Predicting Group Satisfaction in Meeting Discussions

Predicting Group Satisfaction in Meeting Discussions

Catherine Lai
University of Edinburgh
Edinburgh, UK
clai@inf.ed.ac.uk

Gabriel Murray
University of the Fraser Valley
Abbotsford, Canada
gabriel.murray@ufv.ca

ABSTRACT

We address the task of automatically predicting group satisfaction in meetings using acoustic, lexical, and turn-taking features. Participant satisfaction is measured using post-meeting ratings from the AMI corpus. We focus on predicting three aspects of satisfaction: overall satisfaction, participant attention satisfaction, and information overload. All predictions are made at the aggregated group level. In general, we find that combining features across modalities improves prediction performance. However, feature ablation significantly improves performance. Our experiments also show how data-driven methods can be used to explore how different facets of group satisfaction are expressed through different modalities. For example, inclusion of prosodic features improves prediction of attention satisfaction but hinders prediction of overall satisfaction, but the opposite for lexical features. Moreover, feelings of sufficient attention were better reflected by acoustic features than by speaking time, while information overload was better reflected by specific lexical cues and turn-taking patterns. Overall, this study indicates that group affect can be revealed as much by how participants speak, as by what they say.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; *Machine learning approaches*; • **Human-centered computing**;

KEYWORDS

Group satisfaction, sentiment, multimodal interaction, speech and language processing, social signal processing, affective computing

ACM Reference Format:
Catherine Lai and Gabriel Murray. 2018. Predicting Group Satisfaction in Meeting Discussions. In *Workshop on Modeling Cognitive Processes from Multimodal Data (MCPMD'18), October 16, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Modeling group affect is an important part of understanding multi-party interaction. In particular, estimating group satisfaction is important for developing strategies for computer-aided decision making and robot interactions with groups, as well as helping understand the cognitive states of individual participants. However, there are multiple ways that a group may be satisfied with an interaction which may, in turn, be reflected by different aspects of a spoken dialogue. In this paper, we investigate how different spoken language features can be used to detect varying aspects of group satisfaction in multi-party meetings.

Previous computational work on dialogue satisfaction has generally focused on predicting dyadic call-center conversations. Such studies have often highlighted the potential of speaker activity patterns, particularly turn-taking behavior, for predicting user satisfaction for both spoken dialogue systems and human-human conversations [12, 30, 36]. In fact, Chowdhury et al. [11] find that turn-taking features perform better than acoustic and lexical features for human call center satisfaction prediction. While little work has been done on predicting satisfaction in dialogues with more than two participants, previous analyses also suggests that turn-taking patterns are indicative of how well a multi-party meeting is going [10, 18]. In particular, the analysis of meeting ratings in Lai et. al [20] found that participants have a more positive attitude when there is less silence, fewer barge-ins, more very short utterances, and more unpredictable turn-taking.

In terms of understanding group interaction, previous work has often focused on predicting group task performance [14, 19, 21], or detecting emergent leadership and leadership styles [4, 17, 29]. Most such work has focused on developing multi-modal models of non-verbal interaction. For example, Avci and Aran [2] identify an HMM-based turn-taking influence measure and group looking features as predictive cues of group performance. Similarly, Dong and Pentland [13] find that more active and balanced group discussion improved performance in a social dilemma task. Beyan et al. [3] show that acoustic features can be used to detect autocratic and democratic leadership styles using multiple kernel learning, although speaker activity features were found to be better discriminators than prosodic features.

Other related work has used acoustic and lexical features to automatically detect sentiment and subjectivity in meetings. For example, Raaijmakers et al. [27] use multi-modal features to detect subjectivity expressed during AMI corpus meetings. However, the subjectivity or sentiment expressed during a meeting may differ markedly from a participants private views. In that sense, our task of satisfaction prediction based on post-meeting ratings is more similar to that of Murray [23], who predicts the sentiment levels found in private meeting summaries authored by each participant.

In the following, we examine the utility of acoustic, lexical, and turn-taking features for predicting group satisfaction ratings from the AMI meeting corpus, where our outcomes of interest are taken from individual participant questionnaires. Our general approach is to use machine learning methods to understand the factors involved

MCPMD'18 , October 16, 2018, Boulder, CO, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

in successful group interactions. We address three tasks: predicting overall satisfaction with the meeting, and the more specific tasks of predicting whether participants felt that each member received sufficient attention during the meeting, and whether participants felt they were overloaded with information. We explore different regression models and perform feature ablation studies to identify what aspects of spoken interaction are likely to reflect different aspects of meeting satisfaction.

2 CORPUS AND QUESTIONNAIRES

2.1 The AMI Corpus

We examine 120 scenario-based meetings from the AMI meeting corpus [9].[1] This consists of 30 groups of 4 participants engaged in a series of 4 meetings. The meeting briefs were about designing a remote control and each participant was given a specific role (project manager, user interface designer, marketing expert and industrial designer) associated with specific information and materials. While the groups were engaged in an artificial role-playing scenario, the speech was spontaneous (unscripted), and each group had freedom in their design solutions and decision-making processes. Audio recordings of each meeting were manually transcribed, time aligned at the word level, and segmented into Dialogue Acts (DAs).

2.2 AMI Meeting Satisfaction Ratings

We use ratings from the post-meeting questionnaire described in [20, 26]. These questionnaires ask participants to rate various aspects of the meetings related to leadership, process satisfaction, cohesiveness, and information processing. After each meeting, individual participants rated their agreement with 16 statements about the meeting, on a 1 ('not at all') to 7 ('very') scale.

We expect that different aspects of spoken language will bear upon different aspects of meeting satisfaction. Thus, to explore these differences, we focus on the following three questions:

- Q7: *Overall Satisfaction:*
  'All in all, I am very satisfied.'
- Q16: *Attention Satisfaction:*
  'All team members received sufficient attention.'
- Q15: *Information Overload:*
  'There was too much information.'

Q15 and Q16 ask participants to rate quite different aspects of the meeting. In fact, based on previous work, we expect feelings about the levels of attention paid to participants (Q16) to be reflected in the distribution of participant speaking time and turn-taking structure [20]. However, satisfaction related to cognitive load (Q15) might be better reflected by vocal characteristics [5, 37]. We also wanted to get an idea of how analyses of specific aspects of satisfaction may differ from one based on a more general satisfaction rating (Q7).

We sum individual ratings per group to obtain group satisfaction measures. We chose to focus on group measures as a starting point for characterizing the group as a whole, as well as the attitudes of specific individuals. We leave exploration of individual satisfaction and other questionnaire items for future work.

[1] http://corpus.amiproject.org/

3 PREDICTION FEATURES

3.1 Turn-taking Features

We consider a number of turn-taking features which were calculated using spurts (contiguous speech segments separated by at least 500ms silence [31]) and dialogue act segments. The segment times were induced from manual transcription word timings. Immediately preceding segments were identified as having the maximum start time before the segment in question (similarly for following segments), thus allowing for overlaps.

Dominance. We calculate Turn-Taking Freedom (i.e., predictability of turn-taking) and Participation Equality as described in [20]. We also record the proportion of active meeting time of the participants who spoke the most and the least, the proportion of dialogue acts uttered by those speakers, and the proportion of dialogue act transitions that involve speaker changes.

Overlap. We measure the following meeting averages over spurts and DA segments: segment duration, minimum time between segment transitions, times from segment start (and end) to the start of any barge-in, overlap duration, and uninterrupted speech duration. We also record total overlap and uninterrupted speech durations. Additionally, we measure the rate of speaker changes between segments, and separate barge-in rates for Very Short Utterances (VSUs) [15] with durations less than 0.5s and 1s, and for all segments.

Pause. We include the total pause duration in seconds and as a proportion of total meeting time, as well as the mean and standard deviation over pause durations, and the maximum pause duration calculated over spurts and DAs.

Activity. We note the total number of DAs and spurts, the total number of laughs and words, laugh rate, and total meeting duration.

Individual Turn-Taking. To investigate the potential for role based effects, we record participant specific turn-taking information: the number of laughs, non-words, and words; total speaking time, number of DAs, number of VSUs that barged into a speaker segment and the times a speaker barged onto another, speaker change rate, number of overlapped segments, total overlapped and uninterrupted speaking time, and mean time from/to the previous/next segment.

3.2 Acoustic Features

We extract acoustic features corresponding to the Interspeech 2010 Paralinguistic Challenge feature set, using openSMILE [16]. This feature set includes a number of standard spectral representations of speech which are generally used to capture segmental aspects of the signal but have also been used for emotion recognition: 15 Mel-Frequency Cepstral Co-efficients (**MFCC**); 8 Line Spectral Pair frequencies (**LSP**); Log power of Mel-Frequency Bands 0-7 (**LMFB**), and associated rate of change (delta) measurements. The feature set also includes several prosodic (i.e. suprasegmental) features: speech wave amplitude based **loudness**; Fundamental frequency ($\mathbf{F_0}$), i.e. pitch, in terms of smoothed F0 envelope, F0 contour, and voicing probability; and voice quality (**VQ**) in terms of pitch-period jitter, differential jitter, and shimmer, which indicate, for example, the tenseness/laxness of the vocal tract.

Moving average smoothing is applied to frame level features before calculating aggregate statistics. In the following experiments, we only look at meeting level standard deviation features to get an

idea of how variability in these features relates to meeting satisfaction and to abstract away from individual speaker patterns.

3.3 Lexical Features

We extract a number of transcript-based lexical features.

Psycholinguistic. Words are scored for their concreteness, imageability, typical age of acquisition, and familiarity.[2] We also derive SUBTL scores for words, which indicate how frequently they are used in everyday life as based on a large corpus of television and movie subtitles [8].

Dependency Parse Features. All sentences are parsed using spaCy's dependency parser.[3] We extract the branching factor of the root of the dependency tree, the maximum branching factor of any node in the dependency tree, sparse bag-of-relations features, and the type-token ratio for dependency relations.

Sentiment. We use the SO-Cal sentiment lexicon [33], which associates positive and negative scores with sentiment-bearing words, and sum these scores over the meeting.

GloVe Word Vectors. Words are represented using GloVe word embeddings,[4] with vectors summed over sentences. We then average the sentence vectors over the meeting. The first five dimensions of the document vectors are used as features, in order to keep the feature dimensionality low given our relatively small number of observations.

Lexical Cohesion. We measure cohesion using the average cosine similarity of adjacent GloVe sentence vectors in a document.

Sentence Rates. We include the average number of words per sentence, and average number of sentences per meeting.

Part-of-Speech Tags. We use a sparse bag-of-tags representation from the spaCy POS tagger for the most frequent tags, as well as the type-token ratio for tags.

Bag-of-Words. Finally, we use a bag-of-words representation for the most common 200 non-stopwords in the dataset, and also calculate the type-token ratio for words. We also record the number of **filled pauses**.

4 EXPERIMENTAL SETUP

In this section we describe the machine learning models used, and evaluation methods.

4.1 Regression Models

In these experiments, we examine the performance of three regression methods, which handle regularization in different ways, to see if they produce consistent results for different types of features. All models were trained using the Scikit-Learn Python package [25]. We use default Scikit-Learn training parameters except where noted below.

Bayesian Ridge Regression (BRR) [22] is a linear regression approach which penalizes large model weights by associating them with a spherical Gaussian prior. This provides some robustness against feature collinearity and over-fitting. The variances of the weight prior and model noise parameters are estimated from the training data jointly with the model weights, assuming Gamma prior distributions.

Random Forest Regression (RFR) [7] is an ensemble method in which predictions are averages over a number of regression tree estimators where each estimator is built from a bootstrap sample of the training data. In our experiments, the estimators are limited to 5 features as another means to limit over-fitting. The number of regression tree estimators was tuned using 10 fold cross-validation on the training data/fold using values between 50-500 estimators.

Support Vector Regression (SVR) [32] attempts to fit a function that deviates from the target by at most ϵ while minimizing the norm of the estimated weights ('flattening' the model). A penalty parameter C mediates regularization by weighting the cost of deviations larger than ϵ. We use an RBF kernel and tune C using 10 fold cross-validation on the training data/fold for values between 10^{-3} and 10^{2}.

Table 1: Results for Q7: Overall Satisfaction (MSE) for Random Forest Regression (RFR), Support Vector Regression (SVR), and Bayesian Ridge Regression (BRR) models: Mean baseline 7.09

Feature set	RFR	SVR	BRR
turn	6.51	6.54	6.36
acoustic	6.39	7.23	6.70
lex	6.67	7.05	6.76
acoustic+lex	6.19	6.95	6.49
acoustic+turn	**6.16**	6.64	6.34
lex+turn	6.40	**6.53**	***6.10***
acoustic+lex+turn	6.23	6.78	6.23

4.2 Evaluation

We evaluate the accuracy of our models using Mean Squared Error (MSE). Given the small sample size, we employ leave-one-out cross-validation to obtain test predictions. However, this makes it difficult to assess training data related variability. Thus, we subsequently use repeated 10-fold cross-validation to investigate how much the results vary due to the training data selection (Section 5.5). We scale and center all input features based on the interquartile range and median for each feature in the training set. We compare performance of the three regression methods with respect to models built using just turn-taking, acoustic, and lexical features, and combinations of those modalities.

5 RESULTS

In this section, we present the results for acoustic, lexical, turn-taking, and combined models for our three questions using leave-one-out cross-validation. We also present feature ablation experiments (Section 5.4), estimate the variability of the results (Section 5.5), and look at individual feature effects (Section 5.6).

5.1 Question 7: Overall satisfaction

Table 1 shows results for predicting overall satisfaction (Q7). We see that the best unimodal model varies by regression method. However, combined acoustic and lexical models generally perform better than purely lexical models. When we add turn-taking features we obtain

[2] http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm
[3] https://spacy.io/
[4] https://nlp.stanford.edu/projects/glove/

Table 2: Results for Q16: Attention Satisfaction (MSE), Mean Baseline: 9.06

Feature set	RFR	SVR	BRR
turn	8.72	9.03	8.84
acoustic	7.20	7.05	6.85
lex	7.68	7.95	7.74
acoustic+lex	6.94	7.38	6.90
acoustic+turn	**6.93**	**7.01**	***6.65***
lex+turn	7.88	7.82	7.43
acoustic+lex+turn	7.13	7.40	6.76

Table 3: Results for Q15: Information Overload (MSE), Mean baseline 14.87

Feature Set	RFR	SVR	BRR
turn	12.83	12.63	12.91
acoustic	13.75	12.72	14.39
lex	12.83	13.49	12.88
acoustic+lex	12.58	12.18	11.92
acoustic+turn	12.53	11.86	11.82
lex+turn	12.42	12.14	12.31
acoustic+lex+turn	**12.13**	**11.66**	**11.65**

best results from lex+turn for BRR and SVR models, with the BRR model producing the best results overall. Nevertheless, the RFR model appears to be better able to make use of acoustic features, particularly in conjunction with turn-taking features. This suggests that acoustic features are useful for modeling overall satisfaction, though a different modeling approach may be required to make use of acoustic features in conjunction with lexical and turn-taking features.

It is also interesting to note that turn-taking models mostly perform better than unimodal lexical and acoustic models. This supports the idea that turn-taking patterns are an important predictor of group affect. However, other conversational modalities are clearly required to understand meeting satisfaction.

5.2 Question 16: Attention Satisfaction

Acoustic features appear more useful for estimating group satisfaction with the amount of attention everyone received (Q16). In Table 2 we see that acoustic features generally perform better than lexical features for all three regression methods. Moreover, using a combination of acoustic and lexical features performs better than either feature type alone. Models using just turn-taking features generally perform poorly. However, again, adding turn-taking features to other feature sets generally helps performance. The best performance is obtained via the combination of acoustic and turn-taking features. This suggests that perceived attention is related to the manner of speaking as well as the amount of talk-time.

5.3 Question 15: Information Overload

Table 3 shows the results for predicting Information Overload (Q15). As for overall satisfaction, the best unimodal model varies for the different regression types, although we see that turn-taking and lexical models generally perform better than acoustic models. However, the best results for each regression type is the combined acoustic, lexical and turn-taking model. This suggests the different modalities provide complementary information in this task

Although all models perform better than the mean value baseline, the MSEs are significantly higher than what was observed for the overall and attention satisfaction. So, it seems that more sophisticated approaches are necessary to explain the variance here, particularly with respect to modeling lexical content. The following sections further explore which specific aspects of speech are important for this task.

Table 4: Acoustic Feature Ablation (BRR). Negative values indicate worse performance when the feature type is removed (i.e. full model MSE < ablated model MSE).

Ablation_Features	Q7	Q16	Q15
MFCC	**-0.16**	**-0.32**	**-0.18**
LMFB	**-0.02**	**-0.01**	0.07
LSP	0.03	**-0.03**	**-0.13**
F_0	0.07	**-0.02**	0.28
voice quality	0.13	**-0.03**	**-0.67**
loudness	0.01	0.02	0.01

5.4 Feature Ablation

Beyond understanding the relative utility of acoustic, lexical and turn-taking features for predicting group satisfaction, we would also like to explore the predictiveness of specific feature types. This is particularly important in the current task given the small sample size relative to the total number of extracted features. In the following, we investigate the usefulness of specific feature types via ablation. We remove feature subsets from the combined acoustic, lexical, and turn-taking feature set and report the difference between the original and modified model MSEs. For brevity, we only report BRR results as it generally provided the best overall results for our three questions in the previous experiments.

5.4.1 Acoustic Features. Table 4 shows the difference in performance when acoustic feature subsets are removed. Interestingly, we can see that including speech prosody features (F_0, voice quality, loudness) produce worse performance for predicting overall satisfaction (Q7), though their inclusion does help predict attention satisfaction (Q16). This, again, suggests that quality of participation is reflected in *how* speakers speak. However, when it comes to overall satisfaction, the relevant speech aspects are not captured by our prosodic measures. Nevertheless, voice quality features appear to be important for predicting information overload (Q15). This is consistent with previous work arguing that cognitive load is reflected in, for example, variation in the tenseness in the vocal tract [37]

5.4.2 Turn-taking features. The ablation results in Table 5 suggest that overlap and pause features are more useful for predicting overall satisfaction than attention satisfaction. Conversely, activity measures are predictive of attention satisfaction. However, removing these features does not have as great an impact on the results as removing MFCC features. This supports the importance of acoustic features for understanding attention satisfaction.

Table 5: Turn-taking Feature Ablation (BRR)

Ablation_Features	Q7	Q16	Q15
individual TT	**-0.11**	**-0.03**	**-0.16**
dominance	**-0.05**	**-0.02**	0.10
pause	**-0.05**	0.01	0.08
overlap	**-0.02**	0.04	0.00
activity	0.01	**-0.02**	0.04

Table 6: Lexical feature ablation (BRR)

Ablation_Features	Q7	Q16	Q15
parse	**-0.07**	**-0.03**	0.13
psycholinguistic	**-0.04**	0.02	0.02
sentiment	**-0.01**	0.00	0.01
coherence	**-0.01**	0.00	**-0.02**
sentence rates	0.00	0.00	0.00
filled pause	0.00	0.00	**-0.00**
glove	0.00	0.00	**-0.01**
part-of-speech	0.00	0.02	0.21
bag-of-words	0.16	0.04	**-0.65**

Table 7: Results from removing feature sets that caused decreased performance in the different modalities (BRR).

Ablation Modality	Q7	Q16	Q15
none	6.23	6.76	11.65
turn ablation	6.23	6.70	11.41
acoustic ablation	5.89	6.73	11.31
lexical ablation	6.05	6.60	11.35
all	***5.57***	***6.52***	***10.60***

The results also show that inclusion of individual turn-taking features generally improved performance, particularly for predicting information overload and overall satisfaction. These features are quite specific to the AMI meeting structure. So, while we wouldn't expect models including these features to generalize directly to other types of meetings, they do indicate that role specific dependencies are important for predicting satisfaction. Thus, work on identifying assigned versus emergent group leaders, for example, is likely to be important for understanding meeting satisfaction.

5.4.3 Lexical Features. Lexical feature ablation results (Table 6) show that the inclusion of more abstract lexical features (parse, psycholinguistic, sentiment, coherence) are beneficial for predicting overall satisfaction. Removing bag-of-words features improves performance, although including aggregated sentiment scores helps somewhat. This indicates that abstraction over affective lexical content is necessary for this aspect of satisfaction.

Only the parse features were helpful for predicting attention satisfaction, which supports the idea that non-lexical features are more important for monitoring attention satisfaction. However, specific lexical content features (GloVe and bag-of-words) were found to be important for predicting information overload, suggesting specific lexical content is important for this question.

5.4.4 Performance of Post-Ablation Models. Table 7 shows MSE results after ablation of features from specific modalities with respect to the full feature set. The experiments show that ablation in single modalities generally results in improved performance. Using only selected features from all modalities gives us our overall best results for all three questions. Thus, for these sorts of machine learning models features from some modalities can obscure the usefulness of other modalities for the task.

5.5 Estimating Performance Variability

The results reported above are obtained using leave-one-out cross-validation. To investigate the variability of our results, we instead use repeated 10-fold cross-validation, randomizing the folds each time. Figure 1 shows MSE for Bayesian Ridge Regression models trained using turn, acoustic, lexical, and combined feature sets (full and ablated) over 100 repetitions. We performed pairwise t-tests over the distributions of results to identify significant differences between models. We use $p < 0.05$ (Bonferroni corrected) as the threshold for statistical significance.

As discussed above, we see that individual modalities have different relationships with each aspect of satisfaction we investigated. However, the usefulness of specific modalities can change with ablation. For example, the turn-taking model is significantly more predictive than the acoustic model for overall satisfaction (Q7). However, the performance of the ablated acoustic model is not significantly different from either the full or ablated turn-taking models. Similarly, the full acoustic model is not significantly different from the lexical model but the ablated version is.

Ablation does not make a significant difference to the turn-taking model for overall satisfaction, although it does for the other modalities, which follows from the fact that the most turn-taking features are kept in the ablated model. Similarly, ablation of the acoustic model does not improve performance for the attention satisfaction, while ablation does not significantly improve the lexical model for information overload prediction. This supports the idea that acoustic features are more indicative for the attention satisfaction, while lexical features are more useful for detecting information overload. However, ablation of the less predictive modalities helps for each of our questions. Overall, we see that using a subset of features from all modalities significantly improves performance, particularly for prediction of overall satisfaction.

The graph of actual versus predicted values for the ablated models in Figure 2 shows these models capture continuous differences between meetings, although there is clearly a lot of variance still unaccounted for. This shows that using features from all modalities helps improve prediction performance, as long as we perform judicious feature selection.

5.6 Individual Feature Effects

We further investigate the predictiveness of individual features by looking at the estimated feature coefficients from our Bayesian Ridge Regression models. We take the mean over estimated coefficient values for each of the models trained in the leave-one-out cross-validation experiments described above. We take effect sizes (coefficient magnitude) to be significant when they are more than three standard errors away from zero. Of the features that passed

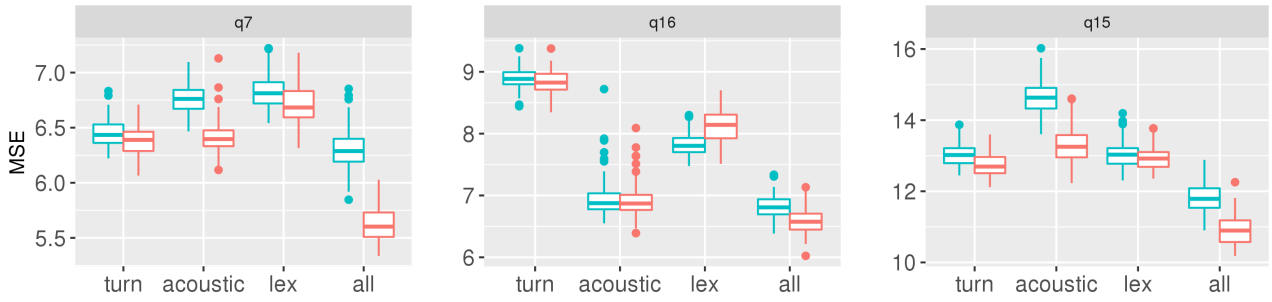

Figure 1: MSE from 10-fold cross-validation repeated 100 times with random folds. Results are shown for full (left, blue) and ablated (right, red) feature sets for different modalities.

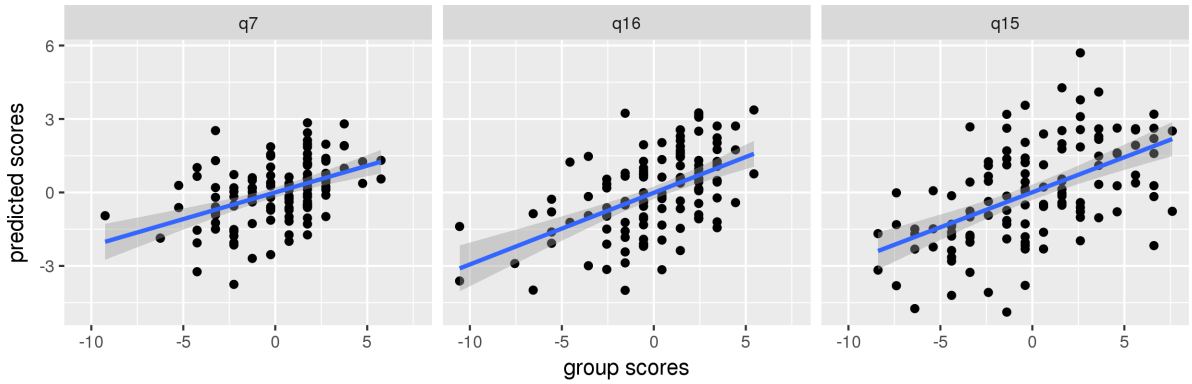

Figure 2: Scatterplots of group scores vs best model predictions with linear model fit. Group scores are centered to the mean value for each question: $q7 = 22.25$**,** $q16 = 21.56$**,** $q15 = 12.39$

ablation, only very few had effect sizes that where not significantly different from zero.[5]

In Table 8, we show the top 10 features with positive and negative effect sizes. For information overload, the top features are dominated by lexical content and features describing the project manager's (PM) turn-taking behaviour. In particular, higher frequency of the word *'sorry'* is associated with a higher information load, as is the attribution verb *'said'*. This indicates that further analysis is warranted into how these sort of discourse markers are related to break downs of communication or task structure and to cognitive load.

Information overload appears to also be positively correlated with overlaps and barge-ins on project manager turns and is decreased when the project manager keeps the floor more. This is in contrast with the overall satisfaction, where large amounts of uninterrupted speech from the project manager appears to decrease satisfaction. In this vein, we see a positive effect for increased turn-taking freedom (i.e. less predictability of who will speak next) for overall satisfaction. Having a long pause also seems to be associated with increased satisfaction, as does use of more imageable and concrete vocabulary.

[5]Features with non-significant effects: Overall satisfaction: δ mfcc(8), total silence duration, SUBTL score; Attention Satisfaction: DOBJ, number of laughs; Information Load: δ mfcc(6), *'hand'*, *'nt'*, *'room'*, *'second'*, *'start'*, *'um'*

The top features for the attention satisfaction model highlight more acoustic and syntactic features. We expect that voicing probability acts as a proxy for the amount of speaking time. In line with this, we see more evenness in the proportion of speaking time associated with each speaker is correlated with positive scores. It also appears that certain complex syntactic structures may be useful cues for this aspect of satisfaction. It is possible that participants use more complicated syntactic structures when they feel they are being attended to. However, a more fine-grained analysis of their use (and similarly for the highlighted acoustic features) is required to understand when and why they appear and how this is related to participant satisfaction.

6 DISCUSSION

In general, this study is consistent with a long line of research that has found that multi-modal approaches for understanding speaker affect are better than unimodal ones [34, 35]. However, the current work differs from much of past work in looking at multi-party rather than dyadic conversation, and in the types of satisfaction ratings collected. Thus, it is hard to directly compare results with previous work. The most directly comparable work is [12] who frame call-center user satisfaction prediction as a classification task. Unlike that study we find that including modalities beyond turn-taking improves performance. Interestingly, that work focused

Table 8: Features with most positive and negative effects for the best Bayesian Ridge Regression models. SMALLCAPS indicates SpaCY based parse features, while *'italics'* indicates specific word features. For acoustic features, δ indicates the first derivative of the named feature. Numbers in parentheses indicate coefficient for spectral feature types. For individual turn-taking features: PM = Project Manager, UI = User Interface Designer

	Q7: Overall Satisfaction		Q16: Attention Satisfaction		Q15: Information Load	
			Top Positive Effects			
1	spurt max. pause duration	0.13	voicing probability	0.12	*'sorry'*	0.36
2	DATIVE	0.12	δ differential jitter	0.11	*'having'*	0.27
3	imageability	0.11	PARATAXIS	0.11	*'kind'*	0.24
4	mfcc(6)	0.10	mfcc(6)	0.11	*'said'*	0.22
5	DA max. pause dur	0.09	CSUBJ	0.10	*'decision'*	0.17
6	δ mfcc(3)	0.09	DOBJ||XCOMP	0.10	*'means'*	0.17
7	concreteness	0.08	δ voicing probability	0.10	PM overlap	0.16
8	PUNCT	0.08	min speaker DA proportion	0.09	differential jitter	0.16
9	turn-taking Freedom	0.08	δ mfcc(14)	0.09	*'maybe'*	0.15
10	δ mfcc(12)	0.08	lmfb(7)	0.09	PM barged into	0.15
			Top Negative Effects			
1	ADVMOD||XCOMP	-0.17	ADVMOD||XCOMP	-0.20	δ shimmer	-0.28
2	PM no. words	-0.13	δ lsp(0)	-0.12	*'let's'*	-0.20
3	XCOMP	-0.10	INTJ	-0.12	mfcc(1)	-0.18
4	familiarity	-0.10	UI barged onto	-0.10	δ jitter	-0.18
5	DOBJ||XCOMP	-0.09	shimmer	-0.09	PM mean time from prev.	-0.17
6	ATTR	-0.09	NPADVMOD	-0.08	PM uninterrupted duration	-0.17
7	PM uninterrupted duration	-0.09	PM no. words	-0.08	PM stay rate	-0.16
8	PM sum duration	-0.09	lsp(6)	-0.08	*'think'*	-0.16
9	NEG	-0.08	CONJ	-0.07	*'nice'*	-0.16
10	cohesion	-0.08	max. speaker proportion	-0.07	*'possible'*	-0.15

on prosodic features, whereas we found other acoustic features to have more predictive power. The current work also includes more abstract lexical features than in that study. This again suggests more work needs to go into identifying aspects of spoken interaction relevant to multi-party affect.

Unlike most previous work, the current study examines how spoken language features relate to different aspects of satisfaction. The cross-validation results indicate that the current findings are representative of the types of meetings in this particular corpus. However, the generalizability of these findings to other types of multi-party spoken interaction is yet to be seen. A good test case would be the ELEA corpus [29] where participants are not given roles and there are clearer measures of task success. However, the ELEA corpus does not include participant satisfaction ratings. Thus, further data collection appears to be necessary to test the generalizability of our approach. The GAP corpus [6] is an ongoing data collection effort using the same winter survival scenario as the ELEA corpus, and it does contain participant satisfaction ratings. Future work will extend our analysis to the ELEA and GAP corpora.

The relatively small amount of available group data is a challenge for developing data-driven systems in general. Collecting more multi-party spoken dialogue and participant data to fill the gap is a long term project. A promising avenue for data collection may be to look at games that can be played in text or audio modality, e.g. Settlers of Catan [1]. However, recent work by Murray and Oertel [24] has found both domain adaptation and data augmentation strategies substantially improved prediction of group performance with the ELEA corpus. Thus, we expect these approaches could be similarly harnessed for detecting group affect in the near future.

The small number of group ratings led us to focus on shallow regression models with varying regularization components. From these, Bayesian Ridge Regression appeared the most promising. However, the results above indicate that modeling more complex interactions between and within modalities could improve performance for this task. Thus, we would like make use of neural network based feature learning methods to learn appropriate multimodal feature representations given the raw audio and lexical input [34]. However, given the small amount of labeled data we have for this task, it is unlikely this will succeed without employing transfer learning techniques. A potential source could be, for example, using bottleneck-style features from a recurrent neural network based turn-taking prediction model [28].

In the current work, we focused on regression analysis to see how aspects of group satisfaction were reflected by different modalities. However, to increase comparability with studies, we also plan to carry out a similar analyses using classification instead of regression. In particular, we plan to look at the properties and separability of groups whose members are very satisfied or dissatisfied. We also plan to investigate prediction of individual participant satisfaction rather than aggregating at the group level, to provide insight into cases where a single group consists of members with substantially differing levels of satisfaction.

7 CONCLUSION

We investigated using acoustic, lexical and turn-taking features for predicting group meeting satisfaction as a regression task. Models using just acoustic or lexical features perform better than the baseline and the addition of turn-taking features consistently improved performance. An approach using selected features from all modalities produced the best overall results. That is, features from all three modalities can be helpful for inferring the cognitive state of participants after a meeting. However, each aspect of group satisfaction was reflected by different features from each modality. For example, feature ablation experiments indicate that more abstract lexical features were helpful for predicting overall satisfaction, while specific lexical cues were important for predicting information overload. A large range of acoustic features were identified as predictive for attention satisfaction, while voice quality seemed much more important for predicting information overload.

In general, it appears that a greater focus on extracting affective lexical content from spoken interactions appears is warranted for this task, as is further examination of potential interactions between features from different modalities in expressing participant affect. A major constraint for this is the relatively small number of observations. The discussion above pointed out potential avenues for making up for this using data augmentation and domain adaptation. Future work will also look at how well the current approach generalizes to the other AMI ratings, as well as understanding leadership and satisfaction in other scenarios such as the ELEA and GAP corpora.

REFERENCES

[1] Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In *LREC 2016*.

[2] Umut Avci and Oya Aran. 2016. Predicting the performance in decision-making tasks: From individual cues to group interaction. *IEEE Transactions on Multimedia* 18, 4 (2016), 643–658.

[3] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2018. Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features. *IEEE Transactions on Multimedia* 20, 2 (2018), 441–456.

[4] Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. 2016. Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 317–324.

[5] Hynek Bořil, Seyed Omid Sadjadi, Tristan Kleinschmidt, and John HL Hansen. 2010. Analysis and detection of cognitive load and frustration in drivers' speech. In *Proceedings of Interspeech 2010*.

[6] Mckenzie Braley and Gabriel Murray. 2018. The Group Affect and Performance (GAP) Corpus. In *Proceedings of the ICMI Group Interaction Frontiers in Technology (GIFT) Workshop*. ACM.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[8] Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods* 41, 4 (2009), 977–990.

[9] Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation* 41, 2 (2007), 181–190.

[10] J. Carletta, S. Garrod, and H. Fraser-Krauss. 1998. Placement of Authority and Communication Patterns in Workplace Groups The Consequences for Innovation. *Small Group Research* 29, 5 (1998), 531–559.

[11] Shammur Absar Chowdhury, Evgeny Stepanov, Morena Danieli, and Giuseppe Riccardi. 2017. Functions of Silences towards Information Flow in Spoken Conversation. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*. 1–9.

[12] Shammur Absar Chowdhury, Evgeny A Stepanov, and Giuseppe Riccardi. 2016. Predicting User Satisfaction from Turn-Taking in Spoken Conversations. In *Proceedings of Interspeech 2016*. 2910–2914.

[13] Wen Dong, Bruno Lepri, Taemie Kim, Fabio Pianesi, and Alex Sandy Pentland. 2012. Modeling conversational dynamics and performance in a social dilemma task. In *Communications Control and Signal Processing (ISCCSP), 2012 5th International Symposium on*. IEEE, 1–4.

[14] Wen Dong and Alex Sandy Pentland. 2010. Quantifying group problem solving with stochastic analysis. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ACM, 40.

[15] J. Edlund, M. Heldner, S. Al Moubayed, A. Gravano, and J. Hirschberg. 2010. Very short utterances in conversation. In *Proceedings of Fonetik*. 11–16.

[16] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[17] Sebastian Feese, Amir Muaremi, Bert Arnrich, Gerhard Troster, Bertolt Meyer, and Klaus Jonas. 2011. Discriminating individually considerate and authoritarian leaders by speech activity cues. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 1460–1465.

[18] Daniel Gatica-Perez. 2006. Analyzing Group Interactions in Conversations: a Review. *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems* (2006), 41–46.

[19] Dineshbabu Jayagopi, Dairazalia Sanchez-Cortes, Kazuhiro Otsuka, Junji Yamato, and Daniel Gatica-Perez. 2012. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 433–440.

[20] Catherine Lai, Jean Carletta, and Steve Renals. 2013. Modelling Participant Affect in Meetings with Turn-Taking Features. In *Proceedings of WASSS 2013, Grenoble, France*.

[21] Bruno Lepri, Nadia Mana, Alessandro Cappelletti, and Fabio Pianesi. 2009. Automatic prediction of individual performance from thin slices of social behavior. In *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 733–736.

[22] David JC MacKay. 1992. Bayesian interpolation. *Neural computation* 4, 3 (1992), 415–447.

[23] Gabriel Murray. 2016. Uncovering hidden sentiment in meetings. In *Canadian Conference on Artificial Intelligence*. Springer, 64–72.

[24] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM on International conference on multimodal interaction*. ACM.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[26] Wilfried M. Post, Mirjam Huis in 't Veld, and Sylvia van den Boogaard. 2007. Evaluating meeting support tools. *Personal and Ubiquitous Computing* 12, 3 (March 2007), 223–235.

[27] Stephan Raaijmakers, Khiet Truong, and Theresa Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of EMNLP 2008*. Association for Computational Linguistics, 466–474.

[28] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs. In *Proceedings of Interspeech 2018*.

[29] Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia* 14, 3 (2012), 816–832.

[30] Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: assessing the quality of ongoing spoken dialog interaction by expertsâĂŤand how it relates to user satisfaction. *Speech Communication* 74 (2015), 12–36.

[31] Elizabeth Shriberg, Andreas Stolcke, and Don Baron. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Seventh European Conference on Speech Communication and Technology*.

[32] Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.

[33] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37, 2 (2011), 267–307.

[34] Leimin Tian, Johanna Moore, and Catherine Lai. 2016. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In *Spoken Language Technology*. IEEE, 565–572.

[35] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291.

[36] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1998. Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech & Language* 12, 4 (1998), 317–347.

[37] Tet Fei Yap, Julien Epps, Eliathamby Ambikairajah, and Eric HC Choi. 2015. Voice source under cognitive load: Effects and classification. *Speech Communication* 72 (2015), 74–95.