigate if there is complementarity among them. Figure 7 shows the classification performance of some combinations using the log-likelihood classifier for each of the three tasks. The combination of Fraction of Silence (FS) and Group Speaking Length (GSL) improves the classification accuracy to 81.3% in the distributed setting (Task 1). The combination of Fraction of Silence (FS) and Fraction of Overlap (FO) improves the classification accuracy to 83.3% in the collocated case (Task 2). When Group Speaking Length (GSL), Group Speaking Turns (GST), and Group Unsuccessful Interruptions (GUI) were added the accuracy improved to 87.5%. For the combined dataset (Task 3), the combination of Fraction of Silence (FS), Group Speaking Length (GSL), and Group Unsuccessful Interruptions (GUI) improved the classification accuracy to 81.3%.

5. DISCUSSION AND CONCLUSIONS

In this work, we hypothesised and verified that '*brainstorming interactions often have different group dynamics compared to decision-making meetings*' and '*that such differences can be reasonably captured using automatically extracted nonverbal behavior*'. Our nonverbal cues, obtained using privacy-sensitive sociometric sensors, characterized the entire group by the aggregation (both temporal and person-wise) of their nonverbal behavior. We could discriminate these interactions with an accuracy of up to 87.5% and 81.3% in the collocated and distributed setting respectively. The group adaptation step helps in obtaining good performance and also tackling inter-group differences in individuals and relationship among individuals (as the mean behavior is subtracted out). In the future, we would like to use more data and an expanded feature set to include prosodic cues and temporal aspects of cues to explore generative models that would characterize brainstorming and decision-making interactions better.

Acknowledgments: This research was supported by the Swiss NCCR IM2. Taemie Kim would like to thank Anges Chang for her support in the experiments.

6. REFERENCES

[1] T. Choudhury et al. The sociometer: A wearable device for understanding human networks. In *CSCW'02 Workshop: Ad hoc Communications and Collaboration in Ubiquitous Computing Environments*, 2002.

[2] J.M. DiMicco et al. Using visualizations to review a group's interaction dynamics. In *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 2006.

[3] D. Gatica-Perez. Automatic Nonverbal Analysis of Social Interaction in Small Groups: a Review. In *Image and Vision Computing, Special Issue on Human Behavior*, volume 27, Dec 2009.

[4] J. Grudin. Computer-supported cooperative work: History and focus. *Computer*, 27(5):19–26, 1994.

[5] J.A. Hall et al. Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, 131(6):898–924, 2005.

[6] T. Kim et al. Meeting mediator: enhancing group collaboration using sociometric feedback. In *Proc. ACM 2008 conference on CSCW*, pages 457–466. ACM, 2008.

[7] M.L. Knapp and J.A. Hall. *Nonverbal communication in human interaction*. Holt, Rinehart and Winston NY, 1978.

[8] P.R. Laughlin and A.L. Ellis. Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Exp. Social Psychology*, 22(3):177–189, 1986.

[9] J.E. McGrath. *Groups: Interaction and Performance*. 1984.

[10] D.S. Wilson et al. Cognitive cooperation: when the going gets tough, think as a group. *Human Nature*, 15(3):225–250, 2004.

[11] D. Wyatt, T. Choudhury, and J. Bilmes. Conversation detection and speaker segmentation in privacy sensitive situated speech data. In *Proc. Interspeech*, 2007.