CHARACTERIZING CONVERSATIONAL GROUP DYNAMICS USING NONVERBAL BEHAVIOUR

Dinesh Babu Jayagopi[1,2], *Bogdan Raducanu*[3] *and Daniel Gatica-Perez*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
[3]Computer Vision Center, Barcelona, Spain
{djaya, gatica}@idiap.ch, bogdan@cvc.uab.es

ABSTRACT

This paper addresses the novel problem of characterizing conversational group dynamics. It is well documented in social psychology that depending on the objectives a group, the dynamics are different. For example, a competitive meeting has a different objective from that of a collaborative meeting. We propose a method to characterize group dynamics based on the joint description of a group members' aggregated acoustical nonverbal behaviour to classify two meeting datasets (one being cooperative-type and the other being competitive-type). We use 4.5 hours of real behavioural multi-party data and show that our methodology can achieve a classification rate of upto 100%.

***Index Terms*—** Competitive and cooperative meetings, group dynamics, nonverbal cues

1. INTRODUCTION

Automatic nonverbal analysis of face-to-face group meetings to characterize group behaviour is a relevant research problem in social computing. The potential applications of such an analysis include identifying irresponsible behaviour and leadership skills and monitering team cohesiveness. From a human resource perspective, analyzing group behaviour could signal the need for a team-building exercise or a leadership change. Tracking teams could also indicate what the teams are mostly engaged in - cooperative or competitive behaviour.

Group meetings have different dynamics depending on the group's objective [1]. Competitive meetings like debates, whose primary objective is that of 'resolving' or winning an argument, demand a different response from the members vis-a-vis that of colloborative meetings like brain storming sessions, whose primary objective is to cooperate and accomplish a task together.

The research question addressed in the paper is as follows: Can cooperative and competitive meetings be discriminated from each other using only nonverbal behaviour? Various works have analyzed face-to-face meetings group conversations [2], attempting to characterize *individual* social attributes like dominance [3, 4],status [5], roles [6, 7] and personalities [8]. The works differ widely in the cues and the models they employ. Very few works however have attempted to characterize *groups as a whole*. In [9] four short conversations were characterized in terms of their interactivity and centralization. The work used very few meetings and was not exhaustive in the analysis of the features. Also, the attempt to classify the meeting into monologues and dialogues was not validated using ground-truth. In our work, by segmenting each participants' audio signal into 'speech' and 'silence' and then extracting some 'intutive' and 'computationally simple' features, we obtain aggregated acoustical nonverbal behaviour of the group as a whole. Such features capture the competitiveness and allow for comparing groups of varying size and meetings of different durations. Our best methods have an accuracy of 100%, showing the effectiveness of our feature set.

2. OUR APPROACH

In order to test our hypotheses that '*competitive meetings have different group dynamics as compared to cooperative meetings*' and '*the difference in group dynamics can be captured using nonverbal behaviour.*', we adopt the following methodology (Figure 1): We choose two group meeting datasets differing in their objective (and competitiveness), one from the Augmented MultiParty Corpus (AMI) corpus [10] and The Apprentice [11] (a subset of US reality TV show) dataset (example snapshots in Figure 2). While the groups in AMI Corpus need to cooperate to design a remote control, the groups in the Apprentice dataset need to debate and help the boss decide the person who is to be fired from the team. The *individual* nonverbal behaviour description is obtained by extracting speech activity and then computing features like speaking time, turns and interruptions, which characterize the floor occupation of individuals. Then *group* nonverbal behaviour is inferred by either aggregating these features (for example 'how much this group talks per unit time') or comparing the *individual* nonverbal behaviour with others' behaviour (for example 'do every participant take equal number of turns or interruptions?'). The meeting type is predicted using a simple 'likelihood ratio' based classifier as well as a Support Vector Machine (SVM) with a linear and quadriatic kernel.

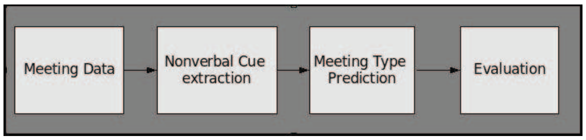

Fig. 1. Block Diagram of our work

2.1. Meeting datasets

The AMI meeting dataset (cooperative meetings)

The teams in the AMI meeting dataset consisted of 4 participants, who were given the task of designing a remote control over a series of meeting sessions. Each participant was assigned distinct roles: 'Project Manager', 'User Interface specialist', 'Marketing Expert', and 'Industrial Designer'. During each session, the team was required to carry out certain tasks, such as a presentation on particular subjects related to the task, or a discussion about a particular aspect. To encourage natural behaviour, the meetings were not scripted and the teams met over several sessions to achieve the common goal.

Fig. 2. *Top*: Snapshot from an AMI meeting, showing the participants from two side-view camera view. *Bottom*: Snapshot of an Apprentice meeting - highlighting the high-status leader (Trump) - *bottom left* and a long-shot of the board-room meeting - *bottom right*

The Apprentice meeting dataset (competitive meetings)

The teams in The Apprentice meeting dataset have a variable number of participants (5 to 11). The group has a well-defined hierarchy, with Donald Trump being the person with highest status and the objective of the group is to fire one of the members. On one side of the meeting room we have the 'candidates board' and on the other side we have the 'executive board'. The executive board is formed by Trump together with other persons (usually two) which will help him make the decision regarding the candidate who will be fired. The data collected for our study belongs to the 6th season of a TV show, which was aired in early 2007.

Figure 2 shows a snapshot of both meetings.

2.2. Nonverbal cue extraction

For the AMI data we extract the following vocalic cues from the four close-talk microphones attached to each of the participants. Firstly, we extract speaking energy and speaking status.

Speaking energy: The starting point is to compute the real-valued speaker energy for each participant using a sliding window at each time step. A window of 40 ms was used with a 10 ms time shift.

Speaking status: From the speaking energy, a binary variable was computed by thresholding the energy values. This indicates the speaking / non-speaking (1/0) status of each participant at each time step.

For the Apprentice dataset, we had only one audio channel available as we used the show broadcast. Due to the recording conditions (background music for the whole duration of each meeting), for our study we decided to manually produce the speaker segmentation for each participant.

Figure 3 summarises the cue extraction process.

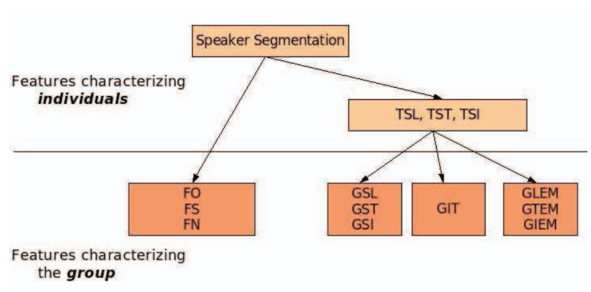

Fig. 3. Nonverbal Cue Extraction

From the speaking status of all the participants, the following three sets of features were computed: Let T be the total number of frames in a meeting, S be the number of frames when no participant speaks, N be the number of frames when only one participant talks and O be the number of frames when more than one participant talks.

- **Fraction of Overlapped Speech (FO):** $FO = \frac{O}{T}$.
- **Fraction of Silence (FS):** $FS = \frac{S}{T}$.
- **Fraction of Non-overlapped speech(FN):** $FN = \frac{N}{T}$.
- Speaking Length (TSL(i)): This feature considers the total time that participant i speaks according to his speaking status.
- Speaking Turns (TST(i)): We define a turn as a continuous period of time for which the person's speaking status is 'true'. TST is accumulated over the entire meeting for participant i.

- Successful Interruptions (TSI(i)): The cumulative number of frames that participant i starts talking while another participant j speaks, and j finishes his turn before i does, i.e. only interruptions that are successful are counted.

As a second set of cues, from speaking length, speaking turns and interruptions of each of the participants, the following additional features are computed to characterize the joint group behaviour.

- **Group Speaking Length (GSL):** This feature measures how much the participants speak put together (in aggregation) per unit time($\frac{\Sigma_i TSL(i)}{MeetingDuration}$).
- **Group Speaking Turns (GST):** This feature measures how many turns the participants put together take per unit time($\frac{\Sigma_i TST(i)}{MeetingDuration}$).
- **Group Speaking Interruption (GSI):** This feature measures how many interruptions the participants put together make per unit time($\frac{\Sigma_i TSI(i)}{MeetingDuration}$).
- **Group Speaking Interruption-to-Turns Ratio (GIT):** This feature measures the ratio of total interruptions the participants make to the total turns they take($\frac{\Sigma_i TSI(i)}{\Sigma_i TST(i)}$).

A third set of cues is derived from the hypothesis that co-operative meetings tend to be more 'egalitarian' with respect to the use of speaking floor, and that a description of group behaviour based on the distribution of nonverbal cues can be used cleverly for this purpose.

Let **TSL** denote the vector composed of P elements, whose elements are $TSL(i)$. i denotes the participants who are P in number. Employing an analogous notation for **TST** and **TSI** and then normalizing **TSL**, **TST**,and **TSI**, these vectors are first ranked and then compared with that of the uniform (i.e. egalitarian) distribution (a vector of the same dimension with values equal to $\frac{1}{P}$). The comparison is done using the Bhattacharya distance (a bounded distance measure useful to compare probablility distributions). The distance function always returns a value between 0 and 1. For our case 0 signifies a egalitarian meeting and 1 corresponds to a one-man show. This results in 3 features.

- **Group Speaking Length Egalitarian Measure (GLEM)**
- **Group Speaking Turns Egalitarian Measure (GTEM)**
- **Group Speaking Interruptions Egalitarian Measure (GIEM)**

Figure 4 shows the empirical distribution of the two features - GIT and GTEM. As one can observe, these two features are discriminative.

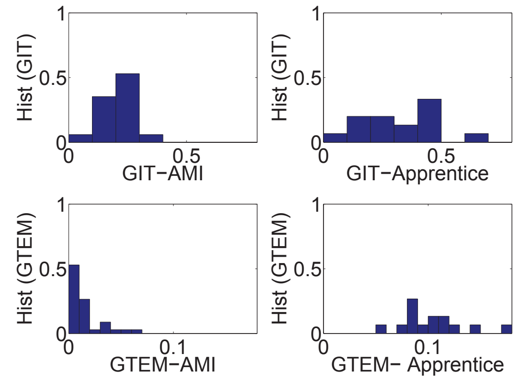

Fig. 4. Empirical distibution for **GIT** and **GTEM** in the two meeting datasets

2.3. Meeting Type Prediction

We used two supervised models to predict the meeting type. The first is a classifier using a log-likelihood ratio, with an independence assumption among the features. Let A denote the event $MeetingType = AMI$ and B the event $MeetingType = The\ Apprentice$. Also, let $(f_1, f_2, ...f_N)$ denote the feature set and f_1, f_2, $...f_N$ the individual features. Then the log-likelihood ratio is given as follows (expanding using Bayes' theorem and cancelling the common terms)

$$log(\frac{P(A|(f_1, f_2, ...f_N))}{P(B|(f_1, f_2, ...f_N))}) \quad (1)$$

$$= log(\frac{P(f_1|A).P(f_2|A)..P(f_N|A).P(A)}{P(f_1|B).P(f_2|B)..P(f_N|B).P(B)}) \quad (2)$$

The probabilities $P(f_n|A)$ or $P(f_n|B)$, where f_n is an individual feature, are estimated by fitting a Gaussian to each of the classes and the ratio of the priors are inferred from the data. The second model is an SVM classifier, employing a linear and a quadriatic kernel, using $(f_1, f_2, ...f_N)$ as features.

3. EVALUATION

We used 34 five-minute AMI meeting segments where there is full-agreement of multiple human annotators on the most dominant person (in order to control the variable - presence of a dominant leader in the apprentice meetings). All these meetings had 4 participants and the total data was approximately 170 minutes.

The Apprentice data set is formed of 15 meetings. These meetings have an average duration of 6 minutes and a total duration of 90 minutes. The number of participants has a median of 7.

Our final dataset consists of 49 meetings (34 from AMI and 15 from Apprentice). In order to evaluate the models we adopt a leave-one-out cross-validation strategy to classify the meetings and report the classification accuracy (Table 1).

Features	Accuracy(%) (likelihood)	Accuracy(%) (SVM-lin)	Accuracy(%) (SVM-quad)
FO	63.2	69.4	69.4
FS	67.3	69.4	69.4
FN	69.3	69.4	69.4
GSL	65.3	69.4	67.3
GST	69.3	69.4	69.4
GSI	63.2	69.4	69.4
GIT	85.7	83.6	83.6
GLEM	61.2	67.3	69.4
GTEM	**93.8**	**93.8**	**98.0**
GIEM	71.5	69.4	75.5
GIT,**GTEM**	**95.9**	**98.0**	**100**

Table 1. Accuracy (%) of speech activity based nonverbal cues for predicting the meeting type.

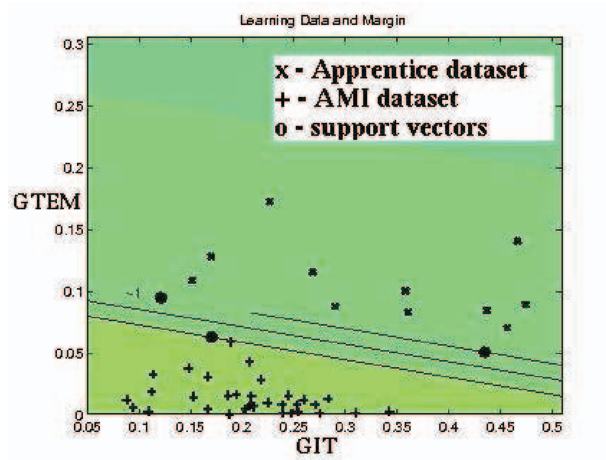

Fig. 5. Classification using SVM in the feature space of **GIT** and **GTEM**

While interpreting the results, it is to be noted that due to the difference in the number of samples between the two datasets, if an algorithm always predicts AMI as the meeting class always would perform with an accuracy of 69.4%. Also, a random prediction would give an accuracy of 50%.

Features like Fraction of Overlapped Speech(**FO**), Fraction of Silence(**FS**), Fraction of Non-Overlapped Speech(**FN**), Group Speaking Length(**GSL**), Group Speaking Turns(**GST**) and Group Speaking Interruptions(**GSI**) were not discriminative. Though we expected that in competitive meetings, the interruption rate (**GSI**) and the proportion of overlap (**FO**) would be more, our classification results did not show that. On the other hand, meetings could be discriminated when using the proportion of interruptions in the turns (**GIT**) and the distribution of turns and interruptions among participants (**GTEM** and **GIEM**). Figure 5 illustrates how an SVM with a linear kernel in the joint space of **GIT** and **GTEM** classifies the two meeting datasets. Also, it was interesting to observe that the features derived from speaking length were not as effective, although they were the best for other tasks like predicting the most dominant person in a meeting [3].

The distribution of speaking turns, how egalitarian it is, captures the competitiveness among the group members very effectively. Also, along with a slightly complementary feature 'the proportion of interruptions in the turns', this feature predicts the meeting type almost perfectly.

4. CONCLUSION

In this paper we investigated the problem of characterizing types of group meetings using nonverbal turn taking behaviour. Specifically, we attempted to classify two datasets, differing in their goals and level of competitivess. We verified our hypotheses that '*competitive meetings have different group dynamics as compared to cooperative meetings*' and that '*the difference in the group dynamics can be captured using nonverbal behaviour*'. Our methods could classify the meetings with an accuracy of upto 100% which is encouraging and suggests that the characterization of entire group by the aggregation (both temporal and person-wise) of their nonverbal behaviour is promising. One limitation that needs to be overcome in future work is the moderate size of the datasets we used. Future work will investigate all possible feature combinations and prosody-based characterization of joint behaviour.

Acknowledgments: This research was partly supported by the EU project AMIDA, the Swiss NCCR IM2.

5. REFERENCES

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

[2] D. Gatica-Perez, "Analyzing group interactions in conversation: a review," in *Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Heidelberg, Sep. 2006.

[3] D. Jayagopi et al., "Modeling dominance in group conversations using nonverbal activity cues," *Special issue - IEEE Transactions on Audio, Speech and Language Processing, Mar 2009.*

[4] R.J. Rienks and D. Heylen, "Automatic dominance detection in meetings using easily detectable features," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, Jul. 2005.

[5] D. Jayagopi et al., "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *Proc. ICMI*, Chania, Greece, Oct. 2008.

[6] A. Vinciarelli, "Speakers Role Recognition in Multiparty Audio Recordings Using Social Network Analysis and Duration Distribution Modeling," *IEEE Transactions on Multimedia, Oct 2007.*

[7] W. Dong et al., "Using the influence model to recognize functional roles in meetings," in *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, New York, NY, USA, 2007, pp. 271–278, ACM.

[8] F. Pianesi et al., "Multimodal Recognition of Personality Traits in Social Interactions," in *Proc. ICMI*. Chania, Greece, 2008.

[9] K. Otsuka et al., "Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns," in *Proc. ACM CHI Extended Abstract*, Montreal, Apr. 2006.

[10] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," in *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh,UK, Jul. 2005.

[11] The Apprentice, "http://www.nbc.com/the_apprentice/," .