Computational Modeling of Face-to-Face Social Interaction Using Nonverbal Behavioral Cues

THÈSE N^{O} 4986 (2011)

PRÉSENTÉE LE 6 MAI 2011

À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE L'IDIAP

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Dinesh Babu JAYAGOPI

acceptée sur proposition du jury:

Dr P. Pu Faltings, présidente du jury
Dr D. Gatica-Perez, directeur de thèse
Prof. A. Nijholt, rapporteur
F. Pianesi, rapporteur
Prof. J.-Ph. Thiran, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

Abstract

The computational modeling of face-to-face interactions using nonverbal behavioral cues is an emerging and relevant problem in social computing. Studying face-to-face interactions in small groups helps in understanding the basic processes of individual and group behavior; and improving team productivity and satisfaction in the modern workplace. Apart from the verbal channel, nonverbal behavioral cues form a rich communication channel through which people infer - often automatically and unconsciously - emotions, relationships, and traits of fellow members.

There exists a solid body of knowledge about small groups and the multimodal nature of the nonverbal phenomenon in social psychology and nonverbal communication. However, the problem has only recently begun to be studied in the multimodal processing community. A recent trend is to analyze these interactions in the context of face-to-face group conversations, using multiple sensors and make inferences automatically without the need of a human expert. These problems can be formulated in a machine learning framework involving the extraction of relevant audio, video features and the design of supervised or unsupervised learning models.

While attempting to bridge social psychology, perception, and machine learning, certain factors have to be considered. Firstly, various group conversation patterns emerge at different time-scales. For example, turn-taking patterns evolve over shorter time scales, whereas dominance or group-interest trends get established over larger time scales. Secondly, a set of audio and visual cues that are not only relevant but also robustly computable need to be chosen. Thirdly, unlike typical machine learning problems where ground truth is well defined, interaction modeling involves data annotation that needs to factor in inter-annotator variability. Finally, principled ways of intergrating the multimodal cues have to be investigated.

In the thesis, we have investigated *individual* social constructs in small groups like dominance and status (two facets of the so-called vertical dimension of social relations). In the first part of this work, we have investigated how dominance perceived by external observers can be estimated by different nonverbal audio and video cues, and affected by annotator variability, the estimation method, and the exact task involved. In the second part,

we jointly study perceived dominance and role-based status to understand whether dominant people are the ones with high status and whether dominance and status in small-group conversations be automatically explained by the same nonverbal cues. We employ speaking activity, visual activity, and visual attention cues for both the works.

In the second part of the thesis, we have investigated *group* social constructs using both supervised and unsupervised approaches. We first propose a novel framework to characterize groups. The two-layer framework consists of a individual layer and the group layer. At the individual layer, the floor-occupation patterns of the individuals are captured. At the group layer, the identity information of the individuals is not used. We define group cues by aggregating individual cues over time and person, and use them to classify group conversational contexts - cooperative vs competitive and brainstorming vs decision-making. We then propose a framework to discover group interaction patterns using probabilistic topic models. An objective evaluation of our methodology involving human judgment and multiple annotators, showed that the learned topics indeed are meaningful, and also that the discovered patterns resemble prototypical leadership styles - *autocratic*, *participative*, and *free-rein* - proposed in social psychology.

Key words: Small group, face-to-face interactions, nonverbal cues, automatic social inference, group conversational context, cooperative behavior, competitive behavior, brainstorming, decision-making, group behavior discovery, topic models.

Résumé

La modélisation informatique des interactions face-à-face à partir de manifestations non verbales du comportement constitue un problème émergent et pertinent en sociologie informatique. Etudier les interactions directes de petits groupes permet de mieux comprendre les processus fondamentaux qui régissent les comportements individuels et de groupe, ainsi que d'améliorer la productivité et la satisfaction de groupe en milieu professionel. En plus du discours, le comportement non verbal constitue un riche moyen de communication par lequel les gens déterminent (souvent automatiquement et inconsciemment) les émotions, les rapports ainsi que la personnalité des membres du groupe.

Dans le domaine de la psychologie sociale et de la communication non-verbale, il existe déja un solide ensemble de connaissances concernant l'étude de petits groupes et la nature multimodale des manifestations non verbales. Toutefois, ce n'est que récemment que la communauté du traitement du signal multimodal a commencé à s'attaquer au problème. Une tendance récente consiste à analyser les interactions à l'aide de plusieurs capteurs, dans le cas de conversations de groupe en face-à-face, et à établir des inférences automatiquement sans l'intervention humaine d'un expert. Ces problèmes peuvent être formulés dans un cadre d'apprentissage automatique, impliquant l'extraction de primitives auditives et visuelles pertinentes, ainsi que la conception de modèles d'apprentissage avec ou sans étiquettes.

En tentant de lier psychologie sociale, perception sociale et apprentissage automatique, certains facteurs doivent être pris en considération. Tout d'abord, différents motifs de conversation de groupe apparaissent à différentes échelles de temps. Par exemple, un changement de locuteur se manifeste temporellement de manière locale, alors qu'une relation de domination ou l'émergence de tendances de groupe sont observées sur une échelle de temps plus étendue. Par ailleurs, outre la pertinence de l'ensemble de signaux visuels et auditifs choisis, leur tractabilité informatique revêt une importance capitale. De plus, contrairement aux tâches habituelles traitées en apprentissage automatique dans lesquelles la vérité de terrain est disponible, la modélisation des interactions implique de prendre en compte la variabilité des annotations provenant de différents annotateurs. Enfin, des méthodes d'intégration de signaux mutimodaux doivent être explorées.

Dans cette thèse, nous nous sommes intéressés aux concepts sociaux de domination et de statut, qui sont deux facettes de la dimension dite verticale des relations sociales. Dans une première partie, nous avons étudié comment la domination peut être estimée à l'aide de différentes manifestations sonores et visuelles et affectée par la variabilité inter-annotateur, par la méthodes d'estimation et par la tâche exacte en question. Dans une seconde partie, nous étudions conjointement la domination et le statut basé sur le rôle afin de comprendre si les personnes dominantes sont celles qui ont un statut élevé et si la domination et le statut dans les conversations en petits groupes peuvent être automatiquement expliqués par les mêmes éléments non verbaux. Pour ces deux tâches, nous avons employé l'activité de parole, l'activité visuelle et l'attention visuelle.

Dans la deuxième partie de la thèse, nous avons exploré les mêmes concepts sociaux de manière à la fois supervisée et non supervisée. Nous proposons tout d'abord un cadre pour caractériser les groupes. Cette approche consiste en deux niveaux. Au niveau individuel, l'implication de chaque individu dans la conversation est déterminée. Au niveau du groupe, l'identité des individus n'est pas utilisée. Les groupes sont définis en regroupant les signaux individuels de chaque personne sur une période de temps et en les utilisant pour classifier la nature des conversations du groupe : coopératif vs compétitif et brainstorming vs prise de décision. Nous proposons ensuite une approche pour découvrir les motifs d'interactions de groupe basée sur des modèles probabilistes appelés "topic models". Une évaluation objective de notre méthodologie basée sur le jugement humain et faisant intervenir plusieurs annotateurs révèle que les motifs appris sont en effet significatifs et également que les tendances découvertes s'apparentent à des prototypes de style de leadership proposés en psychologie sociale : leadership autocratique, leadership laisser-faire, ou leadership démocratique.

Mots clés : Petits groupes, interactions face-à-face, comportement non verbal, inférence sociale automatique, contexte conversationnel de groupe, comportement coopératif, comportement compétitif, brainstorming, prise de décision, découverte de comportement de groupe, modèles à topic.

Acknowledgements

First and foremost I would like to thank my supervisor Daniel Gatica Perez, for choosing me to work on this interesting thesis. I am greatly indebted to his perfect mentoring, and support - both technical and personal. Interactions with him are always positive and fruitful. I would always cherish this 4 years experience and take back plenty of wisdom that he has happily shared with me.

I also would like to thank my wife Kavitha. Her constant encouragement and sacrifices need special mention. I attribute a lot of my character and scholarship to my parents. Thanks to them for that. I also thank my lovely brother and all my relatives who have supported me.

I am grateful to the contribution of my thesis committee - Fabio Pianesi, Anton Nijholt, Jean-Philippe Thiran, Daniel Gatica Perez, and Pearl Pu - for being part of my thesis and providing constructive feedbacks to improving it.

Collaborating with Sileye, Hayley, Jean-Marc, Chuohao, Bogdan, Taemie, and Dayra was a great learning opportunity. Sileye and Hayley supported me a lot in the beginning of my PhD. I would also like to thank all the group members of the social computing group - Radu, Kate, Joan, Paco, Dayra, Oya, Hari, Gokul, and Minh-Tri. The group's diversity and talent is amazing. I had lots of opportunities to interact with Jean-Marc's group. Learnt a lot in their reading groups.

Idiap is a great environment. Thanks to Herve for creating such a place. The support staff at Idiap -Nadine, Sylvie, Chris, Ed, Vincent, Valerie, Frank, Bastien, Norbert, Cedric, and Tristan - are very effective. Thank you guys. Also thanks to EPFL support staff. They are very professional indeed.

I should also thank my office-mates - Alex, Majid, Venki, Kate, Stefan, Radu, Remi, Chris, CC - for adding life to the office. Thanks to Indian friends in Martigny - Joel, Shakeela, Venki, Abhilasha, Anindya, Jagan, Hari, Gokul, Sriram, Harsha, Ramya, Murali, Lakshmi, Saheer, Francina; and Lausanne - Prakash, Viswa, Perumal, Arvind for adding life to life. It was fun to share apartment with Deepu and

Sriram. Azhagu and his family, Michelle, Muneer, and Patricia have been a great support for me and my wife.

Thanks also to Mathew, Alex, Laurent, Harsha, Hari for helping with the thesis. Mathew has been a source of knowledge and experience for the past 4 years. Fabio helped with the rehearsal of thesis defence. Thanks to 'apple' Gokul for much needed support in coding.

I take this opportunity to acknowledge some people who have shaped my life. My uncle Mr. Murthy who inspired me to think big. Mr. Arun Kumar, my maths teacher challenged me to think and generalize, rather than memorize. Professor P.V. Ramakrishna who inspired students to achieve. Dr. Mala John, Prof. Rajgopal, Dr. Ganesh Murthy, and Dr. Shanmukh were some of my well-wishers. Akash, Dina, Anusha, Sevvel, Jayanthi, Swarna, Rat, Oswin, Suresh, Mathi, Divya, Anand, Prakash, Megha, Gokul, Arun, Chandra are some of my good and supportive friends.

Doing PhD has its ups and downs. I express my sincere apologies if I had hurted someone's feelings or expectations during the last four years. Hope my thesis is atleast a drop in the ocean of knowledge created by numerous research scholars over hundreds of years.

I also would like to acknowledge my funding sources - US Video Analysis and Content Extraction (VACE) project and the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2).

Contents

1 Introduction **1**

1.1 Objective . . . 2

1.2 Motivation . . . 2

1.3 Related work . . . 4

1.3.1 Group interaction in social psychology . . . 4

1.3.2 Nonverbal cues in social psychology . . . 5

1.3.3 Automating nonverbal cue extraction . . . 5

1.3.4 Computational modeling of group interaction . . . 6

1.4 Contribution . . . 9

1.5 Organization . . . 10

2 Computational Modeling of Dominance **11**

2.1 Related work . . . 13

2.1.1 Dominance in social psychology . . . 13

2.1.2 Dominance in social computing . . . 15

2.2 Our approach . . . 17

2.3 Meeting data and dominance tasks . . . 18

2.3.1 Meeting data . . . 18

2.3.2 Annotating the data . . . 19

2.3.3 Analysis of the annotations . . . 20

2.4 Audio and visual nonverbal cues for dominance modeling . . . 22

2.4.1 Audio cues . . . 22

2.4.2 Visual activity cues . . . 24
2.4.3 Visual attention cues . . . 27
2.5 Models for dominance estimation . . . 31
2.5.1 Unsupervised model . . . 31
2.5.2 Supervised model . . . 32
2.5.3 Experimental protocol . . . 32
2.6 Classifying the Most-Dominant person . . . 33
2.6.1 Full-agreement data set . . . 33
2.6.2 Majority-agreement data set . . . 39
2.7 Classifying the Least-dominant person . . . 42
2.7.1 Full-Agreement data-set . . . 43
2.7.2 Majority-agreement data-set . . . 46
2.8 Discussion and Conclusion . . . 47

3 Beyond Dominance: estimating status 53

3.1 Related work . . . 55
3.1.1 Related work on role modeling . . . 55
3.1.2 Related work on status modeling . . . 56
3.2 Experimental setup: Meeting data and tasks . . . 58
3.2.1 Dominance Task: Estimate the most-dominant person . . . 58
3.2.2 Status Task: Estimate the project manager . . . 58
3.3 Nonverbal cues . . . 58
3.4 Estimation and evaluation method . . . 60
3.5 Results . . . 61
3.5.1 Audio cues . . . 62
3.5.2 Visual activity cues . . . 63
3.5.3 Visual attention cues . . . 64
3.5.4 Centrality measures . . . 66
3.6 Discussion and Conclusion . . . 67

4 Classifying group conversational context 69

4.1 Related Work . . . 70

4.2 Our Approach . . . 72

4.2.1 Individual nonverbal cue extraction . . . 72

4.2.2 Group nonverbal cue extraction . . . 74

4.2.3 Group conversational context classification . . . 75

4.3 Classifying cooperative vs competitive interaction . . . 76

4.3.1 Meeting datasets . . . 76

4.3.2 Experiments and results . . . 77

4.4 Classifying brainstorming vs decision-making interaction . . . 81

4.4.1 Meeting dataset . . . 81

4.4.2 Experiments . . . 83

4.4.3 Results . . . 84

4.5 Discussion and Conclusion . . . 86

5 Mining group nonverbal conversational patterns 89

5.1 Related Work . . . 90

5.2 Our Approach . . . 91

5.3 Low level Cue extraction, Bag-of-NVPs, and the Topic model . . . 93

5.3.1 Low level nonverbal cue extraction . . . 93

5.3.2 Bag-of-NVPs generation: . . . 95

5.3.3 Latent Dirichlet Allocation (LDA) topic model . . . 98

5.3.4 From interaction slices to group characterization . . . 99

5.4 Meeting data . . . 100

5.5 Experiments and results . . . 100

5.5.1 Bag-of-NVPs over varying slice duration . . . 101

5.5.2 LDA based pattern discovery . . . 102

5.6 Discussion and Conclusion . . . 109

6 Conclusions and Future Directions 113

A Objective evaluation: Human annotation **117**

Curriculum Vitae **133**

List of Figures

1.1 Thesis overview: shows the problem of social inference in humans, social psychology, and computational modeling. Social psychology literature has studied human perception using nonverbal behavior. Computational methods broaden the scope of social inference modeling by automating cue extraction; and jointly studying multiple behavioral cues and social constructs using machine learning frameworks. 3

2.1 Flow diagram of our approach. 17

2.2 Plan view of the meeting room set up. 18

2.3 Examples of the seven camera views available in the AMI meeting room. The top row shows the right, centre and left cameras, while the bottom row shows the view from each of the close up cameras. 19

2.4 Illustration of compressed domain features. (a) Shows the original image. (b) Shows the direction of motion vectors. (c) Shows the residual coding bitrate at different pixel locations (red means high magnitude). (d) Shows the locations where skin color was detected in red color. 25

2.5 (a). Shows the top view of the meeting room. (b). Shows the side camera views and the estimated visual focus of participants using side-view camera views. Each of the participants is labeled and their focus of attention is displayed above their head (T stands for Table and S stands for Slide-screen). Colored rectangle around the head shows the head location and colored arrows shows the head pose of each of the participants. The white transparent box placed on participant A shows that her speaking status is 'true'. 28

2.6 Flow diagram showing our experimental protocol. 33

2.7 Scatter plots of the total speaking and visual activity length, where the red crosses show the data points belonging to the positive class and the black circles show the negative class in each case. . . . 37

2.8 Comparison of the best performance values for the most-dominant estimation tasks. A:Audio, V. Act: Visual Activity, V. Att: Visual Attention, A/V: Audio-Visual. . . . 43

2.9 Comparison of the best performance values for the least-dominant estimation tasks. A:Audio, V. Act: Visual Activity, V. Att: Visual Attention, A/V: Audio-Visual. . . . 47

3.1 Venn diagram showing overlapping and non-overlapping subsets of most-dominant and high-status person data. . . . 62

3.2 Histogram plots of normalised Total Speaking Length for both the most-dominant (MD) and project manager (PM) task. . . . 64

3.3 Histogram plots of normalised Total Speaking First after another participant (TSF) for both the most-dominant (MD) and project manager (PM) task. . . . 65

4.1 Block Diagram of our work. . . . 73

4.2 Nonverbal Cue Extraction. . . . 73

4.3 *Top*: Snapshot from an AMI meeting, showing the participants from two side-view camera view. *Bottom*: Snapshot of an Apprentice meeting - highlighting the high-status leader (Trump) - *bottom left* and a long-shot of the board-room meeting - *bottom right*. 77

4.4 Normalized histograms of GIT and GTDM in the two meeting datasets. . . . 79

4.5 Classification using SVM in the feature space of GIT and GTDM. . . . 80

4.6 *Top-left*: Snapshot from the AMI meeting, showing the participants from the center-view camera. *Top-right*: Distribution of speaking length, speaking turns, and successful interruptions among the participants. *Bottom-left*: The evolution of the Group-Interruption-to-Turns Ratio with time. *Bottom-right*: The evolution of the Group Turn Distribution Measure with time. . . . 80

4.7 Sociometric badge developed by Human Dynamics group, MIT Media Lab (Olguín and Pentland, 2008). . . . 82

4.8 Example of an interacting group wearing sociometric badges around the neck. . . . 82

4.9 Performance of the group cues on classifying the brainstorming and decision-making meetings during collocated setting (Task 1). 84

4.10 Performance of the group cues on classifying the brainstorming and decision-making meetings during distributed setting (Task 2). 85

4.11 Performance of the group cues on classifying the brainstorming and decision-making meetings (Task 3). 85

4.12 Performance of combination of group features on predicting the brainstorming and decision-making meetings. 86

5.1 Overview of the group NVP discovery process using topic models. 93

5.2 Diagram showing the features to characterize individual and group behavior (generic-based and leadership-based) extracted in our approach. See main text for details. 94

5.3 Example joint histograms for each of the Speaking Distribution NVPs other than *Silence*. 96

5.4 Latent Dirichlet Allocation (LDA) model 100

5.5 Empirical distribution of Speaking Distribution patterns at different time scales (from 30-seconds to 5-minute). x-axis of each of the sub-figure is the classes and y-axis is the probability of the particular class. 102

5.6 Empirical distribution of leadership patterns at two different time scales (2-minute and 5-minute). x-axis of each of the sub-figure is the classes and y-axis is the probability of the particular class. '0' corresponds to the case when there is silence, 'L' (resp. 'NL') when leader (resp. someone else) has maximum feature value. 103

5.7 Leadership styles by Lewin et al. The blue envelope shows the emphasis (in terms of power) that is placed on the various group members. 104

5.8 Speech segmentation of two sample 5-minute meeting slices for each of the three topics - *autocratic*, *participative* and *free-rein*. The four participants are marked 1, 2, 3,and 4 along the y-axis. The position marked 1 corresponds to the leader (project manager) in all cases. 105

5.9 Topic distribution over groups at 5-minute scale (DL combination). 106

5.10 Topic evolution for selected groups at 5-minute scale (DL combination). The topics are color coded - *autocratic* in red, *participative* in light-blue, *free-rein* in yellow. The x-axis represents time. The y-axis represents meeting sessions. 107
5.11 Topic distribution over groups at 2-minute scale (DL combination). 109
5.12 Three snapshots of a group interaction - at 2-minute, 3-minute, 4-minute - with the top left panel showing the center view camera, the top right showing the speech segmentation evolution w.r.t time in x-axis and the participants in the y-axis, the bottom left panel showing the low level cues for each of the participant, and the bottom right panel showing the topic distribution - red being *autocratic*, blue being *participative* and green being *free-rein* for the intervals 0-2 min, 1-3 min, and 2-4 min. This meeting slice corresponds to group 5, which is *participative* at both 2-minute and 5-minute time scales.110
5.13 Speech segmentation of two sample 5-minute meeting slices for each of the three topics - *Leader-domination*, *Group Interaction*, *Monologue*. The x-axis indicates time. The four participants are marked 1, 2, 3, and 4 along the y-axis. The position marked 1 corresponds to the leader (project manager) in all cases. 111
5.14 Speech segmentation of two sample 5-minute meeting slices for each of the three topics - *Laid-back monologue*, *Monologue with brief exchanges*, *Interaction hot-spot*. The x-axis indicates time. The four participants are marked 1, 2, 3, and 4 along the y-axis. The position marked 1 corresponds to the leader (project manager) in all cases. 112

List of Tables

2.1 Dominance tasks and corresponding data-sets. 21

2.2 Glossary of feature abbreviations 31

2.3 Performance of **Audio** cues for **Most**-dominant person with **Full**-agreement data. 34

2.4 Performance of **Visual Activity** cues for **Most**-dominant person task with **Full**-agreement data. 37

2.5 Performance of **Visual Attention** cues for **Most**-dominant person task with **Full**-agreement data. 38

2.6 Performance of **Audio-Visual** cues with **Most**-dominant person task with **Full**-agreement data. 38

2.7 Performance of **Audio** cues for **Most**-dominant person task with **Majority**-agreement data. 40

2.8 Performance of **Visual Activity** cues for **Most**-dominant person task with **Majority**-agreement data. 41

2.9 Performance of **Visual Attention** cues for **Most**-dominant person task with **Majority**-agreement data. 42

2.10 Performance of **Audio-Visual** cues for **Most**-dominant person task with **Majority**-agreement data. 42

2.11 Performance of **Audio** cues for **Least**-dominant person task with **Full**-agreement data . 44

2.12 Performance of **Visual Activity** cues for **Least**-dominant person task with **Full**-agreement data. 45

2.13 Performance of **Visual Attention** cues for **Least**-dominant person task with **Full**-agreement data. 45
2.14 Performance of **Audio-Visual** cues with supervised model for **Least**-dominant person task with **Full**-agreement data. 46
2.15 Performance of **Audio, Visual, and Audio-Visual** cues for **Least**-dominant classification task with **Majority**-agreement data. 48

3.1 Glossary of feature abbreviations. 60
3.2 Performance of **Audio** cues for estimating the most-dominant person and the project manager. 63
3.3 Performance of **Visual Activity** cues for estimating the most-dominant person and the project manager. 64
3.4 Performance of **Visual Attention** cues for estimating the most-dominant person and the project manager. 66
3.5 Performance of **Centrality** measures for estimating the most-dominant person and the Project Manager. 66

4.1 Glossary of abbreviations for the group cues. 75
4.2 Accuracy (%) of speaking activity based nonverbal cues for classification of group conversational context. In the caption, GNB stands for Gaussian Naive Bayes classifier and SVM-lin is the short form of SVM using a linear kernel. 78

5.1 LDA based topic discovery at 5-minute scale (DL combination). 104
5.2 Evaluation: Confusion matrix between the ground-truth and the model output 105
5.3 LDA based discovery at 2-minute scale (DL combination). 108
5.4 LDA based discovery at 5-minute scale (OL combination). 108
5.5 LDA based discovery at 5-minute scale (OGD combination). 110

Chapter 1

Introduction

Computational modeling of face-to-face interaction using nonverbal behavioral cues is an emerging and relevant problem in social computing. Studying face-to-face interactions provide insights into the functioning of small groups. With teams becoming ubiquitous in business, government, and non-governmental organizations, the need to understand group dynamics and connecting them to group productivity and satisfaction has become more and more relevant. Though verbal communication plays a significant role, the nonverbal channel too conveys a wealth of information about group dynamics (Knapp and Hall, 1978). Also, nonverbal analysis is privacy-sensitive as 'what is spoken' is never made use of. Recent technological trends in sensing, signal processing, and machine learning have enabled automatic sensing, cue extraction, and modeling of social interactions.

Social psychology literature has studied small groups and nonverbal behavior for more than half a century. Researchers have tried to understand various issues related to formation of small groups; structure in small groups- status, norms, roles, cohesion; and performance -role of leaderhip, productivity, and decision-making (Levine and Moreland, 1990; Poole *et al.*, 2004). Nonverbal cues have been known to be key in social inference of emotions, expectancies, relationships, and traits of human subjects (Hassin *et al.*, 2005). Often expression and perception of nonverbal behavior are known to be automatic and unconscious (Hassin *et al.*, 2005).

With the new framework of automatic modeling of social interactions, both individual and group behavior could be understood and modeled, by employing multimodal nonverbal cues that can be robustly extracted. Analyzing and modeling social interaction helps in understanding human behav-

ior and retrieving meeting recordings using queries related to behavior. Computationally efficient method allow the possibility to support online group collaboration. A recent trend is to analyze social interaction in the context of group conversations, using multiple sensors like cameras and microphones and make inferences automatically without the need of a human expert. These problems can be formulated in a machine learning framework involving the relevant audio and video feature extraction and supervised or unsupervised learning models.

When attempting to bridge social psychology and machine learning, certain factors have to be considered. Firstly, unlike typical machine learning problems where ground truth is well defined, group interaction modeling involves data annotation that needs to factor in inter-annotator variability. Secondly, a set of audio and visual cues that are not only relevant but also robustly computable need to be chosen. Thirdly, principled methods to combine these features have to be investigated. Finally, various group conversation patterns emerge at different time-scales. For example, turn-taking patterns evolve over shorter time scales (Gatica-Perez, 2006), whereas dominance or group-interest trends get established over larger time scales.

1.1 Objective

The primary objective of this thesis is to design and develop computational models for a few fundamental social constructs in small group interaction including dominance, status, group conversational context, and leadership styles. The setting for these problems is face-to-face conversations using multimodal nonverbal cues. Our work places emphasis on automatic cue extraction, joint modeling of nonverbal cues, joint understanding of social constructs, and characterization of both individual and group behavior. Figure 1.1 illustrates our work in the thesis. There exists a solid body of knowledge about the multimodal nature of these phenomena in social psychology. However, the problem has only recently begun to be studied in the multimodal processing community.

1.2 Motivation

We foresee three types of applications that could be developed when the computational nonverbal modeling of face-to-face interaction matures as a research field:

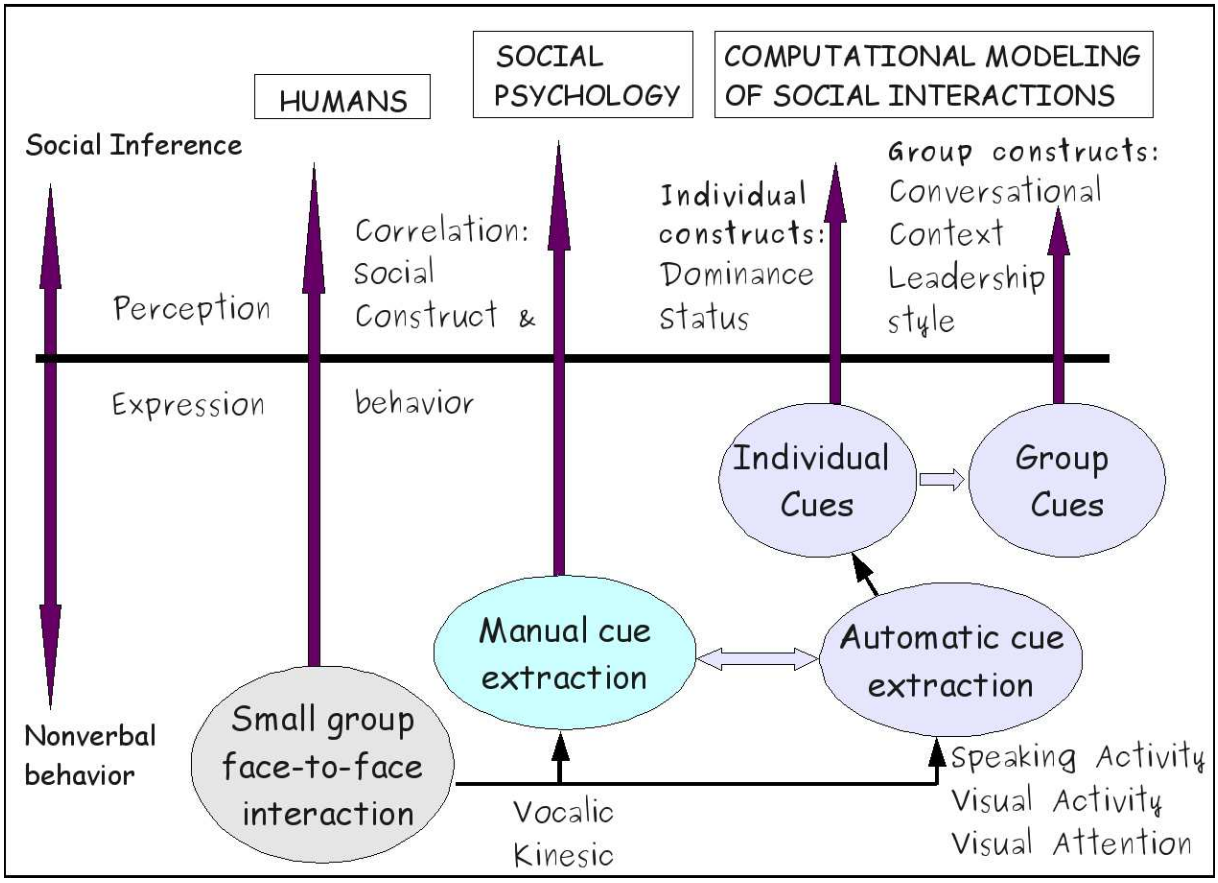


Figure 1.1. Thesis overview: shows the problem of social inference in humans, social psychology, and computational modeling. Social psychology literature has studied human perception using nonverbal behavior. Computational methods broaden the scope of social inference modeling by automating cue extraction; and jointly studying multiple behavioral cues and social constructs using machine learning frameworks.

1. **Behavior-based media retrieval**: The potential applications of automatic nonverbal analysis of face-to-face group interactions in workplace include identifying leadership skills and monitoring team cohesiveness. From a human resource perspective, analyzing group behavior could signal the need for a team-building exercise or a leadership change. Tracking teams could also indicate what types of behavior teams are mostly engaged in - for example cooperative or competitive behavior.

2. **Behavior-based support of individuals and groups**: Social inference machines could be part of relevant applications including self-assessment, training, and educational tools (Pentland, 2005), and of systems to support group collaboration (DiMicco *et al.*, 2004). There is support from the social psychology literature about the fact that people who display cues like verbal flu-

ency, well modulated voice, etc are often seen as more competent and become more influential, whereas displaying dominant behavior might be an ineffective strategy to gain influence as compared to (Ridgeway, 1987). Also, nonverbal self-accuracy (how aware are we about our own nonverbal behavior) is not uniform for all nonverbal cues, and cognitive activity can reduce this accuracy even further. This motivates the need for measuring human behavior automatically for self-assessment.

3. **Tools for social psychology research**: Barring few exceptions, the nonverbal cues studied in the social psychology literature have been manually coded. This process is highly labour-intensive and expensive. With the availability of ubiquitous and infrastructure-based sensors and automatic extraction of nonverbal cues, the cue extraction process can be easily automated. As more interaction data becomes available, computational models to extract behavior using multiple cues could become common in the future.

1.3 Related work

In this section, we first review the literature on small-group and nonverbal behavior research in social psychology. Then, we review the state-of-the-art on automatic nonverbal cue extraction and computational modeling of social interactions.

1.3.1 Group interaction in social psychology

Groups have been traditionally looked as vehicles for influencing members, performing tasks, and improving member self-understanding (Arrow *et al.*, 2000), thanks to some of the pioneering works were done by Lewin, Bales, and Mc. Grath. Lewin's work studied the need for groups, importance of the member-member relations and of member-group relations (Lewin and Lewin, 1948). Bales developed a systematic method of observing and describing groups emphasizing that the mental processes of individuals take place in systematic contexts which can be measured and hence allow for explanation and estimation of behavior (Bales, 1950). McGrath gave special emphasis to temporal processes in group interaction and task performance (McGrath, 1984). A large volume of work has followed investigating issues related to composition of groups; structural issues such as status, norms, roles, and cohesion; and performance issues such as group decision-making, productivity, and lead-

ership in small groups. Some of the more recent reviews on small group literature include (Levine and Moreland, 1990; Poole *et al.*, 2004). Overall the field is clearly active and of particular importance for our work are the connections between small groups and nonverbal communication in the workplace (Remland, 2006). More specifically, the vertical aspects relating to power, dominance, status, hierarchy, and related concepts and nonverbal behavior (Hall *et al.*, 2005).

1.3.2 Nonverbal cues in social psychology

The history of the empirical study of nonverbal behavior begins with Charles Darwin ((Darwin, 1965), work of 1872 reprinted), where he studied the expression of emotions in man and animals. The multifunctionality of nonverbal expression as a *symptom* (of the expresser's state), as a *symbol* (of a socially shared meaning category), and as an *appeal* (a social message toward others) is well known (*Methodological issues in studying nonverbal behavior* in (Harrigan *et al.*, 2008)). Nonverbal cues have been documented extensively in the study of relationship of individuals in dyads, groups, and group as a whole (Knapp and Hall, 1978; Manusov and Patterson, 2006; McNeill, 2000; Hassin *et al.*, 2005). Both expression and perception of many of these cues are often automatic and unconscious (Hassin *et al.*, 2005). Nonverbal cues include among others *vocalic* - prosody, speaking turns, laughter - and *kinesic* - gestures, moves, gaze - (Dunbar and Burgoon, 2005b). Various nonverbal cues correlated with social constructs like dominance, status, and power (Hall *et al.*, 2005) and individual constructs like personality have been extensively studied (Rotter, 1966; John and Srivastava, 1999). Turn-taking patterns, gazing, smiling, touching, and various body positions can be used to infer social verticality in human relations (Hall *et al.*, 2005). Our work takes inspiration from the nonverbal behavior literature to extract relevant cues, robustly and automatically.

1.3.3 Automating nonverbal cue extraction

Various nonverbal cues have been automatically extracted in the signal processing and computer vision literature. So far, extraction of turn-taking patterns has been the most robust, which involve recognizing 'when someone speaks' using close-talk or distant microphones (Basu, 2002). Prosodic cues describing 'how someone speaks', using cues such as energy, pitch frequency, rate of speech have been studied to infer 'affect' and group interest (Wrede and Shriberg, 2003), (Harrigan *et al.*, 2008).

Automatic facial expression analysis to infer emotional states has also been extensively studied (Tian *et al.*, 2005). Visual attention cues have been investigated for the meeting space using both head-pose (Otsuka *et al.*, 2007; Murphy-Chutorian and Trivedi, 2009; Ba and Odobez, 2010); and eye-gaze (Gorga and Otsuka, 2010). Audio-visual laughter detection (Petridis and Pantic, 2008) and smiling has also been investigated (Kumano *et al.*, 2009). Head-gestures like nodding, shaking have been studied (Kapoor and Picard, 2001; Morency *et al.*, 2005; Otsuka *et al.*, 2007). Fidgeting was investigated in (Chippendale, 2006). Some of the nonverbal cues used to infer social constructs have been referred to as 'honest signals', due to their uncontrollable nature both from the expressor and the perceiver's viewpoint (Pentland, 2008). For an overiew of the audio-visual technology for a conversation scene analysis system (both offline and real-time) developed in NTT laboratories, Japan, the reader is referred to (Otsuka and Araki, 2010). In this thesis, we extract automatically three types of nonverbal cues relevant to the study of verticality 1. The turn-taking cues, based on speech activity 2. Visual activity cues, extracted in the compressed domain and 3. Visual attention cues, based on head pose and use them to model social verticality.

1.3.4 Computational modeling of group interaction

In this subsection, we summarize the existing literature on individual and group behavior modeling in group conversations using both supervised and unsupervised approaches. As our work and review mainly concerns small groups, we do not discuss works that relate to dyadic interactions as opposed to small group interaction, although some of the pioneering works by Pentland et al. need mention, as they showed the predictive power of robustly extractable nonverbal cues ('honest signals') in dyadic relations (Pentland, 2008). These speech and physical activity based cues, characterized in terms of emphasis, activity, influence, and mimicry have been shown to estimate job performance, negotiation outcomes, dating outcomes, etc. This research group also pioneered the use of wearable sensors, also called sociometers (Choudhury and Pentland, 2002) for recording interactions, as opposed to infrastruture based recordings.

Supervised learning approaches for modeling individual and group behavior

Regarding individual behavior modeling, attempts have been made to estimate *dominant behavior*, certain *personality traits*, and certain *roles* that individuals are involved in. *Dominance* can be defined as a personality trait or behavior involving the motive to control others, the self-perception of oneself as controlling others, and/or as a behavioral outcome (success in controlling others or their resources) (Hall *et al.*, 2005). In (Rienks and Heylen, 2005) *dominant behavior* was estimated by computing speaking turns based features (like speaking time, turns, successful interruptions) using manual annotations of speaking turns and Support Vector Machines (SVM) on meetings from the M4 (MultiModal Meeting Manager) meeting corpus (McCowan *et al.*, 2005). Personality traits, specifically *extraversion* (sociable, assertive, playful) vs *intraversion* (aloof, reserved, shy) were estimated using support vector regression and applied to sequences of the MS (Mission Survival) Corpus (Pianesi *et al.*, 2008a). Using an influence model *functional roles* in meetings related to tasks and socio-emotional roles were estimated (Dong *et al.*, 2007) on the MS Corpus (Pianesi *et al.*, 2007). The work in (Lepri *et al.*, 2009) estimated *individual performance* from interaction slices. The above three works employed speaking activity cues, prosodic cues, and visual fidgeting cues. In (Vinciarelli, 2007; Garg *et al.*, 2008) *adhoc roles* in broadcast video and the AMI (Augmented Multiparty Intercation) corpus (Carletta et al., 2006) were estimated using Dynamic Bayesian Networks (DBN) and turn-taking information. Recently, *emergent leadership* was modeled using turn-taking patterns and employing score-level fusion techniques (Sanchez-Cortes *et al.*, 2010).

Regarding group behavior modeling, *group activities* have been characterized employing layered sequential approaches [either Hidden Markov Models (HMM) or DBN], where the first layer modeled the individuals' behavior, and the second layer the activity (monologue, presentations, or discussions) in (Zhang *et al.*, 2006; Dielmann and Renals, 2007) or conversational regimes (convergence or monologue, dyad-link and divergence) in (Otsuka *et al.*, 2007). While (Zhang *et al.*, 2006; Dielmann and Renals, 2007) employed speaking-activity and motion-activity in terms of blobs (region of image pixels) as the features, (Otsuka *et al.*, 2007) employed speaking-activity and visual gaze. The latter work was also extended to estimate *interpersonal influence* (Otsuka *et al.*, 2006). *Group interest* was investigated by segmenting meetings temporally into high or neutral interest level segments in HMM based supervised framework and fusing audio-visual activity cues in (Gatica-Perez *et al.*, 2005). Recently, group discussion dynamics was studied further with two different corpora (in two different languages) and

the *group performance* was estimated using turn-taking patterns and the 'honest' signals described in the beginning of this subsection. The work employed three types of supervised models - support vector machines, hidden Markov models and the influence model (Dong *et al.*, 2011).

Unsupervised learning approaches for modeling individual and group behavior

Unlike the previous methods, unsupervised approaches do not need labeled training data. Regarding individual behavior modeling, the *pair-wise influence* between participants in a group was estimated using a dynamic Bayesian approach (Basu *et al.*, 2001). The observations were speaking activity features and influence was estimated using a variation of the coupled HMM (Hidden Markov Model) called the influence model. On the Augmented Multi-Party Interaction with Distance Access (AMIDA) corpus, the *remote participant* in a remote meeting was estimated (Sanchez-Cortes *et al.*, 2009). In another study, on a corpus collected from a TV show, the task was to predict the *participant who would be fired* from the group and who had the *highest status* (Raducanu and Gatica-Perez, 2010). Unlike most other works, the group was competitive in nature i.e. the participants had to ensure that someone else was fired out of the job. The above two works employed turn-taking cues. In all the cases excepting the influence model, the best single features for the estimation tasks were investigated.

Regarding group behavior modeling, various prosody related cues correlated with interest *hot-spots*, where the interest level of the meeting participants was perceived to be high was studied in (Wrede and Shriberg, 2003). Other works have also attempted to quantify *interactivity* and *centrality* in meetings (Otsuka *et al.*, 2006). The 'honest' signals described in the beginning of this subsection were found to be correlated with *team performance* (**?**) and *expertise* (Waber and Pentland, 2009). Recent findings indicate the existence of a general collective intelligence factor in groups that explains a group's performance on a wide variety of tasks. The research shows that this factor is not strongly correlated with the average or maximum individual intelligence of group members, but instead correlated with the average social sensitivity of group members, the equality in distribution of conversational turn-taking, and the proportion of females in the group (Woolley *et al.*, 2010).

We defer the detailed review of related works on dominance, status, role, online support of groups, and discovering human activity to subsequent chapters. Few recent thesis that our work relate to include (Rienks, 2007; Lepri, 2009; Dong, 2010).

1.4 Contribution

The contributions of this thesis are

- We conduct an original and systematic study of vocalic and kinesic nonverbal cues for perceived dominance estimation in small group meetings, and present a detailed objective evaluation of the performance of single and multimodal cues, and of unsupervised and supervised learning approaches (Jayagopi *et al.*, 2009b). Our vocalic cues are based on speaking activity; and kinesic cues are computationally efficient visual activity cues in the compressed domain and visual attention cues use head pose. Unlike all previous computational work, we analyze the annotation of perceived dominance by multiple human observers and are thus able to analyze the implications that the variation of human perception has on the performance of the automatic approaches. Our source of data for this work is the publicly available AMI meeting corpus.
- We propose a novel investigation of automatic estimation of both perceived dominance and role-based status in small-group conversations (Jayagopi *et al.*, 2008b). While some social psychology literature has found common ground for the nonverbal display and interpretation of both constructs, and recent computational literature has started to investigate models for automatic estimation of dominance or roles in conversations, no computational attempt has previously been made to study these two dimensions of social verticality jointly. We use the same set of vocalic and kinesic cues as in the dominance study. Our source of data for this work is the AMI meeting corpus.
- We propose a novel framework for characterizing group nonverbal behavior as compared to individual behavior and then automatically classify group conversational context in a supervised framework. We characterize group conversational behavior by measuring speaking patterns and the overlap-silence patterns of the group as a whole. Specifically, we address two tasks: classifying cooperative vs competitive interactions (Jayagopi *et al.*, 2009a) and the task of classifying brainstorming vs decision-making interactions (Jayagopi *et al.*, 2010). Our source of data for the first task is the AMI meeting corpus and conversational data from a TV show. For the second task, we used a dataset recorded at MIT Media Lab using privacy-sensitive sociometers .
- We address the largely unexplored problem of discovering group nonverbal patterns proposing an unsupervised framework based on probabilistic topic models (Jayagopi and Gatica-Perez,

2009, 2010). We define a new group behavioral descriptor on time slices of group conversational data that is robust to several factors occuring in realistic interactions. we show that the topics discovered by our model are meaningful using ground-truth produced from external observers of the interaction. We also propose new topic-based ways of characterizing groups by aggregating group behavior over multiple interactions.

1.5 Organization

This thesis is organized as follows:

- In Chapter 2, we investigate the problem of modeling dominance in small group face-to-face interactions using multimodal nonverbal cues. We systematically study both single and multiple cues using single and multiple modalities.
- In Chapter 3, we investigate both dominance and role-based status estimation in small-groups using multimodal nonverbal cues.
- In Chapter 4, we study the problem of automatically classifying group converational contexts using nonverbal behavior. We address two tasks: discriminating cooperative vs competitive interactions and discriminating brainstorming vs decision-making interactions.
- In Chapter 5, we explore the problem of discovering group nonverbal patterns in an unsupervised fashion using probabilistic topic models.
- Chapter 6 provides a final discusson about the achievements and limitations of the thesis and discusses future directions.

Chapter 2

Computational Modeling of Dominance using Nonverbal Cues

Certain people are consistently successful at dominating conversations and their results. In fact, within a few minutes of interaction among unacquainted individuals, a dominance order or a participation hierarchy often emerges (Rosa and Mazur, 1979). A concept largely studied in social psychology, dominance is one of the basic mechanisms of social interaction and has fundamental implications for communication both among individuals and within organizations (Burgoon and Dunbar, 2006). While dominant behavior could bring benefits to the person displaying it in certain contexts, in others it could negatively affect the social dynamics of a group, impacting its cohesiveness and effectiveness, and eroding social relationships. Furthermore, displaying *dominant cues* like loud speech or pointing, as opposed to *task cues* like verbal fluency or well-modulated voice tone, is an ineffective strategy to gain influence (Ridgeway, 1987).

The automatic modeling of dominance patterns in groups is a key problem in social interaction analysis from sensor data (Pentland, 2005; Gatica-Perez, 2006), which spans research in audio and visual processing, information fusion, human-computer interaction, and ubiquitous computing. The analysis of face-to-face multiparty conversations to extract patterns of dominance (Basu *et al.*, 2001; Rienks and Heylen, 2005) is challenging, given the complex nature of real communication, and the difficulty to model, accurately and efficiently, the behavior of multiple interacting individuals. Auto-

matic dominance estimators from audio-visual media could be part of relevant human-centered applications including self-assessment and training (Pentland, 2005; Pianesi *et al.*, 2008b), and systems to support group collaboration (DiMicco *et al.*, 2006; Nijholt *et al.*, 2006; Kulyk *et al.*, 2006; DiMicco and Bender, 2007; Kim *et al.*, 2008).

A solid body of work in psychology has documented the multimodal nature of dominance (Dunbar and Burgoon, 2005a), and in particular the role that nonverbal communicative cues (not involving the spoken words) play in the expression and perception of dominant behavior. Although speech is the main modality in conversations (Tusing and Dillard, 2000; Schmid Mast, 2002), substantial information is conveyed in the visual modality through body movement, postures, gaze, and gestures. It is known that, in terms of *vocalic* and *kinesic* cues, dominant individuals behave more actively (i.e., talk and move more, more often and with larger ranges, and receive more attention) than non-dominant people (Dunbar and Burgoon, 2005a; Burgoon and Dunbar, 2006). Some of these activity cues can be automatically extracted from data, and initial work (Basu *et al.*, 2001; Rienks *et al.*, 2006; Rienks and Heylen, 2005) mainly investigated perceptual modalities in isolation (where cues were often extracted manually), or proposed dominance recognition approaches that were applied to relatively constrained interaction scenarios or that were limited in their validation.

This chapter presents a systematic study on fully automated modeling of perceived dominance in small group meetings from nonverbal cues. Focusing on the AMI corpus, a data set of face-to-face interactions recorded with multiple cameras and microphones, our work contains several contributions. First, we investigate a number of robustly extracted and efficient activity cues in both audio and visual domain for the characterization of dominant behavior. Our cues include a novel set of visual cues extracted in compressed-domain video. The visual attention cues are extracted by tracking the head and pose jointly. We consider audio-only, visual activity-only, visual attention-only and audio-visual cases to understand the relative power of each of the modalities and the benefits of using them jointly. Second, we study unsupervised and supervised approaches for dominance modeling, which differ in complexity and needs for training data. Third, through the analysis of the variability of human judgment of perceived dominance in our corpus, we define and study a set of dominance estimation tasks (most-dominant person, least-dominant person) that allow us to objectively quantify the difficulty of each of them, as well as the variation in performance as human performance itself varies. Our results highlight a number of relevant issues, including the robustness of basic audio features, the

power of some visual cues, and the overall advantages of relatively simple approaches. To our knowledge, this work constitutes the most detailed study on automatic modeling of dominance in small group meetings from audio and visual cues to date. The work in this chapter is an expanded version of this publication (Jayagopi *et al.*, 2009b).

The chapter is organized as follows. Section 2.1 reviews the literature on dominance in social psychology and on computational approaches related to our work. Section 2.2 presents the components of our work. Section 2.3 describes the data, its annotation process, and the definition of the dominance classification tasks. Section 2.4 presents the audio and visual cues. Section 2.5 presents our models for estimating dominance and describes the experimental protocol. Sections 2.6 and 2.7 present and discuss the results for the studied dominance classification tasks. Section 2.8 summarizes the chapter and provides some concluding remarks.

2.1 Related work

In the next subsections, we summarize the most relevant work in social psychology and social computing related to our own.

2.1.1 Dominance in social psychology

Dominance is a fundamental construct in social interaction (Burgoon and Dunbar, 2006). In social psychology, dominance is often seen in two ways, "as a personality characteristic (trait) and to indicate a person's hierarchical position within a group (state)" (Schmid Mast, 2002) (pp. 421). Although dominance and closely related terms like power, status, and influence have multiple definitions and are often used as equivalent, many social psychologists advocate for a clearer distinction, power being "the capacity to produce intended effects, and in particular, the ability to influence the behavior of another person" (Dunbar and Burgoon, 2005b) (pp. 208), and dominance being a set of "expressive, relationally based communicative acts by which power is exerted and influence achieved", "one behavioral manifestation of the relational construct of power", and "necessarily manifest" (Dunbar and Burgoon, 2005b) (pp. 208-209).

The study of dominance has spanned several decades of work in psychology and is too large to summarize here [for recent accounts, see (Burgoon and Dunbar, 2006; Dunbar and Burgoon, 2005b)].

However, two main threads of work are key to the development of automated dominance modeling approaches, as both justification and inspiration: the existence of specific social cues used by people to express dominance in conversations, and the ability to correctly infer or perceive dominance by observers of an interaction using such cues.

The first thread of work is rich, and has been widely studied. Both verbal and nonverbal cues are indicators of dominance. Being the primary interest of our work, we focus on nonverbal cues, which are known to be effective in predicting behavioral outcomes. Directly related to our work, nonverbal cue categories of interest include vocalic and kinesic (Dunbar and Burgoon, 2005b). Vocalic cues involve amount of speaking time (or length) (Schmid Mast, 2002), speech loudness (or energy), speech tempo, pitch, vocal control, (Dunbar and Burgoon, 2005b), and interruptions (Brody and Smith-Lovin, 1989). Among these, *speaking activity* as measured by speaking length has been shown to be a particularly robust cue to predict dominance (Schmid Mast, 2002). Kinesic cues include body movement, posture, and elevation, and gestures, facial expressions, and eye gaze (Dunbar and Burgoon, 2005b). In particular, it has been found that, regarding body movement, dominant people are normally more active than non-dominant people (the former move more and with a wider range of motion, the latter tend to be more limited in their amount and range of body activity), and that gestures that accompany speech are positively correlated with dominance (Burgoon and Dunbar, 2006; Dunbar and Burgoon, 2005a). This suggests that *visual activity* (and in particular, activity that correlates with speaking activity) are strong cues for predicting dominance. Also, gaze patterns have been observed to be reliable indicators of visual dominance (Hall *et al.*, 2005). Early research by Efran showed that high-status persons receive more visual attention than low-status people (Efran, 1968). Cook et al. showed that people who very rarely look at others in conversations are perceived as weak (Cook and Smith, 1975). The percentage of eye contact, gaze frequency, gaze duration, 'looking-while-speaking', and 'looking-while-listening' have been shown to be correlated with social verticality (Hall *et al.*, 2005), emphasizing the importance of *visual attention* cues. Exline et al. showed that high-power people exhibit a relatively high ratio of looking-while-speaking to looking-while-listening periods (Exline *et al.*, 1975; Dovidio and Ellyson, 1982).

The second thread of work is also crucial: the fact that people can correctly decode dominance (both as participants of an interaction and as external observers) provides support for the expectation of obtaining reliable human annotations and the promise of designing methods for automatic analy-

sis. The literature here is also rich. Almost three decades ago, Dovidio et al. showed that people can systematically decode patterns of visual dominance displayed by others (Dovidio and Ellyson, 1982). It has been also found that participants and external observers present differences in their perception of dominance (Dunbar and Burgoon, 2005b). For automatic approaches, this is important for manual data annotation (first-party vs. third-party) in order to generate ground-truth for training purposes. As Dunbar and Burgoon state: "Perhaps coders' perception of dominance correspond more closely with objective measures of verbal and nonverbal dominance than those of participants themselves... However, the coders' observations are limited to the behaviors in a particular interaction, whereas participants are privy to the ongoing interaction that is part of a continuing relationship. Thus, as with many other findings, whose perception you trust depends on what question is being asked." (Dunbar and Burgoon, 2005b) (pp. 228). We believe the third-party option to be an adequate approach for the questions addressed in this chapter.

2.1.2 Dominance in social computing

Previous research on automatic dominance modeling can be categorized based on the specific group interaction setting, the addressed task, and the technical implementation, including both cues and dominance models. All of the works discussed below studied small groups recorded with multiple cameras and microphones.

For a debating game setting, Basu et al. (Basu *et al.*, 2001) used the influence model (IM) - an unsupervised DBN that models a group as a set of Markov chains, each of which influences the others' state transitions - to determine the degree of influence a person has on the others on a pair-wise basis. Both vocalic cues (manually labeled speaker turns and automatically extracted speaker energy and voicing information) and kinesic cues (region-based motion energy derived from pre-defined regions and skin-color blobs) were used. While promising results were presented, this work neither studied the impact of individual features nor evaluated the performance of the resulting system in a systematic way.

On a small set of meetings from the M4 (MultiModal Meeting Manager) and AMI (Augmented Multi-party Interaction) corpora, Rienks et al. (Rienks and Heylen, 2005) studied a supervised approach based on Support Vector Machines (SVMs). The addressed task was three-way classification of the participants' dominance level (high, normal, low). Audio-only features derived from manually

annotated data were used, and included a combination of nonverbal (e.g. speaker turns, speaking length, floor grabs) and verbal cues (e.g. number of spoken words). In this work, no study of the annotation quality was conducted, and so a clear understanding of the sources of complexity of the data was missing. Furthermore, labeling the data with a predefined number of dominance levels is, to some extent, arbitrary, and a study of the effect of these choices was not conducted. Rienks et al. (Rienks *et al.*, 2006) later extended this approach to a subset of the AMI corpus where the dominance judgments came from the participants themselves.

In a third research line, Otsuka et al. proposed, following the ideas of (Basu *et al.*, 2001), to quantify pair-wise influence from automatically estimated vocalic and kinesic mid-level cues (speaking-turn and gaze patterns, respectively), computed in turn with a complex DBN that integrates low-level features (Otsuka *et al.*, 2006). While the proposed influence model is simple, and the proposed features are conceptually appealing, neither an objective evaluation nor a comparison to previous approaches were conducted in this work.

Our work substantially extends previous research in several ways. First, unlike (Basu *et al.*, 2001; Otsuka *et al.*, 2006) , we conduct a systematic study of both vocalic and kinesic features and dominance models on a common data set, and present a detailed objective evaluation of the performance of single- and multi-modal cues, and of unsupervised and supervised learning approaches. Second, the specific research tasks we study are distinct, and so complementary, to the ones studied in all previous work. Third, unlike (Rienks *et al.*, 2006; Rienks and Heylen, 2005) we introduce a set of novel visual activity cues, distinct from those in (Basu *et al.*, 2001; Otsuka *et al.*, 2006) and computed in the compressed domain with low computational cost. Fourth, unlike (Otsuka *et al.*, 2006), we systematically evaluate several visual attention cues for estimating dominance. Fifth, unlike (Basu *et al.*, 2001; Rienks *et al.*, 2006; Rienks and Heylen, 2005) , we rely on fully automatically extracted features, and in this sense the presented work is closer to 'what is truly achievable using machines'. Finally, unlike all previous work, we analyze the annotation of perceived dominance by multiple human judges and are thus able to analyze the implications that the variation of human perception has on the performance of our automatic approaches.

2.2 Our approach

Figure 2.1 shows a block diagram of the structure of our work:

- **(a,b): Section 2.3.1.** We use meeting data from the publicly available AMI corpus (Carletta et al., 2006), where multiple microphones and video cameras have been used to capture audio and video.
- **(d): Sections 2.3.2, 2.3.3.** We generated a detailed ground truth annotation of the perceived dominance for each individual in the meetings using multiple human judgments. Through a study of the annotator levels of agreement, we define two sub-tasks to observe the effect on the performance of the dominance models when increased variability in the perception of dominance was present.
- **(c): Section 2.4.** From the raw audio and video data, we derive features which are used to characterize certain nonverbal behaviors. Both the audio and video features have been treated similarly for comparison of the two modalities.
- **(e-f): Section 2.5.** Two models were considered for estimating dominance; one unsupervised and one supervised. The supervised approach was used for single as well as multi-modal fusion, which allowed us to study the contributions of the audio and video cues to the dominance estimation performance. We evaluated the performance of the models using both hard and soft evaluation criteria, where the latter accounted for the amount of variability in the annotations.

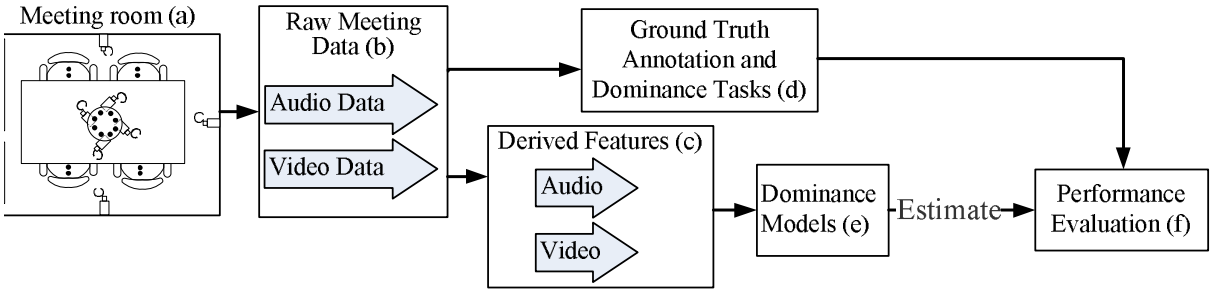


Figure 2.1. Flow diagram of our approach.

In summary, our work studies both the underlying variability in *perceived dominance by external observers*, and systematically analyzes the objective performance of single and multi-modal dominance estimation models for a number of classification tasks.

2.3 Meeting data and dominance tasks

Various corpora have been collected with the explicit goal of studying group interaction (Gatica-Perez, 2009). We chose meetings from the AMI corpus (Carletta et al., 2006) for our study. The AMI corpus is publicly available with group interactions that were task-oriented and not scripted. They were recorded in 'smart-rooms' equipped with audio-visual sensors. These meeting recordings suited our dominance study with external observers. We describe the AMI dataset in detail and the annotations thereafter.

2.3.1 Meeting data

The AMI meetings were carried out in the meeting room shown in Figure 2.2. The room contains a table, a slide screen, and a white board. A circular microphone array containing eight evenly distributed microphones is set in the middle of the table, and one with four microphones is set at the ceiling. Participants were also asked to wear both headset and lapel omni-directional microphones, which were attached via long cables to enable freedom of movement around the room. Three cameras were mounted on the sides and back of the room to capture mid-range and global views, respectively, while 4 additional cameras mounted on the table captured individual visual activity only, as shown in Figure 2.3.

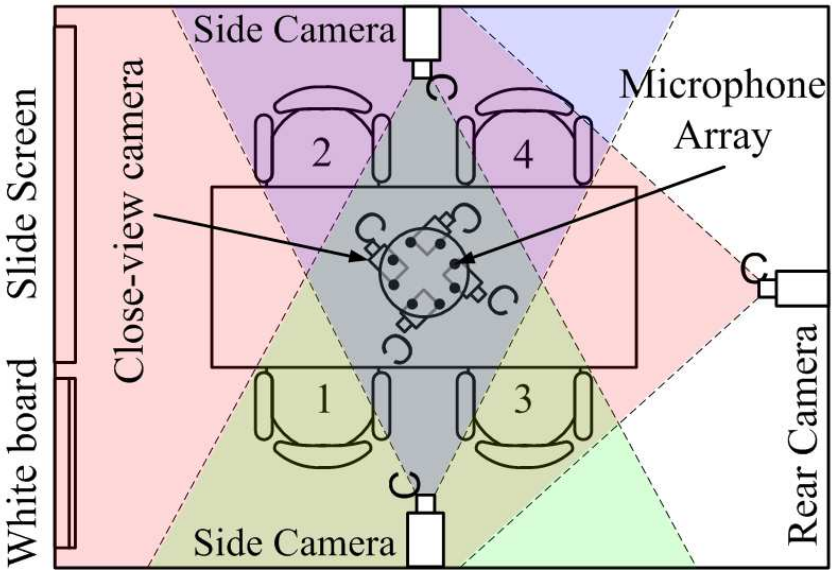


Figure 2.2. Plan view of the meeting room set up.

Figure 2.3. Examples of the seven camera views available in the AMI meeting room. The top row shows the right, centre and left cameras, while the bottom row shows the view from each of the close up cameras.

From the AMI data, a subset of five teams of participants were selected for our meeting data. Each team consisted of 4 participants, who were given the task of designing a remote control over a series of meeting sessions. The level of previous acquaintance among team members varied from being completely unacquainted to knowing each other well. Each participant was assigned distinct roles: 'Project Manager', 'User Interface specialist', 'Marketing Expert', and 'Industrial Designer'. During each session, the team was required to carry out certain tasks, such as a presentation on particular subjects related to the task, or a discussion about a particular aspect of the task. To encourage natural behavior, the meetings were not scripted and the teams met over several sessions so that they achieved the common goal.

2.3.2 Annotating the data

From the AMI data, 11 meeting *sessions* varying from 15 to 35 minutes were divided into five-minute segments for ground truth annotation so that a total of 59 meeting segments were used. The segments were chosen to be 5 minutes long, rather than the original full meetings, since this provided more data points for training and testing. There is also evidence that people often need a relatively small amount of time to make accurate judgments about the behavior of others (Ambady *et al.*, 2000). Our choice is therefore supported by this empirical evidence.

A total of 21 annotators were used and were split into groups of 3 so that each group always annotated the same segments. The annotators were not professional coders or experts in psychology.

They were shown a video with views from the side and rear cameras, which are shown in the top row of Figure 2.3 and listened to the audio also. As the annotators understood the spoken language which is English in the AMI scenario, they had access to the verbal channel as well. For a given meeting, each annotator viewed only one five-minute segment (in other words, an annotator never judged more than one segment of the same session). The annotators were requested to judge a person's dominance based only on the evidence within each meeting. Importantly, annotators were given neither a prior definition of dominance, nor were told what specific verbal or nonverbal cues to look for in order to make their judgments. The annotators were compensated monetarily for their effort. As hiring annotators was hard and costly, we chose to annotate every meeting using 3 annotators and not more.

For each 5 minute meeting segment (simply called *meeting* from here on for convenience), annotators were asked to rank the participants, from 1 (highest) to 4 (lowest), according to their level of perceived dominance. As well as an absolute ranking, annotators were also asked to rank people proportionately by distributing a total of 10 units among the participants, where more units signified higher dominance. To identify segments where the rankings were difficult to allocate, annotators were asked about their confidence in their absolute and proportionate rankings on a seven-point scale. Annotators were also requested to ascertain specific characteristics of each participant such as their degree of activity, timidness, and talkativeness, also on a seven-point scale (Dunbar and Burgoon, 2005b). Finally, they were requested, on completion of the annotations, to provide a free form written description of the personal criteria they used to decode dominance.

2.3.3 Analysis of the annotations

From the human annotations, we wished to discover whether there was significant inter-annotator agreement across all meetings. Initial analysis of the meeting data indicated that 12 out of 59 meetings showed full agreement for all 4 absolute rankings of each meeting. This was clearly not enough for an analysis of dominant behavior for our experiments. Therefore we decided to relax the agreement condition by considering only the task of estimating the *most* dominant or the *least* dominant person. A significant number of the meeting segments (34) showed full agreement of the most dominant person, i.e. all three annotators agreed on the most dominant participant. Furthermore, the corresponding self-reported average confidence for the annotation for these meetings was 1.7 (where 1 represents the highest confidence and 7 represents the lowest). This subset represents almost 3 hours of meeting

data where the agreement and confidence of the annotators was high. An additional observation of interest is that in 24 out of 34 cases, the most dominant person who was chosen by the annotators played the 'project manager' role.

We conducted further analysis and found that in 57 meetings where atleast two out of the three annotators agreed on the most dominant person. These values and the corresponding average self-reported confidence levels are shown in Table 2.1. This subset contains a larger intrinsic variation in the perceived dominance by human judges.

Finally, a similar analysis showed that there were 31 meetings with full agreement of the least dominant person, and 54 meetings where atleast two out of the three annotators agreed on the least dominant person. Similar to the most dominant case, the confidence decreases as the variability of the data-sets increases (see Table 2.1). It is interesting to note that the confidence in the annotation of the least dominant person was always less than that of the corresponding experiment in the most dominant case. Also, the decrease in confidence as the variability of the data set increased was greater for the least dominant case compared to the most dominant case. We speculate that the behavior of less dominant people might be more difficult to observe since they tend to speak and move less than dominant people (Dunbar and Burgoon, 2005b).

Following the analysis of the annotations, we decided to define a number of dominance classification tasks, one for each of the different subsets discussed above. These are summarized in Table 2.1 below. Within each dominance task there are two sub-tasks that correspond to meetings where there is (i) Full agreement among annotators who labeled the same meeting (denoted Full in the following), and (ii) Majority where at least 2 out of the 3 annotators agreed (denoted Maj in the following).

Dominance Estimation Task	**Sub-Tasks**	**Average Annotator Confidence**	**Number of Meetings**	**Proportion of Total Meetings (%)**
Most	Full-agreement	1.74	34	57.6
	Majority-agreement	1.85	57	96.6
Least	Full-agreement	2.11	31	52.5
	Majority-agreement	2.4	54	91.5

Table 2.1. Dominance tasks and corresponding data-sets.

2.4 Audio and visual nonverbal cues for dominance modeling

In order to measure the dominant behavior of people in meetings, we followed the social psychology literature and hypothesized that activity and attention levels are correlated with dominance. Here we chose to represent activity in terms of audio and visual cues; and visual attention using head-pose direction cues. From the audio sources, we adapted existing analysis techniques to characterize the speaking activity of the meeting participants. From the video data, compressed-domain features were extracted from multiple cameras to characterize visual activity and head-pose was extracted to characterize visual attention. More details are described in the following subsections.

2.4.1 Audio cues

Audio cues were extracted from the four close-talk microphones attached to each participant (one per person). Firstly we considered time-varying aspects of the speech.

Speaking Energy: The starting point for audio feature extraction is to compute a speaker energy value for each participant, using a sliding window at each time step as described in (Zhang *et al.*, 2006). Speaking energy was extracted using the root mean square amplitude of the audio signal over a sliding time window for each audio track. A window of 40 ms with a 10 ms time shift was used. For our experiments, the final signal was sub-sampled to a frame rate of 5 frames per second.

Speaking Status: From the speaking energy, a binary variable was computed by thresholding the speaker energy values. This indicates the speaking or non-speaking status of each participant at each time step. The discrepancy between the automatic and the manual segmentation is 4% in terms of frames.

Then we considered features accumulated from the entire conversation. These features provided a simple way of quantifying the relative opportunities that participants had to speak. The following list summarizes the features used for our study.

- **Total Speaking Energy (TSE):** Speaker energy accumulated over the entire meeting. This feature follows the findings in psychology that establish that speaker energy is a manifestation of dominant behavior (Dunbar and Burgoon, 2005b). It is to be noted that the TSE feature captures how much a participant speaks as well as how loud he speaks, and not just how loud he speaks.
- **Total Speaking Length (TSL):** This feature considers the total time that a person speaks accord-

ing to their binary speaking status (Schmid Mast, 2002).

– **Total Speaking Turns (TST):** A speaking turn is the time interval for which a person's speaking status is active. The total number of speaker turns was accumulated over the entire meeting for each participant.

– **Total Speaking Turns without Short Utterances (TSTwoSU):** This is a variation of the TST feature, computed as the cumulative number of turns that a speaker takes, such that the speaker turn duration is longer than one second. The goal is to retain only those turns that are most likely to correspond to 'real' turns, eliminating all short utterances that are likely to be back-channels or other utterances with no content (coughing etc).
– **Average Turn Duration (AvTDur):** This is the ratio of TSL and TST, which is the average duration of the speaker's turns.
– **Total Successful Interruptions (TSI):** This feature encodes the hypothesis that dominant people interrupt others more often (Brody and Smith-Lovin, 1989). The feature is defined by the cumulative number of times that speaker $i \in \{1,2,3,4\}$ starts talking while another speaker $j \in \{l : l \neq i\}$ speaks, and speaker j finishes his turn before i does, i.e. only interruptions that are successful are counted. Though such a definition does not perfectly capture successful interruptions, nevertheless it is a computationally efficient proxy.
– **Total Unsuccessful Interruptions (TUI):** This feature encodes the hypothesis that dominant people do not let others interrupt more often. The feature is defined by the cumulative number of times that while speaker $i \in \{1,2,3,4\}$ is speaking, another speaker $j \in \{l : l \neq i\}$ speaks, and speaker j finishes his turn before i does, i.e. only interruptions that are unsuccessful by another participant are counted.
– **Total Short Unsuccessful Interruptions (TSUI):** This feature encodes the hypothesis that dominant people get backchanneled more often. The feature is defined by the cumulative number of times that while speaker $i \in \{1,2,3,4\}$ is speaking, another speaker $j \in \{l : l \neq i\}$ speaks, and speaker j finishes his turn (which is one second or less) before i does, i.e. only backchannels (or short utterances) by another participant are counted. Again, similar to the interruptions, this definition of backchannels may not correspond perfectly to real backchannels, but again is a good proxy.

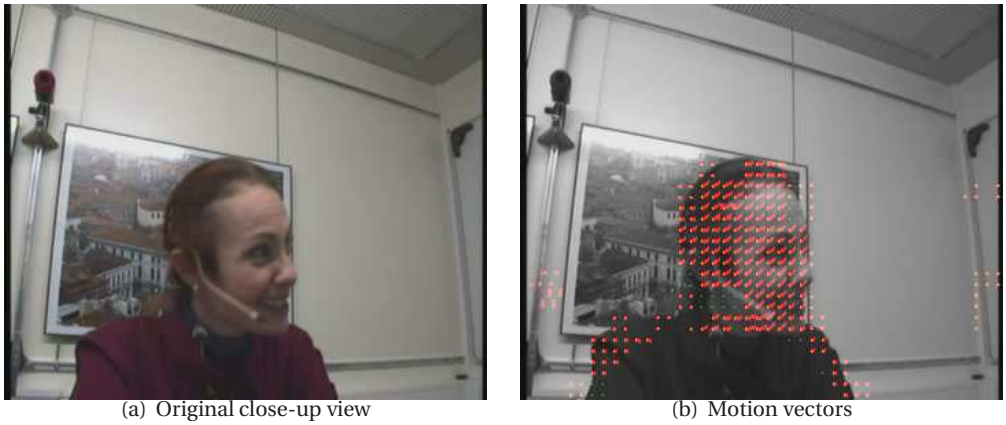
2.4.2 Visual activity cues

In order to capture visual motion activity efficiently, we leverage the fact that meeting videos are already in compressed form to extract visual activity features at a much lower computational cost. These features are generated from compressed-domain information such as motion vectors and block discrete-cosine transform (DCT) coefficients that are accessible at almost zero cost from compressed video (Wang *et al.*, 2003; Yeo and Ramchandran, 2008). In our data set, there is a camera taking a close-up shot of each participant, as shown in the bottom row of Figure 2.3. Each of these video streams has already been compressed by a MPEG-4 encoder with a group-of-picture (GOP) size of 250 frames and a GOP structure of I-P-P-..., where the first frame in the GOP is Intra-coded (I), and the rest of the frames are predicted frames (P) (Coimbra and Davies, 2005).

Figure 2.4 summarizes the various compressed domain features which can be extracted cheaply from compressed video. In particular, we consider the use of the *motion vector magnitude* [Figure 2.4(b)] and the *residual coding bitrate* [Figure 2.4(c)] to estimate visual activity level. Motion vectors, illustrated in Figure 2.4(b), are generated from motion compensation during video encoding; for each source block that is encoded in a predictive fashion, its motion vectors indicate which predictor block from the reference frame (in this case the previous frame for our compressed video data) is to be used. Typically, a predictor block is highly correlated with the source block and hence similar to the block to be encoded. Therefore, motion vectors are usually a good approximation of optical flow, which in turn is a proxy for the underlying motion of objects in the video (Coimbra and Davies, 2005). We use the *motion vector magnitude* as one measure of visual activity in this work.

After motion compensation, the DCT coefficients of the residual signal, which is the difference between the block to be encoded and its prediction from the reference frame, are quantized and entropy-coded. The *residual coding bitrate*, illustrated in Figure 2.4(c), is the number of bits used to encode this transformed residual signal. While the motion vector captures gross block translation, it fails to fully account for non-rigid motion such as lips moving. On the other hand, the residual coding bitrate is able to capture finer motion, since a temporal change that is not well modeled by the block translational model will result in a residual with higher energy, and hence will require more bits to code it. In combination with the motion vector magnitude, the residual coding bitrate provides complementary evidence for visual activity.

For each meeting participant, we detect when they are in view. To do this, we implement a Gaussian



(a) Original close-up view (b) Motion vectors

(c) Residual coding bitrate (d) Detected skin blocks

Figure 2.4. Illustration of compressed domain features. (a) Shows the original image. (b) Shows the direction of motion vectors. (c) Shows the residual coding bitrate at different pixel locations (red means high magnitude). (d) Shows the locations where skin color was detected in red color.

Mixture Model (GMM) based on skin-color block detector (McKenna *et al.*, 1998) that can detect face and hand regions. This works in the compressed domain with chrominance DCT DC coefficients and motion vector information, and produces detected *skin-color blocks* such as in Figure 2.4(d). We then threshold the number of skin-colored blocks in the close-up view to detect when a participant is seated. If a participant is not detected in a frame of the close-up view, he is assumed to be presenting at the projection screen, which is a reasonable assumption in the meeting data. We also assume that a person who is presenting is by default visually active.

If the participant is visible in the close-up view, we measure his visual activity by using either or both of motion vector magnitude and residual coding bitrate. To meaningfully compare motion vector magnitudes and residual coding bitrate, we normalize the quantities. Consider computing a normalized visual activity from motion vector magnitude for participant i in frame t. We first calculate the average motion vector magnitude, $v_{i,t}$, over all blocks in each frame. For each participant in each meeting, we find the median of the average motion vector magnitude, $\tilde{v}_i$, over all frames where the participant is in the close-up view. We also compute the average of the medians, $\bar{v}$, of all the participants. Normalization is then performed where the visual activity level for participant i in frame t, $V_{i,t}^M$ using motion vector, is computed by normalizing as follows:

$$V_{i,t}^M = \begin{cases} \frac{v_{i,t}}{2\bar{v}} & v_{i,t} < 2\bar{v} \\ 1 & v_{i,t} \geq 2\bar{v} \end{cases} \tag{2.1}$$

The visual activity level from the residual coding bitrate, $V_{i,t}^R$, is also normalized in a similar fashion.

We use the average of visual activity from motion vector magnitude, $V_{i,t}^M$, and from residual coding bitrate, $V_{i,t}^R$, as another estimate of visual activity. This allows us to approximate both rigid and non-rigid local motion. The combined estimate of visual activity for the participant i in frame t, $V_{i,t}^C$, is given by:

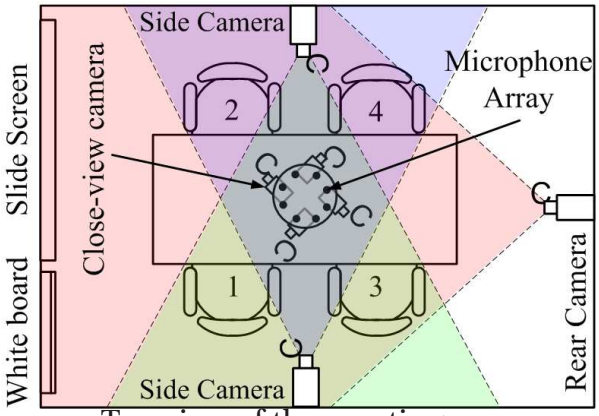
$$V_{i,t}^C = \frac{1}{2}\left(V_{i,t}^M + V_{i,t}^R\right) \tag{2.2}$$

After raw visual activity extraction in order to facilitate the comparison between audio and visual cues, visual cues are derived in an analogous fashion to those for audio cues as described in Section 2.4.1. More specifically, the following cues were derived from the raw motion activity values:

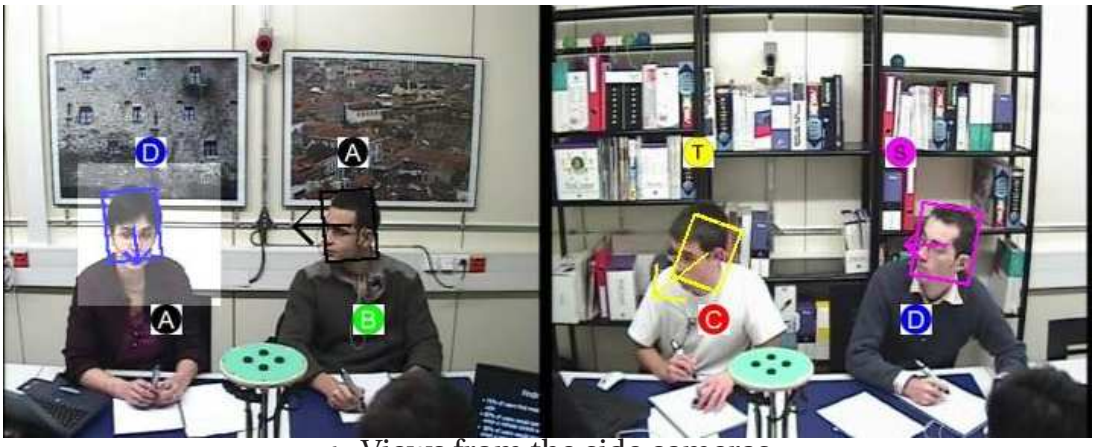
- **Visual Activity.** A binary variable computed from compressed-domain video that indicates whether a participant is visually active or inactive at each time step (extracted at 25 frames per second). Three variations were tested, based on Motion Vectors (called Vector in the following discussion), Residual Coding Bitrate (Residue), and the average of both features (Combo).
- **Total Visual Activity Length (TVL).** The accumulated motion activity for a person can be of three types, depending on whether it is estimated from the motion vectors, the residual coding bitrate, or their combination.
- **Total Visual Activity Turns (TVT).** This feature quantifies the number of times someone is continuously moving without breaks. This is analoguous to the total speaking turns feature defined in Subsection 2.4.1.
- **Total Visual Activity Interruptions (TVI).** This captures when one person starts and remains visually active while another stops. While there may not be a meaningful notion of visual activity interruption in daily life, our hypothesis is that visual activity is correlated with speech activity such that speaker interruptions might be reflected in TVI as well. It is similar to the TSI feature defined in Subsection 2.4.1.

2.4.3 Visual attention cues

In our work, head pose is used to infer visual attention. We apply the work by Ba and Odobez (Ba and Odobez, 2010) to estimate the joint focus state of all participants. Visual attention is estimated using a DBN, by modeling the relationship between people's visual attention, their head pose, their speaking status, and other contextual cues related to the group activity. These contexual cues include slide-screen activity and conversational events like silence or monologue or dialogue or discussion. Head pose was estimated by jointly tracking the head and head-pose using side-view cameras (as illustrated in Figure 2.5(b)). There were seven visual attention targets defined in the AMI meetings, i.e. the four participants, the slide-screen, the white-board, the table and one unfocused label. The accuracy of automatic Visual Focus Of Attention (VFOA) estimation reported in (Ba and Odobez, 2010) was around 52%. Note that unlike other settings (Otsuka *et al.*, 2006), as the AMI meetings had objects that distract the visual attention of participants like the slide-screen, the white-board, and the table, and so the task of VFOA estimation was more difficult. The seating arrangement was also not circular as in (Otsuka *et al.*, 2006), rather it was rectangular with 2 people facing each other, making the VFOA



(a) Top view of the meeting room



(b) Views from the side cameras

Figure 2.5. (a). Shows the top view of the meeting room. (b). Shows the side camera views and the estimated visual focus of participants using side-view camera views. Each of the participants is labeled and their focus of attention is displayed above their head (T stands for Table and S stands for Slide-screen). Colored rectangle around the head shows the head location and colored arrows shows the head pose of each of the participants. The white transparent box placed on participant A shows that her speaking status is 'true'.

estimation when the focus of attention is certain seat positions (seats numbered 1 and 2 in Figure 2.5(a)) more difficult than the others.

From the visual attention of individual participants, along with the speech activity cues, we computed a number of features that capture the gazing behavior of participants as follows:

Overall attention cues

– **Total Received Visual Attention (TRVA):** This feature encodes the hypothesis that dominant or high status people are looked at longer (Efran, 1968). The feature is defined by the cumulative

number of frames that a participant i is looked at by the other participants (regardless of their identity).

- **Total Looking-At-Others Length (TLOL):** This follows the hypothesis that dominant or high status people look at others longer. The feature is defined by the cumulative number of frames that a participant i looks at other participants (regardless of their identity).
- **Total Looking-At-Others Turns (TLOT):** This follows the hypothesis that dominant or high status people look at others more often, by inverting the hypothesis of Cook et al, that weak people rarely look at others (Cook and Smith, 1975). The feature is defined by the cumulative number of times a participant i looks at other participants (regardless of their identity).

While-Speaking attention cues

These three cues follow the three cues above, computed only when the participants speak.

- **Total Received Visual Attention while speaking (TRVAwS):** This feature follows the hypothesis that dominant or high status are looked at longer while speaking. The feature is defined by the cumulative number of frames that a participant i is looked at by the other participants while speaking (regardless of their identity).
- **Total Looking-At-Others Length while speaking (TLOLwS):** This feature follows the hypothesis that dominant or high status people look at others longer while speaking. (Exline *et al.*, 1975). The feature is defined by the cumulative number of frames that a participant i looks at other participants while speaking (regardless of their identity).
- **Total Looking-At-Others Turns while speaking (TLOTwS):** This follows the hypothesis that dominant or high status people look at others more often while speaking. The feature is defined by the cumulative number of times a participant i looks at other participants while speaking (regardless of their identity).

While-not-Speaking attention cues

These three cues follow the three cues above, computed instead when the participants do not speak.

- **Total Received Visual Attention while not speaking (TRVAwNS):** This feature follows the hypothesis that dominant or high status are looked at longer while not speaking. The feature is

defined by the cumulative number of frames that a participant i is looked at by the other participants while not speaking (regardless of their identity).

– **Total Looking-At-Others Length while not speaking (TLOLwNS):** This feature follows the hypothesis that dominant or high status people look at others longer while not speaking. (Exline *et al.*, 1975). The feature is defined by the cumulative number of frames that a participant i looks at other participants while not speaking (regardless of their identity).

– **Total Looking-At-Others Turns while not speaking (TLOTwNS):** This follows the hypothesis that dominant or high status people look at others more often while not speaking. The feature is defined by the cumulative number of times a participant i looks at other participants while not speaking (regardless of their identity).

Visual Dominance Ratio

The Visual Dominance Ratio (VDR) was defined in (Dovidio and Ellyson, 1982) as the ratio between the total looking-while-speaking periods to the total looking-while-listening periods for dyadic interactions. We generalize it to multi-party conversations, by approximating 'looking while listening' as 'looking while someone else is speaking' and 'looking while not speaking' and hence define the following two ratios. The new ratios are called Multi-Party Visual Dominance Ratios (MVDR) (Hung *et al.*, 2008b).

– $MVDR_1$**:** Defined as the following ratio

$$MVDR_1 = \frac{\text{Total Looking at others -while- speaking}}{\text{Total Looking at others-while-someone-else-speaks}} \tag{2.3}$$

– $MVDR_2$**:** Defined as the following ratio

$$MVDR_2 = \frac{\text{Total Looking at others -while- speaking}}{\text{Total Looking at others-while-not-speaking}} \tag{2.4}$$

Table 2.2 provides a summary of all the audio and video cues and their associated acronyms.

Glossary of Feature Acronyms	
'Audio Activity'	
Total Speaking Energy	TSE
Total Speaking Length	TSL
Total Speaking Turns	TST
Total Speaking Turns without Short Utterances	TSTwoSU
Average Turn Duration	AvTDur
Total Successful Interruptions	TSI
Total Unsuccessful Interruptions	TUI
Total Short Unsuccessful Interruptions	TSUI
'Visual Activity'	
Total Motion Length	TVL
Total Motion Turns	TVT
Total Motion Interruptions	TVI
'Visual Attention'	
Total Received Visual Attention	TRVA
Total Looking At Others Length	TLOL
Total Looking At Others Turns	TLOT
Total Received Visual Attention while speaking	TRVAwS
Total Looking At Others Length while speaking	TLOLwS
Total Looking At Others Turns while speaking	TLOTwS
Total Received Visual Attention while not speaking	TRVAwNS
Total Looking At Others Length while not speaking	TLOLwNS
Total Looking At Others Turns while not speaking	TLOTwNS
Multi-Party Visual Dominance Ratios	$MVDR_1$ and $MVDR_2$

Table 2.2. Glossary of feature abbreviations

2.5 Models for dominance estimation

In this work, we use a simple unsupervised model and a supervised model based on SVMs (Burges, 1998) as prototypical models for dominance estimation. Our goal was to understand the relative predictive power of single cues for the dominance estimation task using the unsupervised model, and to explore whether cue fusion, using an SVM, could be useful. Though we experimented with other models, including a Gaussian Naive Bayes Classifier and a logistic regression classifier (Mitchell, 1997), we report the results using the Support Vector Machine for brevity reasons. Also, the cue fusion results using the different models were comparable.

2.5.1 Unsupervised model

In this model, audio or visual cues are accumulated over the duration of the meeting. The model is rule-based and computes either the largest or smallest accumulated value of each feature, depending

on whether we are estimating the most or least dominant person, respectively. That is, we hypothesize that someone is likely to be the most dominant if they speak, move, or grab the floor the most out of all the participants in the meeting. While this model is simple, it showed promising performance in our preliminary work (Hung *et al.*, 2007). Similarly, we use the smallest accumulated value of the feature to identify the least dominant person in the meeting. We evaluate the model by comparing the label of the person who is estimated automatically with that of the ground truth annotated data.

2.5.2 Supervised model

We also use a supervised method to investigate both single and multi-modal cue fusion. This allowed us to observe which cues were complementary or correlated more closely, and led to interesting findings about the comparative importance of the activity cues for robust dominance estimation. In order to make the cues comparable across meetings, we normalized them before fusion. The supervised approach uses a two-class SVM classifier to discriminate between the 'most' and 'non-most' dominant participants in each meeting. A second two-class SVM is trained to discriminate between the 'least' and 'non-least' dominant person. A linear kernel was employed for both experiments. For each task, the SVM score produced for each person's features are ranked. The rankings are then used to determine which participant is assigned the most (resp. least) dominant person label, by considering the point which is furthest from (resp. closest to) the class boundary. This procedure generates exactly one most (resp. least) dominant person per meeting. Note that as stated in Section 2.2, this is different from the work in (Rienks and Heylen, 2005; Rienks *et al.*, 2006) where each person independently was labeled as 'high', 'middle' or 'low' in terms of dominance level. The model was evaluated using a leave-one-out approach for each combination of input features.

2.5.3 Experimental protocol

Figure 2.6 shows a summary of the experiments that we carried out. As shown in Figure 2.6(a), the experiments were split into two tasks: the estimation of the most dominant and the least dominant person.

For each of the tasks, we considered the set of experimental conditions illustrated in Figure 2.6 (b-c). Firstly, we considered each modality separately for both the supervised and unsupervised ap-

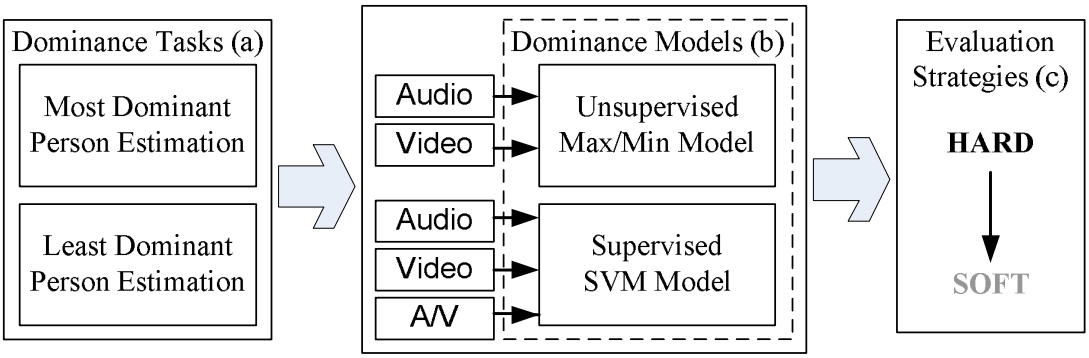


Figure 2.6. Flow diagram showing our experimental protocol.

proaches. The supervised approach also allowed us to compare the performance of audio-visual feature fusion with combining features from the same modality. For each dominance task, we also considered different evaluation criteria, which accounted for increasing variability in the ground truth annotations, where hard (EvH) or soft (EvS) scoring criteria were used [Figure 2.6 (c)]. The criteria themselves are explained in more detail in subsections 2.6.1 and 2.6.2. For each of the two dominance tasks that we investigated, we consider two sub-tasks; full and majority agreement, as illustrated in Table 2.1. It is important to note that for each model and evaluation criterion, the overall performance is calculated based on the estimation for each meeting rather than for each participant. The results are reported as classification accuracies, and discussions regarding the statistical significance of the results are summarized in Section 2.8. In order to compute statistical significance we have used the standard binomial test throughout the thesis.

2.6 Classifying the Most-Dominant person

2.6.1 Full-agreement data set

For this dataset, computing the classification accuracy is straightforward, which is the percentage of meetings where the estimated most-dominant person and the ground-truth matches and the hard and soft evaluation criteria become equal.

Audio cues

Table 2.3 shows the results obtained using audio cues. Using the unsupervised model with single features, the total speaking length (TSL) was most effective at 85.3% classification accuracy. This re-

sult is important not only because of the simplicity of this automated technique but also because it confirms the findings in social psychology (Schmid Mast, 2002; Dunbar and Burgoon, 2005b) about speaking time being a strong cue for dominance perception by humans. The total speaking energy (TSE) also performed well (with an accuracy of 82.4%). While the total number of speaking turns (TST) did not perform as well, removing short utterances (TSTwoSU), performed as well as TSE. Finally, while the total number of successful interruptions (TSI) did not perform as well, Total Short Unsuccessful Interruptions (TSUI) did perform well (with 76.5% classification accuracy). This result could suggest that dominant people do not interrupt much, but receive significant feedback (e.g. backchannels) in a cooperative scenario like the AMI meetings. All these audio cues performed significantly better than chance (which would result in 25% classification accuracy).

Dominance Model	Features	Class. Acc.(%)
Unsupervised	TSL	**85.3**
	TSE	82.4
	TST	61.8
	TSTwoSU	82.4
	AvTDur	73.5
	TSI	64.7
	TUI	70.6
	TSUI	76.5
Supervised	TSE, TST	88.2
	TSL, TSE, TST	88.2
	TSL, TST, TSI	88.2
	TSE, TST, TSI	**91.2**
Random Guess	None	25.0

Table 2.3. Performance of **Audio** cues for **Most**-dominant person with **Full**-agreement data.

A selection of the results with the supervised model using multi-dimensional audio cues is also shown in Table 2.3.

A closer look at the meetings where TSL or TSE failed indicated that in some cases speaking turns or successful interruptions predicted the most dominant person correctly. This suggested that using the features jointly might improve performance. In practice, fusing these features in the supervised learning setup proved beneficial. We observe that although TST is not very discriminative as a single feature, it helps when combined with TSE alone or with TSE and TSL, yielding a 3% accuracy improvement. The best feature combination (TSE, TST, TSI) yield an absolute performance improvement of 6% with respect to the performance obtained with TSL, with 91.2% accuracy.

A direct comparison of these results with the existing literature on automatic dominance detection is not possible as the addressed tasks , the data sets, and the experimental protocols used in each case are different. However, a few observations are still pertinent. First, both our results and (Rienks *et al.*, 2006) suggest that benefits can be obtained with audio fusion. Second, both speaking length and number of turns appear in our work and in (Rienks *et al.*, 2006) as part of the best performing feature combinations, an important difference being that, unlike (Rienks *et al.*, 2006), in our case all features are fully automatic. Third, the best performance figure obtained for our two-class task (around 90%) is considerably higher than the best reported performance obtained for the three-class problem in (Rienks *et al.*, 2006) (around 70%). Hypothetical reasons for this include the larger number of classes but also the fact that the data in (Rienks *et al.*, 2006) was not separated using any knowledge about the variability in perceived dominance. We study the case of variability in the human judgments in Section 2.6.2.

Visual activity cues

Table 2.4 shows the results obtained with visual cues. Regarding single cues in the unsupervised setting, the total visual activity length (TVL), which quantifies how much people move, is consistently the best visual feature (76.5% accuracy), and seems to be the most robust. Motion turns (TVT) quantify how often people move. In practice, we observe that these features are generally 'noisy', presenting spikes of very short duration. However, removing short turns and leaving only those that should correspond to *intentional* motion (and that likely correspond to conversational activity too) results in the same performance as TVL. This is an interesting finding that also seems to be supported by evidence in social psychology (Burgoon and Dunbar, 2006). It was interesting to observe that, for TVL and TVT, the residual bitrate option performed slightly better than using the motion vectors; for TVT, the combination worked the best. The motion vector and residue cues capture different information. The former, being derived from block motion compensation in video compression, is better at capturing translational motion. The latter is related to the amount of non-rigid motion in the close-view cameras, including finer visual activity that is usually not captured by motion vectors. In contrast, TVI is not an effective cue: the results indicate that the notion of visual activity interruption (i.e., overlap) does not hold for video as clearly as it does for audio. As with audio cues, all the results with single video cues are considerably better than a random guess.

Compared to single audio cues, the best results with single visual cues degrade by 8.8% (76.5% vs. 85.3%). This is interesting since from the free-form verbal descriptions of how annotators perceived dominance, we found that about half of them mentioned the use of how much a person talks. In addition, annotators mentioned audio or language-based cues more often than those related to visual activity. Despite this, it is remarkable that without using the audio at all, the most dominant person can still be correctly estimated in more than 75% of the cases with easily computable visual cues. Furthermore, it is interesting to note that the use of compressed-domain cues, as compared with similar visual activity cues extracted in the pixel domain, did not lead to any classification performance loss (for more details, please refer to (Yeo and Ramchandran, 2008)). Also note that TVL performed better than some single audio cues. To illustrate the dependencies between audio and visual activity cues, Figure 2.7(a) plots the values of TSL and TVL for all meetings in the full-agreement data set. The red crosses correspond to the positive examples (most-dominant) and the black circles to the negative ones. The figure indicates that there is a significant degree of correlation (correlation coefficient $= 0.58$ with $p < 0.01$)between the visual activity and speaking activity, but that the discrimination seems to be higher for the audio case.

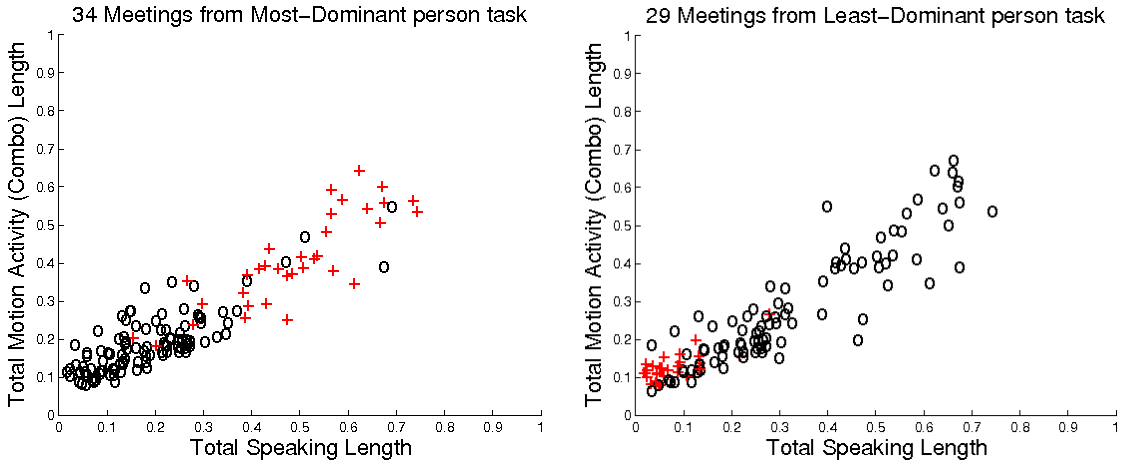
For the multiple feature case, a small selection of the best performing combinations is also shown in Table 2.4. The combination of the two best performing single features (TVL and TVT) did not improve performance over the single cues. We find that cue fusion was not very useful for the visual activity cues, implying that the cues were not that complementary. Overall, the best achieved performance with visual cues and supervised learning is 14.7% worse than the corresponding best performance for audio cues (76.5% vs. 91.2%), as can be seen by comparing Tables 2.3 and 2.4.

Visual attention cues

Table 2.5 shows the results obtained with visual attention cues. Regarding single cues in the unsupervised setting, Multi-Party Visual Dominance Ratios (MVDR) was the best cue (with an accuracy of 79.4%), reaffirming why the cue is popular in social psychology literature as an estimator of dominance (Hall *et al.*, 2005). The numerator of the MVDR, Total Looking-At-Others Length while speaking (TLOLwS), and the Total Received Visual Attention while not speaking (TRVAwNS) also performed well (with an accuracy of 70.6%), showing that dominant people look at others longer while speaking and they get attention while they are not speaking. Total Received Visual Attention (TRVA) was the fourth

Dominance Model	Features	Class. Acc.(%)
Unsupervised	TVL (Vector)	73.5
	TVL (Residue)	**76.5**
	TVL (Combo)	73.5
	TVT (Vector)	67.6
	TVT (Residue)	70.6
	TVT (Combo)	**76.5**
	TVI (Vector)	52.9
	TVI (Residue)	52.9
	TVI (Combo)	44.1
Supervised	TVL, TVT(Motion)	64.7
	TVL, TVT(Bitrate)	**73.5**
	TVL, TVT(Combo)	70.6

Table 2.4. Performance of **Visual Activity** cues for **Most**-dominant person task with **Full**-agreement data.



(a) Most-dominant task (full agreement)

(b) Least-dominant task (full agreement)

Figure 2.7. Scatter plots of the total speaking and visual activity length, where the red crosses show the data points belonging to the positive class and the black circles show the negative class in each case.

best with 67.6% accuracy, which implies that dominant people receive more attention than others. The denominator of MVDR, Total Looking-At-Others Length while not speaking TLOLwNS using the minimum option had an accuracy of 52.9%, in other words estimating the 'one who looks at others the least while speaking' was also a reasonably good feature. It is interesting to observe that this feature, when combined with TLOLwS in the form of a ratio, MVDR, becomes the best estimator of dominance.

For the multiple feature case, a small selection of the best performing combinations is also shown in Table 2.5. The combination of TRVA and TRVAwNS improves the classification accuracy to 82.3%. Overall, the best achieved performance with visual attention cues and supervised learning is 8.9%

Dominance Model	Features	Class. Acc.(%)
Unsupervised	TRVA	67.6
	TLOL	26.5
	TLOT	55.8
	TRVAwS	11.8
	TRVAwS(min)	29.4
	TLOLwS	70.6
	TLOTwS	64.7
	TRVAwNS	70.6
	TLOLwNS(min)	52.9
	TLOTwNS	41.2
	$MVDR_1$	**79.4**
	$MVDR_2$	**79.4**
Supervised	TRVA,TLOLwS	79.4
	TLOLwS,TRVAwNS	79.4
	TRVA,TRVAwNS	**82.3**
	TRVA,TLOLwS,TLOLwNS	79.4
	TRVA,TLOLwS,TRVAwNS	**82.3**

Table 2.5. Performance of **Visual Attention** cues for **Most**-dominant person task with **Full**-agreement data.

worse than the corresponding best performance for audio cues (82.3% vs. 91.2%), compare Tables 2.3 and 2.5. In contrast, the best performance with attention cues was 5.8% better than the best performing visual activity cue (82.3% vs. 76.5%).

Audio-Visual fusion

A selection of results obtained with visual and audio-visual cue fusion are shown in Table 2.6. For the visual cue fusion, we could not improve the performance of 82.3%. The visual activity cues pull down the performance when combined with the visual attention cues. Interestingly, TSE and TST when combined with TLOLwS, a visual attention cue gave the best classification accuracy (91.2%). Note that the difference in performance between the best methods are not statistically significant at the 5% level using a standard binomial test, as the number of data points is small. Nevertheless these results show that such features and feature combinations are worth exploring.

Fusion	Feature	Class. acc. (%)
Visual	TRVAwNS,TVT(Residue)	**79.4**
	TRVA,TLOLwS,TLOLwNS,TVT(Residue)	**79.4**
Audio-Visual	TSE, TST, TVL(Residue)	85.3
	TSE, TST, TLOLwS	**91.2**

Table 2.6. Performance of **Audio-Visual** cues with **Most**-dominant person task with **Full**-agreement data.

2.6.2 Majority-agreement data set

The second classification task addressed involves the 57-meeting set where at least 2 annotators agree, which corresponds to almost all the data (96%). This data set inherently has more variability with respect to human perceptions of dominance (as further suggested by the lower confidence self-reported by the annotators as discussed in Section 2.3). The evaluation of this task is therefore aimed at analyzing the performance of models and cues in more challenging conditions.

For evaluation, we used two different ways of computing classification accuracy. Let N denote the total number of meetings, and A_i and B_i be the most-dominant-person ground truth labels corresponding to the 'most-voted' (two votes) and 'least-voted' (one vote) cases, respectively, for meeting i, $1 \leq i \leq N$. Furthermore, let n be the number of times the automatically estimated most dominant person is A_i, and m be the number of times the estimated most dominant person is B_i. A first, hard evaluation criterion, (called EvH for short) computes the classification accuracy as n/N, and a second, soft criterion (called EvS), computes classification accuracy as $(n+m)/N$. The hard criterion assumes that there is only one correctly labeled most-dominant-person for each meeting - the one corresponding to the majority vote by the annotators - and is obviously the correct way to evaluate performance on the full-agreement data set, as done in the previous section. In contrast, the soft criterion assumes that both the 'most-voted' and the 'least-voted' most-dominant-person labeled by the annotators for a given meeting are correct, and thus the estimation of either of them is considered as correct. This evaluation is clearly less stringent, but it is nevertheless important to observe the ability of the algorithms to estimate either of the two people perceived by annotators as being most-dominant.

Audio cues

Table 2.7 presents a selection of the classification accuracy results obtained for audio cues. For single cues and the unsupervised model, TSL and TSE are the best performing features for EvH (77.2% and 73.7%, respectively). TSTwoSU is the third best performing feature. For EvS, TSL and TSTwoSU were the best performing with an accuracy of 84.2% and 82.5%, respectively. TST and TSI are not as effective. Interestingly, these findings are consistent with the ones obtained for the full-majority data set (compare to Table 2.3). A consistent decrease in performance (8.1% for TSL) is observed for all

cues which suggests that the inclusion of the data that is intrinsically more ambiguous with respect to perceived dominance results in a more challenging task. On the other hand, the results obtained with the soft criterion, which assumes that more than one person can be most-dominant, brings the performance of most features back to the same level they had for the full-agreement data set, which indicates that in several cases the methods guessed the 'least-voted' person as being most dominant. Selected results for the supervised model and fused audio cues also appears in Table 2.7. The selection shown is a subset of those in Table 2.3 and includes the best performing cases. We observe that, using the EvH criterion, a few feature combinations performed at the same level, but not better, than the best single cue. On the other hand, using the EvS criterion, we observe that the same feature combinations were capable of slightly improving performance (a best performance of 86.0% for the same feature combination that performed the best for full-agreement data). Overall, the supervised approach brought a slight improvement (although not statistically significant) over the much simpler unsupervised case.

Dominance Model	Feature	Class. Acc. %	
		EvH	EvS
Unsupervised	TSL	**77.2**	**84.2**
	TSE	73.7	79
	TST	54.4	64.9
	TSTwoSU	71.9	82.5
	AvTDur	63.2	73.7
	TSI	57.9	64.9
	TUI	63.2	71.9
	TSUI	63.2	75.4
Supervised	TSL, TSE, TST	**77.2**	**86.0**
	TSE, TST, TSI	75.4	**86.0**
	TSL, TST, TSI	**77.2**	84.2

Table 2.7. Performance of **Audio** cues for **Most**-dominant person task with **Majority**-agreement data.

Visual activity cues

Table 2.8 shows selected results obtained with visual cues.

Compared to the results obtained for the full-agreement case (Table 2.4), many observed trends hold: TVL and filtered TVT are the best performing single cues. TVI is a poor estimator, and overall the visual-only features perform worse than their audio counterpart. Furthermore, similarly to the audio-only results in this section, we observe a general decrease in performance with respect to the

Dominance Model	Feature	Class. Acc. %	
		EvH	EvS
Unsupervised	TVL (Vector)	63.2	77.2
	TVL (Residue)	66.7	**80.7**
	TVL (Combo)	64.9	**80.7**
	TVT (Vector)	61.4	75.4
	TVT (Bitrate)	64.9	77.2
	TVT (Combo)	**70.2**	**80.7**
	TVI (Vector)	47.3	63.1
	TVI (Bitrate)	47.3	59.6
	TVI (Combo)	47.4	61.4
Supervised	TVL, TVT (Vector)	63.2	77.2
	TVL, TVT (Combo)	63.2	77.2
	TVL, TVT (Residue)	**66.7**	**78.9**

Table 2.8. Performance of **Visual Activity** cues for **Most**-dominant person task with **Majority**-agreement data.

full-agreement data set when using the EvH criterion (for the best performing single visual cues, the absolute degradation is 6.3%). The results obtained with the EvS criterion for the best visual cues brings the performance back to the same level they had for the full-agreement case. Finally as also shown in Table 2.4, supervised learning and multiple visual cues did not improve performance over the simple unsupervised, single-cue model.

Visual attention cues

Table 2.9 shows selected results obtained with visual attention cues. Regarding single cues in the unsupervised setting, Multi-Party Visual Dominance Ratios (MVDR) was the best cue (with an accuracy of 73.7% for EvH), followed by Total Looking At Others Length while speaking (TLOLwS) and Total Received Visual Attention while not speaking (TRVAwNS) . The top ranked cues are consistent with the results in the full-agreement case. Finally, supervised learning and multiple visual cues slightly improved performance over the simple unsupervised, single-cue model, with the combination of TRVA and TRVAwNS performing at 75.4% accuracy for EvH.

Audio-visual cues

The results for the best combinations appear in Table 2.10. All visual activity features have been derived with the 'residue' option. We observe that audio-visual fusion did not improve, but equalled the performance over audio-only under both evaluation criteria. The overall best results are summa-

Dominance Model	Feature	Class. Acc. %	
		EvH	EvS
Unsupervised	TRVA	61.4	73.7
	TLOL	22.8	33.3
	TLOT	47.3	54.4
	TRVAwS	17.5	29.8
	TRVAwS(min)	22.8	31.5
	TLOLwS	63.2	73.7
	TLOTwS	57.9	70.2
	TRVAwNS	63.2	75.4
	TLOLwNS(min)	47.4	63.2
	TLOTwNS	40.4	49.1
	$MVDR_1$	71.9	**80.7**
	$MVDR_2$	**73.7**	**80.7**
Supervised	TRVA,TLOLwS	70.2	80.7
	TLOLwS,TRVAwNS	70.2	80.7
	TRVA,TRVAwNS	**75.4**	**86.0**
	TRVA,TLOLwS,TLOLwNS	73.7	82.4
	TRVA,TLOLwS,TRVAwNS	71.9	82.4

Table 2.9. Performance of **Visual Attention** cues for **Most**-dominant person task with **Majority**-agreement data.

rized in Figure 2.8.

Fusion	Feature	EvH	EvS
Visual	TRVAwNS,TVT(Residue)	68.4	77.2
	TRVA,TLOLwS,TLOLwNS,TVT(Residue)	68.4	77.2
Audio-Visual	TSE, TST, TVL(Residue)	73.7	**84.2**
	TSE, TST, TLOLwS	**77.2**	**86.0**

Table 2.10. Performance of **Audio-Visual** cues for **Most**-dominant person task with **Majority**-agreement data.

2.7 Classifying the Least-dominant person

In this section, we discuss our results for the least-dominant person classification task. The experiments that were carried out were identical to the most-dominant case so the discussion in this section will be rather concise. We first conducted experiments on the least dominant person task with full-agreement data (31 meetings), and then on the majority-agreement data (54 meetings). For the unsupervised model, the person that corresponds to the lowest proportion of the feature among all participants is classified as least dominant. The supervised model is trained on the least vs. non-least dominant classes defined in the annotations.

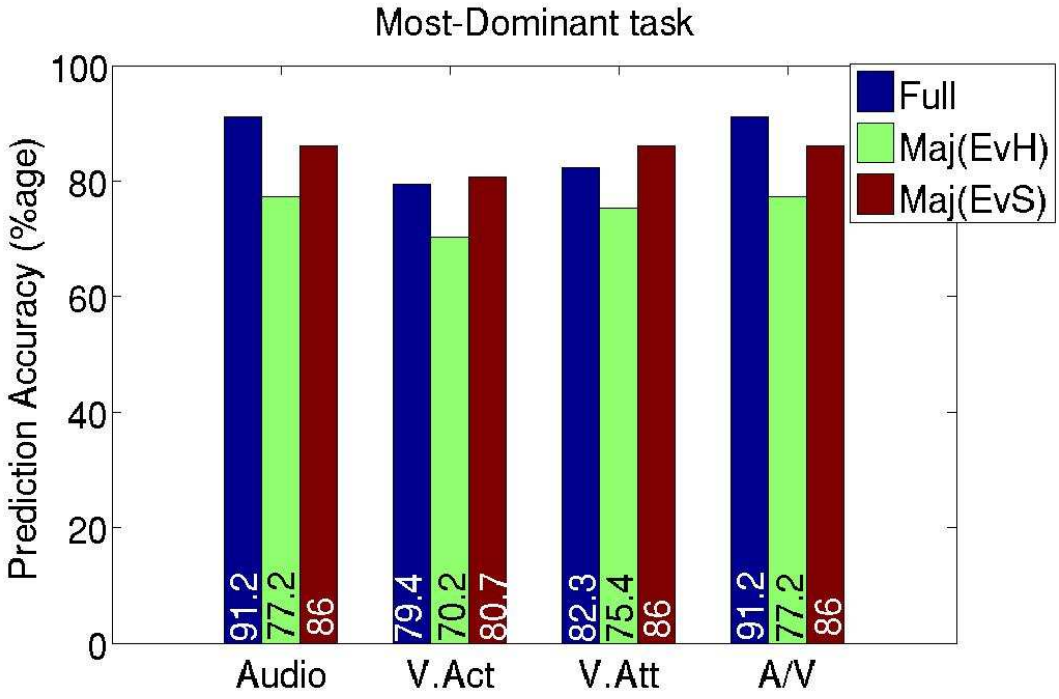


Figure 2.8. Comparison of the best performance values for the most-dominant estimation tasks. A:Audio, V. Act: Visual Activity, V. Att: Visual Attention, A/V: Audio-Visual.

2.7.1 Full-Agreement data-set

Audio cues

The classification accuracy of the audio cues under the unsupervised and supervised schemes are shown in Table 2.11. The highest performance of 83.9% was achieved by both the TSL an TSTwoSU. Supervised fusion of these cues improved the accuracy to 87.1%.

Like the equivalent case in Section 2.6.1, the TSI feature performed the worst for the unsupervised case. It was also interesting to see the increase in performance between the TST and TSTwoSU features. This suggests that the short turns added noise to the TST features. This was similarly observed for the corresponding set of results in Table 2.3 for the most dominant person task.

Unlike the most dominant case, here there is a significant reduction in performance for TSE compared to TSL. We speculate that this is because the total energy is much lower and therefore more sensitive to noise (i.e. the signal-to-noise ratio is lower). TSL showed a slight decrease in performance for estimating the least dominant person, compared to estimating the most dominant person. These

results suggest that a similar trend will also be observed with the visual cues; less dominant people are less active, so their measured activity will be more sensitive to noise. In addition, we note that some annotators did comment on how it was more difficult to rank passive participants than active ones.

Dominance Model	Feature	Class. Acc. (%)
Unsupervised	TSL	**83.9**
	TSE	67.7
	TST	71.0
	TSTwoSU	**83.9**
	AvTDur	67.7
	TSI	42.0
	TUI	71.0
	TSUI	71.0
Supervised	TSE, TST	80.6
	TSL, TSTwoSU	**87.1**
	TSE, TSTwoSU	83.9
	TSL, TSE, TST	77.4
	TSL, TST, TSI	77.4
	TSE, TST, TSI	80.6
Random Guess	None	25.0

Table 2.11. Performance of **Audio** cues for **Least**-dominant person task with **Full**-agreement data

Visual activity cues

Table 2.12 shows some selected results from our experiments using only the visual cues for the full-agreement data-set. While in the equivalent results of the most-dominant task in Table 2.4, both TVL(Residue) and TVT(Combo) had the best performance, for the least-dominant task, only TVL(Vector) performed the best. This is likely to be caused by the removal of the shorter turns, which account for noisy measurements of the visual activity. However, TVT might also eliminate significant amounts of true activity for the most passive person. We also found that the TVI feature performed less well in general. Overall, the visual features are less discriminative than the audio ones, and also less effective compared to the most-dominant task. In terms of statistical significance, the decrease in performance between the best audio and video performance for the full-agreement case was not statistically significant at 5% level using a standard binomial test.

Dominance Model	Method	Class. Acc.(%)
Unsupervised	TVL(Vector)	**54.8**
	TVL(Bitrate)	45.2
	TVL(Combo)	48.4
	TVT(Vector)	41.9
	TVT(Bitrate)	41.9
	TVT(Combo)	48.4
	TVI(Combo)	32.3
	TVI(Combo)	41.9
	TVI(Combo)	38.7
Supervised	TVL, TVT(Combo)	41.9
	TVL, TVT(Bitrate)	35.4
	TVL, TVT(Combo)	**48.4**

Table 2.12. Performance of **Visual Activity** cues for **Least**-dominant person task with **Full**-agreement data.

Visual attention cues

Table 2.13 shows the results obtained with visual attention cues. Regarding single cues in the unsupervised setting, Total Received Visual Attention while not speaking (TRVAwNS) and Total Received Visual Attention (TRVA) are the two best cues (with an accuracy of 77.4%) showing that less-dominant people receive less attention in general and also while they are not speaking. The Multi-Party Visual Dominance Ratios (MVDR) and the numerator of MVDR, Total Looking At Others Length while speaking (TLOLwS) also performed well (with an accuracy of 71.0%).

Dominance Model	Features	Class. Acc.(%)
Unsupervised	TRVA	**77.4**
	TLOL	19.4
	TLOT	22.6
	TRVAwS	54.8
	TLOLwS	71.0
	TLOTwS	58.1
	TRVAwNS	**77.4**
	TLOLwNS(max)	51.6
	TLOTwNS	19.6
	$MVDR_1$	71.0
	$MVDR_2$	71.0
Supervised	TRVA,TLOLwS	80.6
	TLOLwS,TRVAwNS	**83.9**
	TRVA,TRVAwNS	77.4
	TRVA,TLOLwS,TLOLwNS	80.6
	TRVA,TLOLwS,TRVAwNS	80.6

Table 2.13. Performance of **Visual Attention** cues for **Least**-dominant person task with **Full**-agreement data.

For the multiple feature case, a small selection of the best performing combinations is also shown in Table 2.13. Combining Total Received Visual Attention while not speaking (TLOLwS) and Total Received Visual Attention while not speaking (TRVAwNS) improved the performance to 83.9%. The best result obtained using audio cues and visual attention cues were the same.

Audio-Visual Fusion

Although the fusion of visual attention and visual activity cues did not improve the performance, similar to the most dominant full-agreement case, the fusion of audio and visual attention cues help in achieving the best accuracy of 90.4% (refer Table 2.14). Interestingly, the best feature combination for both most-dominant and least-dominant tasks were different. While the combination of TSE, TST, and TLOLwS was the best audio-visual option for the most dominant task (accuracy of 91.2%), the combination of TSL, TSTwoSU, TRVA was the best audio-visual option for the least dominant task (accuracy of 90.4%).

Fusion	Feature	Class. acc. (%)
Visual	TLOLwS,TRVAwNS,TVT(Residue)	79.4
	TLOLwS,TRVAwNS,TVL(Vector)	**80.6**
Audio-Visual	TSL, TSTwoSU, TVT(Residue)	87.1
	TSL, TSTwoSU, TVL(Vector)	87.1
	TSL, TSTwoSU, TRVA	**90.4**

Table 2.14. Performance of **Audio-Visual** cues with supervised model for **Least**-dominant person task with **Full**-agreement data.

2.7.2 Majority-agreement data-set

For this task, there was a total of 54 meetings, which accounted for 91.5% of the total data. We show a selection of performance results for this task in Table 2.15. The best achieved results are also shown in Figure 2.9.

Firstly, it was interesting to see that TSL was not the feature that gave the best performance, though it was ranked second behind TSTwoSU. This observation suggests that the adding annotator variability and having proportionately less observations in the captured signal leads to a greater need for noise removal. Furthermore, we found that the shorter turns were not a discriminative feature for estimating dominance as it is likely that for the least-dominant person, they would represent a larger proportion of a person's total speaking turns than for the most dominant person. The combination of

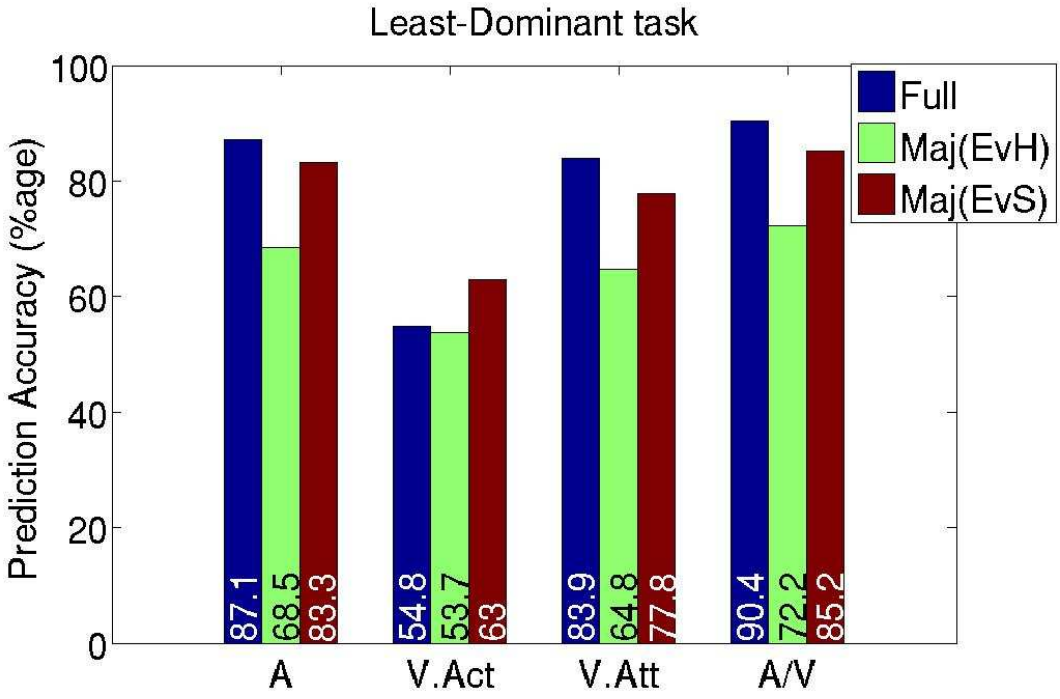


Figure 2.9. Comparison of the best performance values for the least-dominant estimation tasks. A:Audio, V. Act: Visual Activity, V. Att: Visual Attention, A/V: Audio-Visual.

TSL, TSTwoSU, and TRVA gave the best performance of 72.2% with EvH and 85.2% with EvS respectively, similar to the full-agreement case, reiterating that visual attention cues are complementary to audio cues.

2.8 Discussion and Conclusion

Overall, our study has investigated how dominance can be estimated by different audio and video cues, and affected by annotator variability, estimation method, and the specific classification task involved. Our investigation suggests the following:

Audio cues. When taking the cue which performed best in all categories, the audio cues always gave the highest classification accuracy. We observed that TSL gave the best results as a single feature, though was second best for the task of estimating the least-dominant person when the data set had majority agreement. In addition, TSTwoSU was found to be more robust to annotator variability by obtaining the highest performance in both most and least dominance tasks. There was a marked

Dominance Model	Features	Class. Acc. %	
		EvH	EvS
Unsupervised	TSL	59.3	75.9
	TSTwoSU	**68.5**	**83.3**
	TVL(Vector)	53.7	63.0
	TVL(Residue)	50.0	61.0
	TVL(Combo)	46.3	59.3
	TVT(Combo)	42.6	57.4
	TRVA	61.1	75.9
	TLOLwS	50.0	67.7
	TRVAwNS	61.1	75.9
Supervised	TSL, TSE, TST	63.0	83.3
	TSE, TST, TSI	63.0	77.8
	TSL, TST, TSI	59.3	77.8
	TVL, TVT (Vector)	51.8	61.1
	TVL, TVT (Combo)	48.1	61.1
	TVL, TVT (Residue)	48.1	64.8
	TRVA,TLOLwS	64.8	77.8
	TLOLwS,TRVAwNS	64.8	79.6
	TSL, TSTwoSU, TVL(vector)	70.4	**85.2**
	TSL, TSTwoSU, TRVA	**72.2**	**85.2**

Table 2.15. Performance of **Audio, Visual, and Audio-Visual** cues for **Least**-dominant classification task with **Majority**-agreement data.

improvement in performance between the TST and TSTwoSU features, indicating that much of the noise in the TST feature was caused by the shorter turns, which were not discriminative for our task. TSI performed badly in general, suggesting that interruptions may not be a good cue for dominance estimation in cooperative scenarios like AMI meetings. One point to note, however, is that this cue was derived using a coarse measure, which did not quantify the quality or actual intention of the interruption.

Visual activity cues. We found that their performance was never able to improve upon those of the best audio cues. However, it was interesting to see that a comparison of the performance of the single audio and video cues for the most-dominant case shows that the gap between modalities in some cases is small (see Figure 2.8) even though the visual activity cues are coarse and fast to compute and the resulting features are noisy. For the least-dominant task as shown in Figure 2.9, the visual activity cues were comparatively worser. These cues are particularly interesting in applications when it is not possible or ethical, due to privacy reasons, to listen to the conversations at all.

Visual attention cues. We found that TRVA, TLOLwS, TRVAwNS, and MVDR were the best single cues, for both the most-dominant and the least-dominant tasks. These cues were slightly better than

the visual activity cues for the most-dominant case, and much better for the least-dominant case. Cue fusion of visual attention cues helped in improving the classification accuracy in both tasks. The MVDR, a popular cue in social psychology literature, was effective in estimating dominance. In scenarios where there are no distracting objects like laptop, slide-screen, and white board the visual attention features could be expected to perform even better. It is however to be noted that the computational cost of obtaining visual attention cues are much higher that both the visual and speech activity cues.

Audio-Visual Cues. In terms of audio-visual cue fusion, we found that for both tasks, the best performing cue was either an audio cue or a combination of audio and visual attention cues. The combination with visual activity cue did not help. This can be explained by the overall lower performance of the visual activity cues and the fact that they are connected with the speaking activity (Fig. 2.7). One observation we must make here is that the audio signal was extracted from close-talk headset microphones while the video signal was captured from a much further distance from the participants. It would be important to see how the results using audio cues would change if more challenging audio data from far-field microphones was used. Parallel work using a single distant microphone to extract the total speaking length has shown that there is indeed a decrease in performance (Hung *et al.*, 2008a) although not too drastic.

Full and Majority Agreement Data. From the two evaluation criteria that were used for the data sets with majority agreement, we found a systematic drop in performance when comparing the performance of the hard evaluation criterion with the Full-agreement case. However, it was interesting to observe that with the soft criterion, the performance in some cases was equivalent to that of the corresponding Full-agreement case.

Supervised and Unsupervised Models. It was interesting to observe that while the best performance of 91.2% (and 90.4%) for the estimation of the most (and least resp.) dominant person was obtained using the SVM method, the best performance with the unsupervised model and a single cue was already 85.3% (and 83.9%). This is an interesting result since the unsupervised model does not require training data and has a much lower computational overhead compared to the supervised model.

Most and Least Dominant Tasks. It was interesting to observe that there was a consistent drop in performance between the two tasks as shown in Figure 2.8 and Figure 2.9. Closer inspection also shows that there is a more significant decrease in performance between the visual activity cues for the least dominant task compared to that of the most dominant. This is an interesting finding that

highlights the inherent increase in uncertainty when trying to identify people who have a lower level of activity. While the most dominant person in a meeting might be considered the most active and therefore more observable, finding the least-dominant person is closer to identifying the most passive or someone with the least observable cues. This seems to be reflected in the self-reported annotator confidence values (see Table 2.1).

Evaluation advantages and limitations. Our work has produced novel evaluation resources (data annotation, research tasks, and corresponding data sets) that build upon and enrich the publicly available AMI meeting corpus. Finally, as the size of the data set is relatively small, many of the observed performance differences between the best cues are not statistically significant at conventional levels although the difference between the best cues and random performance are statistically significant. In this view, the results presented here need to be interpreted with care, specially from the view of generalization. While the social psychology literature has validated, over multiple studies, the robustness of certain nonverbal cues for dominance perception (Schmid Mast, 2002), similar work to ours would have to be done in other scenarios to thoroughly validate such cues in automatic systems, using larger and varied data sets.

Possible Extensions. One of the limitations of our work is its reliance on high-quality audio (derived from close-talk microphones) to extract cues. How the results generalize when using single distant microphones have been recently studied (Hung *et al.*, 2008a). The results suggest that the most-dominant person classification performance degrades, as compared to the head-set microphones, but the degradation is not drastic. In the second place, a related dominance problem is to estimate dominant cliques (or subsets of people) rather than dominant individuals, since there are occasions when multiple people can be perceived as similarly dominant. We performed an initial investigation about this subject in (Jayagopi *et al.*, 2008a). In the third place, cue fusion with many other learning techniques both supervised and not could also be investigated (Aran and Gatica-Perez, 2010). In the fourth place, modeling annotators and therefore generalizing the majority voting principle is also a promising research direction. As a fifth direction, the nonverbal communication literature also refers to various cues related to other cues for dominance (e.g. postures and gestures) and this would be interesting to explore. The role of prosodic cues like pitch frequency, speaking rate to predict dominance is also a interesting research direction. An open question is how much improvement (if any) could be obtained with features that might be significantly more expensive to compute. Finally, the

performance measures considered in this paper are simply a few of the various possible options. In the future, it would be interesting to examine the effect of various cues on the speed of detecting dominance and define performance as a tradeoff between complexity and classification accuracy.

Chapter 3

Beyond Dominance: estimating status with nonverbal cues

As stated in the previous chapter, the understanding in the workplace of fundamental constructs related to power, hierarchy, dominance, and status called the vertical dimension of social interaction by Hall et al. (Hall *et al.*, 2005)) would open doors to tools to support research in social and organizational psychology and for personal self-assessment (Pentland, 2005).

In this chapter we go beyond the study of dominance by adding another aspect of verticality in group interaction, namely status. As stated in Chapter 2, dominance can be defined as "expressive, relationally based communicative acts by which power is exerted and influence achieved" (Dunbar and Burgoon, 2005b) (p. 208), or as "a personality trait involving the motive to control others, the self-perception of oneself as controlling others, and/or as a behavioral outcome (success in controlling others or their resources)" (Hall *et al.*, 2005) (p. 898). On the other hand, status can be defined as "an ascribed or achieved quality implying respect or privilege, which does not necessarily include the ability to control others or their resources)" (Hall *et al.*, 2005) (p. 898). In the workplace, status often corresponds to a person's position in a group or in the organization's hierarchy, and it is often defined by a formal role (e.g. a project manager or a team leader). Dominance and status are related constructs: dominant people often occupy high positions in an organization; conversely, high-status people are often allowed (even expected) to use dominant behavior with their subordinates. At the

same time, these two concepts do not always coincide, and can even contradict each other: for example, a high-status manager could have an intrinsic non-dominant personality, or fail to control or influence his team (Hall *et al.*, 2005).

Both dominance and status structure nonverbal behavior in important ways (Leffler *et al.*, 1982; Dunbar and Burgoon, 2005b; Hall *et al.*, 2005). From a rich amount of work in social psychology and communication, it is known that several audio and kinesic cues (Dunbar and Burgoon, 2005b; Leffler *et al.*, 1982) are related to dominance and status. For instance, both dominant and high-status people are often more vocally and kinesically expressive than their counterparts, and that both types of people often receive more visual attention. Less clear, however, is whether these cues are correlated in similar amounts with the expression and perception of each construct, and whether automatically extracted cues - likely to be imperfect - would be useful for the estimation of both types of social patterns.

This chapter addresses two questions. First, can perceived dominance and role-based status in small-group conversations be automatically explained by the *same* nonverbal cues? While some social psychology literature has found common ground for the nonverbal display and interpretation of both constructs, and recent computational literature has started to investigate models for automatic estimation of dominance (Rienks and Heylen, 2005; Jayagopi *et al.*, 2009b) or roles (Zancanaro *et al.*, 2006; Dong *et al.*, 2007; Vinciarelli, 2007) in conversations, no attempt has been made to jointly study these two dimensions of social verticality using common data and nonverbal cues. Second, is it possible to estimate these two aspects of verticality from relatively brief observations and using fully automatic nonverbal cues? Although significant evidence in cognitive science support 'thin-slice' explanations for many aspects of social cognition, and such approaches have started to be used with success in computational methods (Pentland, 2005), the question remains essentially open for the two concepts we investigate here.

We present a comparative study of the discriminative power for perceived dominance and assigned status estimation of a number of automatic nonverbal cues (extracted from multiple audio and visual sensors) that characterize speaking activity, visual activity, and visual attention. Many of the investigated cues have empirical support in social psychology for either or both status and dominance. Using five hours of five-minute slices of the AMI corpus, our work shows that (1) although dominance and status might be related in terms of the associated nonverbal behavior, they are in practice better explained by different nonverbal cues; and (2) the best single nonverbal cues can correctly estimate the

person with highest dominance or role-based status with reasonable accuracy. The material in this chapter was originally published in (Jayagopi *et al.*, 2008b).

The chapter is organized as follows. Section 3.1 summarizes the related work. Section 3.2 details the data and the research tasks. Section 3.3 describes the nonverbal cues used in our study. Section 3.4 presents the estimation model. Section 3.5 presents and discusses the results. Section 3.6 offers some concluding remarks, some of the challenges involved for future work, and discussion

3.1 Related work

As the literature on dominance was reviewed in the previous chapter, in this section we review the literature on role and status modeling in social psychology and computational literature.

3.1.1 Related work on role modeling

Formal role, as defined by Hare in the social psychology literature is "that is associated with a position in a group (or status) with rights and duties to one or more group members[...] that members perform consciously" (Hare, 1994). Informal roles also emerges during the interaction. A notable study on informal roles is the work by Bales on Interaction Process Analysis - IPA, a framework to study small groups by classifying individual behavior in a two-dimensional role space consisting of a Task and of a Social-Emotional area (Bales, 1970). The Task Area consists of roles relating to the facilitation and coordination of the tasks the group is involved in, for example orienter, information seeker, etc. The Socio-Emotional Area concerns the relationships between group members and roles oriented towards harmonising or destabilizing the functioning of the group, for example attacker, supporter, etc.

The computing literature on automatic role recognition is quite diverse in the types of roles that have been investigated. Banerjee et al. using simple speech-based turn-taking features and a decision tree classified the roles of the meeting participants as discussion participator, presenter, information provider, and information consumer (Banerjee and Rudnicky, 2004). The accuracies reported at every window of 1 second duration was of the order of 50%. Vinciarelli studied the problem of role recognition in multiparty audio recordings of radio bulletins using features based on social network and duration distribution analysis (Vinciarelli, 2007). The six roles studied were domain-specific and included an anchorman among others. Unlike our work, the conversations in this case are often dyadic,

making the task easier when compared to the role recognition in meetings. The reported performance was of approximately 85 % frame-based classification accuracy on programs of 12-minute average duration each, more than twice the duration we analyze in this work. On the AMI corpus, the recognition of roles of the team members - Project Manager (PM), Marketing Expert (ME), User Interface Expert (UI), and Industrial Designer (ID) - was attempted by Salamin et al (Salamin *et al.*, 2009). Frame-level accuracies reported using features based on social network and duration distribution analysis were of the order of 70%. The meetings were of 20 minutes duration on an average. Another role recognition problem was addressed by Zancanaro et al. (Zancanaro *et al.*, 2006) and Dong et al. (Dong *et al.*, 2007). Instead of organizational roles, the authors targeted the recognition of two types of functional roles, studied by Bales, in meetings: 'task-based' functional roles, which included Orienteer, Giver, Seeker, Procedural Technician, and Follower; and 'socio-emotional' roles, which included Attacker, Supporter, Protagonist, and Neutral. Each analyzed meeting was 25-minute long in average, a much longer temporal support than we address here. In their work, the authors explored the use of SVMs (Zancanaro *et al.*, 2006) and the Influence Model (Dong *et al.*, 2007). In both (Zancanaro *et al.*, 2006; Dong *et al.*, 2007), the authors reported 60-70% frame-based classification accuracy for the two role classification tasks. In a different line of work, Educational role, as Professor, PhD Student and Graduate student, were classified (Laskowski *et al.*, 2008) on the ICSI meeting corpus, obtaining a best frame-level accuracy of the order of 60%. Raducanu and Gatica-Perez (Raducanu and Gatica-Perez, 2010) addressed the problem of analysis of competitive meetings making use of "The Apprentice" reality TV show, which features a competition for a real, highly paid corporate job. Their analysis centered around two tasks regarding a person's role in a meeting: estimating the person with the highest status, and estimating the fired candidates on the whole meeting data. The reported estimation accuracies were of the order of 85%. Valente and Vinciarelli (Valente and Vinciarelli, 2010) studied roles in TV debates (composed of a moderator and two groups of participants) and used the information as prior for a speaker diarization system. Most of the above works employed only acoustic cues, except (Zancanaro *et al.*, 2006) and (Dong *et al.*, 2007) which also made use of body fidgeting cues.

3.1.2 Related work on status modeling

The social psychology literature on status in small groups concerns mainly with the emergence and measurement of status using nonverbal behavior. The design of status systems was either exper-

imentally manipulated though role-play (playing roles of manager-subordinate, teacher-student etc) or measured later after the interaction (including self-report or observed by external observers) (Hall *et al.*, 2005).

The participation hierarchy that indicates status differentiation quickly emerges in a discussion even when unacquainted individuals are placed together. This rapid structuring of status hierarchy might happen through some subtle forms of signalling through eye glances or turn-taking (Rosa and Mazur, 1979). It has also been suggested that the basis for status formation is in the expectation about task performance. Influential leaders display more task related cues (like verbal fluency and modulated voice) than dominance cues (like pointing and glaring) (Ridgeway, 1987).

The relationship between nonverbal cues and status is clear for some cases. People with high status speak more often than others, are more likely to criticize, command, or interrupt others, and are spoken to more often than others (Levine and Moreland, 1990). They have higher visual dominance ratio, lean forward less, use fewer verbal facilitators (expressions "such as mm-hmm" and "yeah"), and speak louder (Hall and Friedman, 1999). On the other hand, some nonverbal cues have contradictory relationship with status. Weak or dependent people are sometimes found to gaze more, but sometimes more powerful or higher-status people gaze more. (Hall and Friedman, 1999).

As compared to the work on role modeling in computational literature, the work on status modeling has been rather limited. Sachez-Cortes et al. (Sanchez-Cortes *et al.*, 2010) explored the problem of emergent leadership in newly formed small-groups using turn-taking cues and fusing cues at the score level. Varni et al. (Varni *et al.*, 2010) also studied the emergence of leadership, albeit in a novel active music listening scenario, by modeling the synchronization aspect of affective behavior within a small group. The cues employed include trajectories of body-parts, velocity, acceleration, gesture features from video; loudness, spectral features, beat tracking, melodic contour, phrasing from audio.

Our work differs significantly from most existing works. As compared to works in social psychology literature, we extract and study nonverbal cues automatically. Also, we compare the effectiveness of the cues to estimate both dominance and status on a publicly available AMI corpus. As compared to works in computational literature, we attempt a novel task and report estimation accuracies on slices of interaction of 5 minutes duration, as compared to other works that report accuracies at either frame-level or much larger interaction duration. Finally, our feature set is truly multimodal, unlike most existing works, and uses speaking activity, visual activity, and visual attention features.

3.2 Experimental setup: Meeting data and tasks

Our objective in this work is to study and model social verticality in task-oriented small groups. We chose meetings from the Augmented Multi-party Interaction (AMI) corpus (Carletta et al., 2006) because every meeting had a 'project manager' who we assume has the higher status. For a more detailed description about the AMI corpus the reader should refer to Chapter 2.4.1.

3.2.1 Dominance Task: Estimate the most-dominant person

As described in Chapter 2, we performed dominance annotation on 59 five-minute meetings. Except 2 meetings, 57 meetings had majority agreement (two or three annotators agreed) on the most-dominant person. We use these 57 meetings for our experiments. The data is approximately 5 hours of interaction.

3.2.2 Status Task: Estimate the project manager

In order to study dominance and status together, we use the same 57 meetings for this task. Similar to the most dominant person task, we define the project manager task. As each participant was assigned distinct roles in the AMI corpus: 'Project Manager', 'User Interface specialist', 'Marketing Expert', and 'Industrial Designer', the ground truth is given. In fact, out of the 57 meetings, 37 meetings were such that the Project Manager (PM) was also judged to be the most-dominant person on whom the majority of the annotators agree. This suggests that in many cases (around 65 % of the cases), the project manager also displayed a dominant behavior.

3.3 Nonverbal cues

Various nonverbal behaviors that indicate dominance and status or role have been reported in the literature (Burgoon and Dunbar, 2006; Dovidio and Ellyson, 1982; Dunbar and Burgoon, 2005b; Hall *et al.*, 2005; Leffler *et al.*, 1982; Ridgeway, 1987; Schmid Mast, 2002). We employ speech activity, visual activity, and visual focus of attention for estimating the most dominant person and the project manager. We extract the same cues defined in chapter 2. We define one more audio cue as follows:

Total number of times speaking first after another speaker (TSF): This feature encodes the hypothesis that dominant or high status people respond to others first (Ridgeway, 1987; Leffler *et al.*, 1982). The feature is defined by the cumulative number of times that participant i speaks first (before other participants by backchannelling or successfully interrupting), after another participant j started talking.

Additionally we also extract two measures based on centrality. The Social Network Analysis literature has studied interaction among people in social environments (Wasserman and Faust, 1994). Various network centrality measures exist for different relationships. Wasserman et al. (Wasserman and Faust, 1994) discuss measures in which the centrality or status of positions are recursively related to the centrality or status of the positions to which they are connected.

Such measures of centrality can be readily applied where relational data exists. We applied two such measures on some of the relational features. We use an eigenvector-like measure based centrality (Bonacich and Lloyd, 2001), which we refer to as $Centrality^1$, and another measure of centrality as defined below, called $Centrality_i^2$ **:**

$$Centrality_i^2 = \frac{K-1}{\sum_{j=1}^{K} d_{ij}}, \forall i = 1, 2, 3, ..K \tag{3.1}$$

where K is the number of participants (the number of nodes in the social network), and d_{ij} is the distance between nodes i and j. Maximizing $Centrality^2$ is equal to minimizing $\sum_{j=1}^{K} d_{ij}$.

We investigated whether centrality measures could be used to estimate status or dominance, using it on two representative relational data (arranged as a matrix):

The two relational data matrix considered are defined as follows:

- **Total 'number of times speaking first after another speaker' matrix (TSF matrix) :** Each matrix element a_{ij} is defined by the cumulative number of times that a participant i speaks first (before other participants), after another participant j started talking.
- **Total 'number of times looking at others' matrix (VFOA matrix) :** The matrix element a_{ij} is defined by the cumulative number of times that a participant i looks at j.

We approximate d_{ij} as a_{ij}^{-1}, which means that the larger the interaction between people the smaller the distance between them.

Table 3.1 provides a summary of all the audio and video cues and their associated acronyms. We

have reproduced this table to facilitate reading.

Glossary of Feature Acronyms	
'Audio Activity'	
Total Speaking Energy	TSE
Total Speaking Length	TSL
Total Speaking Turns	TST
Total Speaking Turns without Short Utterances	TSTwoSU
Average Turn Duration	AvTDur
Total Successful Interruptions	TSI
Total Unsuccessful Interruptions	TUI
Total Short Unsuccessful Interruptions	TSUI
Total Speaking First	TSF
'Visual Activity'	
Total Motion Length	TVL
Total Motion Turns	TVT
Total Motion Interruptions	TVI
'Visual Attention'	
Total Received Visual Attention	TRVA
Total Looking At Others Length	TLOL
Total Looking At Others Turns	TLOT
Total Received Visual Attention while speaking	TRVAwS
Total Looking At Others Length while speaking	TLOLwS
Total Looking At Others Turns while speaking	TLOTwS
Total Received Visual Attention while not speaking	TRVAwNS
Total Looking At Others Length while not speaking	TLOLwNS
Total Looking At Others Turns while not speaking	TLOTwNS
Multi-Party Visual Dominance Ratios	$MVDR_1$ and $MVDR_2$
'Centrality measures'	
Eigen-vector like Centrality measure	$Centrality^1$
Centrality measure defined in Equation 3.1	$Centrality^2$

Table 3.1. Glossary of feature abbreviations.

3.4 Estimation and evaluation method

Estimating the most-dominant or the project manager and its evaluation are done as follows. Firstly, the audio cues, visual activity cues, and visual attention cues are accumulated over the duration of the meeting (as explained in Section 3.3). Then, depending on whether the relation of the feature to the task is assumed to be direct or inverse, either the largest or smallest accumulated value of each feature is taken. It is to be noted that unless specified otherwise, the largest value is chosen and whenever the smallest value is chosen, '(min)' appears next to the feature name like TBI(min).

That is, we hypothesize that someone is likely to be more dominant if they speak, move, look, or grab the floor the most out of all the participants in the meeting. We evaluate the method by comparing the predicted person with that of the ground truth for both tasks, and computing the classification accuracy as percentages. It is important to note that we estimate outcomes for full meetings, rather than for frames unlike works such as (Salamin *et al.*, 2009). For the dominance task, when there is full agreement on the most dominant person, computing the estimation accuracy is straight-forward. When there is majority agreement, a weighting scheme is used to compute the accuracy in order to accomodate the judgments of all the three annotators. Let N denote the total number of meetings, and A_i and B_i be the most-dominant-person ground-truth labels corresponding to the 'most-voted' (two votes) and 'least-voted' (one vote) cases, respectively, for meeting i, $1 \leq i \leq N$. Furthermore, let n be the number of times the automatically predicted most dominant person is A_i, and m be the number of times the predicted most dominant person is B_i. We compute the classification accuracy as $(2/3 * n + 1/3 * m)/N$. We have also experimented with other evaluation methods in the previous chapter on the same dataset. With this evaluation, the maximum achievable performance is less than 100%. In our case it is of 86.5%. It is important to note that the dominance models considered are unsupervised and therefore do not involve any training.

3.5 Results

We conducted experiments using audio cues (see Section 3.5.1), visual activity based cues (see Section 3.5.2), and visual attention based cues (see Section 3.5.3) on the two tasks - most-dominant person and the project manager. In the tables of this section, the column titled MD gives the classification performance in percentages, for the most dominant person task on the 57 meetings set. The classification performance for the project manager task is shown in the column titled PM. It is important to note that, though the tasks are independent, the ground truth for both tasks have overlaps i.e. 65% of the project managers are also the most dominant. We also report the results on the overlapping and non-overlapping subsets of meetings, corresponding to the columns titled PM $=$ MD (37 meetings) and PM $\neq$ MD (20 meetings). The results on the subsets helps us understand how specialized these features are for each of the tasks. Figure 3.1 illustrates these overlapping and non-overlapping subsets of most-dominant and high-status person data.

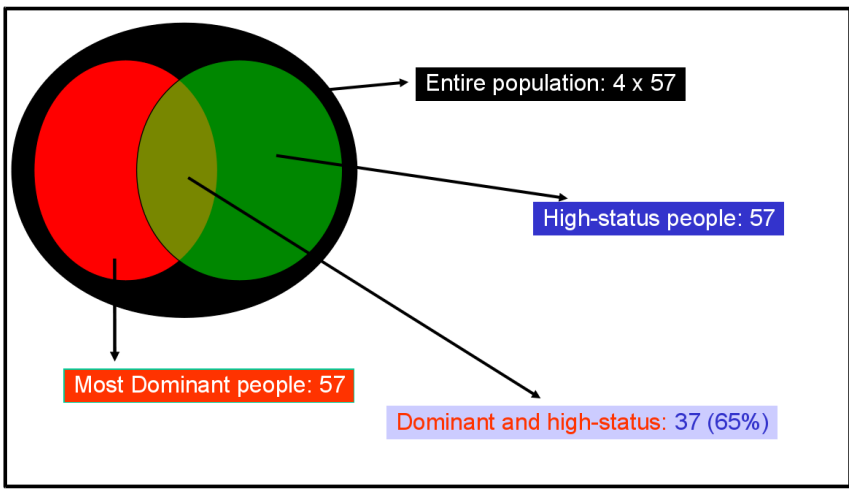


Figure 3.1. Venn diagram showing overlapping and non-overlapping subsets of most-dominant and high-status person data.

3.5.1 Audio cues

Table 3.1 shows the results obtained using audio cues. For the most-dominant person task, the total speaking length (TSL) and total number of speaker turns removing short turns (TSTwoBC) were most effective in classifying the most dominant person with a classification accuracy of around 70%. Social psychology literature (Schmid Mast, 2002) supports the results that speaking time is a very strong cue for dominance perception by humans. It is to be noted that the same cues estimate the most dominant person on a cleaner dataset, with full-agreement on the most-dominant person with an accuracy of 85% (see Section 2.5). The total speaking energy (TSE) also performed well. For the project manager task, the total number of speaker turns (TST) and the total number of times speaking first after a speaker (TSF) were the best indicators (with a classification accuracy of 63.2% and 66.7%). Also, it is interesting to observe that including the short utterances (of duration around 1 sec) is useful to estimate the project manager and not the most-dominant person. For PM $\neq$ MD case, TSL and TSE totally failed as a predictor of the status. This highlights some of the differences between dominance and status.

Successful interruption cue performed better than random, similar to the results obtained in the previous chapter. Total Unsuccessful Interruptions and Total Short Unsuccessful Interruptions performed slightly better for the MD task, but slightly worser for the PM task. It is important to notice that in the AMI data, groups were gathered with volunteers, and each person was randomly assigned

a role. So it might be the case that the people assigned the PM manager does not have a naturally 'interrupting' personality.

Features	MD (57)	PM = MD (37)	PM ≠ MD (20)	PM (57)
TSL	**70.8**	75.7	0	49.1
TSE	67.3	70.3	0	45.6
TST	52.0	73.0	45.0	63.2
TSTwoSU	**70.2**	**78.4**	10	54.4
TSI	51.5	56.8	30.0	47.4
TUI	63.2	56.8	0	36.8
TSUI	63.2	54.1	0	35.1
TSF	50.3	75.7	**50.0**	**66.7**

Table 3.2. Performance of **Audio** cues for estimating the most-dominant person and the project manager.

Figure 3.2 shows the histogram of speaking length for both the most-dominant task and the project manager task. We observe that TSL is more discriminant for the dominance task. Similarly, Figure 3.3 shows the histogram of TSF. It is interesting to observe the difference between the histograms of the project manager and the others, showing that the manager responds first more often than the others, as he has the role of anchoring the meeting. This can be seen from the mean of the TSF feature for the project manager being higher as compared to others.

3.5.2 Visual activity cues

Table 3.2 shows the results obtained with visual activity cues. As in Chapter 2, we experimented with the three options , Motion Vectors (called Vector in the following discussion), Residual Coding Bitrate (Residue), and the average of both features (Combo).

For the MD task, the Total Visual activity Length (TVL) that quantifies how much people move, and Total Visual activity turns (TVT) that quantifies how often people move (removing the very short turns that we assume to be noise), performed relatively well, with a classification accuracy of 62.6%. The social psychology literature supports the value of similar features (Burgoon and Dunbar, 2006). All the three options - motion vector, residual bitrate, and their combination performed similarly. Compared to the speaking length, the visual activity length was 8.2% worser for the MD task. But for the PM task, the difference was not much. For the meetings where PM $\neq$ MD, the TVL cues were much better than TSL. The Total Visual activity Turns (TVT), both bitrate and combo, have some ability at

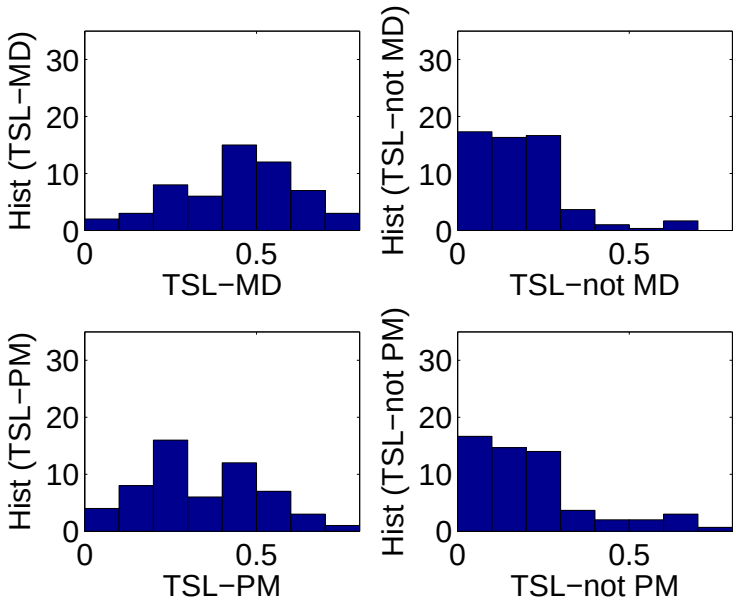


Figure 3.2. Histogram plots of normalised Total Speaking Length for both the most-dominant (MD) and project manager (PM) task.

estimating the project manager, similar to their audio counterparts, the Total Speaking Turns (TST) cues (a classification accuracy of 52.6%).

Features	MD (57)	PM = MD (37)	PM ≠ MD (20)	PM (57)
TVL(Vector)	59.6	59.5	30.0	49.1
TVL(Bitrate)	**62.6**	62.2	15.0	45.6
TVL(Combo)	61.4	62.2	25.0	49.1
TVT(Vector)	59.1	59.5	25.0	47.4
TVT(Bitrate)	**62.6**	**70.3**	20.0	**52.6**
TVT(Combo)	61.4	**70.3**	20.0	**52.6**
TVI(Vector)	46.2	54.1	**40.0**	49.1
TVI(Bitrate)	49.7	59.5	25.0	47.4
TVI(Combo)	49.1	64.9	30.0	**52.6**

Table 3.3. Performance of **Visual Activity** cues for estimating the most-dominant person and the project manager.

3.5.3 Visual attention cues

Table 3.3 shows the results obtained with visual attention cues. We systematically explored being-looked-at (passive) and looking-at (active) cues, as single events as well as jointly with speech activity

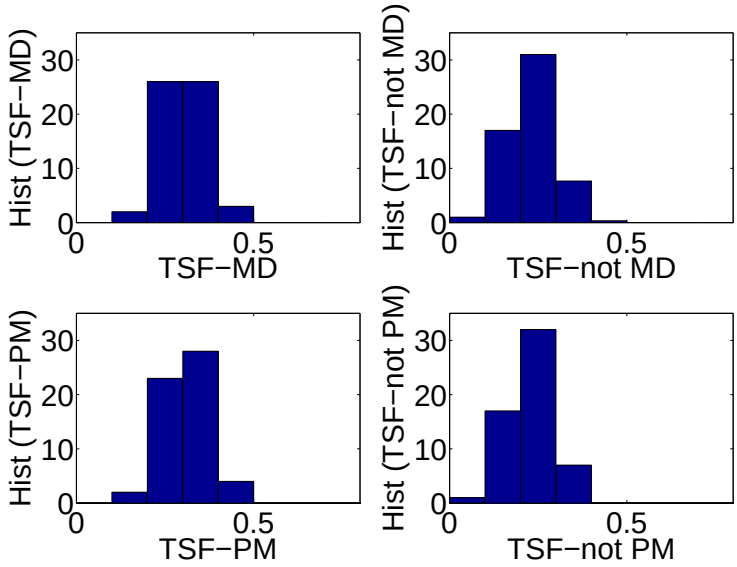


Figure 3.3. Histogram plots of normalised Total Speaking First after another participant (TSF) for both the most-dominant (MD) and project manager (PM) task.

and silence i.e while not speaking. Various popular hypotheses in social psychology literature could be verified.

The hypothesis that dominant or high status people are looked at longer (Efran, 1968) was verified as the Total Received Visual Attention feature (TRVA) performed significantly better than chance. TRVA while not speaking (glancing while someone else speaks), seems to carry more information about both dominance and status than TRVA while speaking. The hypothesis that dominant or high status people look at others more often was also verified with the TLOT feature (Cook and Smith, 1975). Also, 'looking-at-others while speaking' correlates with both tasks, as seen by the TLOLwS feature. The 'looking-at-others while not speaking', correlates negatively (using the min option) with both tasks, as seen by the TLOLwNS feature. The best performing features were the MVDR ratios for the dominance task (67.3%) and the 'looking-at-others while speaking' turns (TLOT) for the Project Manager task (59.6%). The second fact suggested that in our data the project manager frequently observes at his team members, while he is speaking. The visual attention cues were slightly better than the visual activity cues for the dominance task.

Features	MD (57)	PM = MD (37)	PM ≠ MD (20)	PM (57)
Overall attention cues				
TRVA	58.5	62.2	15.0	45.6
TLOL	24.0	24.3	20.0	22.8
TLOT	45.0	62.2	30.0	50.9
While-Speaking attention cues				
TRVAwS	24.0	27.0	20.0	24.6
TLOLwS	59.6	67.6	15.0	49.1
TLOTwS	55.6	73.0	**35.0**	**59.6**
While-not-Speaking attention cues				
TRVAwNS	60.2	64.9	15.0	47.4
TLOLwNS(min)	47.4	48.6	25.0	40.4
TLOTwNS	38	59.5	**35.0**	50.9
MVDR				
$MVDR_1$	66.7	73	10.0	50.9
$MVDR_2$	**67.3**	**75.7**	10.0	52.6

Table 3.4. Performance of **Visual Attention** cues for estimating the most-dominant person and the project manager.

3.5.4 Centrality measures

In Table 3.4 , we observe that the most central person, as predicted using both the measures, has significant correlation with the most-dominant person and the project manager. The $Centrality^1$ measure is consistently better than the $Centrality^1$ measure for both TSF and VFOA matrix choices. The $Centrality^2$ measure using the TSF matrix, predicts the manager with an accuracy of 68.4%, which makes it the best performing feature for the project manager task. It is also interesting to observe that this measure performs well even for the other three tasks, i.e. MD task, PM = MD, and PM ≠ MD.

Features	MD (57)	PM = MD (37)	PM ≠ MD (20)	PM (57)
$Centrality^1$				
using TSF matrix	49.7	70.3	40.0	59.6
using VFOA matrix	**56.1**	64.7	20.0	49.1
$Centrality^2$				
using TSF matrix	50.3	**75.7**	**55.0**	**68.4**
using VFOA matrix	48.5	56.8	30.0	47.4

Table 3.5. Performance of **Centrality** measures for estimating the most-dominant person and the Project Manager.

3.6 Discussion and Conclusion

Overall our study suggests the following:

Summary of results. In this chapter we investigated the problem of automatic estimation of the most-dominant and the high-status person using multimodal nonverbal cues. We employed automatic nonverbal cues - speaking activity based audio cues, visual activity cues, and visual attention cues - for doing the estimation. The best accuracies for both the tasks were of the order of 70%. At the level of human perception, we found that 65% of the time an 'assigned' project manager was also perceived as the most dominant. This was also revealed in the results as some of the nonverbal cues had comparable classification accuracies for both the tasks. It was interesting to observe that certain cues reveal the dominance behavior aspect better, whereas certain others capture the status better. Though the audio modality was the best, the visual attention based cues and the visual activity based cues are promising. Centrality measures, used in social network analysis, also correlate well with both tasks. Our study verifies some of the hypotheses related to the nonverbal cues, for both the dominance and the status tasks. Total Speaking Length and Total Speaking Turns without Short Utterances are the best nonverbal cues to estimate the most dominant person. The hypothesis that high-status people respond first (by back-channeling or attempting to grab the floor) was supported. Dominant or high-status people are active, as verified by the motion length and motion turns. Finally, received visual attention, looking at others while speaking, and the visual dominance ratios also indicate status and dominance.

Limitation. The study shows that some of the most difficult cases are when high-status people do not show dominant behavior through the measured nonverbal cues. Estimating in these cases is a very interesting open issue. As mentioned in Chapter 2, the size of the dataset is a limitation for this work as well. Also, it would be interesting to study the estimation accuracies when instead of head-set microphones, single distant microphone or array microphone data is used. Though the AMI corpus served as useful source of non-scripted group interaction data, a limitation of the dataset for our problem studied is in the 'assigned' nature of status, rather than measured as 'perceived status' or reflecting 'real' status in a status-differentiated group (for instance, a small group consisting of a supervisor and subordinates).

Possible extensions. One way of extending the work on verticality aspects of our thesis would be to jointly study 'power', a third facet of the vertical dimension, along with dominance and status. Studying these three social constructs together could involve collecting a new dataset. Another way of extending the work would be to study status in a real-life scenario, as compared to this 'role-assigned' scenario. With respect to the nonverbal cues that could be studied, prosodic cues could be interesting to study. Studies have suggested that dominance is correlated with prosodic cues such as pitch frequency, speaking rate (Tusing and Dillard, 2000). The need to model the prosodic cues would become even more to study dominance in non-cooperative settings like debates.

Chapter 4

Classifying group conversational context using nonverbal cues

Chapter 2 and 3 were concerned with modeling two individual social constructs, dominance and status. As compared to previous two chapters, our work departs from modeling behavior of individuals to groups. In this chapter we propose a novel framework to characterize group behavior.

With teams becoming ubiquitous in workplaces, the need to understand what influences group behavior and how it eventually affects performance and satisfaction in task-oriented groups is at the crux of understanding groups. Recent results have emphasized the importance of groups, by establishing that 'collective intelligence' of groups exceeds 'individual intelligence' (Woolley *et al.*, 2010). Studying group behavior in face-to-face interaction is the first step to understand how organisations function (Olguín and Pentland, 2010).

Various factors like leadership style (e.g. participative vs autocratic), group cohesiveness (e.g. close friends vs strangers), and goal at hand (e.g. cooperation vs competition) influence group conversational behavior. The automatic analysis of group interactions could potentially quantify the effect of such hidden factors on the group dynamics and to infer these factors from potentially huge collections of group conversation recordings in an automated and help data-driven manner. Automatically inferring group conversational context- that could potentially include the goal of the group, the type of the interaction (task-oriented vs casual), and the type of members (close friends vs strangers) -

would simplify and improve social inference. In some tasks like inference of social verticality, the group conversational context (cooperative vs competitive) moderates perceived verticality by external observers. In cooperative interactions, verticality is correlated with the one who speaks the most, whereas in competitive interactions it is correlated with the one who successfully interrupts the most (Jayagopi *et al.*, 2009b), (Raducanu and Gatica-Perez, 2010).

Automatic recognition of group interaction context is a useful module for Computer-Supported Cooperative Work (Grudin, 1994). With the advent of ubiquitous and mobile sensing platforms, novel ways of collecting and visualizing group interaction behavior have been explored as briefly discussed in Chapter 1 (DiMicco *et al.*, 2006; Nijholt *et al.*, 2006; DiMicco and Bender, 2007; Kim *et al.*, 2008; Pianesi *et al.*, 2008b) with the primary objective of influencing the group's behavior. Such applications would greatly benefit from the knowledge of the interaction context i.e. awareness about the interaction type, e.g. a cooperative vs competitive interaction, or a brainstorming vs decision-making phase.

Group meetings have different dynamics depending on the group's objective (McGrath, 1984). Competitive meetings like debates, whose primary objective is that of resolving or winning an argument, demand a different response from the members vis-a-vis that of colloborative meetings like brainstorming sessions, whose primary objective is to cooperate and accomplish a task together. Cooperative group tasks, further more, may be ordered on a continuum anchored by intellective and judgmental tasks (Laughlin and Ellis, 1986).

Our novel framework to characterize group conversational behavior defines a novel set of group nonverbal cues from individual cues. At the group level, there is no information about the identity of the individuals. Our research goal is to infer group conversational context, which in this case is group's objective, by quantifying group nonverbal dynamics. Specifically we address two problems 1. Classifying cooperative and competitive interactions and 2. Classifying brainstorming and decision-making interactions. The results of this chapter resulted in two publications (Jayagopi *et al.*, 2009a) and (Jayagopi *et al.*, 2010).

4.1 Related Work

Below, we briefly review some related works in social computing and social psychology.

The literature on modeling groups in social computing can be classified into two categories. The first category addresses offline modeling to understand groups (Gatica-Perez, 2009). We reviewed this category of literature in Section 1.3. The second category addresses novel ways of collecting and visualizing such behavior online or offline (DiMicco *et al.*, 2006; Nijholt *et al.*, 2006; Kulyk *et al.*, 2006; DiMicco and Bender, 2007; Kim *et al.*, 2008; Pianesi *et al.*, 2008b; Bachour *et al.*, 2010) with the objective of influencing the group's behavior. We review this category here. The objective of this body of research has been to directly improve human-human communication either offline or online. (DiMicco *et al.*, 2006) presented a visualization system to understand turn-taking and behavioral patterns of the participants. (Kulyk *et al.*, 2006) visualized gaze patterns also along with the turn-taking patterns. (Nijholt *et al.*, 2006) explored the possiblility of 3D virtual representation of meetings emphasizing turn-taking, gaze, and influence. (Kim *et al.*, 2008) used real-time visualized summaries of turn-taking information on mobile phones. (Bachour *et al.*, 2010) used a table that is interactive, to show the participants how much each of them speak. (Pianesi *et al.*, 2008b) provided meeting support by giving the participants an automatic multimedia feedback on their relational behavior, like a 'team-coach'. Some of the above works also present user studies about how acceptable and useful such systems are for the individuals and the group as a whole. As our work makes use of computationally not so demanding cues, our conversational context inference could simplify and enhance collecting and visualizing group behavior.

Next, we review the literature that relates to the group conversational context that we have chosen to investigate in this chapter i.e. Cooperative and competitive; and brainstorming and decision-making mainly in the social psychology literature.

Cooperative and competitive behavior among individuals in a group is well documented (Bornstein, 2003). Evidence on laboratory experiments like prisoner's dilemma show that individuals exhibit competitive behavior even if cooperation is a better strategy. Group members tend to pursue self-interest and strive to outperform the rest. Cooperative-compared with competitive-intergroup relations has been found to lead to better task performance and satisfaction in groups that make decision in a 'participative' fashion (Oostrum and Rabbie, 1995). Cooperation and competition as we see are fundamental constructs in group behavior understanding.

Laughlin and Ellis postulated that cooperative group tasks may be ordered on a continuum anchored by intellective and judgmental tasks (Laughlin and Ellis, 1986). According to them, intellective

tasks are defined as tasks for which a demonstrably correct solution exists, as opposed to decision making or "judgmental" tasks where "correctness" tends to be defined by the group consensus. Such a distinction was made to study how the performance of group versus individuals varied depending on the task type. Brainstorming, an intellectual task and Decision-making, a judgmental task are two complementary types of tasks that a task-oriented group can be engaged in.

4.2 Our Approach

We propose the following methodology to classify the group conversational context types (Figure 4.1). Assume that we have labelled group interaction data where the interactions differ in their objectives (e.g. cooperative vs competitive). Our approach uses a layered approach for classification. In the first layer, the *individual* nonverbal behavior description is obtained by extracting speaking activity and then computing features which characterize the floor occupation patterns of individuals. In the second layer, *group* nonverbal behavior is inferred by either aggregating these features (for example 'how much this group talks per unit time') or by comparing the *individual* nonverbal behavior with others' behavior (for example 'does every body take an equal number of turns or interruptions?'). The group conversational context is classified using supervised learning approach using the group behavioral cues as input.

We discuss the main blocks of our framework in the following subsections.

4.2.1 Individual nonverbal cue extraction

Firstly, we extract speaking energy and speaking status.

Speaking energy: The starting point is to compute the real-valued speaker energy for each participant using a sliding window at each time step.

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

Individual cues. From the speech segmentation, we compute Total Speaking Length [$TSL(i)$] defined as the total time that participant i speaks, Total Speaking Turns [$TST(i)$], Total Successful interruptions [$TSI(i)$], and Total Unsuccessful interruptions [$TUI(i)$] defined as the number of turns,

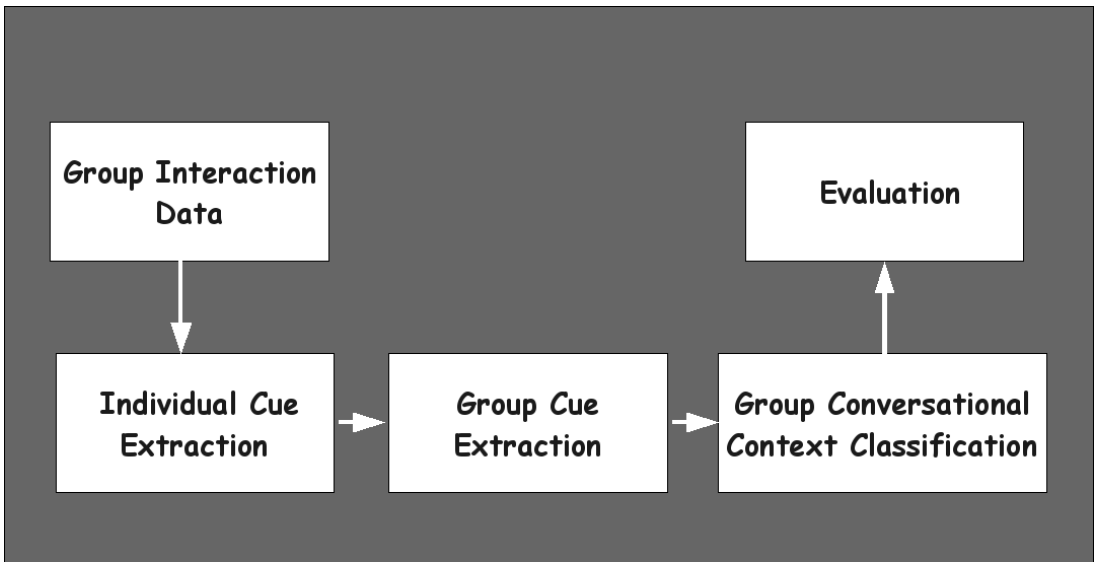


Figure 4.1. Block Diagram of our work.

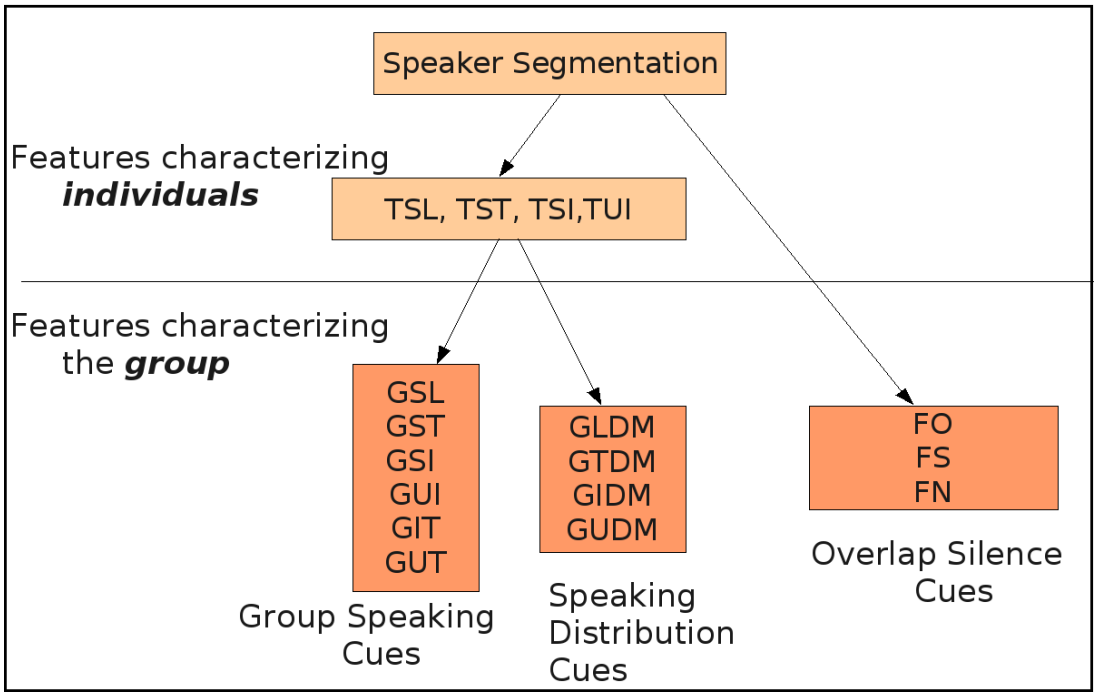


Figure 4.2. Nonverbal Cue Extraction.

successful interruptions, and unsuccessful interruptions accumulated over the entire meeting for every participant i, respectively. These features only take into account individual contributions and so contain the identity of each person. Figure 4.2 summarises the cue extraction process.

4.2.2 Group nonverbal cue extraction

Different groups differ in the way they speak. Some groups speak a lot. Some groups are silent. While some groups are more egalitarian either in nature or due to the performed task, some other groups have status differences leading to differences in the level of participation. Some groups could have lots of overlapped speech due to the nature of the participants or the social situation, while other groups prefer don't. Our group cues capture these differences.

Three types of group cues are extracted. A first set of cues characterize the participation rates of the group by accumulating it over the participants. Let D denote the duration of the meeting. We compute the following six cues from speaking length, turns, and interruptions of each of the participants:

- Group Speaking Length(GSL) $= \frac{\Sigma_i TSL(i)}{D}$
- Group Speaking Turns(GST) $= \frac{\Sigma_i TST(i)}{D}$
- Group Successful Interruptions(GSI) $= \frac{\Sigma_i TSI(i)}{D}$
- Group Unsuccessful Interruptions(GUI) $= \frac{\Sigma_i TUI(i)}{D}$
- Group Successful Interruptions-to-Turns Ratio(GIT) $= \frac{\Sigma_i TSI(i)}{\Sigma_i TST(i)}$
- Group Unsuccessful Interruptions-to-Turns Ratio(GUT) $= \frac{\Sigma_i TUI(i)}{\Sigma_i TST(i)}$

A second set of cues attempts to capture the overlap and silence patterns of a group as a whole. Let $T = D * Fps$ be the total number of frames in a meeting, S be the number of frames when no participant speaks, M be the number of frames when only one participant is speaking, and O be the number of frames when more than one participant talks. Then we define the following three cues:

- Fraction of Silence(FS) $= \frac{S}{T}$,
- Fraction of Non-overlapped Speech(FN) $= \frac{M}{T}$
- Fraction of Overlapped Speech(FO) $= \frac{O}{T}$

A third set of cues characterizes which meeting is more 'egalitarian' with respect to the use of the speaking floor i.e. everyone gets equal opportunities. Let **TSL** denote the vector composed of P elements, whose elements are $\frac{TSL(i)}{\Sigma_i TSL(i)}$ for the ith participant. Employing an analogous notation for **TST**, **TSI**, and **TUI**, these vectors are first ranked and then compared with the uniform (i.e. "egalitarian") distribution i.e. a vector of the same dimension with values equal to $\frac{1}{P}$. The comparison is done using the Bhattacharya distance (a distance measure useful to compare probability distributions and bounded

between 0 and 1). For our case 0 would correspond to a egalitarian meeting and 1 corresponds to a one-man show. This results in four cues:

– Group Speaking Length Distribution Measure (GLDM)

– Group Speaking Turns Distribution Measure (GTDM)

– Group Successful Interruption Distribution Measure (GIDM)

– Group Unsuccessful Interruptions Distribution Measure (GUDM)

These group features do not take into account individual contributions and so do not contain the identity of each person. Table 4.1 summarizes the group cues.

Glossary of Feature Acronyms	
Group Speaking Length	GSL
Group Speaking Turns	GST
Group Successful Interruptions	GSI
Group Unsuccessful Interruptions	GUI
Group Successful Interruptions-to-Turns Ratio	GIT
Group Unsuccessful Interruptions-to-Turns Ratio	GUT
Fraction of Overlap	FO
Fraction of Silence	FS
Fraction of Non-overlapped Speech	FN
Group Speaking Length Distribution Measure	GLDM
Group Speaking Turns Distribution Measure	GTDM
Group Successful Interruptions Distribution Measure	GIDM
Group Unsuccessful Interruptions Distribution Measure	GUDM

Table 4.1. Glossary of abbreviations for the group cues.

4.2.3 Group conversational context classification

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

$$\log(\frac{P(A|f_{1:N})}{P(B|f_{1:N})}) = \log(\prod_{k=1}^{N} \frac{P(f_k|A)}{P(f_k|B)}) + \log(\frac{P(A)}{P(B)}) \tag{4.1}$$

The probabilities $P(f_k|A)$ or $P(f_k|B)$ are estimated by fitting a Gaussian to the data from the respective class and the ratio of the priors are inferred from the data. When this ratio is greater than zero, the test data is assigned to class A. Otherwise to class B.

The second model is an SVM classifier, employing a linear kernel, using $(f_1, f_2, ...f_N)$ as features. This framework for two concrete classification tasks described in the rest of the chapter.

4.3 Classifying cooperative vs competitive interaction

In this section we describe the meeting dataset used for the task of classifying Cooperative vs Competitive interactions and then present the experiments and the results.

4.3.1 Meeting datasets

The AMI meeting dataset (cooperative meetings):

As explained in previous chapters, the teams in the AMI meeting dataset consisted of 4 participants, who were given the task of designing a remote control over a series of meeting sessions. Each participant was assigned distinct roles: 'Project Manager', 'User Interface Specialist', 'Marketing Expert', and 'Industrial Designer'. During each session, the team was required to carry out certain tasks to achieve the common goal.

The Apprentice meeting dataset (competitive meetings):

The data collected for our study, which we call the Apprentice dataset, belongs to the 6th season of a TV show, which was aired in early 2007. Each season starts with two groups of job candidates aspiring to work for Donald Trump, a real business tycoon in the US. Both groups are assigned a task and the team that performs better wins. The winning team receives a reward, while the losing team faces a "boardroom showdown" in order to determine which team member should be fired (eliminated from the show). We use the boardroom recordings as our source of data. On one side of the board room we have the 'candidates board' and on the other side we have the 'executive board'. The executive board is formed by Trump together with other persons (usually two) which will help him make the decision regarding the candidate who will be fired. We chose these interactions because the group's objective is competitive as against the AMI interactions which are cooperative.

The teams in the Apprentice meeting dataset have a variable number of participants (5 to 11). The group has a well-defined hierarchy, with Donald Trump being the person with highest status and the

objective of the group is to fire one of the members. Figure 4.3 shows a snapshot of both meetings.

Figure 4.3. *Top*: Snapshot from an AMI meeting, showing the participants from two side-view camera view. *Bottom*: Snapshot of an Apprentice meeting - highlighting the high-status leader (Trump) - *bottom left* and a long-shot of the board-room meeting - *bottom right*.

4.3.2 Experiments and results

For the AMI data, we extract the speaking activity cues from the four close-talk microphones attached to each of the participants. A window of 40 ms was used with a 10 ms time shift.

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

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

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

Our final dataset consists of 49 meetings (34 from AMI and 15 from Apprentice). In order to evaluate the models we adopt a leave-one-out cross-validation strategy to classify the meetings and report the classification accuracy (Table 4.2).

Features	Accuracy(%) (GNB)	Accuracy(%) (SVM-lin)
GSL	65.3	69.4
GST	69.3	69.4
GSI	63.2	69.4
GIT	85.7	83.6
FO	63.2	69.4
FS	67.3	69.4
FN	69.3	69.4
GLDM	61.2	67.3
GTDM	**93.8**	**93.8**
GIDM	71.5	69.4
GIT,GIDM	91.8	91.8
FN, GTEM	91.8	95.9
GIT,GTDM	**95.9**	**98.0**

Table 4.2. Accuracy (%) of speaking activity based nonverbal cues for classification of group conversational context. In the caption, GNB stands for Gaussian Naive Bayes classifier and SVM-lin is the short form of SVM using a linear kernel.

While interpreting the results, it is to be noted that due to the difference in the number of samples between the two datasets, if an algorithm always labels all test cases as 'AMI meetings' it would perform with an accuracy of 69.4%. Also, a random prediction would give an accuracy of 50%. As unsuccessful interruptions were unavailable for the Apprentice dataset, for this classification task, we did not use Group Unsuccessful Interruptions (GUI) and Group Unsuccessful Interruptions Distribution Measure (GUDM) features. All other features described in Section 4.2.2 were extracted.

The results show that features like Fraction of Overlapped Speech(FO), Fraction of Silence(FS), Fraction of Non-Overlapped Speech(FN), Group Speaking Length(GSL), Group Speaking Turns(GST), and Group Speaking Interruptions(GSI) were not discriminative. Though we expected that in competitive meetings, the interruption rate (GSI) and the proportion of overlap (FO) would be more, our classification results did not show that. On the other hand, meetings could be discriminated when using the proportion of interruptions in the turns (GIT) and the distribution of turns and interruptions among participants (GTDM and GIDM). Figure 4.4 shows the empirical distribution of the two features - GIT and GTDM. As one can observe, these two features are discriminative. Figure 4.5 illustrates how the SVM with a linear kernel in the joint space of GIT and GTDM classifies the two meeting

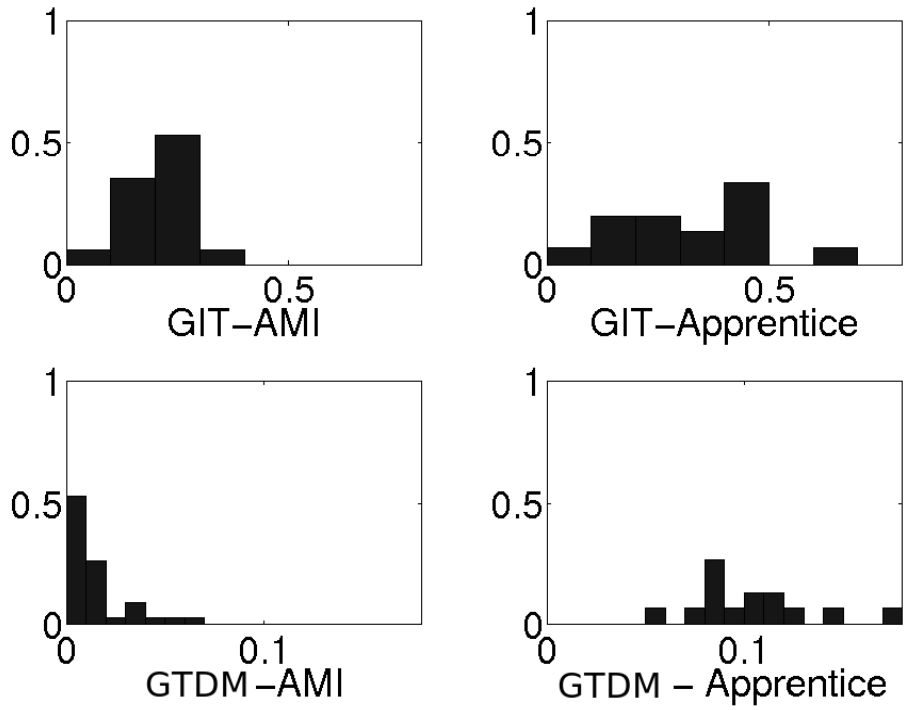


Figure 4.4. Normalized histograms of GIT and GTDM in the two meeting datasets.

datasets. Also, it was interesting to observe that the features derived from speaking length were not as effective, although they were the best for other tasks like estimating the most dominant person in a meeting (Jayagopi *et al.*, 2009b).

To conclude, the distribution of speaking turns which tends to indicate how 'egalitarian' an interaction is, captures the competitiveness among the group members very effectively. Also, along with a slightly complementary feature (the proportion of interruptions in the turns), this feature classifies the meeting type with very high accuracy. Although the dataset is small, this framework is quite interesting and promising to characterize group behavior.

Figure 4.6 shows a snapshot from a demo, with an image from the center-view camera, some individual cues, and some group cues visualized.

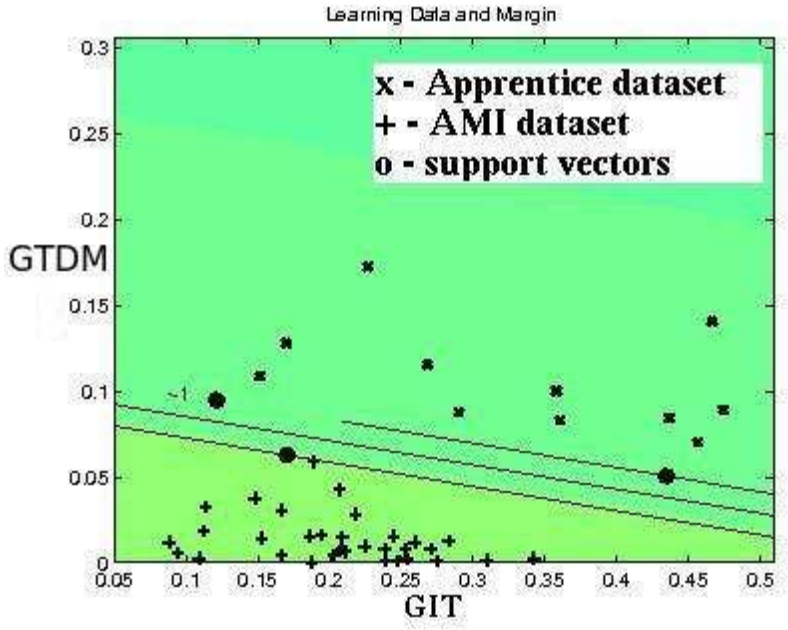


Figure 4.5. Classification using SVM in the feature space of GIT and GTDM.

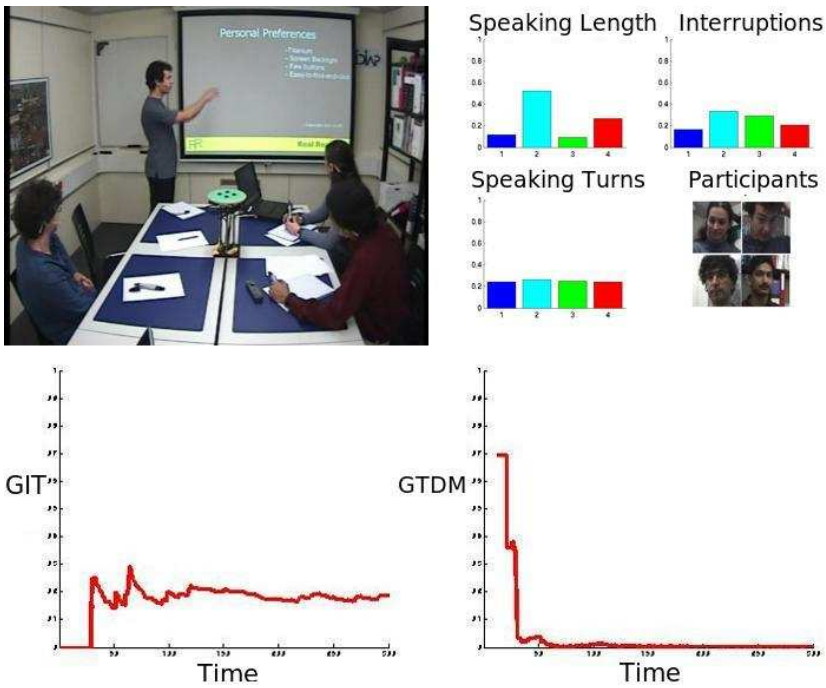


Figure 4.6. *Top-left*: Snapshot from the AMI meeting, showing the participants from the center-view camera. *Top-right*: Distribution of speaking length, speaking turns, and successful interruptions among the participants. *Bottom-left*: The evolution of the Group-Interruption-to-Turns Ratio with time. *Bottom-right*: The evolution of the Group Turn Distribution Measure with time.

4.4 Classifying brainstorming vs decision-making interaction

Next, we investigate the second problem of classifying brainstorming vs decision-making interactions on a larger dataset, recorded using a mobile recording platform. Also, the dataset has the same group of people participating in both types of interaction, allowing an important dimension to be controlled in the discrimination study. In this section we describe the meeting dataset used for this task and then present the experiments and the results.

4.4.1 Meeting dataset

The dataset was collected from 24 groups of four members each. Each participant wore a sociometric badge (Figure 4.7) - a wearable electronic badge with multiple sensors collecting interaction data, developed at Human Dynamics Group, MIT Media Lab by Daniel Olguín Olguín (Olguín and Pentland, 2008). By interacting with other badges it can collect proximity data, other badges in direct line of sight, movement data, and speech features. Speech features collected by the badge include pitch, tone, volume, etc. Due to privacy concerns, content of speech or any other features that may identify the speaker was not collected. The microphone of the sociometric badges collected speech variation data sampled at 50Hz, which is immediately processed on the badge so that only the processed data is saved on its SD (Secure Digital) card. The badges communicate with each other via 2.5GHz radio which allows synchronization error to be less than 0.003 msec. Figure 4.8 shows an interacting group wearing sociometric badges.

The interaction task given to subjects was based on a modification of the game "Twenty-Questions", replicating Wilson's experiments (Wilson *et al.*, 2004). Each round consisted of two phases. In the first phase, each group was given a set of ten yes/no question-and-answer pairs, related to the object that the group has to guess correctly. For example one question could be 'Is it used for entertainment' and the answer could be 'No'. The groups were given 8 minutes to collaboratively brainstorm as many ideas that satisfy the set of question-and-answers. We label these interactions as 'brainstorming'. Then in the second phase, groups were given 10 minutes to ask the remaining ten questions of the Twenty-Question game to determine the correct solution. As this problem-solving phase mainly involved the group making decisions about the subsequent questions, we regard and label them as 'decision-making' interactions. In the second phase, groups were asked to select a leader

Figure 4.7. Sociometric badge developed by Human Dynamics group, MIT Media Lab (Olguín and Pentland, 2008).

Figure 4.8. Example of an interacting group wearing sociometric badges around the neck.

among themselves that would be the question-asker who communicates with the experimenter.

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

throughout the two measured rounds.

The dataset we used for our experiments was 9.8 hours of group conversational recordings and was collected by Taemie Kim, Human Dynamics Group, MIT Media Lab. We used the data from both collocated (i.e. face-to-face) and distributed (i.e. remote) settings to understand which group nonverbal cues were the most effective in each of the two settings.

4.4.2 Experiments

For this dataset, the speaking status was obtained by thresholding the speech variation data collected by the sociometer. The speaking status was downsampled to 10 frames per second. As described in Section 4.4.1, we have 24 participant groups, solving two "Twenty-questions" games, one in collocated and the other in distributed settings. Each game involved a brainstorming phase followed by a decision-making phase. In order to model the difference between brainstorming and decision-making interactions, we define the following four datasets and three binary classification tasks.

1. Dataset A - consists of 24 brainstorming meetings in collocated scenario.
2. Dataset B - consists of 24 decision-making meetings in collocated scenario.
3. Dataset C - consists of 24 brainstorming meetings in distributed scenario.
4. Dataset D - consists of 24 decision-making meetings in distributed scenario.

Based on the datasets we define three classification tasks.

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

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

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

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

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

4.4.3 Results

We first analyze the performance of single cues. Figure 4.9 shows for Task 1 (collocated setting). Random performance for all the tasks is 50%. Though we experimented with two different classifiers, as described in Section 2.3, we report the results using the Gaussian Naive Bayes classifier only as the results are similar when a linear SVM is employed. Fraction of Silence (FS), Group Speaking Length (GSL), and Group Unsuccessful Interruptions (GUI) were the top performing cues with a performance of 81.3%, 81.3%, and 79.1% respectively. Figure 4.10 shows the performance of the group cues for Task 2 (distributed setting). Fraction of Silence (FS), Fraction of Overlap (FO), and Group Speaking Length (GSL) were the top performing cues with an accuracy of 79.2%. For Task 3, a similar trend was observed. Fraction of Silence (FS), Group Speaking Length (GSL), and Fraction of Overlap (FO) gave the best classification result with an accuracy of 80.2%, 78.1%, and 74% (Figure 4.11). All these results are statistically significant compared to the random performance at 5% level using a standard binomial test. The results suggest that some of the investigated features indeed have discriminating power. Also, it is interesting to observe the following trend: Most groups have higher Fraction of Silence during brainstorming and higher Group Speaking Length and Fraction of Overlap while making decisions. A possible reason may be that during brainstorming groups tend to have higher cognitive load and hence speak less as compared to decision-making interactions

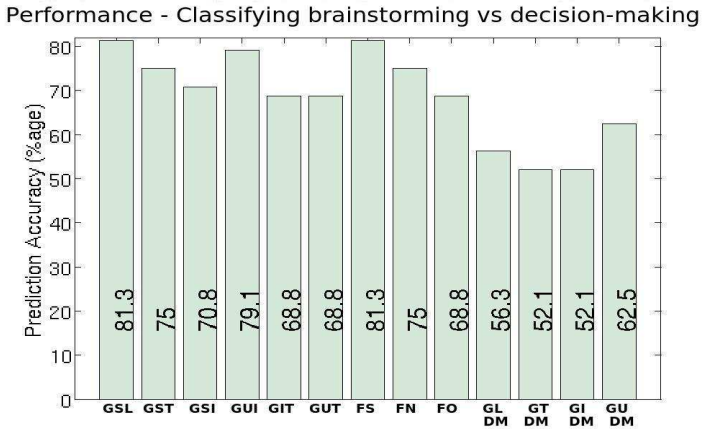


Figure 4.9. Performance of the group cues on classifying the brainstorming and decision-making meetings during collocated setting (Task 1).

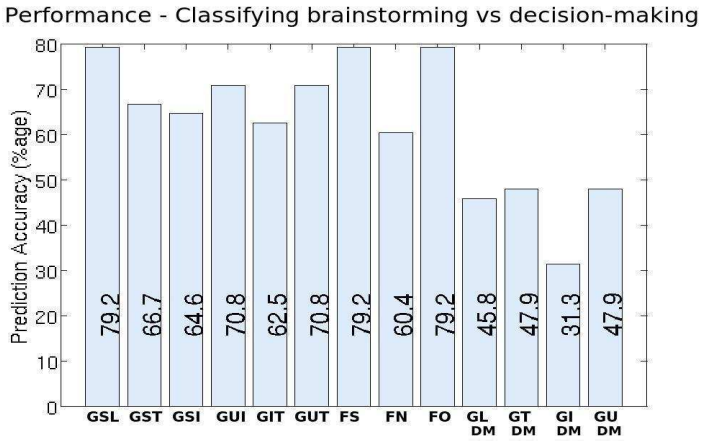


Figure 4.10. Performance of the group cues on classifying the brainstorming and decision-making meetings during distributed setting (Task 2).

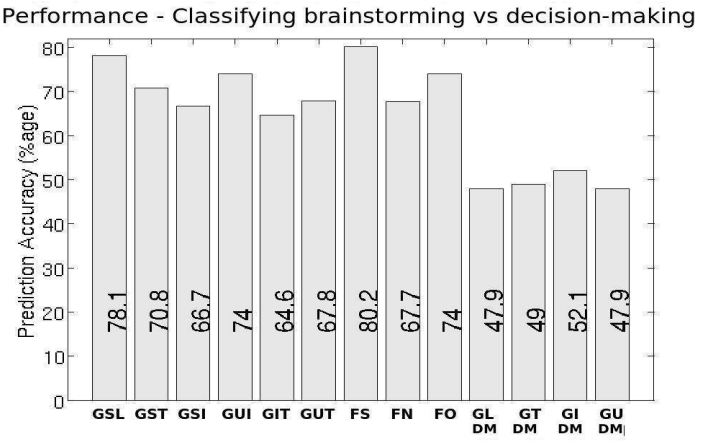


Figure 4.11. Performance of the group cues on classifying the brainstorming and decision-making meetings (Task 3).

Later, we also combined the cues to investigate if there is complementarity among them. Figure 4.12 shows the classification performance of some combinations using the Gaussian-Naive Bayes classifier for each of the three tasks. The combination of Fraction of Silence (FS) and Fraction of Overlap (FO) improves the classification accuracy to 83.3% in the collocated case (Task 1). When Group Speaking Length (GSL), Group Speaking Turns (GST), and Group Unsuccessful Interruptions (GUI) were added the accuracy improved to 87.5%. The combination of Fraction of Silence (FS) and Group Speaking Length (GSL) improves the classification accuracy to 81.3% in the distributed setting (Task 2). For the combined dataset (Task 3), the combination of Fraction of Silence (FS), Group Speaking Length (GSL), and Group Unsuccessful Interruptions (GUI) improved the classification accuracy to 81.3%.

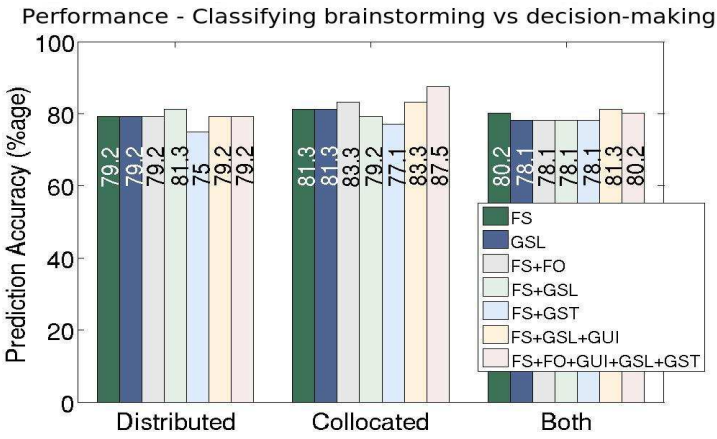


Figure 4.12. Performance of combination of group features on predicting the brainstorming and decision-making meetings.

To conclude, we could discriminate these interactions with an accuracy of up to 87.5% and 81.3% in the collocated and distributed setting respectively. The group adaptation i.e z-normalization step helps in improving performance and also tackling inter-group differences (as the mean behavior is subtracted out).

4.5 Discussion and Conclusion

Overall our study suggests the following.

Summary of results. In this chapter we investigated the problem of characterizing group conversational context using nonverbal turn taking behavior. Specifically, we presented a supervised learning approach that works at two layers, with the first layer capturing individual behavior and the second layer capturing group behavior. We apply our framework for two classification problems 1. Classifying cooperative vs competitive interactions 2. Classifying brainstorming vs decision-making interactions. Our methods produce an accuracy of up to 98% for the first problem and 87% for the second problem, which is encouraging and suggests that the characterization of entire group by the aggregation (both temporal and person-wise) of their nonverbal behavior is promising. The most effective features for classifying cooperative vs competitive interactions were : Group-Interruptions-to-Turns Ratio (GIT) and the Group Turn Distribution Measure (GTDM), whereas for classifying brainstorming vs decision-making, Fraction of Silence (FS), Fraction of Overlap (FO), and Group Speaking Length (GSL) were the best.

Limitations. As the size of the data set is relatively small, many of the observed performance differences between the best cues are not statistically significant at 5% level although the difference between the best cues and random performance are statistically significant. Our work shows the promise of characterizing group behavior using just an instance of cooperative and competitive interaction; and brainstorming and decision-making interaction. More such studies need to be done with varied and larger datasets to understand the generalizability of the results, despite the fact that collecting such data is a rather intensive and expensive task which involves mobilization of participants, and many a times who do not already know each other.

Possible Extensions. Future work should use more data and an expanded feature set to include prosodic cues and temporal aspects of cues to explore generative models that would characterize brainstorming and decision-making interactions better. Also, with more data showing statistical significance with cue fusion as compared to single cues would be possible. In the second place, other group conversational contexts, apart from group's objective or other group objectives could also be interesting to study. As more of these contexts are studied and understood, an online detection of group interaction contexts in real situations would also be a possibility in the future. In the third place, future work could also investigate how to build a general model of social verticality that works in both competitive and cooperative scenarios. Finally, investigating the group behavior of 'better' performing groups in both brainstorming and decision-making scenarios could be an interesting study.

Chapter 5

Mining group nonverbal conversational patterns

The methods to investigate communicative behavior in small groups have mostly used manual coders and self-reported data. As discussed in previous chapters, with the advent of cheap audio and video sensors and improved perceptual processing methodologies, computational models of social interactions are beginning to appear, particularly using nonverbal cues (Gatica-Perez, 2009). The methods studied so far in the computational literature have mostly used supervised learning approaches. In this work we propose an unsupervised discovery approach to automatically mine group communicative behavior patterns in conversation, in a principled, robust, and data-driven fashion.

This chapter presents a novel framework to address the problem of automatically discovering group conversational patterns from nonverbal cues extracted from brief observations (or slices) of interaction. In Chapter 4, we showed the advantages of characterizing the behavior of a group by descriptors of the joint individual behavior. Characterizing the group as a whole allows the study of specific group constructs like cooperation vs competition (Jayagopi *et al.*, 2009a). In this chapter, we propose and analyze a novel descriptor of interaction slices - a bag of group nonverbal patterns. This group descriptor captures the behavior of the group as a whole and integrates its leader's position in the group. We then propose the use of principled probabilistic topic modeling (Steyvers and Griffiths, 2007) on the group descriptors, we are able to discover group interaction patterns in an unsupervised

way. We have used the AMI meeting corpus as our data. We have also carried out an objective evaluation of our framework using human judgment with multiple annotators.

The specific contribution of this work is as follows. First, we address the largely unexplored problem of discovering group nonverbal patterns in an unsupervised fashion. Second, we define a new group behavioral descriptor on slices of group conversational data that is robust to several factors occuring in realistic interactions. Third, we study interaction slices of varying duration to understand the discovery process at different time scales. Fourth, we propose the use of topic models, and more specifically Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) and propose new topic-based ways of characterizing groups by aggregating group behavior over multiple interactions. Finally, we show that the topics discovered by our model are meaningful using ground-truth produced from external observers of the interaction. The material presented in this chapter was published originally in (Jayagopi and Gatica-Perez, 2009, 2010).

The chapter is organized as follows: Section 5.1 reviews the literature on automatic modeling of behavior in small groups. Section 5.2 introduces our approach. Section 5.3 describes the cue extraction process, the definition of the NVPs, and the LDA model. Section 5.4 introduces the data set used in the experiments. Section 5.5 presents and discusses the experimental results. Section 5.6 summarizes the findings of our work and provides concluding remarks.

5.1 Related Work

This work addresses discovery of group behavior in face-to-face interactions using infrastructure based sensors and topic models. We have already reviewed the literature on analyzing behavior of individuals or groups using infrastructure based sensors in Chapter 1, and most of them have employed supervised approaches or correlation based unsupervised methods. In this section, we review the relevant literature on topic models and describe few applications of topic models to discover 'human-related' activities.

Topic models are tools to cluster and retrieve documents, originally proposed in the text modeling literature. Probabilistic Latent Semantic Analysis (PLSA) proposed by Hofmann, represents a document as a mixture of topics, where each topic is a probability distribution over words (Hofmann, 1999). Later, LDA extended PLSA to represent topics as being sampled from a Dirichlet distribution,

of which PLSA represents a special case (Blei *et al.*, 2003). LDA, thereafter, was further extended notably in two directions. One, to make it represent topics in a hierarchical fashion called Hierarchical Dirichlet Process (Teh *et al.*, 2006) and the other to include the authorship information, called Author Topic Model (ATM) (Rosen-Zvi *et al.*, 2004). Adapting the topic models to applications other than text modeling, involves defining a bag-of-words that suits the application. We next review couple of such works that adapt topic models to new 'human-related discovery-type' applications.

Indoor daily routines, like commuting and office work, were discovered using LDA in (Huynh *et al.*, 2008) using wearable sensors and accelerometer data with applications in elderly care, office space management etc. Farrahi et al. (Farrahi and Gatica-Perez, 2008, 2010) discovered outdoor routines using location and proximity data recorded using mobile phones. The experiments used both LDA and ATM. The work has applications in understanding large-scale human mobility patterns and epidemiology.

The PLSA model was used for human action discovery (Niebles *et al.*, 2008). Both normal and abnormal scene-level activity patterns were discovered through co-occurrence analysis of low-level relying on low-level features like location and velocity and their statistics and topic models (Li *et al.*, 2008; Varadarajan and Odobez, 2010; Xiang and Gong, 2008). (Li *et al.*, 2008) employed PLSA, while (Xiang and Gong, 2008) employed a hierarchical version of PLSA. (Varadarajan and Odobez, 2010) proposed a novel Probabilistic Latent Sequential Motif Model to represent multiple activities. Such discoveries have applications in outdoor surveillance of humans and other moving objects.

5.2 Our Approach

Different individuals have different speaking, gesturing, and gazing styles. Group dynamics evolve out of these individual styles constrained by social rules. While some groups speak or interrupt a lot, others tend to be more silent. While some groups are more egalitarian either in nature or due to the performed task, some other groups have status differences leading to differences in the level of participation.

In order to capture such differences in a data-driven fashion, we first define group descriptors (bag-of-NVPs) and then cluster them. So our approach consists of two stages. First, analogous to how topics could be infered from a text collection by representing documents in a corpus as histograms of words

(so-called bags-of-words), we propose to discover the group behavior patterns by characterizing the group dynamics in terms of bag-of-group NVPs or bag-of-NVPs for short. In a second stage, we use the Latent Dirichlet Allocation (LDA) topic model to discover topics by considering co-occurence of NVPs i.e. NVPs that tend to co-occur get clustered as NVPs belonging to the same topic. It is important to note that the topics discovered by LDA are not to be confused with the actual topic that the group discusses. We hypothesize that there is enough structure in the behavioral patterns that by clustering them by a method that exploits co-occurrence, we would observe meaningful 'group behavior topics'. Following our analogy with text, in our analysis and discussion, we interchangeably use 'words' and 'NVPs' to refer to the group nonverbal behavior descriptors.

Figure 5.1 shows the overview of our work. First, we extract low-level nonverbal cues from interaction slices of small-group meetings. We then quantize these cues to produce a bag-of-NVPs. Finally, we mine the collection of bags-of-NVPs using a probabilistic topic model to discover joint patterns of group conversational behavior. We experiment with meeting slices of different duration, to study the effect on the bag representation and the discovery process.

Various nonverbal cues are known to be correlated with interpersonal relations (Hall *et al.*, 2005). Building our group behavioral descriptor as a bag-of-NVP has the following advantages:

- it facilitates fusion of individual cues;
- through aggregation over people and time, the cues are made more robust compared to low-level individual cues;
- the use of group NVPs facilitates the eventual comparison of groups of varying sizes;
- it allows for the usage of principled methods for unsupervised learning.

The proposed bag-of-NVPs includes two types of patterns: *generic* group patterns and *leadership* patterns. The generic group patterns are descriptors about the group as a whole without taking the identity of the interactions into account. The leadership patterns are descriptors about the "leader" in the group, assuming that such a role is played by a team member (a situation that is pervasive in the workplace). In other words, the generic group patterns can describe any group, whereas the leadership patterns apply to those groups with a leader. In our study, such a split allows us to consider the effect of the predominant person of the group. Though in this work we consider conversational patterns alone for our bag-of-NVPs, this framework can be easily extended to include various other multimodal descriptors - like gazing or 'looking-while-speaking' patterns as well.

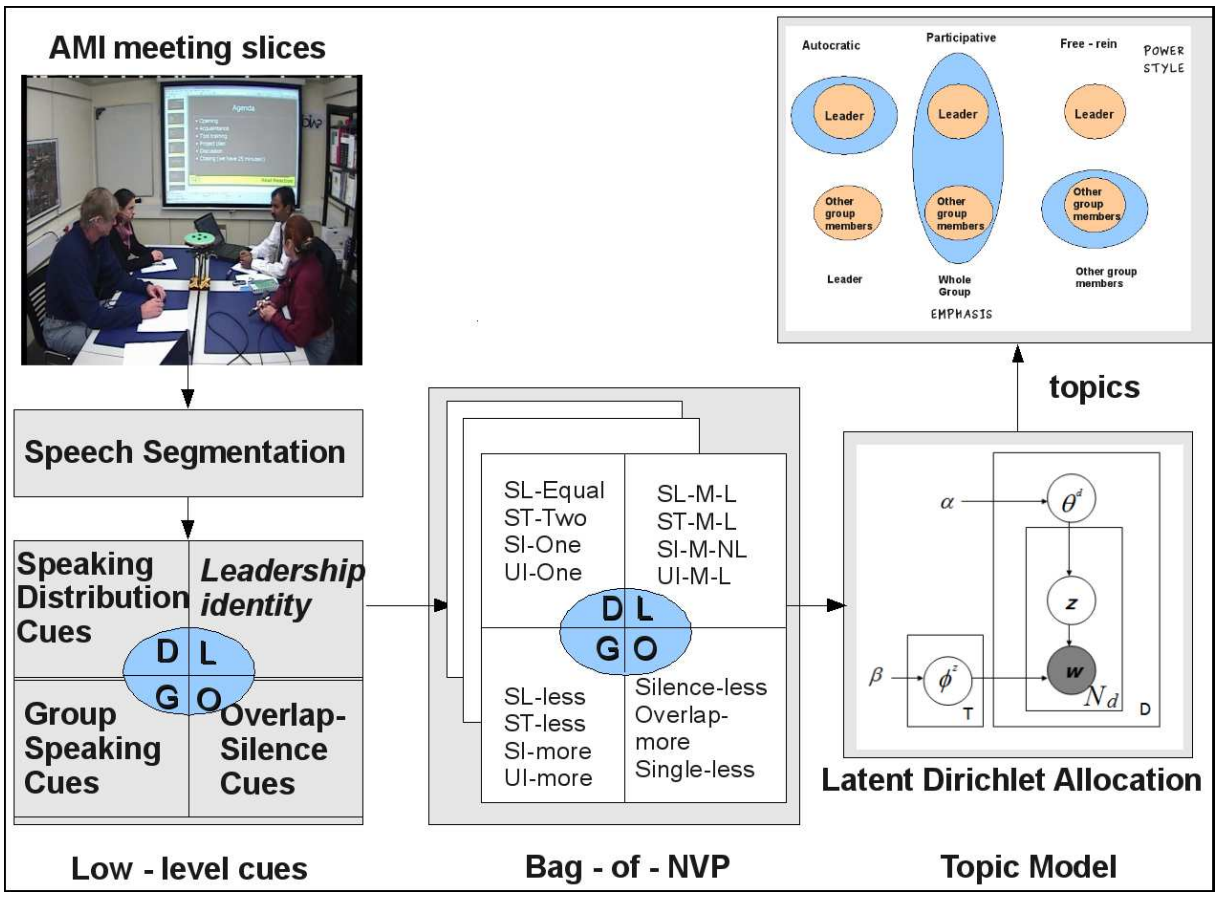


Figure 5.1. Overview of the group NVP discovery process using topic models.

5.3 Low level Cue extraction, Bag-of-NVPs, and the Topic model

5.3.1 Low level nonverbal cue extraction

We extract the following speaking activity based cues (see Figure 5.2). For each interaction slice from a given group conversation recorded with close-talk microphones, we first perform a binary speech vs silence segmentation for the N_p group members at each time step (five frames per second) (Dines *et al.*, 2006).

As in Chapter 4, the individual cues involve extracting for the i th participant: Total Speaking Length [TSL(i)], Total Speaking Turns [TST(i)], Total Successful Interruptions [TSI(i)], and Total Unsuccessful Interruptions [TUI(i)]. where $i = 1, 2, ..N_p$.

The group cues are of three types:

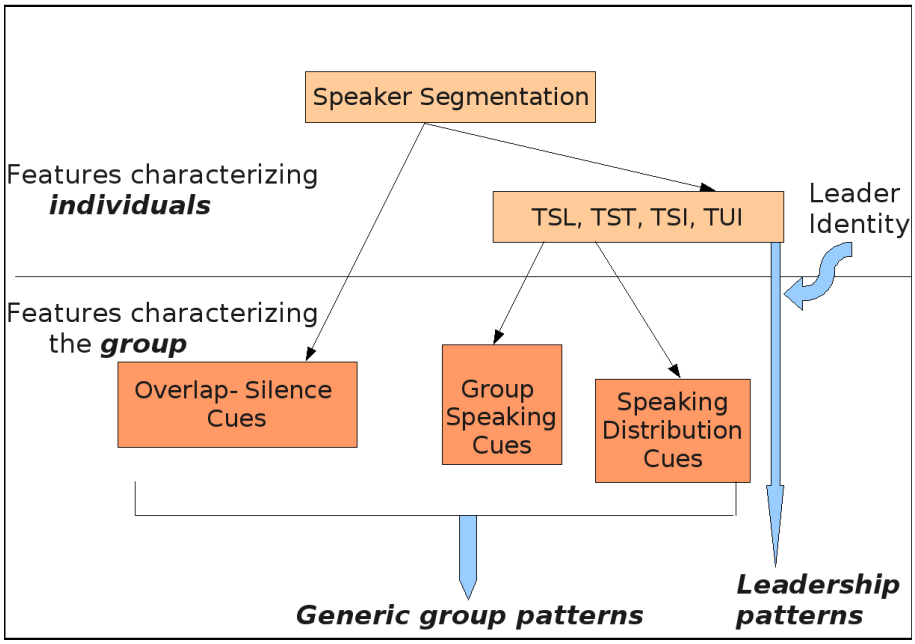


Figure 5.2. Diagram showing the features to characterize individual and group behavior (generic-based and leadership-based) extracted in our approach. See main text for details.

1. *Speaking distribution cues*

   Let **TSL** denote the vector composed of N_p elements, whose elements are TSL for each participant after normalization (elements sum up to one). We employ an analogous notation for **TST**, **TSI** and **TUI**.

2. *Overlap-Silence cues*

   As in Chapter 4, from the speaking status of all the participants, we extract Fraction of Overlapped Speech (FO), Fraction of Silence (FS), Fraction of Non-overlapped speech (FN).

3. *Group Speaking cues*

   As in Chaper 4, from speaking length, turns and interruptions of each of the group members, the following additional features are computed to characterize their joint group behavior: Group Speaking Length (GSL), Group Speaking Turns (GST), Group Successful Interruptions (GSI), Group Unsuccessful Interruptions (GUI), Group Successful Interruptions-to-Turns Ratio (GIT), Group Unsuccessful Interruptions-to-Turns Ratio (GUT).

5.3.2 Bag-of-NVPs generation:

As we can observe, extracting the group cues so far has followed the same procedure as in Chapter 4. From here onwards the two frameworks diverge in their approach. In this framework, we then quantize these group cues to produce a bag-of-NVPs. Our bag model includes two types of patterns. The generic group patterns characterize the group conversational behavior whereas the leadership patterns characterize the leader‘s conversational behavior.

Generic group patterns The generic group patterns themselves are of three types - *Speaking Distribution patterns* describe whether all the group members get equal opportunities to occupy the floor etc. *Overlap-Silence patterns* capture the behavior about the competition to capture the floor and finally the *Group Speaking patterns* capture the fact whether a particular group speaks, interrupts, etc, more or less compared to the average level. We explain the construction of each of the patterns in the following.

Speaking Distribution patterns: We quantize each of the vectors **TSL**, **TST**, **TSI**, **TUI** directly into one of the five classes - *Silence*, *One*, *Two*, *Rest*, *Equal* - to describe a group. The class depends on whether silence (‘0’), one-person (‘1’), two-person (‘2’), three or more (‘3’) or all people (‘4’) share most of the probability mass for a particular nonverbal cue. We expect egalitarian groups to belong to class ‘4’. The goal is to map a joint cue over an interaction slice (e.g. speaking length) into a prototypical case (e.g. an interaction pattern in which all people talk about the same time, one person spoke most of the time, etc) where people identity is not important, and therefore makes the description generic. The actual rule is described as follows: Let $SortedVector$ represent the input vector corresponding to an individual nonverbal cue after sorting it in descending order. The output class is ‘1’ if the first element of $SortedVector$ satisfies the condition $SortedVector(1) > 2 * \frac{1}{N_p}$. The output class is ‘2’ if $SortedVector(1) + SortedVector(2) > 3 * \frac{1}{N_p}$. and the output class is ‘4’ if $SortedVector(N_p) > \Delta$, where Δ represents a small interval like 0.05 or 0.1 (representing the minimum probability mass value that a person should have so that the interaction belongs to class ‘4’). Finally, the output class ‘3’ is used as a catch-all class. Figure 3 shows an example histogram ($SortedVector$) for each of the classes other than silence for a group with $N_p = 4$.

The 20 words corresponding to the egalitarian speaking patterns are *SL-Silence*, *SL-One*, *SL-Two*, *SL-Rest*, *SL-Equal*; *ST-Silence*, *ST-One*, *ST-Two*, *ST-Rest*,*ST-Equal*; *SI-Silence*, *SI-One*, *SI-Two*, *SI-Rest*, *SI-Equal*; and *UI-Silence*, *UI-One*, *UI-Two*, *UI-Rest*, *UI-Equal*.

Overlap-Silence patterns: We quantize each of Fraction of Overlapped Speech, Fraction of Silence, Fraction of Non-overlapped speech into one of two classes - *more* and *less*. This quantization depends on the relative value of the considered group conversation to the average value computed over the entire conversation dataset. If the current value is more than the average, we quantize it as *more*. Otherwise, we quantize as *less*. The 6 words corresponding to the Overlap-Silence patterns are *Overlap-more*, *Overlap-less*, *Silence-more*, *Silence-less*, *Single-more*, *Single-less*.

Group Speaking patterns: We quantize each of Group Speaking Length, Group Speaking Turns, Group Speaking Interruption, Group Speaking BackChannels, Group Speaking Interruption-to-Turns Ratio, Group Speaking Backchannels-to-Turns Ratio into one of two classes - *more* and *less*, similar to the extraction of Overlap-Silence patterns explained in the previous paragraph. The 12 words corresponding to the Group Speaking patterns are *GSL-more*, *GSL-less*, *GST-more*, *GST-less*, *GSI-more*, *GSI-less*, *GUI-more*, *GUI-less*, *GIT-more*, *GIT-less*, *GUT-more*, *GUT-less*.

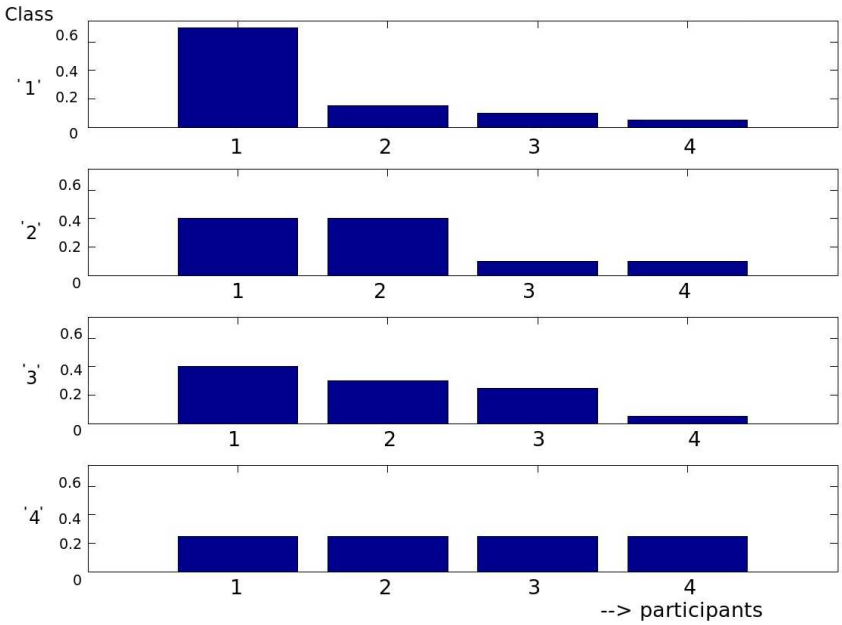


Figure 5.3. Example joint histograms for each of the Speaking Distribution NVPs other than *Silence*.

Leadership patterns As discussed in Section III, very often there are meetings with a designated leader (e.g. a manager). Social verticality in groups has been shown to be correlated to floor occupation related nonverbal cues (Hall *et al.*, 2005). Previous works have shown that the person with the highest speaking time correlates with the most dominant person (Jayagopi *et al.*, 2009b), highest number of speaking turns correlates with role-based status (Jayagopi *et al.*, 2008b) and highest

number of successful interruptions signals real status and power (Raducanu *et al.*, 2009). In order to capture the leader 's position in the group, we add three more words to the NVP vocabulary for each of the 4 sets of features to indicate whether the designated leader ('L') or someone else ('NL') is the one who has the maximum. When the interaction slice is silent, we mark the class as silence ('Silence'). For example the presence of *SL-M-L* means that in this time slice, the leader has the maximum speaking length and the presence of SL-M-Silence means that no one speaks in this interaction slice. Together with the words that characterize the generic group patterns, these words describe the position of the leader. The 12 words corresponding to the leadership patterns are SL-M-Silence, SL-M-L, SL-M-NL; ST-M-Silence, ST-M-L, ST-M-NL; SI-M-Silence, SI-M-L, SI-M-NL ; UI-M-Silence, UI-M-L, UI-M-NL. Please note that SL-M-L is not equivalent to SL-One. While SL-M-L says the leader speaks the most, SL-One says there is one person dominating the discussion. Consider this typical scenario where a leader is challenged by another participant. In this case the leader could speak the most (pattern SL-M-L appears). But the discussion involves two people, hence pattern SL-Two (instead of SL-One) also co-occurs.

The overall size of the NVP-bag vocabulary is 50 and each document (i.e. group interaction slice) contains exactly 12 words. A significant advantage of our representation is that it is robust to the number of participants and hence allows the comparison of groups of different sizes. Also, the framework easily allows the possibility of increasing the size of the vocabulary by considering more nonverbal cues that are of behavioral interest, in a similar fashion.

Robustness of Bag-of-NVPs By construction, the bag-of-NVPs is tolerant of minor variations in the observed low-level cues. So, the bag-of-NVPs are robust with respect to slight variation in individual cues, relative proportion of the group cues, and number of participants. We illustrate this using simple examples. Consider a group of four particpants interacting for five minutes (300 s), and let the speaking turns of individual particpants be distributed as follows: (40, 10, 10, 6). The group speaking turns for the four participants is 66/300. Let us now assume that the average group speaking turns estimated from the corpus is 40/300. Then this group interaction is mapped to ST-more. Also, it is mapped to ST-One, showing that there is one person dominating the interaction as he has more than 60% of the turns. Now, consider the following perturbations in

1. Individual cues: Even when we perturb the individual cues to say (35, 10, 10, 6), this interaction still is mapped to ST-more NVP.

2. Relative proportion of the group cues: If we perturb the pace of the interaction, resulting in more turns (1.5 times) for each of the participant obtaining (60, 15, 15, 9) as compared to (40, 10, 10, 6). These cues are again mapped to ST-One, which means that there is still one person dominating. These egalitarian cues capture the status hierarchy independent of the pace of the interaction.
3. number of participants: Consider the scenario of adding another participant and let the speaking turns then be (38, 8, 8, 8, 4), this interaction would still be mapped to ST-more and ST-One NVPs.

As the example shows, the bag is insensitive to situations, like the above, which occur often in group conversations.

5.3.3 Latent Dirichlet Allocation (LDA) topic model

Topic models, as mentioned in Section 5.1, are probabilistic generative models that were originally used in text modeling. In Latent Dirichlet Allocation (Blei *et al.*, 2003), a text document is modeled as a distribution over topics, and a topic as a multinomial distribution over words. The topics discover patterns based on word co-occurrence.

Let there be D documents in a corpus and let a document contain N_d words. Let V denote the total number of unique words in the corpus. The probability of a given word w_n assuming T topics is $p(w_n) = \sum_{t=1}^{T} p(w_n|z_n = t)P(z_n = t)$, where z_n is a latent variable indicating the topic from which the n^{th} word was drawn. Each document is generated by choosing a distribution over topics $p(z = t) = \theta_t^{(d)}$. Each topic is characterized by a word distribution $p(w|z = t) = \phi_w^{(t)}$ over the vocabulary of words V. In LDA, $p(\theta)$ is a Dirichlet(α) and $P(\phi)$ is a Dirichlet(β), where α and β are hyperparameters (see Figure 5.4). Given α and β, the joint distribution of the set of all words **w**, topics for each of the words **z**, θ, ϕ, in a given document is given by

$$p(\mathbf{z}, \mathbf{w}, \theta, \phi|\alpha, \beta) = \prod_{i=1}^{N_d} p(w_i|z_i, \phi)p(z_i|\theta)p(\theta|\alpha)p(\phi|\beta) \tag{5.1}$$

where z_i is the topic assignment of the i^{th} word.

We first infer the posterior distribution over **z** for a given document (**w** is given) by marginalizing over θ and ϕ, then estimate parameters θ and ϕ using word-topic and document-topic counts. Later we interpret the T topics using the top words (with highest probability) and the documents as mixture of

these topics (Griffiths and Steyvers, 2004; Steyvers and Griffiths, 2007). To estimate $p(\mathbf{z})$, we use Gibbs sampling (a Markov Chain Monte Carlo (MCMC) type method (MacKay, 2003)) where we sample sequentially each component, z_i, conditioned on the rest of the components, $\mathbf{z}_{-i}$.

$$p(z_i = t|\mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) = \frac{p(\mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{z}_{-i}, \mathbf{w}_{-i}|\alpha, \beta)} \tag{5.2}$$

The numerator of equation 2 can be further expanded as

$$p(\mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\mathbf{w}|\mathbf{z}, \beta)p(\mathbf{z}|\alpha) \tag{5.3}$$

By integrating over ϕ, we can derive $p(\mathbf{w}|\mathbf{z}, \beta) = \int p(\mathbf{w}|\mathbf{z}, \phi)p(\phi|\beta)d\phi$. The assumption of a Dirichlet prior for $p(\phi|\beta)$ and the Dirichlet distribution being the conjugate prior for multinomial distribution $p(\mathbf{w}|\mathbf{z}, \phi)$, helps us obtain $p(\mathbf{w}|\mathbf{z}, \beta)$ in closed form. By integrating over θ we can obtain $p(\mathbf{z}|\alpha)$, the second term in equation 3. Following a similar procedure the denominator of equation 2 can also be obtained. After a burn-in period, this procedure of sampling sequentially all the components of $\mathbf{z}$ yields a stationary distribution which corresponds to the probability distribution $p(\mathbf{z})$. For more details about implementing the Gibbs sampling procedure for an LDA topic model the readers should refer to (Steyvers and Griffiths, 2007; Heinrich, 2005).

5.3.4 From interaction slices to group characterization

Using the notations in the preceding subsection, any meeting slice can be represented by its topic distribution $p(z|d)$. When multiple slices of interaction are available for a particular chosen group g, $d \in D_g$, the aggregated group description can be expressed as

$$\begin{aligned} p(z|g) &= \sum_d p(z, d|g) \\ &= \sum_d p(z|d, g)p(d|g) \\ &= \frac{1}{|D_g|} \sum_{d \in D_g} p(z|d) \end{aligned} \tag{5.4}$$

This distribution can then be used to characterize and compare groups.

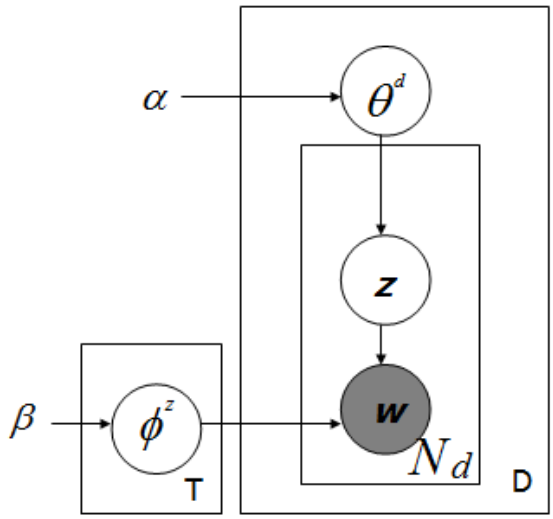


Figure 5.4. Latent Dirichlet Allocation (LDA) model

5.4 Meeting data

We use 37 meetings from the AMI corpus consisting of 10 different sets of participants (i.e. groups which do not have any member who is common). As mentioned in Chapter 2, each group consisted of four participants, who were given the task of designing a remote control over a series of meeting sessions. The level of previous acquaintance among the group members varied from being completely unacquainted to knowing each other well. Each participant was assigned distinct roles: 'Project Manager', 'User Interface specialist', 'Marketing Expert', and 'Industrial Designer'. Each group met over four sessions each of 20-30 minutes so that they achieved the common goal. For 3 groups, the data from one of the four meeting session could not be used (due to recording issues).

5.5 Experiments and results

The 37 meetings constitute 17 hours of recorded data. From this large pool of conversational data, we sampled meeting slices of various durations. We used the audio from the head-set microphones to compute our low-level cues and the bag-of-NVPs. First, we analyze the distribution of our bag features at various time-scales to understand the effect of the time-slice duration on the bag features. Later we report and analyze the topics using certain combinations of the bag features. Though we

experimented with all the possible combinations with the four sets of patterns discussed in Section IV B - *Speaking Distribution*, *Overlap-Silence*, *Group Speaking*, and *Leadership* patterns, due to brevity reasons in this section we report the results with only those combinations that bring new and different insights to understand conversational group behavior. Also, our method discovered topics for the selected combinations at two representative time-slice durations - one short (2-minute) and another long (5-minute) to understand the difference in the topics discovered at these two different time scales. We report results on topic discovery for multiple time scales only for the first combination (the Speaking Distribution-Leadership (DL) combination). For the rest of the combination, we report the discovery results only at 5-minute scale to keep the discussion brief and interesting.

5.5.1 Bag-of-NVPs over varying slice duration

We visualize the distributions of the *Speaking Distribution* patterns and the *Leadership* patterns among the various classes. The distributions of *Overlap-Silence* and *Group Speaking* patterns are not considered because they are equally distributed among the two classes - *more* and *less* - and it is related to the way features are constructed.

Figure 5.5 visualizes the distributions of the Speaking Distribution patterns of **TSL**, **TST**, **TSI** and **TUI** among the five classes ('0' to '4') at different time scales. It is interesting to observe that the group interactions look more like a monologue at finer time scales (e.g. 1-minute) and like a discussion at coarser time scales (e.g. 5-minute), (looking at the probability mass of classes 1, and 4 for speaking length and speaker turns). A gradual transition between these patterns can be observed as the slice duration increases. Also, successful interruptions are not very common at fine time scales, as seen by the significant probability mass at class 0. 1-person, 2-people, 3-people or all participants interrupting are more or less equiprobable at 5-minute scale. Single person getting backchannels looks common at all scales (as the probability mass at class 1 is quite significant).

Figure 5.6 shows the distribution of leadership patterns at two different time scales. If all the four participants had equal status (egalitarian groups) the probability mass at 'L'(resp. 'NL') would be close to 0.25 (resp. 0.75). Qualitatively, the distribution shows that the average statistics of AMI data are close to uniform at some time scales, though individual leaders could have different styles, which we discover using the LDA model.

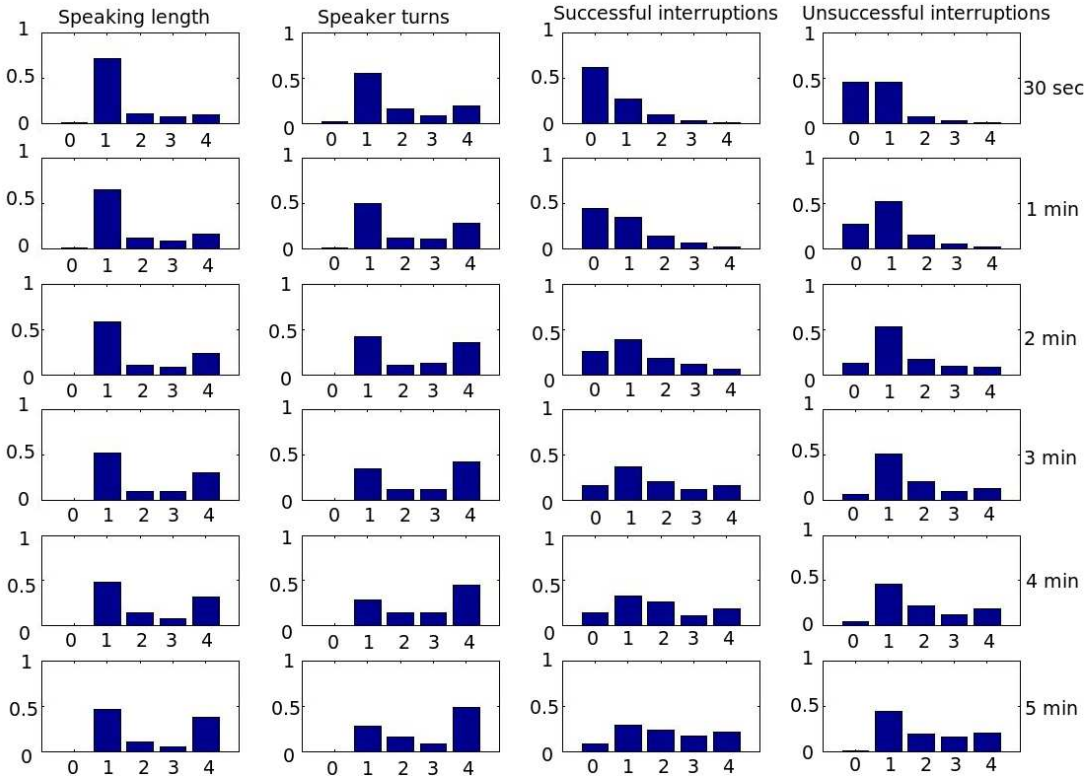


Figure 5.5. Empirical distribution of Speaking Distribution patterns at different time scales (from 30-seconds to 5-minute). x-axis of each of the sub-figure is the classes and y-axis is the probability of the particular class.

5.5.2 LDA based pattern discovery

In our LDA experiments, we use 5-minute and 2-minute scales as representative examples and consider meeting slices from the 37 AMI meetings with overlap. The number of documents for 5-minute slices is 873 and 2-minute slices is 947. We set Δ (introduced in Section IV B) as 0.05. Steyvers et al. explain the role of the parameters α and β of the LDA model in (Steyvers and Griffiths, 2007). For text collections, they use symmetric Dirichlet distribution for α and β, with each of the $\alpha = 50/T$ and $\beta = 0.01$. For our application and corpus, we also used a symmetric Dirichlet distribution with α set to 3 and β set to 0.01. Several other tested values $\alpha = 1, 2, 4, 5$ or $\beta = 0.1, 1$ returned similar results.

LDA-based pattern discovery at 5-minute scale: We first present results for our group descriptor that contains both Speaking Distribution and Leadership patterns (DL combination). We applied our LDA-based discovery procedure varying the number of topics T; we report the results using $T = 3$ topics. Though we fixed the number of topics as three, the number of topics can be increased to get a more detailed understanding of group behavior topics. Table 5.1 shows the resulting top seven words

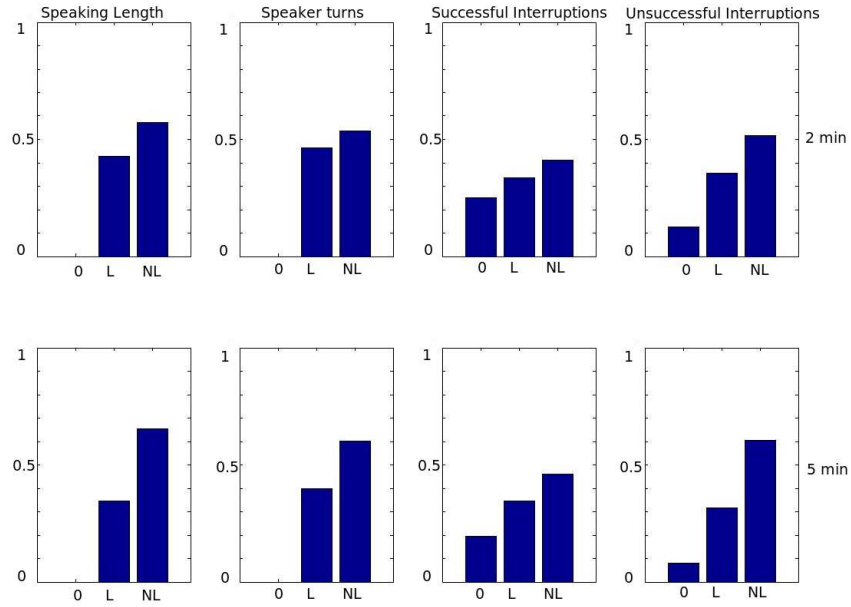


Figure 5.6. Empirical distribution of leadership patterns at two different time scales (2-minute and 5-minute). x-axis of each of the sub-figure is the classes and y-axis is the probability of the particular class. '0' corresponds to the case when there is silence, 'L' (resp. 'NL') when leader (resp. someone else) has maximum feature value.

for each of the topics. Looking at the top words of Topic 1 (SL-M-L, ST-M-L, SI-M-L, UI-M-L terms which means that the leader speaks and interrupts the most, and gets the interrupted unsuccessfully the most), it resembles a meeting where the leader is dominant or autocratic (talks more, more often, and interrupts more) and hence the title *autocratic*. Topic 2 seems to characterize an egalitarian or participative meeting (top words being *ST-Equal*, SL-Equal, SI-Equal - all participants speak and interrupt equally), whereas Topic 3 represents a meeting where there is a single dominant person who, interestingly, is not the leader (top words being SL-One, SI-One, UI-M-NL, SL-M-NL, ST-M-NL, SI-M-NL - meaning someone other than the leader speaks and interrupts the most). Based on manual inspection these patterns for the project managers of AMI meeting slices discovered for $T = 3$ topics seem to resemble the three classic leadership styles of Lewin et al. (Lewin, 1946) as illustrated in Figure 5.7. The three styles - *autocratic* (when the decisions are determined by the leader), *participative* (when the leader encourages group discussion and group decision making), and *free-rein* (when the group or an individual has complete freedom to decide without leader participation), differ according to the emphasis (in terms of power) it places on the leader, the whole group, or the rest of the group. The speech segmentation of two examples from each of the three topics are visualized in Figure 9.

Topic 1 - LDA		**Topic 2 - LDA**		**Topic 3 - LDA**	
$P(z) = 0.32$		$P(z) = 0.33$		$P(z) = 0.34$	
'Autocratic'		**'Participative'**		**'Free-rein'**	
Word	$P(w\|z)$	**Word**	$P(w\|z)$	**Word**	$P(w\|z)$
SI-M-L	0.14	ST-Equal	0.16	UI-M-NL	0.15
ST-M-L	0.13	SL-M-NL	0.14	SL-One	0.14
UI-M-L	0.11	ST-M-NL	0.13	SL-M-NL	0.13
SL-M-L	0.10	UI-M-NL	0.11	SI-M-NL	0.13
SI-Two	0.08	SL-Equal	0.10	UI-One	0.13
ST-Rest	0.06	SI-M-NL	0.08	ST-M-NL	0.12
ST-Two	0.05	SI-Equal	0.07	ST-One	0.11

Table 5.1. LDA based topic discovery at 5-minute scale (DL combination).

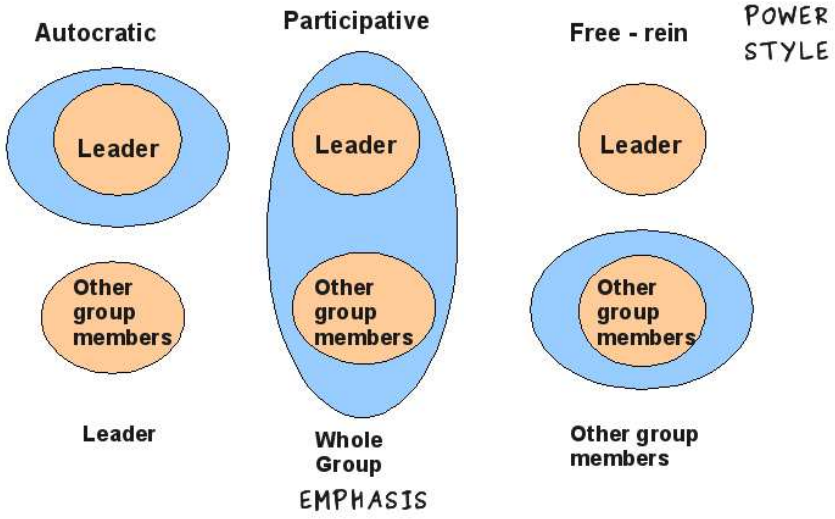


Figure 5.7. Leadership styles by Lewin et al. The blue envelope shows the emphasis (in terms of power) that is placed on the various group members.

Objective evaluation To evaluate how meaningful the discovered topics are we carried out human annotations. We adopted the following protocol, as the cost of annotating the whole corpus is extremely large. For each of the three topics- *autocratic*, *participative* and *free-rein*, we ranked the meeting slices according to $P(z|d)$ and picked the top 8 documents. Each of these 24 meeting slices were annotated by 3 independent annotators. In the protocol, an annotator annotates a particular group only once to avoid potential biases by observing the same group for the second time. The ground-truth is the class that the majority of the annotators agreed. The instructions given to the annotators appear in the appendix.

On this data, we see that the prediction accuracy of our model for the *autocratic* class is 62.5%,

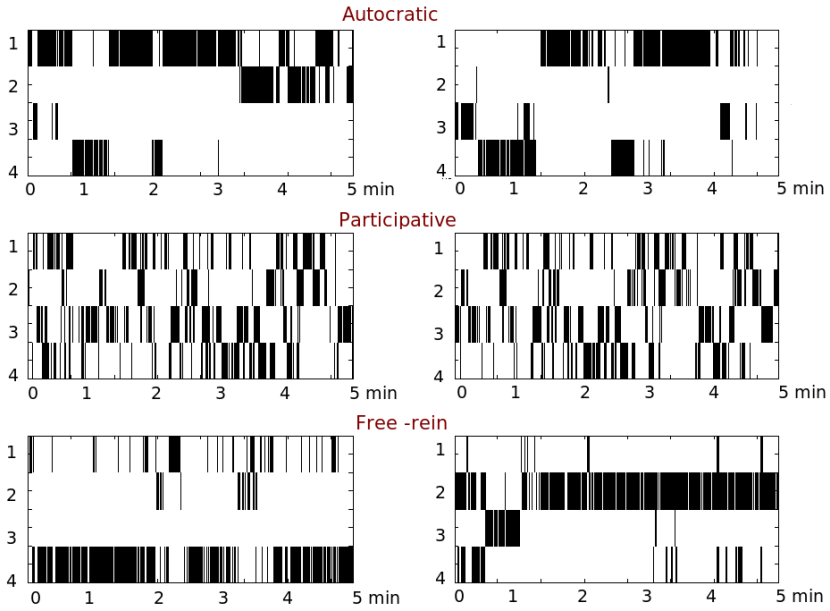


Figure 5.8. Speech segmentation of two sample 5-minute meeting slices for each of the three topics - *autocratic*, *participative* and *free-rein*. The four participants are marked 1, 2, 3,and 4 along the y-axis. The position marked 1 corresponds to the leader (project manager) in all cases.

		MODEL OUTPUT		
		'Autocratic'	**'Participative'**	**'Free-rein'**
	'Autocratic'	5	3	0
GROUND TRUTH	'Participative'	0	8	0
	'Free-rein'	0	2	6

Table 5.2. Evaluation: Confusion matrix between the ground-truth and the model output

participative class is 100%, and *free-rein* is 75%. The confusion matrix is shown in Table 5.2. The results suggest that leaders in the AMI corpus do not show a strong autocratic nature, as seen by the prediction accuracy as well as the top words of the *autocratic* topic. While *free-rein* case has words like SL-One, ST-One as top words, the autocratic case has only SI-Two and ST-Rest words as top words (which implies that though the leader speaks the most, he lets others to participate as well).

Characterizing groups Using the above representation and Eq (4) in Section IV D, we estimate the topic distribution $p(z|g)$ for each of the 10 groups of participants and show it in Figure 10. As one can observe, different groups have different signature distribution of topics. For example, groups 1, 2 seem to have a leader who is less participative as compared to the leader in groups 5, 9, 10.

It is also interesting to visualize the topic evolution of several groups with respect to time (Figure 5.10). The topic shown is the topic with the maximum probability for that meeting slice. Each of the

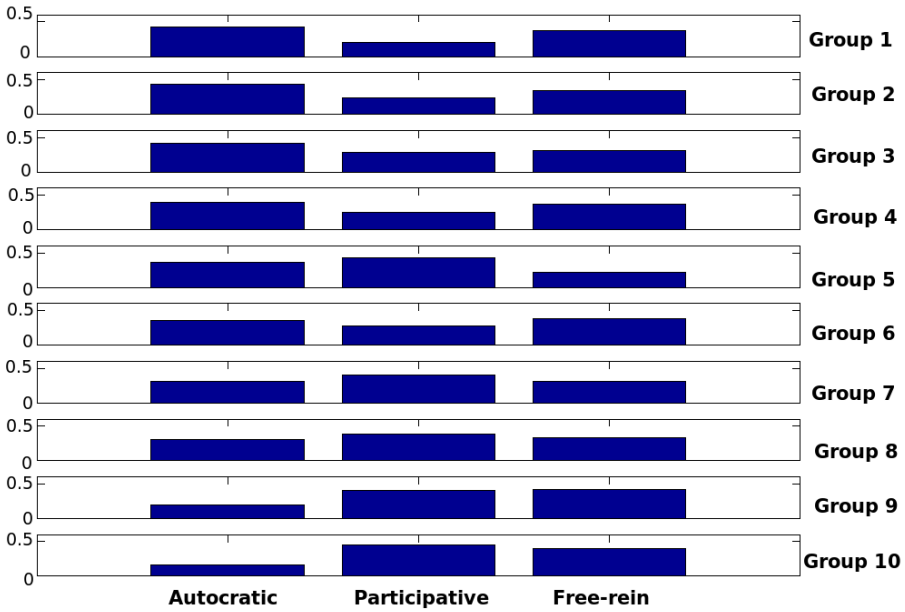


Figure 5.9. Topic distribution over groups at 5-minute scale (DL combination).

six meeting slices have an overlap of four minute with the next meeting slice. The x-axis represents time and the y-axis is the session number (explained in Section V). It is interesting to observe that while the leader in group 1, 2, 6 does not show *participative* style, group 5 does not show *free-rein* style and group 10, 9 does not show *autocratic* style. Also, *autocratic* topic seems more common in the beginning and the end of the meeting session, whereas the *participative* topic appears more often during the middle.

LDA-based pattern discovery at 2-minute scale: The same experiments were repeated with $T = 3$ topics on 2-minute meeting slices (see Table 5.3). We observe that the same three topics emerge, with some differences. For the case of the *free-rein* topic, the top four words are also present in the 5-minute case as well. A new word SI-Silence becomes significant at the 2-minute scale. For the other two topics, we observe that the words in *autocratic* and *participative* topics are also similar to those of the 5-minute case (SL and ST related words are the same).

Figure 5.11 shows the topic distribution for the 10 groups of participants at 2-minute scale. As compared to the 5-minute case, the distribution seems to be more balanced across the three topics. This suggests qualitatively that the interaction styles (as defined here in terms of discovered topics) seem to be captured more strongly over longer intervals of time. Such a conclusion is only qualitative

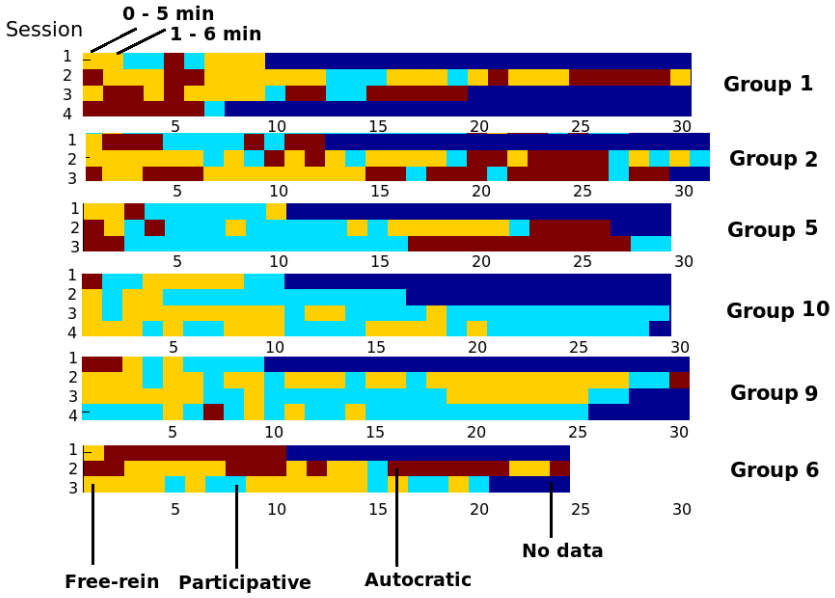


Figure 5.10. Topic evolution for selected groups at 5-minute scale (DL combination). The topics are color coded - *autocratic* in red, *participative* in light-blue, *free-rein* in yellow. The x-axis represents time. The y-axis represents meeting sessions.

due to the fact that the 'interaction styles' are intrinsically sensitive to time granularity. Nevertheless, in a few cases some trends are stable. For instance, groups like group 5, which are more *participative* than other groups at both 5-minute and 2-minute scales, make a more egalitarian group, as compared to for instance group 1 which looks *autocratic* at both scales. Figure 5.12 shows some snapshots of automatic group behavior discovery.

LDA-based pattern discovery for alternative bags of nonverbal behavior

Next we analyze the Overlap-Silence Leadership (OL) combination to understand the relationship between the leader behavior and the competition to occupy the floor. For space reasons we discuss only the 5 min results.

Table 5.4 shows the resulting top seven words for each of the $T = 3$ topics. The first topic corresponds to the case when the leader dominates (talks more, more often, interrupts more and gets unsuccessfully interrupted the most - indicated by words like SL-M-L, ST-M-L, SI-M-L, UI-M-L) but the group also has many silent frames, showing that the leader might not be leading to an interactive group behavior. The second topic characterizes a group which is interactive with presence of overlapping frames, and less cases of silence (indicated by words like Overlap-more and Silence-less). The third topic characterizes a presentation type meeting slice, where there is a single person who is not

Topic 1 - LDA		**Topic 2 - LDA**		**Topic 3 - LDA**	
$P(z) = 0.32$		$P(z) = 0.35$		$P(z) = 0.32$	
'Autocratic'		**'Participative'**		**'Free-rein'**	
Word	$P(w\|z)$	**Word**	$P(w\|z)$	**Word**	$P(w\|z)$
ST-M-L	0.14	UI-M-NL	0.14	SL-One	0.17
SI-One	0.13	SL-M-NL	0.12	SL-M-NL	0.14
SL-M-L	0.12	ST-M-NL	0.11	ST-M-NL	0.12
UI-M-L	0.11	ST-Equal	0.11	ST-One	0.10
ST-Two	0.11	SI-M-NL	0.11	SI-Silence	0.10
SL-Two	0.10	SL-Equal	0.07	SI-M-Silence	0.10
UI-One	0.07	ST-Rest	0.05	UI-One	0.09

Table 5.3. LDA based discovery at 2-minute scale (DL combination).

Topic 1 - LDA		**Topic 2 - LDA**		**Topic 3 - LDA**	
$P(z) = 0.32$		$P(z) = 0.34$		$P(z) = 0.33$	
'Leader-domination'		**'Group Interaction'**		**'Monologue'**	
Word	$P(w\|z)$	**Word**	$P(w\|z)$	**Word**	$P(w\|z)$
Silence-more	0.18	SL-M-NL	0.18	Overlap-less	0.18
Single-less	0.16	Silence-less	0.16	ST-M-NL	0.17
ST-M-L	0.15	SI-M-NL	0.15	UI-M-NL	0.15
UI-M-L	0.13	UI-M-NL	0.14	Single-more	0.14
SL-M-L	0.12	Overlap-more	0.13	SL-M-NL	0.13
SI-M-L	0.12	ST-M-NL	0.11	SI-M-NL	0.09
Overlap-less	0.06	Single-more	0.09	Silence-less	0.09

Table 5.4. LDA based discovery at 5-minute scale (OL combination).

the leader talking most of the time and there is not much of interaction among the group members (indicated by words like Single-more, Overlap-less). Overall, the patterns extracted with this bag are different than the ones extracted using the DL combination. The speech segmentation of two examples from each of the three classes are visualized in Figure 5.13.

Finally, we analyzed the Overlap Silence- Group Speaking- Speaking Distribution(OGD) combination to understand the common topics by clustering the generic group patterns. This combination is useful to analyze groups that do not have a designated leader.

Table 5.5 shows the resulting top 10 words for each of the topics. The first topic corresponds to the case when the group speaks less (is laid-back - indicated by words like Silence-more, Overlap-less, SL-less etc) and there might be a presentation (as there is a single speaker and indicated by words like SL-One, ST-One). The second topic characterizes a group where there are two others who challenge the presenter (the presence of the word SL-One indicates that there is one person who speaks more than

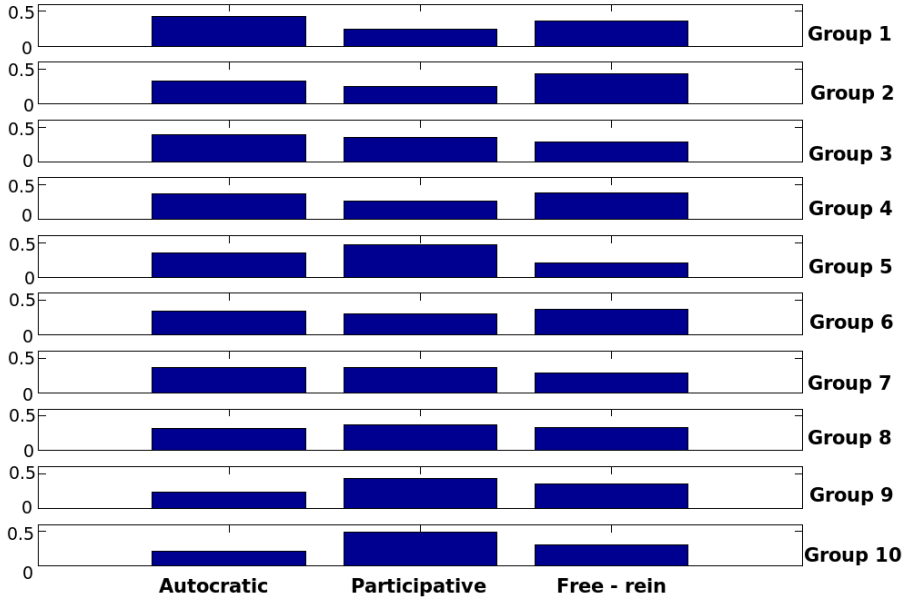


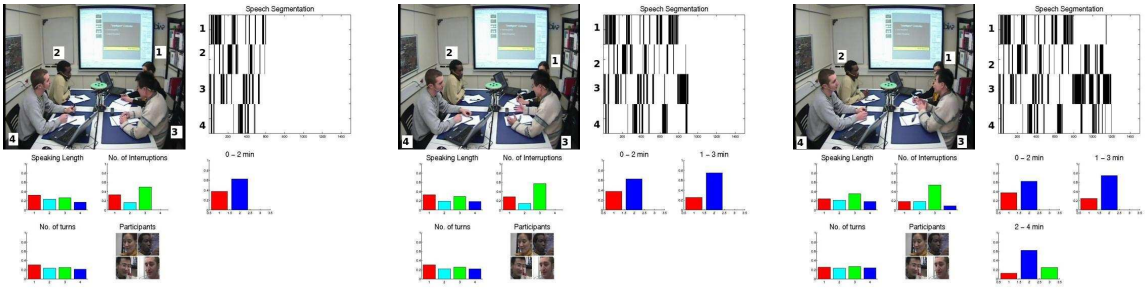
Figure 5.11. Topic distribution over groups at 2-minute scale (DL combination).

half of the total speaking time and ST-Rest indicates that three people get significant speaking turns). The third topic characterizes an interaction *hot-spot* where there is lots of interaction (indicated by the presence of words like ST-more, SL-more, Overlap-more) and everyone is participating (indicated by words like ST-Equal, SL-Equal). The speech segmentation of two examples from each of the three classes are visualized in Figure 5.14.

5.6 Discussion and Conclusion

Overall, our study suggests the following:

Summary of results: Our work has shown a way of discovering conversational group behavior in a data-driven approach. Our method to characterize group behavior by defining group descriptors and then mining them using topic models is promising, allowing for the possibility of learning models to analyze group behavior on large meeting corpora in an unsupervised way, and therefore saving a potentially huge annotation effort (compared to supervised approaches). The proposed bag-of-NVPs described the group in an interpretable and robust fashion, allowing fusion of individual cues, and allowing the comparison of groups of different sizes. The LDA model automatically discovered the



(a) After 2 minutes (b) After 3 minutes (c) After 4 minutes

Figure 5.12. Three snapshots of a group interaction - at 2-minute, 3-minute, 4-minute - with the top left panel showing the center view camera, the top right showing the speech segmentation evolution w.r.t time in x-axis and the participants in the y-axis, the bottom left panel showing the low level cues for each of the participant, and the bottom right panel showing the topic distribution - red being *autocratic*, blue being *participative* and green being *free-rein* for the intervals 0-2 min, 1-3 min, and 2-4 min. This meeting slice corresponds to group 5, which is *participative* at both 2-minute and 5-minute time scales.

Topic 1 - LDA		**Topic 2 - LDA**		**Topic 3 - LDA**	
$P(z) = 0.34$		$P(z) = 0.3$		$P(z) = 0.36$	
'Laid-back monologue'		**'Monologue with brief exchanges'**		**'Interaction hot-spot'**	
Word	$P(w\|z)$	**Word**	$P(w\|z)$	**Word**	$P(w\|z)$
GUT-less	0.12	Single-more	0.14	ST-more	0.10
SL-less	0.11	Silence-less	0.10	Overlap-more	0.10
GIT-less	0.10	SI-less	0.09	ST-Equal	0.09
Silence-more	0.09	UI-One	0.07	GIT-more	0.09
UI-less	0.09	SL-One	0.06	SI-more	0.09
SI-less	0.09	Overlap-less	0.05	UI-more	0.09
Overlap-less	0.08	SI-less	0.05	SL-more	0.07
ST-One	0.07	SL-more	0.05	GUT-more	0.07
SL-One	0.06	UI-less	0.05	SL-Equal	0.06
SI-One	0.05	ST-Rest	0.04	Single-less	0.06

Table 5.5. LDA based discovery at 5-minute scale (OGD combination).

topics based on co-occurence of bag-of-NVPs, and any meeting slices can be described as a probabilistic mixture over the discovered topics. Our method was able to discover group interaction patterns that resemble prototypical leadership styles - *autocratic*, *participative*, and *free-rein*- proposed in social psychology. An objective evaluation of our methodology involving human judgment and multiple annotators, showed that the learned topics indeed are meaningful. Clearly, we don't claim that our method discovers leadership patterns as discussed in psychology, but that the mined results resemble them.

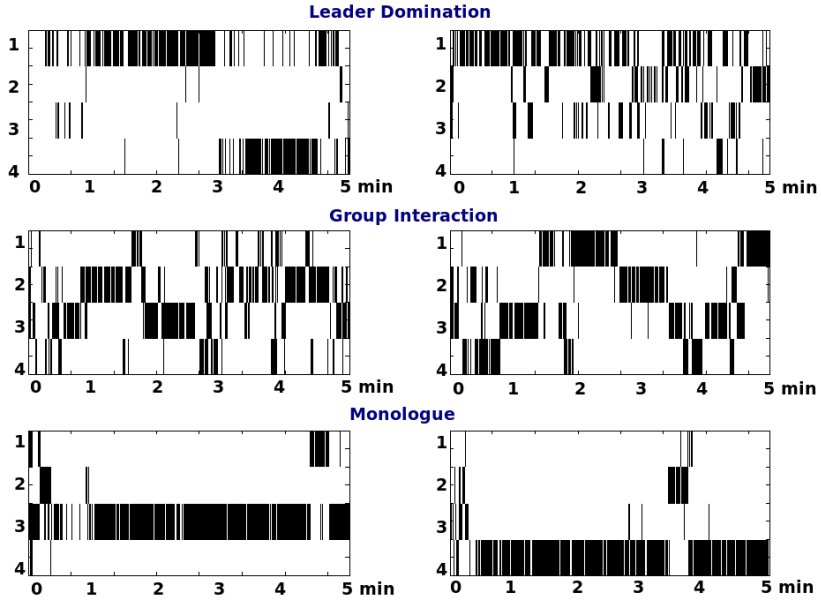


Figure 5.13. Speech segmentation of two sample 5-minute meeting slices for each of the three topics - *Leader-domination, Group Interaction, Monologue*. The x-axis indicates time. The four participants are marked 1, 2, 3, and 4 along the y-axis. The position marked 1 corresponds to the leader (project manager) in all cases.

Limitations and Extensions: One problem not addressed in this work is model selection (i.e., how many topics are needed). In order to evaluate the number of topics and the consistency of the NVP distributions of topics, a variety of other approaches could also be considered (Boyd-Graber *et al.*, 2009). Furthermore, we could investigate other models, for instance to jointly discover group patterns and the groups that best fit them. The current definition of the bag-of-NVPs could also be further extended in the following way. The quantization procedure to generate the bag now depends on the relative feature values of the considered group conversation compared to the average feature values computed over the entire conversation corpus. By using a large corpus constructed to be statistically representative, such a definition could be further strengthened. Another possibility would be to learn the NVP vocabulary via a more elaborate quantization procedure, e.g. as currently investigated in computer vision for visual representation problems (Boureau *et al.*, 2010). Though in this paper we defined and analyzed group conversational patterns derived only from the audio modality, the bag approach can be extended to include multimodal features - e.g. combining prosodic cues and visual attention-based cues, among others.

In terms of applications, our work has the potential to be used for retrieval of group conversational segments where semantically meaningful group behaviors emerge. Our framework can also help characterize groups by aggregating group behavior over multiple interaction slices. This might help under-

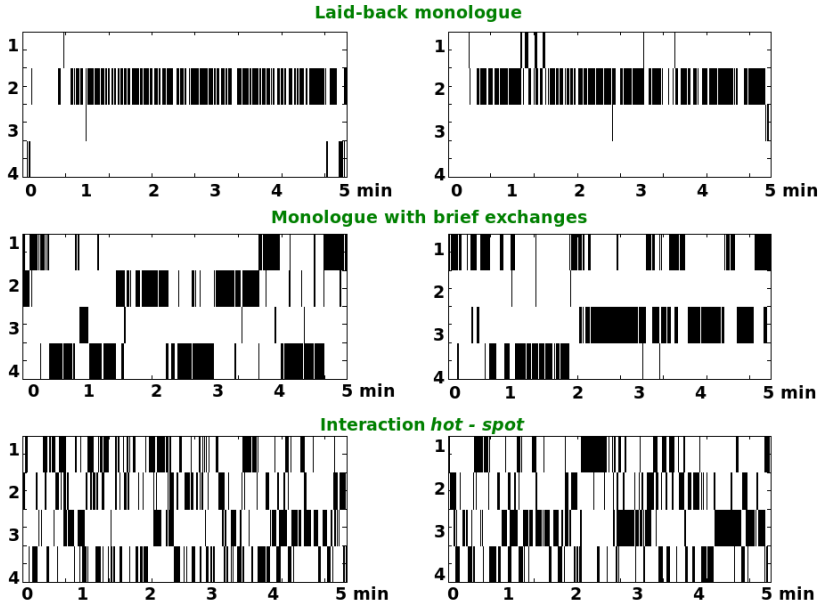


Figure 5.14. Speech segmentation of two sample 5-minute meeting slices for each of the three topics - *Laid-back monologue, Monologue with brief exchanges, Interaction hot-spot*. The x-axis indicates time. The four participants are marked 1, 2, 3, and 4 along the y-axis. The position marked 1 corresponds to the leader (project manager) in all cases.

stand how groups are different from each other in a formal probabilistic sense. We also showed the possibility of visualizing group behavior over time, which could open interesting application options. For instance, in the case of discovering leadership-like styles, we could understand how the manager employs different leadership styles during different phases of a meeting series. Investigating these aspects in further detail could be part of future work.

Chapter 6

Conclusions and Future Directions

In this thesis, we investigate computational frameworks to infer individual and group behavior using automatically extracted nonverbal communication cues. We particularly modeled individual behavioral constructs like dominance and status, two facets of the vertical dimension of human relationships. We also proposed two different frameworks to characterize group behavior. The first framework, a supervised one, aggregates individual behavior over time and individuals, and was used to classify the conversational context. The second framework, using unsupervised learning, discretizes these group cues into bags of nonverbal patterns and infers conversational behavior types using a probabilistic topic model.

In chapter 2, we studied the task of estimating the people who are perceived to be the most and the least dominant by external observers on the AMI meeting corpus. The meetings had four participants with different roles. We separated our analysis into full-agreement and majority-agreement cases to understand the variation in performance of various nonverbal cues with annotator variability. We experimented with both audio and visual cues that have support in the social psychology literature. The audio cues we investigated were based on turn-taking patterns using speaking activity. We experimented with two types of visual cues - one with visual activity estimated in a computationally efficient manner in the compressed domain, and the other with visual attention cues estimated using head pose. Our results show that the audio cues and the visual attention cues were the most effective ones for estimating dominance. Total speaking time, Total speaking turns without the short utterances, and Total short unsucessful interruptions in the audio cues; Total visual activity length and

total visual activity turns using the 'Residue' option in the visual activity cues; The Multi-party visual dominance ratio, total time looking-at-Others while speaking, and Total received visual attention in the visual attention cues were also consistently the top performing cues. Cue fusion, in general, helped in improving the classification accuracies as compared to single cues.

In chapter 3, we investigated the task of estimating the most-dominant person and the high-status person in the AMI meeting dataset. It is to be noted that this status was implicitly 'assigned' to one of the four volunteers for participating in the task-oriented interaction. Our results showed that the high-status person need not always be perceived as most dominant. Furthermore, the nonverbal cues to estimate the most-dominant and the high-status person were different. While the Total speaking time and Total speaking turns without the short utterances was effective in estimating the most-dominant person, the Total speaking turns and the Total speaking turns while speaking first were effective in estimating the high-status person. Centrality-based cues were also effective at estimating the project manager, showing that he plays a central role. Also, the task of estimating dominant-managers was easier as compared to non-dominant managers, which makes intuitive sense as one of the definitions of dominance, emphasizes the expressed aspect of it, and defines it as observable communicative acts.

In Chapter 4, we proposed a novel framework to characterize group behavior from turn-taking cues. We defined two layers of behavioral cues. The first layer consists of individual behavioral cues, and the second layer represents the group behavioral cues. Our group cues characterized the floor-occupation patterns of the group as a whole w.r.t overlap-silence patterns, participation rates, and the distribution of turn-taking patterns among group members. We used the group cues to classify two conversational contexts - cooperative vs competitive and brainstorming vs decision-making using a supervised classifier. Our results show that most competitive interactions have higher number of turns as interruptions and higher inequality in distribution of turns as compared to cooperative interactions; most brainstorming interactions had higher proportion of silent frames and lesser overall speech activity, as compared to decision-making interactions. Inferring group conversational context has applications in understanding individual and group behavior and online support of groups.

In Chapter 5, we propose another novel framework which first involved discretizing the group conversational cues previously employed and also encoding the leader's position in the group, resulting in a representation called the bag-of-NVPs, and then generated co-occurence based topics or soft clusters using principled probabilistic topic models. Our method was able to discover group interaction

patterns in the AMI corpus that under close inspection, seem to resemble prototypical leadership styles proposed in social psychology - *autocratic*, *participative*, and *free-rein*. An objective evaluation of our methodology involving human judgment and multiple annotators, showed that the learned topics indeed are indeed meaningful.

While this thesis made progress along several research lines in group nonverbal modeling, it is clear that many issues remain open. Some of the future directions emerging out of this thesis are listed below.

- One way of extending the work on verticality aspects of our thesis would be to study 'power'. Hall et al. (Hall *et al.*, 2005) define power as "the capacity or structurally sanctioned right to control others or their resources does not necessarily imply prestige or respect". This third facet of the vertical dimension could be compared with the other two facets i.e. dominance and status. Experimental design to study such a problem would nevertheless be challenging.
- The interplay between personality and social verticality is another interesting research direction. The literature on automatic modeling of personality perception shows that some of the 'Big-Five' traits (John and Srivastava, 1999) like introversion and extroversion can be reliably estimated from group interactions (Pianesi *et al.*, 2008a). An interesting research question could be 'do introverts behave in a dominant way?'. Understanding the overlap between these two constructs would be interesting.
- With respect to other nonverbal cues that could be studied, prosodic cues and gesturing behavior could be important to study. The relative effectiveness of the prosodic cues for modeling dominance is well known in the human communication research literature (Tusing and Dillard, 2000). Furthermore, the performance of visual activity cues could be improved if the gesture of the participants is tracked and analyzed. An open issue is to assess whether the additional compuational cost would justify the use of these features, in terms of performance improvements.
- Though initial research has shown that dominance affects performance in brain-storming groups (Kim *et al.*, 2008), a general relationship between dominance and performance has not been firmly established. The results in (Kim *et al.*, 2008) showed that dominance had an interesting effect on performance: having a dominant person in the group had a significant negative effect on brain- storming i.e. groups with dominant people tended to generate fewer ideas. The relationship of dominance with team satisfaction or long term stability of groups could also be

potential research directions.

- We could extend our supervised framework for group behavior modeling in four ways. First, include other features like visual attention or prosody. Second, we could study other group conversational contexts, for example casual chatting among peers vs a formal discussion. Third, with larger datasets, we could attempt to do automatic inference of group conversational context in real scenarios. Fourth, we could pursue a study of automatic modeling across contexts and assess the generalizing abilities of our models.
- We could extend our unsupervised framework for characterizing group behavior in three ways. First, again as in the previous case, we could to include cues like gaze patterns and prosody. This would strengthen the bag-of-nonverbal patterns. Second, we could model short temporal patterns which could be a good way of extending the bag-of-NVP framework for modeling the group dynamics better. Third, understanding the evolution of group behavior topics could help understand how the group evolves with time. An interesting question to ask could be: how different is the group behavior in the beginning of the interaction as compared to few minutes later or towards the end of the interaction?
- An important need for research in this domain is an increase of size and variability of the datasets. The availability of large publicly available corpus like the AMI are a great step forward to encourage research in this domain and eventual comparison of research results.
- Privacy is another crucial issue that needs to be addressed while designing experiments and collecting data. Recording and analyzing real scenarios in a privacy-sensitive way is a challenge. Developing mobile, privacy-sensitive recording solutions and extracting and storing only privacy-sensitive cues could be one way forward to investigate interaction in the real world.

Overcoming some of these limitations and embrazing the future advancements in sensing, analyzing, and modeling of group behavior would facilitate the development of robust social inference machines. Such systems no doubt would make team-work in modern workplaces both a productive and a rewarding experience.

Appendix A

Objective evaluation: Human annotation

In this appendix, we provide the instructions given to the external observers for the experimental evaluation in Section 5.5.2.

Lewin et al. (1948) describes three classic leadership styles as illustrated in Figure 5.7. The three styles - 'autocratic' (A), 'participative' (P), and 'free-rein' (FR), differ according to the emphasis (in terms of power) placed on the leader, the whole group, or the rest of the group.

- The **Autocratic** *style corresponds to the case when the leader makes decisions himself.*
- The **Participative** *style refers to the case where the leader includes all the group members in the decision making process.*
- *A leader using a* ***Free-rein*** *style allows (consciously or unconsciously) the group-members to make the decision.*

Kindly look at the meetings assigned to you and answer each of the following questions.

1. *Which of the three categories do you think this meeting belongs to - autocratic, participative and free-rein? Choose only one.*
2. *How confident are you about this decision, on a five-point scale?*
3. *Add any specific comments regarding the annotation of this meeting, if you want.*

The instructions are based on the definition of the categories, but do not provide any information about the specific nonverbal behavior that the annotators should base their decision upon, or about the method that produced the dataset people are supposed to annotate.

Bibliography

Ambady, N., Bernieri, F. J., and Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In Zanna, M.P., editor, *Advances in Experimental social psychology*, volume 32, pages 201 – 272. Academic Press.

Aran, O. and Gatica-Perez, D. (2010). Fusing audio-visual nonverbal cues to detect dominant people in conversations. In *Proceedings of the International Conference on Pattern Recognition (ICPR), Istanbul, Turkey*, pages 3687 – 3690.

Arrow, H., McGrath, J., and Berdahl, J. (2000). *Small groups as complex systems: Formation, coordination, development and adaptation*. Sage Publications, Inc.

Ba, S. O. and Odobez, J. M. (2010). Multi-person visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **33**(1), 101 – 116.

Bachour, K., Kaplan, F., and Dillenbourg, P. (2010). An interactive table for supporting participation balance in face-to-face collaborative learning. *IEEE Transactions on Learning Technologies*, **3**(3), 203 – 213.

Bales, R. F. (1950). *Interaction process analysis: a method for the study of small groups*. Cambridge, Addison-Wesley.

Bales, R. F. (1970). *Personality and interpersonal behavior*. Holt, Rinehart and Winston New York.

Banerjee, S. and Rudnicky, A. (2004). Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of the Eighth International Conference*

on Spoken Language Processing (Interspeech 2004 - ICSLP), pages 2189 – 2192, Jeju Island, South Korea.

Basu, S. (2002). *Conversational scene analysis*. PhD Thesis, Massachusetts Institute of Technology.

Basu, S., Choudhury, T., Clarkson, B., and Pentland, A. (2001). Towards measuring human interactions in conversational settings. In *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR) Workshop on Cues in Communication*, Kauai, USA.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, **3**, 993 – 1022.

Bonacich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, **23**(3), 191 – 201.

Bornstein, G. (2003). Intergroup conflict: individual, group, and collective interests. *Personality and Social Psychology Review*, **7**(2), 129 – 145.

Boureau, Y., Bach, F., LeCun, Y., and Ponce, J. (2010). Learning mid-level features for recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2559 – 2566, San Francisco, USA.

Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS) 22*, pages 288 – 296.

Brody, C. and Smith-Lovin, L. (1989). Interruptions in group discussions: The effects of gender and group composition. *American Sociological Review*, **54**(3), 424 – 435.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**(2), 121 – 167.

Burgoon, J. K. and Dunbar, N. E. (2006). Nonverbal expressions of dominance and power in human relationships. In V. Manusov and M. Patterson, editors, *The Sage Handbook of Nonverbal Communication*. Sage, Beverly Hills, CA.

Carletta et al., J. (2006). The AMI meeting corpus: A pre-announcement. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction*, volume 3869 of *Lecture Notes in Computer Science*, pages 28 – 39. Springer Berlin, Heidelberg.

Chippendale, P. (2006). Towards automatic body language annotation. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 487 – 492, Southampton, UK.

Choudhury, T. and Pentland, A. (2002). The sociometer: A wearable device for understanding human networks. In *Proceedings of Computer Supported Cooperative Work (CSCW) Workshop, Workshop on Adhoc Communications and Collaboration in Ubiquitous Computing Environments*, New Orleans, USA.

Coimbra, M. T. and Davies, M. (2005). Approximating optical flow within the MPEG-2 compressed domain. *IEEE Transactions on Circuits and Systems for Video Technology*, **15**(1), 103 – 107.

Cook, M. and Smith, J. M. (1975). The role of gaze in impression formation. *The British Journal of Social and Clinical Psychology*, **14**(1), 19 – 25.

Darwin, C. (1965). *The Expression of the Emotions in Man and Animals*. University of Chicago Press. (Originally published 1872, London: John Murray).

Dielmann, A. and Renals, S. (2007). Automatic meeting segmentation using dynamic Bayesian networks. *IEEE Transactions on Multimedia*, **9**(1), 25 – 36.

DiMicco, J., Pandolfo, A., and Bender, W. (2004). Influencing group participation with a shared display. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 614–623, New York, USA.

DiMicco, J. M. and Bender, W. (2007). Group reactions to visual feedback tools. In *Proceedings of the International Conference on Persuasive Technology*, pages 132 – 143.

DiMicco, J. M., Hollenbach, K. J., and Bender, W. (2006). Using visualizations to review a group's interaction dynamics. In *Proceedings of the International Conference on Human Factors in Computing Systems*, pages 706 – 711. New York, USA.

Dines, J., Vepa, J., and Hain, T. (2006). The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proceedings of the Ninth International Conference on Spoken Language Processing (InterSpeech 2006 - ICSLP)*.

Dong, W. (2010). *Modeling the structure of collective intelligence*. PhD Thesis, Massachusetts Institute of Technology.

Dong, W., Lepri, B., Cappelletti, A., Pentland, A. S., Pianesi, F., and Zancanaro, M. (2007). Using the influence model to recognize functional roles in meetings. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, pages 271 – 278, New York, USA.

Dong, W., Mani, A., Pentland, A., Lepri, B., and Pianesi, F. (2011). Modeling group discussion dynamics. *In press, submitted to the IEEE Transactions on Autonomous Mental Development.*

Dovidio, J. F. and Ellyson, S. L. (1982). Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly*, **45**(2), 106 – 113.

Dunbar, N. E. and Burgoon, J. K. (2005a). Measuring nonverbal dominance. In V. Manusov, editor, *The sourcebook of nonverbal measures: Going beyond words*. Erlbaum.

Dunbar, N. E. and Burgoon, J. K. (2005b). Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, **22**(2), 207 – 233.

Efran, J. S. (1968). Looking for approval: effects on visual behavior of approbation from persons differing in importance. *Journal of Personality and Social Psychology*, **10**(1), 21 – 25.

Exline, R. V., Ellyson, S. L., and Long, B. (1975). Visual behavior as an aspect of power role relationships. *Nonverbal communication of aggression*, **2**, 21 – 51.

Farrahi, K. and Gatica-Perez, D. (2008). What did you do today? Discovering daily routines from large-scale mobile data. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 849 – 852. Vancouver, Canada.

Farrahi, K. and Gatica-Perez, D. (2010). Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing*, **4**(4), 746 – 755.

Garg, N. P., Favre, S., Salamin, H., Hakkani-Tur, D., and Vinciarelli, A. (2008). Role recognition for meeting participants: an approach based on lexical information and social network analysis. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 693 – 696. Vancouver, Canada.

Gatica-Perez, D. (2006). Analyzing group interaction in conversations: a review. In *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Special Session on Multisensor Fusion for Human-Activity Analysis, invited paper*, pages 41 – 46, Heidelberg, Germany.

Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing*, **27**(12), 1775 – 1787.

Gatica-Perez, D., McCowan, I., Zhang, D., and Bengio, S. (2005). Detecting group interest-level in meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 489 – 492, Philadelphia, USA.

Gorga, S. and Otsuka, K. (2010). Conversation scene analysis based on dynamic bayesian network and image-based gaze detection. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI - MLMI)*. Beijing, China.

Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, **101**(Supplement 1), 5228 – 5235.

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

Hall, J. A. and Friedman, G. B. (1999). Status, gender, and nonverbal behavior: A study of structured interactions between employees of a company. *Personality and Social Psychology Bulletin*, **25**(9), 1082 – 1091.

Hall, J. A., Coats, E. J., and Smith LeBeau, L. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, **131**(6), 898 – 924.

Hare, A. (1994). Types of roles in small groups. *Small Group Research*, **25**(3), 433 – 448.

Harrigan, J., Rosenthal, R., and Scherer, K. (2008). *New Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press, USA.

Hassin, R. R., Uleman, J. S., and Bargh, J. A. (2005). *The new unconscious*. Oxford University Press, USA.

Heinrich, G. (2005). Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 50 – 57, Berkeyley, USA.

Hung, H., Jayagopi, D., Yeo, C., Friedland, G., Ba, S., Odobez, J.-M., Ramchandran, K., Mirghafori, N., and Gatica-Perez, D. (2007). Using audio and video features to classify the most dominant person in a group meeting. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 835 – 838, Augsburg, Germany.

Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. (2008a). Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2197 – 2200, Las Vegas, USA.

Hung, H., Jayagopi, D., Ba, S., Odobez, J.-M., and Gatica-Perez, D. (2008b). Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proceedings of the ACM International Conference on Multimodal interfaces (ICMI)*, pages 233 – 236, Chania, Greece.

Huynh, T., Fritz, M., and Schiele, B. (2008). Discovery of activity patterns using topic models. In *Proceedings of the International Conference on Ubiquitous Computing (UbiComp)*, pages 10 – 19, Seoul, South Korea.

Jayagopi, D. and Gatica-Perez, D. (2009). Discovering group nonverbal conversational patterns with topics. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI-MLMI)*, pages 3 – 6, Cambridge, USA.

Jayagopi, D. and Gatica-Perez, D. (2010). Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia*, **12**(8), 790 – 802.

Jayagopi, D., Hung, H., Yeo, C., and Gatica-Perez, D. (2008a). Predicting the dominant clique in meetings through fusion of nonverbal cues. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 809 – 812, Vancouver, Canada.

Jayagopi, D., Ba, S., Odobez, J., and Gatica-Perez, D. (2008b). Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI), Special Session on Social Signal Processing*, pages 45 – 52, Chania, Greece.

Jayagopi, D., Raducanu, B., and Gatica-Perez, D. (2009a). Characterizing conversational group dynamics using nonverbal behaviour. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 370 – 373, New York, US.

Jayagopi, D., Hung, H., Yeo, C., and Gatica-Perez, D. (2009b). Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing, Special issue on multimodal processing in speech-based interactions*, **17**(3), 501 – 513.

Jayagopi, D., Kim, T., Pentland, A., and Gatica-Perez, D. (2010). Recognizing conversational context in group interaction using privacy-sensitive mobile sensors. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia (MUM)*, pages 8:1 – 8:4, Limassol, Cyprus.

John, O. P. and Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of personality: Theory and research, New York*, pages 102 – 138. New York: Guilford.

Kapoor, A. and Picard, R. W. (2001). A real-time head nod and shake detector. In *Proceedings of the ACM workshop on Perceptive user interfaces*, Orlando, USA.

Kim, T., Chang, A., Holland, L., and Pentland, A. S. (2008). Meeting mediator: enhancing group collaboration using sociometric feedback. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 457 – 466, San Diego, USA.

Knapp, M. L. and Hall, J. A. (1978). *Nonverbal communication in human interaction*. Holt, Rinehart and Winston New York.

Kulyk, O., Wang, J., and Terken, J. (2006). Real-time feedback on nonverbal behaviour to enhance social dynamics in small group meetings. In S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction*, volume 3869 of *Lecture Notes in Computer Science*, pages 150 – 161. Springer Berlin / Heidelberg.

Kumano, S., Otsuka, K., Mikami, D., and Yamato, J. (2009). Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI-MLMI)*, pages 99 – 106, Cambridge, USA.

Laskowski, K., Ostendorf, M., and Schultz, T. (2008). Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In *Proceedings of the ISCA/ACL 9th SIGdial Workshop on Discourse and Dialogue*, pages 148 – 155.

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

Leffler, A., Gillespie, D. L., and Conaty, J. C. (1982). The effects of status differentiation on nonverbal behavior. *Social Psychology Quarterly*, **45**(3), 151 – 161.

Lepri, B. (2009). *Multimodal Recognition of Social Behaviors and Personality Traits in Small Group Interaction*. PhD Thesis, University of Trento.

Lepri, B., Mana, N., Cappelletti, A., and Pianesi, F. (2009). Automatic prediction of individual performance from thin slices of social behavior. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 733 – 736. Beijing, China.

Levine, J. M. and Moreland, R. L. (1990). Progress in small group research. *Annual Review of psychology*, **41**(1), 585–634.

Lewin, K. (1946). Patterns of aggressive behavior in experimentally created social climates. *Twentieth Century Psychology: Recent Developments in Psychology*.

Lewin, K. and Lewin, G. W. (1948). *Resolving social conflicts: selected papers on group dynamics [1935-1946]*. New York, Harper.

Li, J., Gong, S., and Xiang, T. (2008). Global behaviour inference using probabilistic latent semantic analysis. In *Proceedings of the British Machine Vision Conference (BMVC)*. Leeds, UK.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

Manusov, V. L. and Patterson, M. L. (2006). *The SAGE handbook of nonverbal communication*. Sage Publications, Inc.

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D. (2005). Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **27**(3), 305 – 317.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Prentice-Hall Englewood Cliffs, NJ.

McKenna, S. J., Gong, S., and Raja, Y. (1998). Modelling facial colour and identity with gaussian mixtures. *Pattern Recognition*, **31**(12), 1883 – 1892.

McNeill, D. (2000). *Language and gesture*. Cambridge Univ Press.

Mitchell, T. M. (1997). Machine learning. *Mc Graw Hill*.

Morency, L. P., Sidner, C., Lee, C., and Darrell, T. (2005). Contextual recognition of head gestures. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, pages 18 – 24, Trento, Italy.

Murphy-Chutorian, E. and Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **31**(4), 607 – 626.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, **79**(3), 299 – 318.

Nijholt, A., Rienks, R., Zwiers, J., and Reidsma, D. (2006). Online and off-line visualization of meeting information and meeting support. *The Visual Computer*, **22**(12), 965 – 976.

Olguín, D. O. and Pentland, A. S. (2008). Social sensors for automatic data collection. In *Proceedings of the Americas Conference on Information Systems*, Toronto, Canada.

Olguín, D. O. and Pentland, A. S. (2010). Assessing group performance from collective behavior. In *Proceedings of the Computer Supported Collaborative Work, Workshop on Collective Intelligence In Organizations.*, Savannah, USA.

Oostrum, J. V. and Rabbie, J. M. (1995). Intergroup competition and cooperation within autocratic and democratic management regimes. *Small Group Research*, **26**(2), 269 – 295.

Otsuka, K. and Araki, S. (2010). Audio-visual technology for conversation scene analysis. *Web: https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200902sf2.html*.

Otsuka, K., Yamato, J., Takemae, Y., and Murase, H. (2006). Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI) Extended Abstract*, pages 1175 – 1180, Montreal, Canada.

Otsuka, K., Sawada, H., and Yamato, J. (2007). Automatic inference of cross-modal nonverbal interactions in multiparty conversations: who responds to whom, when, and how? from gaze, head gestures, and utterances. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, pages 255 – 262, Nagoya, Japan.

Pentland, A. (2005). Socially aware, computation and communication. *Computer*, **38**(3), 33 – 40.

Pentland, A. S. (2008). *Honest signals: how they shape our world*. MIT Press.

Petridis, S. and Pantic, M. (2008). Audiovisual discrimination between laughter and speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5117 – 5120, Las Vegas, USA.

Pianesi, F., Zancanaro, M., Lepri, B., and Cappelletti, A. (2007). A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, **41**(3), 409 – 429.

Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., and Zancanaro, M. (2008a). Multimodal recognition of personality traits in social interactions. In *Proceedings of the ACM International Conference on Multimodal interfaces (ICMI)*, pages 53 – 60, Chania, Greece.

Pianesi, F., Zancanaro, M., Not, E., Leonardi, C., and Falcon, V. (2008b). Multimodal support to group dynamics. *Personal Ubiquitous Computing*, **12**(3), 181 – 195.

Poole, M. S., Hollingshead, A. B., McGrath, J. E., Moreland, R. L., and Rohrbaugh, J. (2004). Interdisciplinary perspectives on small groups. *Small Group Research*, **35**(1), 3 – 16.

Raducanu, B. and Gatica-Perez, D. (2010). Inferring competitive role patterns in reality TV show through nonverbal analysis. *Multimedia Tools and Applications*, pages 1 – 20.

Raducanu, B., Vitria, J., and Gatica-Perez, D. (2009). You are fired! Nonverbal role analysis in competitive meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, pages 1949 – 1952, Taipei, Taiwan.

Remland, M. (2006). *Uses and consequences of nonverbal communication in the context of organizational life*. The SAGE handbook of nonverbal communication, Sage Publications, Inc.

Ridgeway, C. L. (1987). Nonverbal behavior, dominance, and the basis of status in task groups. *American Sociological Review*, **52**(5), 683 – 694.

Rienks, R. (2007). *Meetings in smart environments : implications of progressing technology*. PhD Thesis, University of Twente.

Rienks, R. and Heylen, D. (2005). Automatic dominance detection in meetings using easily detectable features. In *Proceedings of the Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Edinburgh, UK.

Rienks, R., Zhang, D., Gatica-Perez, D., and Post, W. (2006). Detection and application of influence rankings in small group meetings. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 257 – 264, Banff, Canada.

Rosa, E. and Mazur, A. (1979). Incipient status in small groups. *Social Forces*, **58**(1), 18 – 37.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 487 – 494.

Rotter, J. B. (1966). *Generalized expectancies for internal versus external control of reinforcement*. American Psychological Association Washington, DC.

Salamin, H., Favre, S., and Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings: using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, **11**(7), 1373 – 1380.

Sanchez-Cortes, D., Jayagopi, D., and Gatica-Perez, D. (2009). Predicting remote versus collocated group interactions using nonverbal cues. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI-MLMI), Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*, pages 3:1 – 3:4, Cambridge, USA.

Sanchez-Cortes, D., Aran, O., Schmid-Mast, M., and Gatica-Perez, D. (2010). Identifying emergent leadership in small groups using nonverbal communicative cues. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI-MLMI)*, Beijing, China.

Schmid Mast, M. (2002). Dominance as expressed and inferred through speaking time: a meta-analysis. *Human Communication Research*, **28**(3), 420 – 450.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, volume 427. Lawrence Erlbaum.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**(476), 1566 – 1581.

Tian, Y. L., Kanade, T., and Cohn, J. F. (2005). Facial expression analysis. In *Handbook of Face Recognition*, pages 247 – 275. Springer New York.

Tusing, K. J. and Dillard, J. P. (2000). The sounds of dominance. *Human Communication Research*, **26**(1), 148 – 171.

Valente, F. and Vinciarelli, A. (2010). Improving speech processing through social signals: Automatic speaker segmentation of political debates using role based turn-taking patterns. In *Proceedings of the International Workshop on Social Signal Processing*, pages 29 – 34, Florence, Italy.

Varadarajan, J. and Odobez, J. M. (2010). Topic models for scene analysis and abnormality detection. In *Proceeedings of the IEEE International Conference on Computer Vision (ICCV), Workshop on visual surveillance*, pages 1338 – 1345, Kyoto, Japan.

Varni, G., Volpe, G., and Camurri, A. (2010). A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. *IEEE Transactions on Multimedia*, **12**(6), 576 – 590.

Vinciarelli, A. (2007). Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, **9**(6), 1215 – 1226.

Waber, B. N. and Pentland, A. S. (2009). Recognizing expertise. In *Winter Conference on Business Intelligence*, University of Utah, Utah, USA.

Wang, H., Divakaran, A., Vetro, A., Chang, S. F., and Sun, H. (2003). Survey of compressed-domain features used in audio-visual indexing and analysis. *Journal of Visual Communication and Image Representation*, **14**(2), 150 – 183.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wilson, D. S., Timmel, J. J., and Miller, R. R. (2004). Cognitive cooperation: when the going gets tough, think as a group. *Human Nature*, **15**(3), 225 – 250.

Woolley, A., Chabris, C., Pentland, A., Hashmi, N., and Malone, T. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, **330**(6004), 686 – 688.

Wrede, B. and Shriberg, E. (2003). Spotting hotspots in meetings: Human judgments and prosodic cues. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 2805 – 2808, Geneva, Switzerland.

Xiang, T. and Gong, S. (2008). Video behavior profiling for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **30**(5), 893 – 908.

Yeo, C. and Ramchandran, K. (2008). Compressed domain video processing of meetings for activity estimation in dominance classification and slide transition detection. Technical Report UCB/EECS-2008-79, EECS Department, University of California, Berkeley.

Zancanaro, M., Lepri, B., and Pianesi, F. (2006). Automatic detection of group functional roles in face to face interactions. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 28 – 34, Banff, Canada.

Zhang, D., Gatica-Perez, D., Bengio, S., and McCowan, I. (2006). Modeling individual and group actions in meetings with layered HMMs. *IEEE Transactions on Multimedia*, **8**(3), 509 – 520.

Curriculum Vitae

Dinesh Babu Jayagopi

30 years old, married

Nationality: Indian

Residence: Switzerland

Contact: Idiap Research Institute,
PO Box 592, Martigny-1920
Switzerland
tel: +41 (0) 774377889 (cellphone)
email: jdineshbabu@gmail.com

Objective

To analyze and model human behavior in social and surveillance settings, using novel multimodal features and machine learning techniques.

Research Interest

Machine Learning, Computer Vision, Signal Processing, Multimodal fusion, Social Psychology, Sociology.

Achievements

Idiap Researcher Award for the year 2009.

Ranked 21 in state level mathematical olympiad competition.

Led a team to win third prize at national level quiz competition on metals and material science.

Professional Experience

Jan 2007 to till Date: Working as a research assistant at Idiap Research Institute, Martigny, Switzerland and doing Ph.D. at Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
Thesis Title: Computational modeling of face-to-face social interaction using nonverbal behavioral cues.
Thesis Advisor: Dr. Daniel Gatica-Perez, affiliated to Idiap Research Institute, Martigny and Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne.

Sep 2003 to Oct 2006: Worked as a Senior Research Engineer at Mercedes-Benz Research and Technology India, Bangalore. Was involved in a project for developing vision-based technologies for *accident-free driving*.

Education

Doing Doctorate in Philosophy (Ph.D.) at Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Masters of Science M.Sc.(engg) at Indian Institute of Science, Bangalore. Specialised in System Science and Signal Processing. Thesis: Adaptive filtering using interpolated filter banks.

Bachelors of Technology (B.Tech) at Madras Institute of Technology, Chennai. Specialised in Electronics Technolgy.

Journal Papers

D. Jayagopi, and D. Gatica-Perez. *Mining group nonverbal conversational patterns using probabilistic topic models*, IEEE Transactions on Multimedia, Volume 12, Issue 8, Dec 2010, Pages 790 - 802.

D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. *Modeling dominance in group conversations from nonverbal activity cues*, Special issue on Multimodal processing in speech-based interactions, IEEE Transactions on Audio, Speech and Language Processing, Volume 17, Issue 3, Mar 2009, Pages 501-513.

K. Rajgopal, **J. Dinesh Babu** and S. Venkataraman, *Generalized adaptive IFIR filter bank structures*, Elsevier Signal Processing, Volume 87, Issue 7, July 2007, Pages 1575-1596.

Conference Papers

D. Jayagopi, T. Kim, A. Pentland, and D. Gatica-Perez. *Recognizing conversational context in group interaction using privacy-sensitive mobile sensors*, In Proceedings of the Ninth International Conference on Mobile and Ubiquitous Multimedia (MUM), Limassol, Cyprus, Dec 2010.

D. Saches-Cortes, **D. Jayagopi**, and D. Gatica-Perez. *Predicting Remote Versus Collocated Group Interactions using Nonverbal Cues*, In Proceedings of the International Conference on Multimodal Interfaces (ICMI-MLMI), Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing, Cambridge, USA, Nov 2009.

D. Jayagopi, and D. Gatica-Perez. *Discovering Group Nonverbal Conversational Patterns with Topics*, In Proceedings of the International Conference on Multimodal Interfaces (ICMI-MLMI), Cambridge, USA, Nov 2009.

D. Jayagopi, B. Raducanu and D. Gatica-Perez. *Characterizing Conversational Group Dynamics using Nonverbal Behaviour*, In Proceedings of the International Conference on Multimedia and Expo (ICME), New York, USA, Jun 2009.

D. Jayagopi, S. Ba, J-M. Odobez and D. Gatica-Perez. *Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues*, In Proceedings of the International Conference on Multimodal Interfaces (ICMI), special session on Social Signal Processing, Chania, Greece, Oct 2008.

H. Hung, **D. Jayagopi**, S. Ba, J-M. Odobez and D. Gatica-Perez. *Investigating automatic dominance estimation in groups from visual attention and speaking activity*, In Proceedings of the International Conference on Multimodal Interfaces (ICMI), Chania, Greece, Oct 2008.

D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. *Predicting the dominant clique in meetings through fusion of nonverbal cues*, In Proceedings of the ACM International Conference on Multimedia (ACM MM), Vancouver, Canada, Oct 2008.

H. Hung, **D. Jayagopi**, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez *Using audio and video features to classify the most dominant person in meetings*, published in Proceedings of the ACM International Conference on Multimedia (ACM MM), Augsburg, Germany, Sep 2007.

K. Rajgopal, **J. Dinesh Babu** (2003) *A Delayless IFIR Adaptive Filter Structure With adapted Filterbank* published in the proceedings of Conference on Convergent Technologies for the Asia-Pacific Region: IEEE TENCON 2003, Volume 3 pages 1051- 1054, Bangalore, India.