Modelling Participant Affect in Meetings with Turn-Taking Features

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published in:
Proceedings of WASSS 2013

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abstract
This paper explores the relationship between turn-taking and meeting affect. To investigate this, we model post-meeting ratings of satisfaction, cohesion and leadership from participants of AMI corpus meetings using group and individual turn-taking features. The results indicate that participants gave higher satisfaction and cohesiveness ratings to meetings with greater group turn-taking freedom and individual very short utterance rates, while lower ratings were associated with more silence and speaker overlap. Besides broad applicability to satisfaction ratings, turn-taking freedom was found to be a better predictor than equality of speaking time when considering whether participants felt that everyone they had a chance to contribute. If we include dialogue act information, we see that substantive feedback type turns like assessments are more predictive of meeting affect than information giving acts or backchannels. This work highlights the importance of feedback turns and modelling group level activity in multiparty dialogue for understanding the social aspects of speech.

Index Terms: Turn-taking, dialogue, affect, groups, meetings, dialogue acts, satisfaction, cohesion, leadership.

1. Introduction
Humans are social animals. The success of meetings is not just a matter of the correct information flow, it also requires that participants to build consensus and rapport. Understanding the relationship between what happens in a meeting and affective outcomes like satisfaction, cohesiveness and perception of leadership is useful not just for meeting analysts but also for getting the design right in a range of technologies. Being able to characterize what makes for an affectually good meeting has implications for the turn-taking strategies of conversational agents attempting to convey information while building group cohesion. Similarly, understanding participant affect has implications for how technology can be used to facilitate meetings. For example, we would like to know if turn-taking patterns change information flow in a way that influences user satisfaction or productivity in computer mediated environments like video-conferencing.

Modelling meeting affect adds more broadly to our knowledge of how speech is used in social signalling. Most previous work on this has focused on involvement (similarly, interest and engagement) [1, 2, 3] or dominance [4, 5]. While understanding these social components is important in its own right, the implications of these studies for higher level meeting interpretation is unclear. For example, does dominant behaviour lead to perception of good leadership? Does active participation (i.e. involvement) from more individuals lead to greater group cohesiveness? In general, progress in this area has been hampered by the lack of direct automatic measures of meeting behaviour at the group level and the cost of employing methods that require these properties to be hand-annotated. To make some headway on these issues, we use a corpus of role-played meetings to show that automatic measures of interaction can explain differences in affectual outcome.

2. Background
Previous work on social signals in speech and group dynamics suggests that direct measures of how much participants speak or don’t speak, i.e. the turn-taking structure, should play a major role in objective and affectual meeting outcomes. For example, brainstorming groups with fewer inter-speaker pauses produce fewer ideas but are more likely by participants [6]. Greater amounts of silence also correlated with higher disagreeability ratings of others in [7]. Substantive contributions were found to improve ratings in [8]. However, face-to-face meetings, characterized by shorter turns and more overlapping speech, had higher satisfaction ratings than video-conference meetings in [9], even when no difference in the quality of solutions the groups came up with was found. This suggests that more talk and less strict turn-taking in general leads to a more positive outlook, even if it doesn’t lead to better task solutions.

Feedback utterances and their timing have also been identified as potentially important factors for understanding affect in spoken dialogue. However, results on the role of backchannels, for example, have not been consistent across different studies. Participants who interrupted more and produced more backchannels were less liked by external raters in task-oriented dialogues [10]. Similarly, male speed date participants perceived as friendly used less backchannels but more turn overlap in [11]. In [12], however, participants’ production of short utterances correlated with how much participants said they were seeking to avoid disagreements, while backchannel counts were correlated with self-ratings of agreeableness, as well as conscientiousness and openness in ratings by other non-dialogue participants in [13]. Backchannels were not correlated with group likeability measures in [7], although the difference in the top two leadership scores per group was correlated with skew in backchannel distribution across participants.

To develop realistic conversational dialogue strategies we need to know when and how much to talk. Previous work incorporating automatic measures of turn-taking have focused on detection of involvement or dominance rather than evaluative aspects of meeting affect like satisfaction. So, we would like to know if equal turn-taking and less rigid turn-taking structure lead to more positive participant affect, and if so, what measures we can use to characterize this automatically. We also need to consider what sort of participation is appropriate. In that vein, we would like to know whether characteristics of conversational speech, like short feedback and overlap, are beneficial in this re-
I am satisfied with the group's discussion.
There was too much information.
I find the members of the group helpful.
All in all, I am very satisfied.
I find the members of the group trustworthy.
I find the members of the group irritating.
Every team member had sufficient opportunity to make
All team members received sufficient attention.
All available information is being used.
I think my contributions affected the group discussion.

3.3. Turn-Taking Features

The turn-taking measures used in the following are calculated using spurts: segments separated by at least 500ms silence [16], where we use word alignments to mark silence. In the following all continuous group and individual level features are converted to z-scores so that estimated effects are easier to compare. At the group level we look at measures of participation equality and predictability of turn-taking structure. Participation equality \( P_{eq} \) is defined as [17]:

\[
P_{eq} = 1 - \frac{\sum_{i=1}^{N} (T_i - T)^2 / T}{E},
\]

where \( N \) is the number of participants, \( T_i \) total spurt time for participant \( i \), \( T = (\sum_{i=1}^{N} T_i) / N \) (i.e. equal participation). \( E \) represents the maximum possible value of the term under the sum: the average distance from equal participation (so \( E \) represents the case when only one participant speaks for the entire meeting). Values closer to 1 indicate greater equality. Similarly, let \( H(Y|X) \) be the conditional entropy of speaker \( Y \) being the next participant to speak after \( X \) begins their spurt, with \( H_{max}(Y|X) \) representing the maximal possible value for groups of a given size. Turn-taking freedom \( F_{cond} \) is defined as

\[
F_{cond} = 1 - \frac{H_{max}(Y|X) - H(Y|X)}{H_{max}(Y|X)}.
\]

So, \( F_{cond} = 0 \) when turn-taking follows a strict order (i.e. only speaker \( y \) follows \( x \)) and is 1 when every speaker follows everyone else in equal proportion. To examine the role of overlaps and possible interruptions, we measure barge-in rate (\( bargein\cdot rate \)) as the number of times any spurt is overlapped by a later starting spurt. We also measure the proportion of the interval that is silence (\( sil\cdot prop \)).

At the individual level we look at participant speaking time (\( ispk\cdot prop\cdot abs \)) and number of Very Short Utterances (VSUs, \( ivsu\cdot rate \)), both divided by the meeting duration. Here VSUs are defined as spurs that have duration less than 500 ms. These are likely to represent backchannels or other forms of short feedback [19].

3.4. Dialogue Acts

To get an idea of the type of contribution the participants make we include measures of dialogue act usage based on manual annotations of the AMI corpus [20]. This annotation scheme includes 15 dialogue acts involving the giving and elicitng of information, assessments, suggestions, offers, and comments about understanding. The label set also includes backchannels, stalls, and fragments, ‘be positive’ and ‘be negative’ categories. The latter two labels apply to utterances that do not fit into any of the previously mentioned categories and are deemed to have a purely social impact. The remaining utterances are assigned the category ‘other’. To normalize across different meeting lengths we look at the number of instances of each dialogue act category proffered by the individual divided by the meeting duration (i.e. DA rates). We also look at the combined rates for participants other than the current rater.

3.5. Linear Model

We model the ratings averaged over group dynamics and for individual items for each participant in terms of a multilevel linear regression [21]. Indicators for team membership and role,
Table 2: Turn-taking feature fixed effect estimates for aggregate satisfaction, leadership and cohesiveness for the speaker activity model. Items in bold are significantly different from zero (95% confidence interval).

<table>
<thead>
<tr>
<th>Feature</th>
<th>sat</th>
<th>lead</th>
<th>coh</th>
</tr>
</thead>
<tbody>
<tr>
<td>sil.prop</td>
<td>-0.15</td>
<td>-0.07</td>
<td>-0.05</td>
</tr>
<tr>
<td>bargein.rate</td>
<td>-0.02</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>P_eq</td>
<td>0.00</td>
<td>0.08</td>
<td>-0.03</td>
</tr>
<tr>
<td>F_cond</td>
<td>0.27</td>
<td>0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>ispk.prop.abs</td>
<td>0.06</td>
<td>-0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>ivsu.rate</td>
<td>0.14</td>
<td>0.09</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 3: Coefficient estimates for the DA model.

<table>
<thead>
<tr>
<th>Feature</th>
<th>sat</th>
<th>lead</th>
<th>coh</th>
</tr>
</thead>
<tbody>
<tr>
<td>sil.prop</td>
<td>-0.24</td>
<td>-0.11</td>
<td>-0.10</td>
</tr>
<tr>
<td>bargein.rate</td>
<td>-0.17</td>
<td>-0.05</td>
<td>-0.12</td>
</tr>
<tr>
<td>P_eq</td>
<td>0.05</td>
<td>0.10</td>
<td>-0.00</td>
</tr>
<tr>
<td>F_cond</td>
<td>0.24</td>
<td>0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>Assess</td>
<td>0.23</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Backchannel</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>BeNegative</td>
<td>0.01</td>
<td>0.05</td>
<td>-0.00</td>
</tr>
<tr>
<td>BePositive</td>
<td>0.07</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>CmtUnderstanding</td>
<td>0.07</td>
<td>-0.02</td>
<td><strong>0.09</strong></td>
</tr>
<tr>
<td>ElicitAssessment</td>
<td>0.02</td>
<td>0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>ElicitCmtUnd</td>
<td>0.07</td>
<td>0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>ElicitInform</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.09</td>
</tr>
<tr>
<td>ElicitOffOrSug</td>
<td>0.05</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Inform</td>
<td>-0.03</td>
<td>0.04</td>
<td>-0.06</td>
</tr>
<tr>
<td>Offer</td>
<td>0.07</td>
<td>-0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Fragment</td>
<td><strong>0.15</strong></td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>Other</td>
<td>-0.03</td>
<td>-0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Stall</td>
<td>-0.00</td>
<td>-0.06</td>
<td>-0.07</td>
</tr>
<tr>
<td>Suggest</td>
<td>-0.18</td>
<td>-0.16</td>
<td>-0.06</td>
</tr>
</tbody>
</table>

4. Results

4.1. Average Ratings

Estimates for the speaker activity model for aggregate satisfaction, leadership and cohesiveness are shown in Table 2. We consider effects to be significant where the 95% confidence interval of the estimate excludes zero. The results show a positive effect for $F_{cond}$ in each of the rating categories, while silence proportion has a negative effect on satisfaction. There are no clear effects for participation equality or individual speaking time. That is, participants have a more positive attitude towards these meetings when floor-taking is less predictable, but not necessarily when everyone talks an equal amount. Similarly, people are more positive when there is less silence, but the fact that an individual talks more in general does not indicate that they will be more satisfied. Moreover, we don’t see any clear effects for participation roles for any of the group dynamics. The positive effects for individual VSU observed suggests that short feedback utterances are important for modelling meeting affect and that more is better.

Incorporating DA rates, we again see that participants are more positive in meetings with higher $F_{cond}$, while meetings with less barge-ins and silence proportion also have greater satisfaction and cohesiveness (Table 3). Adding this turn-type information, it appears that the type of short utterances we are interested in are substantive and non-overlapping: participants had higher satisfaction and cohesiveness ratings when they made more assessments, not when they used more backchannels, where assessments include most short utterances that do more than just invite continuation, e.g. affirmative cue words. People who made more inform acts didn’t appear to be more positive, so again the substance of an utterance is important. Suggestions appear to have a negative relationship with satisfaction and leadership, implying that people didn’t like it when they had to suggest courses of action to others. Interestingly, we do not see any clear effects for the socially oriented DAs (be positive/negative), although this may be due to the low frequency of this category since they are defined via exclusion of other DA types.

4.2. Individual Items

Results for the individual items paint a similar picture to that above for satisfaction and cohesion, although $F_{cond}$ features less prominently for cohesion than satisfaction. Looking at the leadership and information processing questions, however, gives us a better indication of where sources of overall satisfaction come from (Table 4). On the one hand, silence proportion, elicit inform moves and suggestions had a negative impact on perception of how well the meeting was directed (Q3). On the other hand, turn-taking freedom had a positive effect for Q4 (Decisions were made in a democratic way) and Q12 (Every team member had sufficient opportunity to make his or her contribution). Similarly, participants in teams with higher turn-taking freedom also felt that information usage was better (Q5), while those who produced more elicit inform moves were more negative. In this vein, we see more information elicitations in groups that felt that there was too much information overall (Q15), while fewer backchannels and more inform moves were associated with higher ratings of irritation (Q13).

Overall, it appears that having a less predictable turn-taking structure leads to feelings of sufficient participation, even if speaking time isn’t actually equal. Providing assessments similarly increased the feeling of contributing (Q10) although, interestingly, inform acts did not. Even so, feelings of sufficient participation may not result in the impression of good leadership as much as having less silence and not having to explicitly ask for information or make suggestions.

5. Discussion

Other studies have used turn-taking features to predict dominance with the motivation that this should help understand leadership behaviour [22]. In this study, dominance features (i.e. speaking time/inequality) do not appear to bear on how well directed meetings were perceived to be. However, looking at more specific question processes, we saw greater turn-taking freedom in meetings with a higher ratings of how democratic the decision making process was and how well information was flowing. This suggests that it’s not so important how much people talk so much as the fact that individual were able to take the floor and that turn-taking is generally less regimented. We might also
have expected higher features of group involvement to lead to more positive affect in this regard. However, we did not find that higher participation equality led to higher satisfaction or cohesiveness, nor did we find effects for overall individual speaking time. Some aspects of turn-taking usually associated with face-to-face conversational dialogue were associated with positive affect, i.e. very short utterances, while others were not, i.e. overlapping turns.

These results highlight the importance of identifying different types of turns when modelling the social aspect of meetings. While we can approximate feedback rates with VSUs, what we are really interested in are turns that form actual dialogue contributions (e.g. assessments) rather than continuation signals (e.g. backchannels). Many past studies have excluded backchannels and overlapped utterances altogether seemingly due to difficulties in transcript alignment (see references in [17]). However, recent studies indicate that the distinction between backchannels and contentful cue words is not so easy to delineate manually or automatically [23, 24]. As such, it appears that a greater focus on modelling feedback utterances is necessary for dialogue modelling, especially with respect to floor control [25, 26]. Given the cost and low agreement rates associated with manual annotation of social signals [27], development of automatic measures is crucial for carrying out large scale studies in these areas and for checking whether insights from qualitative studies generalize to other types of dialogue.

The current study makes some steps into understanding the relationship between turn-taking structure and meeting affect. However, we should note that the AMI meetings were also designed so that everyone had to participate to complete the task. In other conversational modes, people might be quite happy if just one person talks a lot. So, to generalize the results we would need to look at dialogue with more participants and where participants can easily opt out of speaking. Moreover, this sort of study is somewhat encumbered by the low stakes nature of the task. In more high stakes cases different leadership styles may result in equally good outcomes, while leaving the groups with different levels of satisfaction. Similarly, differences in the task structure may affect the reception of different acts. For example, in instructor/follower style dialogues, the need for frequent feedback from a follower may actually be a sign of task difficulty or problems with alignment between participants. This may explain the negative correlations with backchannels/affirmatives and (external) likeability reported in [10]. In contrast, the AMI scenarios involve more open ended discussion and opinion sharing, so short feedback is more likely to be an expression of involvement in the decision making process. In this vein, it seems unlikely that the negative impact of actively directing task structure or information flow in our study would be the case for instructor/follower dialogues. Of course, this doesn’t mean that those sorts of acts should be avoided in AMI style meetings - they may be necessary to achieve goals. However, we need further evaluation of the quality of meeting task outcomes to be able to investigate this further.

### 6. Conclusion and Future Work

This study examined the relationship between turn-taking measures and participants affectual evaluations of meetings. Overall, the most broadly applicable predictor was found to turn-taking freedom. In general, participants were more satisfied and cohesive when group turn-taking freedom and individual VSU rates were higher. In fact, turn-taking freedom turned out to be a better indicator than actual speaking time equality of whether individuals felt they had a chance to contribute to the discussion. From a facilitation point of view, it appears that encouraging floor grabbing to be more unpredictable is beneficial for participant satisfaction. However, this needs to be conditionalized on the meeting task.

In general, feedback turns like assessments were more predictive of meeting affect than information giving acts. Given the positive effects for VSUs and assessments, it seems likely that video-conferencing systems that inhibit the production of short feedback, may reduce the satisfaction of participants. Similarly, conversational dialogue systems should include strategies that invite assessments from other participants. Even though we did not find effects for equality of participation, it has been linked to invite assessments from other participants. Even though we did not find effects for equality of participation, it has been linked to floor grabbing to be more unpredictable is beneficial for participant satisfaction. However, this needs to be conditionalized on the meeting task.

### 7. Acknowledgements

This work was supported by the European Union under the FP7 project inEvent (grant agreement 287872).
8. References


