Multimodal Analysis of Group Attitudes Towards Meeting Management

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

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

ABSTRACT

We present experimental results on the task of automatically predicting group members' attitudes about management of their meeting, based on linguistic and acoustic features derived from the meeting recordings and transcripts. The group members' attitudes were gathered from detailed post-meeting questionnaires. A key finding is that features of linguistic content by themselves yield poor prediction performance on this task, but the best results are found by combining acoustic and linguistic features in a multimodal prediction model. When trying to automate the detection of group member attitudes that might be manifested subtly in their language and behaviour, a multimodal analysis is key.

CCS CONCEPTS

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

KEYWORDS

group sentiment, meeting management, multimodal interaction, speech and language processing, social signal processing, leadership

ACM Reference Format:
Gabriel Murray and Catherine Lai. 2018. Multimodal Analysis of Group Attitudes Towards Meeting Management. In *Proceedings of Group Interaction Frontiers in Technology (GIFT'18).* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Meetings are an important and ubiquitous part of working life. Understanding how to successfully manage and direct meetings is a vital step towards improving workplace satisfaction and employee engagement and productivity [1, 9]. Thus, a system for automatic analysis of a group's attitudes towards their management, their own group processes, and towards each other, could be very useful for providing feedback to a group or the group leader. This sort of analysis could also help managers decide how a group is comprised, how team leaders are trained, and how the meetings are structured.

Such automated systems require technology for detecting and modeling participant attitudes towards meeting management. Group interactions very often exhibit expressions of attitudes and opinions from the participants. However, the attitudes that are explicitly expressed in a meeting may differ from a participants actual attitudes towards the meeting, their colleagues, the team leader, or the group's goals. Furthermore, those true underlying attitudes may be complex and multi-faceted. While there has been a good deal of research on automatic detection of explicit sentiment or subjectivity in group discussions, very little work has been done on detecting private sentiment and attitudes that may contrast with or elaborate on the participant's explicit statements to the group. Similarly, work on sentiment analysis often only deals with binary distinctions of positive versus negative sentiment [12, 18], while in reality attitudes also vary in magnitude.

In general, technologies for automated meeting management analysis that take into account participant attitudes require multimodal analyses of group interaction. However, the importance of different modalities for understanding various aspects of group interaction and satisfaction is still unclear. Although sentiment analysis of text documents is a well developed field, multimodal sentiment analysis is a relatively new area of research [10, 24]. In contrast, work on social signal processing (SSP) has focused on applying machine learning to group interaction using only non-verbal features, such as prosody, gestures, and turn-taking [3]. Thus, a clearer understanding of the role of verbal features as cues of participant attitudes is a useful step to developing better multimodal models of meetings in general.

With this in mind, the current work investigates multimodal models for predicting attitudes towards the management and direction of meeting from the AMI corpus [4] based on post-meeting ratings by meeting participants [8, 17]. More specifically, we examine models which make use of *linguistic* content pertaining to lexical, syntactic, and psycholinguistic information derived from the transcripts, as well as *speech* features derived directly from the acoustic signal. We hypothesize that linguistic information is useful when trying to detect complex attitudes that may only subtly manifest themselves in group situations. Moreover, we expect that using linguistically interpretable features will help improve model interpretability generally. This is particularly important when systems are meant to provide feedback to a team or a team leader that could be used to adjust and improve aspects of the group interaction.

The structure of the paper is as follows. In Section 2, we discuss related work on social signal processing and multimodal interaction. Section 3 describes our approach to automatic detection of group attitudes. Section 4 presents the experimental setup, and Section 5 contains the key results. Finally, we conclude in Section 6, along with some thoughts on future work.

2 RELATED WORK

Renals et al. [19] provide an overview of research on multimodal signal processing applied to meetings, including verbal and non-verbal analysis. Much of that work was carried out on the AMI corpus [4], which we also utilize and describe in more detail in Section 3. With respect to that body of work, the analysis of attitudes in meetings is closely related to research on detecting sentiment (or *subjectivity*) in meetings. In practice, this means predicting external annotations

of individual dialogue act segments as positive-subjective, negative-subjective, non-subjective, and so on [12, 18]. However, it is possible that the sentiment a person explicitly conveys *during* a meeting inaccurately or incompletely conveys their actual attitudes about the meeting. For example, it is possible for someone to express positive comments in the meeting but to nonetheless be unhappy about the direction of the meeting, or vice versa. In light of this, we attempt to detect attitudes from within the group that are revealed through post-meeting questionnaires, thus reflecting their private states.

Recent work by Murray [11] has similar goals of detecting meeting participants' true sentiment, based on features of the group interaction. That work estimated sentiment based on analysis of participant summaries that were authored by each participant after each meeting. In our current work, we instead use feedback from post-meeting questionnaires that focus on specific attitudes towards the meeting. These questionnaires have previously been used by Lai et al. [8] to investigate how turn-taking dynamics impact attitudes towards group cohesion, satisfaction, and leadership in these meetings. However, that study focused on individual participant attitudes and did not explore speech cues beyond turn-taking. In our current work, we focus primarily on speech and language features. We also specifically focus on group attitudes about meeting management and direction as a first step to understanding how these these participant ratings relate to other work on leadership in group interaction.

Related work on leadership involves analysis of how leadership emerges during task-based interaction [21, 22]. Those studies utilize the ELEA corpus [20], which consists of recordings of small groups collectively performing a ranking task. The participants are not assigned roles, and so it is possible to analyze how some participants take on a leadership role, and how all members perceive each other in terms of leadership. In contrast, participants in the AMI meeting corpus were assigned roles, with one of them acting as project manager (i.e. leader). However, participants other than the project manager can still play a role in how the meeting direction is managed. Thus, the current study focuses on how spoken language reflects how participants feel about meeting management aggregated at the group level, rather than specifically evaluating the project manager or detecting leadership styles and emergent leaders.

As mentioned above, a good deal of work on social signal processing (SSP) has focused on non-verbal features, such as prosody, gestures, and turn-taking [3]. In fact, SSP has been explicitly defined as focusing on non-verbal or non-linguistic aspects of social interaction [13, 16]. In this work, we investigate the usefulness of linguistic content features in addition to acoustic features from the speech signal. In general, we view non-verbal aspects of interaction as a very rich source of information that provides context and elaboration for understanding language in interaction [5].

3 PREDICTION OF GROUP ATTITUDES ABOUT MEETING MANAGEMENT

The goal of this project is to develop a system to automatically predict group attitudes regarding the management and direction of the meeting. We hypothesize that linguistic features will be useful for this prediction task. In this section, we describe the meeting corpus used, and how the attitudes towards meeting management were measured. We then describe the multimodal features that were extracted for this prediction task.

3.1 CORPUS AND PARTICIPANT QUESTIONNAIRES

For these experiments, we utilize 120 scenario-based meetings from the AMI meeting corpus [4].[1] In the AMI scenario meetings, teams were tasked with designing a remote control and bringing it to market. Each team consisted of four participants, with each participant assigned a unique role: project manager, user interface designer, marketing expert, or industrial designer. Each role was associated with specific information and materials. Each team completed a series of four meetings. While the scenario and assigned roles were pre-determined, the discussion and decisions were not scripted, and the teams and individual members had freedom in how they developed and contributed to the product at each phase. In these experiments, we make use of manual transcripts of each meeting, which have been segmented into dialogue act units (DAs).

Meeting Management Ratings: After every meeting, each participant filled out a post-meeting questionnaire. The design of the post-meeting questionnaire is described in [17]. The full set of questions can be found in [8]. The questions ask participants to rate the meeting experience in terms of process satisfaction, cohesiveness, leadership, and other factors. The individual participants privately rated their agreement with 16 statements about the meeting, on a 1 ('not at all') to 7 ('very') scale. For the purposes of this study, we are concerned with Question 3, which directly addressed meeting management and direction:

- Q3: *The meeting was directed in a good manner.*

To get a snapshot of the overall group attitude about the management of the meeting, we sum the ratings for this criterion over all four participants to get a group rating regarding the management of the meeting. The questionnaire rating is not merely a rating of the project manager (PM), though the PM typically plays a large part in directing the meeting. The AMI scenario dictates that each group member has distinct responsibilities for moving the design and decision-making processes forward through the sequence of meetings. Furthermore, the PM may have strong opinions on this criterion, e.g. being disappointed that other team members did not take on more leadership and responsibility. For that reason, in these experiments we do not omit the PM's score from the aggregated group score.

In the experiments contained herein, the goal is to automatically predict the group score. The group score can range from 4 (all participants give the lowest rating) to 28 (all participants give the highest rating). Figure 1 shows the actual distribution of group scores.

One noticeable aspect of the ratings in Figure 1 is their wide variation, ranging from 14 to 27. There are clearly extreme cases where participants are either very satisfied or dissatisfied with the direction of the meeting. That provides substantial motivation for

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

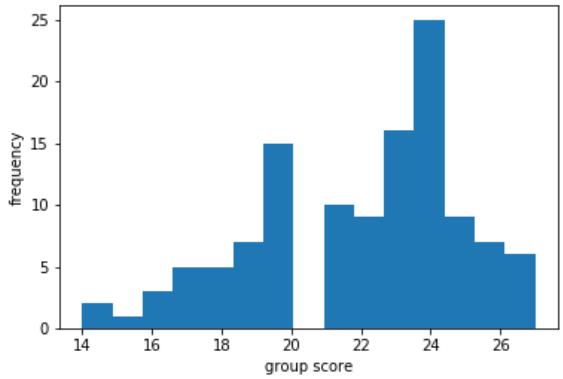

Figure 1: Distribution of Group Scores

our current work on predicting management ratings, given that there are real differences in these ratings across meetings.

3.2 SPEECH FEATURES

We extract acoustic features corresponding to the Interspeech 2010 Paralinguistic Challenge feature set [23], using openSMILE [6]. This feature set includes a number of standard spectral representations of the speech signal: 15 Mel-frequency Cepstral Co-efficients (**MFCC**) and associated delta features; 8 Line spectral pair frequencies (**LSP**); log power of Mel-frequency bands 0-7 (**LMFB**). The feature set also includes features representing speech prosody. These include speaker PCM loudness (**LOUDNESS**); pitch in terms smoothed F0 envelope, F0 contour, and voicing probability (**PITCH**); and voice quality via pitch-period jitter, differential jitter, and shimmer (**VQ**). 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, yielding 76 speech features in total. The features are extracted from the entirety of each meeting. That is, we focus on acoustic variability within the group as a predictor of meeting management satisfaction.

3.3 LINGUISTIC FEATURES

We extract a number of transcript-based lexical, syntactic, and psycholinguistic features, and they are again extracted from the entirety of each meeting.

Psycholinguistic: Words are scored for their concreteness, imageability, typical age of acquisition, and familiarity [26].[2] We also derive SUBTL scores for words, which indicate how frequently they are used in everyday life [2].

Dependency Parse Features All sentences are parsed using spaCy's dependency parser [7].[3] We extract several features, including 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 [25], which associates positive and negative scores with sentiment-bearing words, indicating how positive or negative their sentiment typically is.

GloVe Word Vectors Words are represented using GloVe vectors [15],[4] and the vectors are summed over sentences. We then create a document vector that is the average of the sentence vectors. The first five dimensions of the document vectors are used as features.

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

Sentence and Document Length We include the average number of words per sentence, and average number of sentences per meeting (i.e. document).

Part-of-Speech Tags We use spaCy's part-of-speech tagger, and use a sparse bag-of-tags representation for the most frequent tags, as well as the type-token ratio for tags.

Bag-of-Words 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.

Filled Pauses Finally, we also consider the number of filled pauses such as *uh* and *um* in the discussion.

4 EXPERIMENTAL SETUP

In this section, we briefly describe the predictive models used, and the evaluation metrics.

4.1 MACHINE LEARNING MODELS

We report results for three machine learning models, including two tree-based approaches, Random Forests (RF) and Gradient Boosted Trees (GB). For both RF and GB, the number of estimators was set to 50 and the maximum number of features was set to 20. The third model we compared with was k-nearest neighbours (kNN), with $k = 20$. We use the Scikit Learn implementations for each model [14].

4.2 EVALUATION

For evaluation of all of the systems, we use Mean Squared Error (MSE) scores. To maximize the amount of training data, we used a leave-one-out procedure. After removing meetings that had missing data for these meeting management ratings, we had a total of 120 meetings, giving us 119 in each training fold.

We also present feature-level analysis using feature importance scores for the tree-based models, where each importance score is determined by the average reduction of MSE when the feature is used as a split point in the decision trees.

5 RESULTS

Table 1 shows the MSE results for the machine learning models, as well as for a baseline approach in which the mean score from the training fold is predicted. Gradient Boosted Trees performed best overall, and all three machine learning models exhibited performance better than the baseline. The best GB model is significantly

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

[4] https://nlp.stanford.edu/projects/glove/

better than the baseline, according to a paired t-test on the squared residuals for each set of predictions ($p < 0.05$).

Model	MSE
Baseline (Mean Prediction)	9.02
kNN (k=20)	8.78
Random Forests	7.47
Gradient Boosted Trees	6.96

Table 1: MSE: All Features

We also report results from experiments using feature subsets, including just linguistic features, just speech features, and all of the features. Gradient Boosted Trees were used for these further experiments. Table 2 summarizes the results for these subsets. Interestingly, linguistic features by themselves are not effective for this prediction task, and that particular model performs worse than the baseline. However, the model using the combined linguistic and speech feature set performs best overall. This demonstrates that it is worthwhile to extract linguistic features to accompany other multimodal features when trying to automatically detect participant attitudes in group interaction.

Feature Subset	MSE
Linguistic Only	9.57
Speech Only	7.51
All Features	6.96

Table 2: MSE: Feature Subsets, GB Models

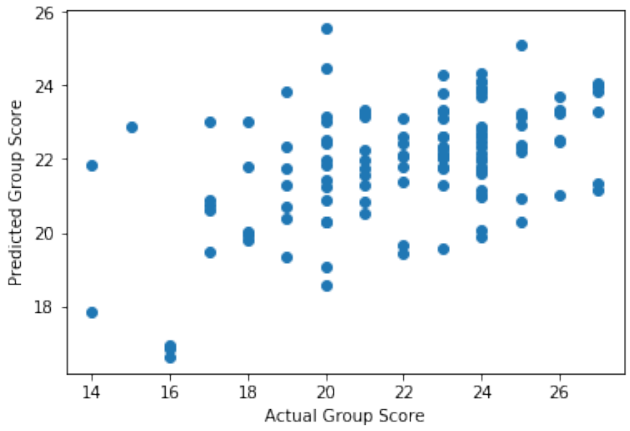

Figure 2: Actual vs. Predicted Group Scores

Figure 2 shows the actual scores vs. predicted scores for each meeting. While there is a strong positive correlation between the actual and predicted scores, we can see a few cases where a meeting score was much lower than predicted.

The results above show that adding linguistic features can improve speech based models of meeting satisfaction. However, given the general nature of the lexical feature set, it is likely that some linguistic features have more relevance to meeting management than others. With this in mind, we examine the usefulness of individual linguistic features. In the following analysis, we focus on linguistic features other than the sparse bag-of features. Figure 3 shows the top 10 features in terms of feature importance for the Gradient Boosted Trees models. Interestingly, sentiment features did not appear in the top 10 features, while GloVe features (the *vdim* features in the graph) appear to be very useful. This suggests some abstraction over lexical information, beyond what is captured in our linguistic features, is required for this task. It also suggests that further work on learning lexical feature representations directly from the transcript is a promising future approach for learning representations of group interaction in general. We also see that type-token ratio is relatively useful, as is the dependency parse feature based on the maximum branching factor of any node (*maxchild_sc*). This means that complexity of lexical content may be a useful measure for characterizing a meeting in terms of its management.

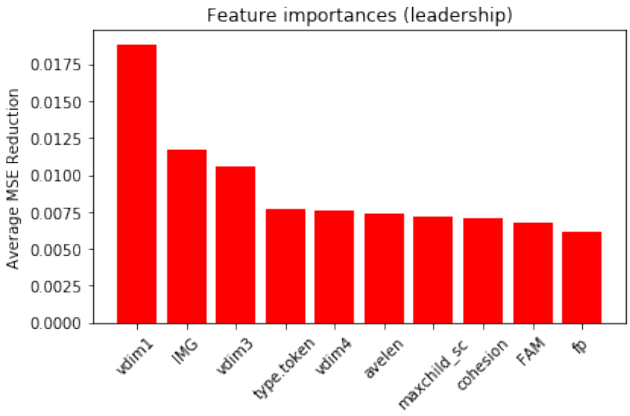

Figure 3: Linguistic Feature Importance (GB)

The psycholinguistic features, particularly concreteness (CNC), familiarity (FAM), imageability (IMG), and age of acquisition (AOA), also appear to be effective features. Figure 4 shows the relationship between IMG and the outcome variable. There is a weak positive correlation between the two variables, showing a slight tendency for meetings with higher management ratings to feature language that is more easily associated with mental images, i.e. has higher specificity. There are similar weak positive correlations between the outcome variable and each of AOA and CNC, whereas the outcome variable has a weak negative correlation with FAM. Taken together, these suggest that meetings with higher ratings for management and direction tend to have language that is more concrete and sophisticated, while meetings with lower management ratings tend to have language that is more vague, abstract, and yet familiar (i.e. generic). However, further detailed analysis of actual instances and their placement with respect to important events in meetings, e.g. decision points, and participant roles is necessary to support this idea more robustly.

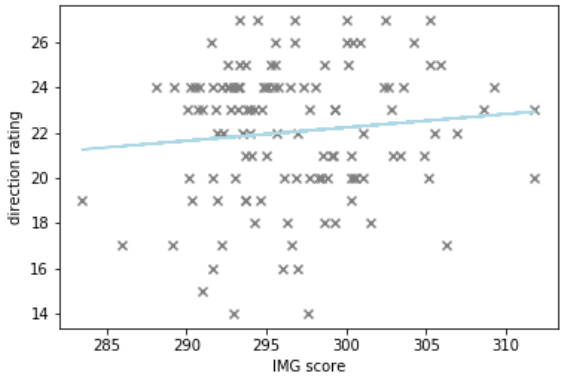

Figure 4: Relationship between Imageability and Management Score

6 CONCLUSION

In this paper, we have shown that multimodal features can be used to predict group members' attitudes about the management and direction of their meeting, at a rate significantly better than baseline performance. Gradient Boosted Trees performed the best overall. Linguistic features on their own did not yield competitive performance, but the best results were found by combining speech and linguistic features. This highlights the advantage of taking a multimodal approach to automatically detecting private affective states that may be manifested subtly in the group interaction.

We initially hypothesized that using linguistic features could improve model interpretability. That is, it would be easier and more effective to explain why a prediction was made in terms of word types, sentiment, and lexical cohesion, than in terms of acoustic features or gestures. However, the results suggest there are dependencies between lexical and acoustic features that need to be taken into account when explaining model predictions. Moreover, feature analysis indicated that the less interpretable word embedding features played an important role in the combined model. Nevertheless, the analysis suggested a number of potential avenues for further investigation with respect to interpretability. In particular, we identified concreteness and imageability of lexical content as a potentially informative cues for group satisfaction with respect to meeting management. Future work will look at the relationship between the more interpretable linguistic features and lexical representations learned from the data such as word embeddings. We also note that the acoustic features used in these experiments only give a very coarse representation of the meeting. Further work will look at combining linguistic features with a richer set of acoustic features at a more fine-grained level using sequence analysis techniques.

Another future direction will be to examine some of the extreme cases where a meeting is rated either very high or very low in terms of management and direction, and supplement the regression experiments described here with classification experiments that attempt to discriminate between the two cases in terms of multimodal features of the discussion. We will also carry out more in-depth feature analysis to determine which factors are highly associated with good or poor meeting management. We also plan to perform more detailed analysis of the effect of roles on predicting group-level attitudes towards meeting management. We will compare the current results with models that omit the ratings of the PM, and also look at cases where the PM had a strongly negative opinion regarding the direction of the meeting. Finally, we will also develop multi-level models for individual attitudes that explicitly take into account participant role and the place of the meeting in the scenario sequence.

Acknowledgement Gabriel Murray was supported by an NSERC Discovery Grant.

REFERENCES

[1] Joseph A Allen and Steven G Rogelberg. 2013. Manager-led group meetings: A context for promoting employee engagement. *Group & Organization Management* 38, 5 (2013), 543–569.

[2] 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.

[3] Judee K Burgoon, Nadia Magnenat-Thalmann, Maja Pantic, and Alessandro Vinciarelli. 2017. *Social signal processing*. Cambridge University Press.

[4] 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.

[5] Elizabeth Couper-Kuhlen and Margret Selting. 2017. *Interactional linguistics: an introduction to language in social interaction*. Cambridge University Press.

[6] 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.

[7] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear* (2017).

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

[9] Nale Lehmann-Willenbrock, Steven G Rogelberg, Joseph A Allen, and John E Kello. 2018. The critical importance of meetings to leader and organizational success. *Organizational Dynamics* 47, 1 (2018), 32–36.

[10] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

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

[12] Gabriel Murray and Giuseppe Carenini. 2011. Subjectivity detection in spoken and written conversations. *Natural Language Engineering* 17, 3 (2011), 397–418.

[13] Maja Pantic and Alessandro Vinciarelli. 2014. Social signal processing. *The Oxford handbook of affective computing* (2014), 84.

[14] 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.

[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[16] Alex Pentland. 2007. Social signal processing [exploratory DSP]. *IEEE Signal Processing Magazine* 24, 4 (2007), 108–111.

[17] Wilfried M. Post, Mirjam Huis in âĂŹt Veld, and Sylvia van den Boogaard. 2007. Evaluating meeting support tools. *Personal and Ubiquitous Computing* 12, 3 (March 2007), 223–235. https://doi.org/10.1007/s00779-007-0148-1

[18] Stephan Raaijmakers, Khiet Truong, and Theresa Wilson. 2008. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 466–474.

[19] Steve Renals, Hervé Bourlard, Jean Carletta, and Andrei Popescu-Belis. 2012. *Multimodal Signal Processing: Human Interactions in Meetings*. Cambridge University Press.

[20] Dairazalia Sanchez-Cortes, Oya Aran, and Daniel Gatica-Perez. 2011. An audio visual corpus for emergent leader analysis. *ICMI-MLMI), Multimodal Corpora for Machine Learning, Nov* (2011), 14–18.

[21] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. 2013. Emergent leaders through looking and

speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces* 7, 1-2 (2013), 39–53.

[22] 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.

[23] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan*. 2794–2797.

[24] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.

[25] 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.

[26] Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers* 20, 1 (1988), 6–10.