Computational Methods for Audio-Visual Analysis of Emergent Leadership in Teams

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service academique.

Thèse n. XYZ 2013
présenté le 1er Mars 2013
à la Faculté de Génie Electriquee
programme doctoral en Génie Electrique

École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur s Sciences
par

Dairazalia Sanchez-Cortes

acceptée sur proposition du jury:

Prof Jean-Philippe Thiran, président du jury
Dr Daniel Gatica-Perez, directeur de thèse
Prof Pierre Dillenbourg, rapporteur
Prof Fabio Pianesi, rapporteur
Prof Marianne Schmid Mast, rapporteur

Lausanne, EPFL, le 1er Mars 2013

Abstract

Face-to-face interactions are part of everyday life, ranging from family to working in teams and to global communities. Social psychologists have long studied these interactions with the aim of understanding behavior, motivations, and emergence of interaction patterns. An organization is environment rich in daily interactions including structured periodic meetings, planning, brainstorming, negotiations, decision-making and informal gatherings and leaders play a key role in many of them. Leader face problems, propose solutions, make decisions, and often are the main source of inspiration of the employees. Identifying emergent leaders at early stages in organizations is a key issue in organizational behavioral research, and a new problem in social computing. The study of this phenomenon requires sensing of natural face-to-face interactions, automatic extraction of behavioral cues and reliable machine learning algorithms to identify emergent leaders. In this thesis we present a computational approach to analyze emergence of leadership in small groups using multimodal audio and visual features.

In the computational framework, we first present an analysis on how an emergent leader is perceived in newly formed, small groups. We present the ELEA (Emergent LEadership Analysis) corpus collected with the aim of analyzing emergence of leaders. We propose to analyze emergent leaders, using a variety of nonverbal cues studied in social psychology and automatically extracted from audio and video streams. Our analysis address how the emergent leader is perceived by his/her peers in terms of speaking and visual active, and its relation with the most dominant person (including external observers' perception). We then propose to investigate which individual nonverbal channel (or combination of features from different channels) provides better inferences of the emergent leader and related concepts

using unsupervised and supervised methods. We use a supervised collective approach which adds relational information to the nonverbal cues and compare its performance, with the performance of supervised (non-collective) and unsupervised methods. We also propose to capture the social visual attention patterns from automatically extracted features from video, in order to analyze who receives or gives the largest amount of visual attention in the group. Finally, with the aim of understanding who receives the largest amount of visual attention while speaking and who has the highest dominance ratio (i.e., many occurrences of looking at others while speaking and few occurrences of looking at others while not speaking). We synchronize the audio and video streams to capture the speaking and attention activity patterns.

We end our analysis exploring the impact of the verbal content (language style) in the interactions and its influence in the perception of emergent leaders. For the language style analysis, we propose to compute word categories extracted from manual transcriptions of the discussions as well as from automatically detected keywords. We propose to use a supervised method to obtain the relevant features, and to use only the top word categories to predict the emergent leader and related concepts in each group. We then propose to differentiate word categories, between highly context-related and context-free, to explore the feasibility to infer the emergent leader in a fully automatic approach from the context-free language style.

This dissertation address an audio and visual analysis of the ubiquitous phenomenon of emergent leadership in a fully automatic computational approach from face-to-face interactions. The nonverbal behavioral analysis is inspired in previous works on social psychology in the context of emergent leadership and related concepts. The automatically extracted nonverbal features are modeled to feed state-of-the-art machine learning techniques in order to infer emergent leaders.

Keywords: social computing, emergent leadership, perceived dominance, nonverbal behavior, audio-visual feature extraction, language style, unsupervised methods, collective classification.

Abstract

Les interactions en face á face font partie de notre vie quotidienne; elles interviennent dans le cadre familial, au travail avec les collègues et dans les réunions entre amis.

Les psychologues sociaux ont longuement étudié ces interactions avec pour objectif de comprendre les comportements, les motivations, et l'émergence de modèles d'interaction. Une organisation est un environnement riche en interactions incluant des réunions périodiques, des séances de brainstorming, de planifications, de négotiations, de prises de décision et de rencontres informelles. Dans ces interactions, le leader joue un rôle clé. En effet, le leader est la personne chargée d'affronter les problèmes, de proposer des solutions, de prendre des décisions et est souvent la principale source d'inspiration pour les employés. L'identification précoce de leaders émergents constitue l'un des problèmes principaux dans la recherche en psychologie comportementale du travail et nous l'abordons comme une nouvelle question de recherche en informatique sociale. L'étude de ce phénomène nécessite l'enregistrement d'interactions en face-à-face naturelles, l'extraction automatique de caractéristiques comportementales et l'utilisation d'algorithmes d'apprentissage par ordinateur dans le but d'identifier les leaders émergents.

Dans cette thése nous présentons une approche informatique pour analyser l'émergence de leaders dans de petit groupes. Nous utilisons des caractéristiques non-verbales et multimodales (auditives et visuelles) d'une maniére complètement automatique. L'analyse du comportement non-verbal est inspirée de la littérature en psychologie sociale traitant de l'étude de leaders émergents et de concepts liés. Les caractéristiques non-verbales extraites automatiquement sont ensuite utilisées par des algorithmes de

pointe d'apprentissage par ordinateur dans le but de recpnnaître les leaders émergents.

Dans ce cadre de travail informatique, nous présentons le corpus ELEA (Emergent LEadership Analysis) que nous avons enregistré dans le but d'analyser l'émergence de leaders. Nous proposons de faire l'analyse des leaders émergents en utilisant différentes caractéristiques non-verbales étudiées dans le domaine de la psychologie sociale et extraites automatiquement à partir d'enregistrements audio et vidéo. Notre analyse aborde comment un leader est perçu par ses pairs au travers de ses comportements visuel et vocal, ainsi que la relation entre la perception des concepts d'émergence de leaders et de domination sociale. Nous proposons donc d'analyser quelle modalité individuelle non-verbale (ou la combinaison de caractéristiques provenant de différentes modalités) offre la meilleure performance dans la reconnaissance automatique de leaders émergents et de concepts liés. Pour ce faire, nous utilisons des techniques d'apprentissage par ordinateur supervisées et non-supervisées. Nous proposons aussi d'encoder des motifs d'attention visuelle à partir de caractéristiques extraites automatiquement de la modalité visuelle dans le but d'analyser qui donne ou reçoit la plus grange quantité d'attention visuelle, ainsi que les motifs synchronisés d'attention visuelle et d'activité de parole. Pour ce faire, il est nécessaire de synchroniser les flux audio et vidéo afin de capturer des motifs d'attention visuelle et d'activité vocale.

Nous étendons notre analyse en étudiant l'impact du contenu verbal (par le biais du style de langage) dans les interactions et leurs influences dans la perception du leader émergent. Pour effectuer cette analyse, nous proposons d'estimer les caractéristiques en relation avec les catégories de mots extraits de transcriptions manuelles de discussions, aussi bien que de mots-clés automatiquement trouvés. Nous proposons d'étudier des techniques d'apprentissages supervisées pour sélectionner les charactéristiques importantes et de reconnaître les leaders émergents (et concepts liés) dans chaque groupe.

Dans l'ensemble, notre travail représente le premier essai d'automatiser complètement la tâche de reconnaissance de leaders émergents, tout en présentant des connexions importantes avec des travaux récents dans le domaine de l'informatique sociale dans le domaine de la reconnaissance du comportement dans le cadre professionnel.

Mots-clés: informatique sociale, émergence de leadership, domination sociale perçue, comportement non-verbal, extraction de caractéristiques audio-visuelles, style de langage, méthodes non-supervisées, classification collective.

To my loved parents Prisca and Alejandro†
for their inmense love and support.
To my dear brothers and sisters.
To my nephews and nices.

Acknowledgements

I want to start saying thanks to my thesis director Dr. Daniel Gatica-Perez, for his valuable support and guidance, wise advises and support all along the four years.

Special thanks to the committee members, Prof. Jean-Philippe Thiran, Prof. Pierre Dillenbourg, Prof. Fabio Pianesi and Prof. Marianne Schmid Mast for their time invested reviewing this thesis, important comments and feedback to improve the document.

Thanks to the collaborators in the SONVB project for their valuable feedback and rich exchange of information during the workshops, Prof. Marianne, Denise, Laurent, Gokul and Daniel.

The moral support, trust and infinite love from my loved family, have been an important factor in every stage of my life. Thanks a lot to my parents for encouraging me to continue, for their unconditional love, love for life and example of tenacity. Thank you Ismael for all your advises and trust, and from the love received form your beautiful and joyful family (Rosy, Ale, Pepe y Karlita). Thanks to my brothers Juan and Alejandro for their charm and nice memories from your families (Mely y Tonio, as well as Tere, Jair and Alelhi). Thank you Daymirey (Mimi) and Esau to be always there for me, and thanks Mimi to come and share amazing holidays (like in the old times). Big thanks to Orquidea for cheering me and the warm conversations that remind me how lucky I am to have you as my sister. Thanks Carolina, Magnolia and Carito for spoiling me (mimarme ;)) and make my days in Mexico, memorable!. Thanks Angelica and Robert, and Blanca and Angel, for your consideration and for sharing great news from your families, and also for receiving with charm and let me enjoy time with my dear nieces and

nephew (Ashbie, Alenka and Angelito). Thanks to all my enthusiasts and loved nephews and nieces, remember that you are always in my thoughts.

Because you made an agreeable environment in every working day, thanks to my lab-mates Paco, Joan, Oya, Laurent, Min-Tri and Gokul, and welcome to Darshan and Dinesh. To the social computing group and from whom I received plenty of advises, feedback, made possible some collaborations and because we could share spare time in the traditional summer and winter activity: Daniel, Oya, Dinesh, Kate, Radu, Paco, Joan, Laurent, Min-Tri, Darshan. Since coffee is an important part of a PhD student, Gracias for the great coffee break times Paco, Joan, Laurent, Ivanna, Leo and Marc. Of course special thanks to my girl friends with whom I enjoyed life out of the work and for the excellent times (including the zumba class), Serena, Sofia and Vicky. Thanks as well to the summer running group, with whom I enjoyed sunsets along the river in Martigny: Serena, Nicolae, Ivanna, Vicky (and sometimes Kay) and Elie. I must include acknowledge to my unique flatmate Radu for his friendship, his patience (during my practical French lessons), and for all the times that he had the dinner ready at my arrival :) (ah! and for organizing the raclette dinners). Děkuji Petr for the fruitful collaboration, your skiing lessons, and the challenging czech lessons, and most importantly for your personal support and love. Thanks also to your charm family for making me feel like at home (Stephania, Vlastik, Zuzan, Adam, Tom and Dan).

Merci beaucoup Michel Salamin! for your friendship and to help me to start and now to improve my French level, as well as for the events that you organized to motivate our integration in Valais (brisolee and the traditional ludge). And also, thanks to my classmates: Nicolae, (sometimes Kenneth), Ivanna, Manuel and Marco for the nice class-times. Thanks to the always present friday beer group at cafe du midi. Also, thanks the gang in Lausanne with whom I shared great times in Lausanne area and Zermatt: Robert, Ganga, Monique and Stano (and welcome to Andrej), Tatiana, Bebe and Andrea.

To my friends that keep cheering me in the distance Noe and Argelia, with whom I shared happy moments and for their comforting words during sad times. Thanks also, Antonio for the spare talks that make me laugh even in stressing times.

Moreover, thanks to Idiap for hosting me and of course the kind personal, Nadine who is always in the better disposition to help in official and unofficial matters, to Silvye, Ed and Chris, admin and accounting team, and the IT guys for their quick responses and kindness. Thanks also to Chantal for her kindness and sincere smile every time that I came to her for help at EPFL.

Finally, special thanks to CONACYT, which award me with a doctoral scholarship, that made my staying at Idiap/EPFL possible.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

Chapter 1

Introduction

1.1 Introduction

On daily basis, we all get involved in multiple social situations with the aim of sharing thoughts and emotions, or to establish relationships in different contexts. The established social interactions are rich communication phenomenons that have been analyzed in social psychology (13, 88). Psychologists have investigated the individual personality and motivations that affect the dynamics in teams (36, 55, 60, 66, 96) as well as the connection between nonverbal behavior and the vertical traits that emerge in groups including power, dominance, influence, competence and leadership (4, 29, 30, 40, 61, 66, 71).

In interactions among the members of a group, the leader is an agent of change, a person whose acts affect other people more than other people's acts (14). When group members meet for the first time, the concept of zero acquaintance in groups emerges (3), and group members use all the verbal and nonverbal behavior available from the other members as basis for their first impression. Leaders play a critical role in teams that has implications on cooperation, cohesion, communication, and coordination towards accomplishment of goals (60, 61, 64). An emergent leader is defined as the person who arises in a group having its leading force from the sympathy of the group (105). The emergent leadership is a key research area in social psychology and there are a number of works that analyze this phenomenon through the verbal and nonverbal communication channels (12, 14, 60, 61, 104).

Social computing is a recent research area in computer science that, among other problems, is examining problems traditionally studied in social psychology, and model-

ing them through automatic means. In the existing literature, automatically extracted audio and visual nonverbal features are used to infer personality traits including extroversion (84), individual variables like dominance (51, 89), and group constructs like influence, cooperation, and competition (15, 51). Our work builds upon the existing body of computational work regarding social interaction sensing and extraction of behavioral cues, and makes an original contribution by focusing on a different aspect of social interaction, emergent leadership.

1.2 Thesis' Objective

In this thesis we investigate computational methods that allow automatic inference of emergent leadership in face-to-face, zero acquaintance teams from audio-visual cues. The study of this problem is novel in computer science, and calls for the integration of knowledge and techniques from social psychology, perceptual processing, and machine learning.

1.3 Motivation

This research adds to recent work done in social computing, that aims to model human patterns in the context of small groups. Since technology is improving fast, automation speed, and acceptable performance of feature extraction and inference methods could be expected in few years. Taking this in consideration, and the relevance of leaders in organizational settings, there are potential applications that could be implemented:

Support of leadership research in social psychology. The emergent leadership topic, previously explored through manual observations and coding, could benefit from the automation of the process of extraction of behavioral cues from audio and video data typically recorded in laboratory settings, at larger scales, and in a relatively short amount of time. The practicality of the approach is to facilitate the work of psychologists, who could concentrate more of their efforts on the analysis and interpretation part, rather than in the coding part.

Leadership skills assessment. Assessment training centers nowadays are either providing theoretic information on the desired leadership skills, or providing personal-judgment feedback based on short dyadic exercising interactions. Most traditional

feedback is based on the general perception from the trainer-judge on the dyadic communication on the defined leader role, involving the theoretic aspects, i.e., listening carefully, sharing information by disclosing own ideas, motivating the other person to speak, and finding a solution that the other agrees with. Although the approach has been beneficial in dyadics, trainers-judges could be missing relevant behavioral information that occurs on group interactions. The general communicative behavior could be captured and processed automatically to provide quantitative measures on speaking turn dynamics like interruptions, control of tone of voice, pauses and possibly the content of the speech (e.g., the use of words that imply achievement or agreement). The computed outcomes could be presented visually and serve as the base for the feedback on the communication patterns and the leadership skills. The assessment could serve as training for leaders in organizations, as well as, to discover and exploit leadership potential of young students and entrepreneurs.

Enacting evidence for hiring decision. In addition to traditional approaches in hiring decisions of new colleagues (typically based personality tests, CVs, references, etc.), having automated scores on behavioral, interactional, and perceived leadership skills, could augment the static data information and provide additional information to support the hiring decision process.

1.4 Summary of Contributions

The main contributions of this thesis are summarized as follows:

First, we propose a computational framework to infer the emergence of leadership and related concepts in teams. Although some works on social psychology have analyzed the emergent leadership phenomenon from manual annotations, there has not been a fully automatic approach in order to infer emergent leaders from face-to-face group interactions using multimodal cues. We propose to automatically extract several nonverbal features from audio and video streams, followed by the computation of unsupervised and supervised methods in order to infer the emergent leader in a group using nonverbal features (98, 100). For the approach, we propose to design and collect a data corpus using novel audio and video commercial portable sensors (98).

Second, we perform a rich multimodal nonverbal behavioral analysis in the emergent leadership context, using automatically extracted features from audio and video.

1. INTRODUCTION

We propose to analyze the emergent leaders nonverbal behavior in several modalities (i.e., audio, visual, audio-visual, verbal). We show the feasibility to infer emergent leadership and related concepts by using only audio nonverbal features (99), that captures speaking activity, turn dynamics and prosody. We propose to analyze the visual activity and motion from the participants in the interaction, and compare the inference using only audio features, only visual features, and the aggregation of audio and visual features (100). We address as well, the analysis using synchronized cues (i.e., speaking activity and visual attention (98)). And, we also propose to analyze the verbal content of the emergent leaders in the interaction, by extracting spoken keywords (102).

Third, we propose a framework to automatically infer the emergent leadership and related concepts using state-of-the-art machine learning techniques. We propose the use of supervised and unsupervised methods. The unsupervised methods, specifically rank-level fusion show a reasonably high performance when audio and visual features are combined, as compared with inferences using only features from one of the modalities (100). The relevance of this finding is that such technique does not require data for training, and accurate inferences can be done in a short amount of time. We proposed the use of support vector machines (SVM), as a supervised approach, which resulted in high performance as compared to baseline inferences. As a second supervised approach, we propose the use of collective classification, which considers relational information in addition to the information from the features. We propose to code the relational information from the speaking turn dynamics in the interacting group. The inferences using a collective approach improve SVM inferences in most of the cases, but not the inferences from the fused-unsupervised method. In addition, the collective classification approach improves performance when one of the labels is annotated, i.e., when we assume that we know a participant who is non-leader in the interaction, and the method predicts the emergent leader from the rest of the participants. We compare the inferences of the proposed methods and show that automatic inferences of the emergent leadership and related concepts are feasible and accurate by using audio and visual features, or their language style.

The following papers have been published in the computer science literature.

Journal papers:

- Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast and Daniel Gatica-Perez. "A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups". *IEEE Transactions on Multimedia*. 14(3-2): 816–832, 2012.

- Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast and Daniel Gatica-Perez. "Emergent Leaders through Looking and Speaking: from Audio-Visual Data to Multimodal Recognition". *Journal on Multimodal User Interfaces*. Published online August, 2012.

Conference and Workshop papers:

- Dairazalia Sanchez-Cortes, Dinesh Babu Jayagopi and Daniel Gatica-Perez. "Predicting Remote Versus Collocated Group Interactions using Nonverbal Cues". *In Proc. International Conference on Multimodal Interfaces (ICMI-MLMI), Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*. Cambridge, USA. November, 2009.

- Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast and Daniel Gatica-Perez. "Identifying Emergent Leadership in Small Groups using Nonverbal Communicative Cues". *In Proc. International Conference on Multimodal Interfaces (ICMI-MLMI)*. Beijing, China. November, 2010.

- Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast and Daniel Gatica-Perez. "An Audio Visual Corpus for Emergent Leader Analysis". *In Proc. International Conference on Multimodal Interfaces (ICMI-MLMI), Workshop on Multimodal Corpora for Machine Learning*. Alicante, Spain. November, 2011.

- Dinesh Babu Jayagopi, Dairazalia Sanchez-Cortes, Kazuhiro Otsuka, Junji Yamato and Daniel Gatica-Perez. "Linking Speaking and Looking Behavior Patterns with Group Composition, Perception, and Performance". *In Proc. International Conference on Multimodal Interaction (ICMI)*. Santa Monica, California. October, 2012. Outstanding Paper Award.

- Dairazalia Sanchez-Cortes, Petr Motlicek and Daniel Gatica-Perez. "Assessing the Impact of Language Style on Emergent Leadership Perception from Ubiquitous Audio". *In International Conference on Mobile and Ubiquitous Multimedia (MUM)*. Ulm, Germany. December, 2012.

1.5 Thesis outline

The thesis is organized as follows. In Chapter 2 we first review related work on the emergent leadership topic in the context of social psychology. In Section 2.3 we present related work on social-computing in the context of the emergent leadership study. We then present the existing publicly available corpora for analysis in small groups.

In Chapter 3 we describe the corpus design and data collection process in order to study the Emergent Leadership phenomenon in small groups. We present the available annotations of the corpus, followed by the description of the subsets used in subsequent chapters, to infer emergent leaders from individual and joint perceptual modalities.

In Chapter 4 we present our framework for nonverbal communicative behavior analysis of leadership perception and an automatic inference of emergent leaders and related variables. We first present a correlation analysis of how the emergent leaders are perceived based on their nonverbal behavior. Then, we present unsupervised and supervised methods to infer the emergent leader in the group from audio nonverbal cues. We also present an analysis of temporal aspects in order to infer emergent leaders.

In Chapter 5, we address the problem of automatically inferring emergent leadership from audio-visual features. We start describing the methodology for the feature extraction, followed by a correlation analysis with the perceived variables. Then, we present the machine learning methods applied to infer the emergent leader in the group and related concepts.

In Chapter 6 we present an analysis on the visual attention dynamics within participants, in order to infer emergent leadership and related concepts. We present an automatic multimodal approach to derive the group social attention, by combining speaking activity and attention features from the audio and video streams respectively. We also present a comparison between the performance of the multimodal approach and the single modalities (i.e., speaking activity and attention) in the emergent leadership and related concepts prediction task.

In Chapter 7, we present an approach to infer emergent leadership in small groups by using their verbal content (i.e., their language style) as opposed to their nonverbal behavior. We describe the process followed in order to extract the language style from manual transcriptions as well as from automatically extracted spoken keywords.

We describe and discuss the results, the method to automatically infer leadership and related concepts.

Finally, in Chapter 8 we present the summary of the contributions of this dissertation, the limitations of our research work, and we propose some extensions of this work.

Chapter 2

Related work

In this Chapter, we first introduce briefly a categorization of nonverbal behavior, and then we present the related work in the context of emergence of leadership in face-to-face interactions. Emergent leadership is a key research area in social psychology and there are a number of works that analyze verbal and nonverbal behaviors displayed by emergent leaders, how is an emergent leader is perceived by observers and how emergent leadership can be measured. Identifying emergent leaders is also becoming relevant in the social computing community, although is has been explored only on asynchronous and/or remote scenarios.

This Chapter is organized as follows. Section 2.2 describes the related work in social psychology in the context of analyzing emergence of leaders in small groups. In Section 2.3 we present the related work from the social computing point of view, centered in the emergent leadership topic. In Section 2.5 we present the existing publicly available corpora. Finally in Section 2.4 we mention related work close to our approach using machine learning techniques.

Parts of this Chapter, were published as part of journal papers in (98) and (100).

2.1 Nonverbal communication

Communication occurs every time over different modalities. In face-to-face communication the words represent the verbal information and the rest (tone of voice, posture, gaze, etc.) is called nonverbal behavior. Nonverbal communication is relevant as it conveys conscious and unconscious information while human interactions occur. Nonverbal

behavior is categorized as follows (63):

- **Kinesics** includes all kinds of behavior related to movement. It includes facial expressions, posture, gestures, and eye gaze patterns. Facial expressions display various emotional states like anger, sadness, surprise, happiness, fear and disgust. Postures are often related to the degree of attention or involvement, the degree of status, or the degree of liking; it also is a key indicator of the intensity of emotional states. Gestures refer to body gesticulations that can be speech-independent or speech-related. Speech-independent gestures are verbal translations of words or phrases that a large community is familiar with, for example gestures like "thumbs-up" for approval or "wish for good luck" using crossed fingers. Speech-related gestures are connected with speech and exemplify something said verbally, for example alternating while moving back and forth, the index and middle fingers while describing a "walking path". Eye gaze refers to eye movements in the general direction of another's face. Dilation and constriction of the pupils are related with indicators of interest, attention, or involvement.

- **Proxemics** refers to distance between people as they interact. The arrangements in spaces for conversational interactions, i.e., the distribution of the furniture, could benefit or affect the interaction. Certain arrangements could motive cooperation or collaboration (sitting side by side) or competition (by sitting opposite). In free spaces, the distance between people in face-to-face interactions also plays an important role that directly affects the eye contact and voice loudness. The distance or personal space area, varies based on the culture, and how much people know each other. Not respecting other's personal space on a first encountering, could cause an uncomfortable and inclusive conflicting interaction. The fact of demanding larger personal spaces could also be a signal of status or power.

- **Haptics (touching)** could be expressed as self-focused or directed to others, and may reflect a person's particular state. The interpersonal touch varies based on the context and the relationship established between the people who interact. Touching is used in professional environments as part of welcoming greetings, as for example handshakes that convey warmth, friendliness and suggest equality. Touching is used in daily communicative situations to express emotions, appreciation or attraction in close relationships, as for example hugs and kisses. Touch is

also used to express support and partnership within team sports, as well as negative emotions like hitting or slapping. Finally, comforting and relaxing aspects are also perceived by patients being touched by nurses in hospital environments.

- **Acoustic (or vocalic)** refers to how something is said. There is a distinction between two types of sounds: variations made with the vocal cords (such as tone, pitch, silence and accent, collectively known as prosody) and sounds that result primarily from physiological mechanisms other than the vocal cords (pharyngeal, oral, or nasal cavities). The prosody is consciously used to complement speech when a person is asking questions or finishing sentences, by raising or lowering pitch. Unconscious expressions of anxiety, anger, disgust, fear or boredom, can be perceived by others, trough changes in the speech rate, fluency, pauses and tone of voice. The vocalic cues, play an important role in the verbal interactions, such that it shapes the responses in the communicative situation.

In everyday communication, we make use of the nonverbal behavior to complement the verbal channel. Its expression varies based on the place, means of communication and people involved in the interaction. On one hand, we use the nonverbal channel in a conscious manner to accompany speech or to express emotions and feelings. On the other hand, most of the behavioral signals prompt unconsciously, triggered by the environment, the situation of the communication and the people involved in the interaction.

To convey the emergence of certain nonverbal behaviors and its patterns in specific situations, psychologists continue analyzing nonverbal behavioral signals in different face-to-face communicative interactions.

2.2 Emergent Leadership in Social Psychology

Psychologists have used nonverbal behavior as an important source of information in the analysis of the vertical dimension in social interactions. The vertical constructs studied include dominance, status, power, and leadership. In the present work we focus on the emergent leader, defined as the person who emerges as the leader in an interaction, rather than being assigned the authority and has his/her base of power from the sympathy of the group, rather than from a higher position (105). Given

that emergent leadership has been measured using different concepts (dominance, influence, leadership, control), we review the literature concerning all of these aspects of verticality.

The initial studies on emergence of leadership and nonverbal behavior date from the mid-seventies. In 1975, Stein (104) conducted a study on perception of emergent leadership using scenarios in which leaderless groups of eight or nine members worked weekly throughout the semester on a research project. Observers were able to identify emergent leadership in small groups from both verbal and nonverbal information using 20 minute edited recordings from the initial 45 minute meetings. Verbal communication was transcribed from videotapes. Nonverbal communication was tested with a visual-only setup and an audio-visual setup, where the audio was filtered such that it provided only acoustic nonverbal information. For emergent leadership, the highest correlation values were obtained with participation cues, particularly the relative amount of time each group member spent talking.

In (12), Baird used visual nonverbal cues to predict emergent leadership in a scenario about reaching consensus on a single statement in a group of five people, in which volunteers from a introductory speech communication course were placed randomly. The videotapes were 20 minutes in length, recorded at different times in the meeting. At the end of the discussion each participant voted for the emergent leader, defined as the most influential member in the group. Interestingly, arm and shoulder movements were found to be the main nonverbal visual cues contributing to participants' perception of leadership. Additionally, gesticulation of shoulders and arms were significantly correlated with eye contact, head agreement, and facial agreement. Where facial/head agreement is an expression/gesture used to expresses approval/disapproval or support.

The relationship between competence and dominance in face-to-face groups was analyzed in (4). Four-person groups of unacquainted people were recorded during 45 minutes while creating an organization and outlining its strategy. A self-dominance report questionnaire was administered, and group members also rated each other on influence, competence, and personality. In addition, external observers rated each member along the same dimensions. The study concluded that, by acting competent, dominant people influence their group more than individuals who are less dominant. In behavioral terms, and in order to attain this influence, dominant people speak the most,

and gain more control over the group and the group decisions. Similarly, Wentworth et al. found that when a person engages with the group by displaying a certain level of knowledge and expertise in problem solving tasks, i.e., providing ideas and guidance, this contributed to attain a leadership position (112). The findings support previous studies in functional roles in teams. According to Benne et al., the leader should be able to approach the members with conscious skills, to make a group work productively, and to maintain a sense of belonging to the group (16).

Other connections between power and dominance, and nonverbal behavior have also been studied in face-to-face interactions. Social power refers to the perception of what a person knows or could do (40, 41). Studies on dyads, revealed that participants in high rank status roles took wider personal space, pointed and touched more, and talked and interrupted more than participants in low status roles, based on a teacher-student scenario (65). Dovidio analyzed the connection between social attention and dominance in conversational scenarios (28). He found that people higher in status or trying to control others, tend to look at others while speaking, but they tend to display disinterest by not looking at others while listening. His findings revealed that a higher visual dominance ratio (i.e., the proportion of looking-while-speaking and looking-while-listening) is a relevant behavioral signal of dominance (28).

Another study (60) investigated the relationship between leadership style and dominance, particularly in sociable and aggressive forms, in the context of three unacquainted people trying to decide on the top five candidates out of a group of ten persons who wanted to rent a room. The 20-minute group discussion was recorded, and responses to questionnaires (first-impression of dominance, socio-emotional and task leadership) were complemented with observations of nonverbal behavior. It was found that although both types of dominance characterize leadership, there was a higher correlation between leadership and social dominance. Aggressively dominant people often attempt to interrupt more, and look at others less while listening (60). Similarly, Schimd Mast found that scores on dominance correlate with total speaking time and average turn duration (71).

The relationship between leadership and personality traits is also of interest to social psychologists. It is showed that cognitive ability and the personality traits of extroversion and openness to experience were predictive of emergent leadership behaviors (61). Groups of four to six participants enrolled in a course took part in a winter survival

simulation, and filled in questionnaires of personality, cognitive ability, teamwork effectiveness, and emergent leadership. The emergent leader was designated as the one receiving the highest rating scores from the group through measures of interpersonal and self-management behavior, as well as task-related behaviors of a leader. The emergent leaders scored higher on cognitive ability and the personality traits of extroversion and openness to experience. In the last decades, new findings in psychology reveal a strong connection between personality traits and linguistic cues in written or spoken forms (69, 81). Even more, word usage cues also provide information in the prediction of successful relationships (44), and electability of presidential candidates (103).

In summary, the literature in psychology has found that human observers can identify emergent leaders in group interactions, and that specific nonverbal and verbal cues do correlate with emergent leadership. These key findings provide the motivation and basic supporting evidence for our automatic approach.

2.3 Emergent Leadership in Social Computing

After a review of the existing literature in social computing, we found only few approaches centered on the computational analysis of emergent leadership. In (22), the emergent leadership phenomenon was analyzed focused on self managed virtual teams with no roles assigned and engaged in collaborative tasks. The interactions of 22 virtual teams (from students registered in semester course in three different universities) were captured with a web-based collaborative technology. The team performance was assessed through manually coded leadership behaviors and messages exchanged within the group members. As main findings, self managed teams showed more emergent leadership behavior focused on performance (producer behavior).

In (109), the emergent leader was found to be popular among his/her team and became the center of the network, such that he/she connects fast with all of them. For their analysis, 25 groups were recruited, with 5 participants each. The participants were students registered in a semester course. Each group was assigned to work on a project during 4 weeks in a virtual learning work space, and the interactions between the participants were coded from the posted messages in a discussion board. To identify the leader, they used social networks measurements, combining degree of centrality, closeness and betweenness.

The leader in a social interactive activity has been explored by means of estimates of the direction of the synchronization in string quartets in (111). For the analysis, players were recorded (audio and video including top and frontal view) while wearing physiological sensors, accelerometers and audio contact microphones on each instrument. The player-leader in the group was derived from scores on pauses, attacks and changes in the dynamics of a performed musical piece in a pair-wise approach.

The few existing approaches (22, 109, 111) are evidence that the emergent leadership phenomenon is a relevant research subject that could be modeled in face-to-face interactions with state-of-the-art machine learning techniques.

2.4 Machine Learning Techniques to Recognize Small-Group Socio-Psychological Constructs

Several machine learning methods have been explored to model different social variables in small group interactions.

Some research works have focused on unsupervised methods, using inferences based on rules or rankings in order to predict roles, dominance and high status (5, 42, 50, 89); Gaussian mixture models to infer group dynamics (101); and probabilistic topic models to infer dominance or type of leaders (48, 49). The inference of the diverse vertical traits is based on automatic extraction of nonverbal behavior, coded from small group interactions (three to four participants). Supervised methods (e.g., support vector machines (SVM)) have also been used in the context of modeling individual and group behavior to infer dominance (50, 92), roles (85, 86, 113) or personality (84). Moreover, boosting algorithms have been used to infer personality traits (110) in small groups. The inference on dominance has been explored using acoustic turn-taking cues, visual cues alone, and combinations. Better results were often obtained using speaking nonverbal cues, confirming previous findings in social psychology (40, 71). One of the main challenges for the use of supervised learning is the need for manually labeled data, which can be time consuming to generate given the need to collect and label data.

In addition to static models, dynamic models have been explored. For example, standard graphical models have been applied to classify roles or group actions (such as Bayesian networks, Conditional Random Fields and Hidden Markov Models (HMMs) (73, 94, 95). Several acoustic and visual communicative cues have been

automatically extracted in conjunction with these methods, inspired by some of the behavioral cues studied in social psychology described in Section 2.2. In the last decade, in order to study informal face-to-face interactions, few models have been proposed. Otsuka et al., proposed a Markov-switching model (78) to infer the structure of a conversation. Similarly, Dong et al., adapted the influence model proposed by Asavathiratham (9), in order to recognize functional roles (15, 27). Based on a N-chain coupled Hidden Markov Model (HMM), it estimates the conditional distribution for a given chain taking a convex combination of the pairwise probabilities. The graphical model predicts functional roles (e.g. the producer of ideas and the seeker) in small groups using speaking cues derived from the conversational scenario. More recently, researchers have used the Granger causality (based on time series) to study the causal effect of dominance (58, 59). The method adds past observations of visual and speaking cues of the dominant participant to the future observations of the other participants in the group. Then the causal density is computed, and it allows further estimation of the casual flow, i.e. how much a dominant participant's behavior affects others, rather than been influenced by others in terms of behavior.

In summary, although existing machine learning techniques and new models have been used to model social constructs, work on research lines continues to grow with the aim of improving performance in three main tasks, dominance, roles, and personality. A detailed discussion on state-of-the-art of features and techniques can be found in extensive surveys on the topic in (7, 33). In our case, we focus on emergent leadership, which represents a new computational issue.

2.5 Existing Data Sets for Small Group Interactions

Most of the corpora that have been collected to study behavior in small groups centered their attention on meeting scenarios where realistic rich interacting patterns can emerge. A detailed look into these corpora reveals a variety of design choices. To promote the interaction between participants, either real or scripted scenarios have been used. The data has been recorded with a wide range of audio-visual sensors, including close-view and mid-view cameras, close talk microphones and microphone arrays. The collected data has been annotated for different aspects, in parallel with the research question in

Table 2.1: Corpora available for small-group interaction study. The audio sensors/microphones include CTM-close-talk; EWM-earset wireless; TTM-tabletop; LAM-lapel; SBM-sociometer badge; ARM-microphone array; ODM-omnidirectional; FCM-four-channel cardioid; OTM-Other distantly placed microphones. Video sensors include CU-close-up; VC-video camera; WC-webcamera; C360-360 degree camera. Personality annotations correspond to LCB-Craig's Locus of Control of Behavior scale, E-BFMS-Extroversion part of the Big Marker Five Scales, NEO-FFI-NEO-Five Factor Inventory, PRF-Personality Research Form.

Corpus	Audio/video	Questionnaires/annotations
VACE (23)	up to 8 EWM, OTMs, 1 OD and 1 FC 10 VC	transcripts, dominant speaker, language metadata, gesture
ICSI (47)	4 to 8 CTM	involvement
ISL (19)	3 to 9 LAM 3 VCs	word tokens, turns, question, non-question, disfluency
AMI-12 (57)	4 CTM, 4 LAM, 1 ARM 4 CU and 3 VC	transcript, addresses, gaze, adjacency pairs (question-answer, statement-agreement)
AMI-40 (93)	1 ARM 4 CU and 3 VC	influence ranking (inter-ranking) dominance
AMI (21)	same as AMI-12 and AMI-40 same as AMI-12 and AMI-40	same as AMI-12 and AMI-40, hand and head gestures
DOME (8)	same as AMI-12 same as AMI-12	same as AMI-12, dominance annotations
M4 (73)	12 microphones (ARM and LAM) 3 VC	transcript, word segmentation, interest level
NIST (32)	3 to 9 CT, LAM and OTMs 5 VC	transcript, speaker segmentation
ATR (20)	1 ARM 1 C360, 1-6 VCs	none
MIT (62)	4 SBM	dominance, questions and ideas, team performance
NTT (78)	4 LAM 3 VC	regime estimates (class + directionality) head direction
MSC-1 (85)	4 CTM, 6 TTM and 7 ARM 5 VC, 4 WC	functional relational roles (task area and socio-emotional)
MSC-2 (70)	4 CTM, 1 ODM same as MSC-1	LCB and E-BFMS, group cohesion, individual and group performance

mind. Table 2.1 summarizes the available corpora focused on small group interactions described in this section.

The Video Analysis and Content Exploitation (VACE) meeting corpus was recorded using real-world scenarios (war games and military exercises) at the U.S. Air Force Institute of Technology (AFIT) (23). The aim was to understand the structure in meetings where the objectives are clearly defined, the roles and hierarchy are known, and the planning activity is present.

Natural weekly discussions of a research group, with known roles and hierarchy, were recorded at the International Computer Science Institute (ICSI) conference room (47). The goal of this corpus was to offer resources to improve automatic speech recognition, transcription, prosody, and dialog modeling.

At Carnegie Mellon University (CMU) another corpus collected real and scripted meetings on scenarios such as project planning, military exercises, games, chatting and discussion (19). The aim of the ISL corpus was to distinguish between different kinds of meetings by characterizing speaking styles.

In the Augmented Multi-party Interaction (AMI-12) corpus, collected at the Idiap smart meeting room (57), the meeting participants had predefined roles and they followed a script. Apart from audio and video resources, a variety of manual annotations that involve verbal, nonverbal and contextual features are available. To study dominance, the Dominance in Meetings (DOME) dataset included dominance annotations on a subset of the AMI corpus, containing 10 hours of Idiap-AMI meetings (8). To analyze participants' influence in project scenario meetings, a part of the AMI corpus was analyzed, containing 40 meetings recorded at the Organization for Applied Scientific Research in Netherlands (TNO-Soesterberg) (93). Several manual annotations are available for this corpus, mostly derived from the audio channel.

Another approach for capturing small group meetings is to use wearable sensors that are able to gather nonverbal signals and proximity data from short distance transmitters. In (62), a corpus was recorded with a wearable sociometer based on two scenarios: brainstorming and problem solving. The aim was to detect social interactions (including dominance) and to promote group collaboration (through real time feedback). For this corpus, nonverbal features and self-reported dominance annotations are available.

Group participants involvement has also been analyzed in business-like meetings. In (20), the ATR speech corpus is presented, which includes recordings of monthly

sessions from a real group project meeting. The main goal of this corpus is to identify the type of participation and the flow of the discourse.

The NTT corpus (78) was collected with the aim of inferring the structure of the meeting and the participants' roles. The corpus contains discussion scenarios in which no roles were assigned. The collected data includes audio, video, and head directions extracted from head-worn sensors.

Among the multimodal corpora in the literature, the closest to our work is the Mission Survival Corpus (MSC-1 and MSC-2) (70, 85). The data comprises small groups performing the winter survival task. The MSC-1 focuses on individual behavior during the decision making process; it includes audio and video recordings of four participants and functional role annotations. The MSC-2 focuses on analyzing performance, group cohesion, and personality, and used the same video recording resources used in MSC-1; in addition they performed online 3D multi-person tracking during the interaction. For audio recordings, they reduced the number of sensors to 4 close-talk microphones and one omni-directional microphone placed on the top of the table. The MSC recordings differ from our corpus in terms of participants, given that participants at MSC-1 knew each other. In terms of settings, both corpora (MSC-1 and MSC-2) used a static setup and all the meetings are recorded in a static location in a smart room.

The aim of the multimodal corpora summarized above is to analyze the multimodal human behavior in diverse settings. Although the recorded scenarios allow the study of several behaviors that emerge in small face-to-face group interactions, *the emergence of leadership* has not been studied in those settings. The existing corpora has one or several of the following limitations that do not allow the study of emergent leadership, 1) participants know each other, 2) participants have roles assigned, 3) participants follow a script or, 4) participants are aware of hierarchy levels within the group. Nevertheless, our review of the collected scenarios, provide us with the learning about several recording settings that can be used in order to have high quality feature extraction. The existing audio and video recording settings, served also as inspiration to design an ad-hoc scenario to computationally analyze emergence of leadership in face-to-face interactions.

2.6 Conclusions

In this chapter we reviewed the existing related work on emergent leadership. First we noticed that most existing works done in the emergent leadership phenomenon in face-to-face groups, have been mostly explored in social psychology from manual annotations of nonverbal behavior. Second, we showed that only few existing works on social computing have addressed the emergent leadership problem from a decision making point of view, in asynchronous scenarios using written forms. And, we also reviewed the existing audio and video recording settings, from which we learnt how to design corpora that is specifically collected to address research topics in small group interactions.

The evidence of works done on small groups, existing deployed recording sensors, and automatic extraction of nonverbal features, as well as existing studies in social psychology on the emergent leadership context, served as a starting point to propose this thesis research work. We aim to address the emergent leadership phenomenon in face-to-face zero-acquaintance groups, from audio and visual cues.

Chapter 3

The Emergent Leader Analysis Data Corpus

In this Chapter we discuss our experience in designing and collecting the ELEA (Emergent LEader Analysis) data corpus called , and describe the use of a portable setup to record small group meetings. The corpus was gathered with the aim of analyzing emergent leadership as a social phenomenon that occurs in newly formed groups. For each group in the corpus, the participants performed the winter survival task. The annotations of the corpus include self-reported personality, concepts related to leadership, and participants' performance in the survival task.

We use a portable recording setup which allows to record a small group meeting anywhere. Although the survival scenario is not completely natural, in the sense that the participants are gathered for the purpose of data collection and are given a task, the meeting they participate in is natural, without any predefined behaviors. As defined in Chapter 1, an emergent leader is defined as the person who stands out from the group during a face-to-face interaction with no hierarchical roles (predefined); furthermore, he/she has the group's sympathy to lead (104).

This Chapter is organized as follows. We first describe the sensors and procedure to collect the corpus in Section 3.1. Section 3.2 explains the annotation (or coding) encoding scheme. Finally, in Section 3.3 we present the different subsets used in the subsequent chapters.

Parts of this Chapter, were published (97, 98, 100).

3.1 Corpus Collection

In this section we describe the setup, and instruments used in order to record the corpus. The colledted ELEA corpus is composed of audio signals, video signals, and questionnaire outputs from the participants and performance on the survival task. For the audio collection we used a fully portable setting and for the video we used two settings, one static and one portable. The full corpus contains approximately 10 hours of audio and video.

3.1.1 Scenario setup

Audio recordings were gathered using the Microcone, a commercial microphone array, designed to record small discussion groups (up to 6 individuals) with audio sample rate of 16kHz (2). As shown in Figure 3.1, the Microcone (dark object at the center of Figure 3.2-top) was placed in the center of the discussion table to capture the interaction. The Microcone automatically segments speakers, and provides audio for prosodic cue extraction. The high quality audio recorded by the device, allows for automatic speech recognition as well.

Figure 3.1: The recording sensors of the portable setup on the ELEA corpus. Two webcameras and the Microcone.

For video recordings, we used two setups, one static setup with six cameras, and one portable setup. The static setup comes from the Idiap smart meeting video resources (75), and is composed of four closeup views, two side views and a center view recording at 25 fps. The portable setup used two webcameras (Logitech®Webcam Pro 9000, see Figure 3.1), with video frame at 30 fps developed by Dr. Oya Aran (at

Idiap Research Institute). Figure 3.2 shows examples from the ELEA corpus from the portable setup and the static setup.

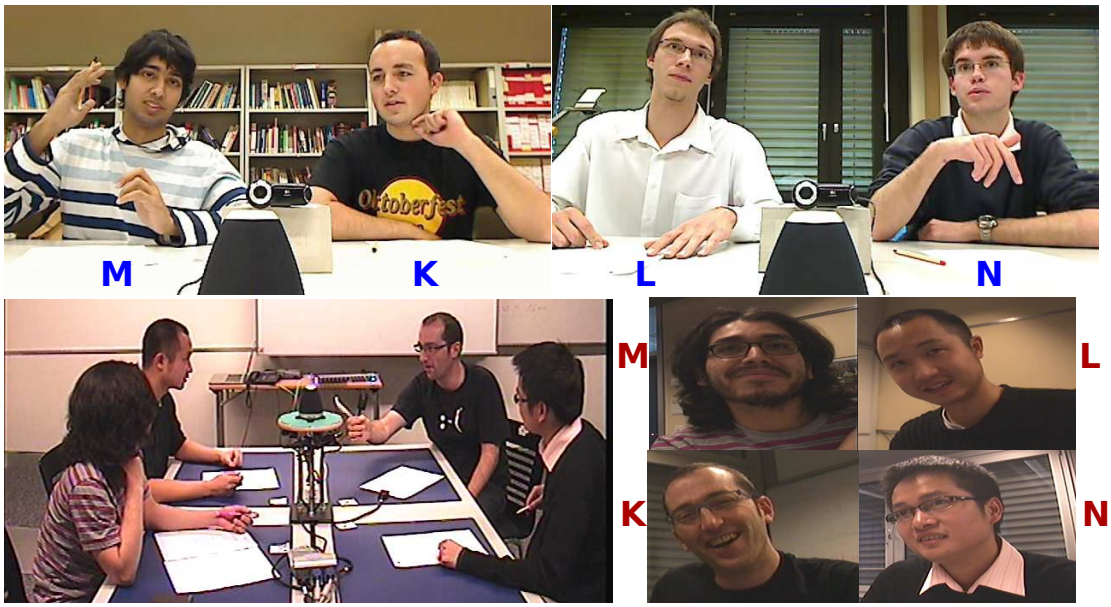


Figure 3.2: The recording setups of the ELEA corpus: Top - Portable setup. Bottom - Static setup (Central and closeup views). K, L, M, N are the participants IDs

3.1.2 Subjects

Potential volunteers were invited to participate in a study on casual social interactions. The invitations were posted in English and French offering a monetary compensation for their participation. Advertisements were placed in two universities, a research center and a business management school in French-speaking Switzerland. After participants contacted the researchers back by phone or email, they were informed of the process and, if they agreed to participate, their cellphone number and email address were requested. Since the participants were not supposed to have previous partnership or work relationship, ad-hoc groups were formed and participants were requested to attend the recordings.

148 participants were recruited (48 females and 100 males). With this population, mixed teams were formed; 28 teams were four-person and 12 teams were three-person. Average age was 25.4 years, with standard deviation 5.5 years.

3.1.3 Trust agreement

On arrival, participants signed a trust agreement. The agreement explained the process of the study, and informed them that audio and video recorded would be used only for research purposes and their identity would be anonymized. The agreement emphasized the participants' right to quit the study at any time. Participants were provided with a copy of the signed agreement, including the researcher's complete names and email addresses for their own records.

3.1.4 Task

There are several tasks that promote group discussion and decision-making. After reviewed the tasks most often used for training in assessments centers, we chose the winter survival task, given that it promotes interactions among the participants in the group and it is the most cited test in studies related to small group performance and leadership (61). The participants in the task were supposed to be survivors of an airplane crash in winter (56). They had 12 items that they had to rank in order of their importance, giving 1 to the item considered the most important to survive, 2 to the second most important, and so on. The task was performed first individually (5 min) and then we asked them to come up with the group ranking (15 min). The average length of the group discussions is 14.61 minutes, ranging from 8 to 19 minutes. Considering that not all the participants might be familiar with the items, we provided them with slides containing a picture and the definition of the item. The slides were consulted only during the individual ranking, to avoid the occlusion of the cameras during the group discussion.

3.1.5 Instruments

We first we administered the NEO-Five Factor Inventory (NEO-FFI) (26), which is a well known measure of the Big Five personality traits: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). We used the self-reported long version of the instrument composed of 60 items. Each item has a score from 1 to 5 ('Disagree totally' to 'Total agreement').

This questionnaire was followed by the Personality Research Form (PRF) (46). This questionnaire yields scoring for personality trait dominance. It consists of 16 true-false items.

Participants were also asked to answer 17 statements that capture how they perceive each participant (including themselves). 16 of the statements, developed by Prof. Marianne Schmid Mast (University of Neuchâtel) were evaluated on a five-point scale. Variables included in these statements are: perceived leadership (PLead: person gets involved, directs the group), perceived dominance (PDom: person is in a position of power, dominates), perceived competence (PCom: person is competent, is experienced) and perceived liking (PLike: person is kind, is friendly). The 16-item questionnaire can be scored from 1 to 5 ('Not at all' to 'Frequently if not always', respectively). Afterwards, participants provided a dominance ranking (RDom), i.e., participants were asked to rank the group, given 1 to the most dominant participant, and 3 or 4 for the less dominant, such that they have to include themselves in the ranking, similarly to previous work in dominance annotation (50).

Finally, participants were asked to provide additional information including age, and experience in practicing outdoor activities and winter sports in a scale from 1-5 ('Not at all'-'Frequently, if not always'). It was optional to provide additional comments to express their feelings during the interaction and about the process itself.

3.2 Annotations

This section describes the coding used to process the collected data and some statistics on the questionnaire data.

3.2.1 Subjects

To keep their identity anonymized, participants chose a letter and to link them with their respective questionnaires and audio/video files, the final identifier is defined as: number of group, participant letter, day and month of recording and a letter indicating the gender. Below, we describe the computations done from each of the questionnaires.

3.2.2 NEO-FFI

From this questionnaire, we computed mean values over the 12 items that correspond to each of the big five traits, taking into account that some items needed to be reversed. For each person we have a vector of five real values between 1.0-5.0, such that each value corresponds to each of the personality traits. Figure 3.3(a) shows the distribution of the self reported personality in the ELEA corpus.

3.2.3 PRF

Since this questionnaire has 16 items in the form true-false, we mapped the values to 1-0. From this questionnaire, each of the items measures dominant personality (8 true and 8 false-reversed items). Apart from the accumulative score, we also estimated the mean value.

3.2.4 Perceived interaction scores

For this 16 items questionnaire, we calculated mean values of 4 items for each of the perceived variables PLead, PDom, PCom, PLike, using the judgment from the other participants (i.e., not herself/himself). We consider as ground truth the annotations from the perceived interactions, such that the emergent leader in the group is the participant with the highest mean value of perceived leadership, and similarly for the related concepts. Figure 3.3(b) shows the distribution of the values for the perceived variables in the ELEA corpus.

3.2.5 Ranked Dominance

We calculated the value per participant as the mean value of the rank assigned from the other participants (i.e. excluding herself/himself), normalized considering the number of particpants in the group.

3.2.6 Survival task performance

Although there is no unique solution for the winter survival task, there is a ranking provided by experts, who justify the item rank order with more chances for survival. We used the survival experts' ranking list to code some variables related to performance

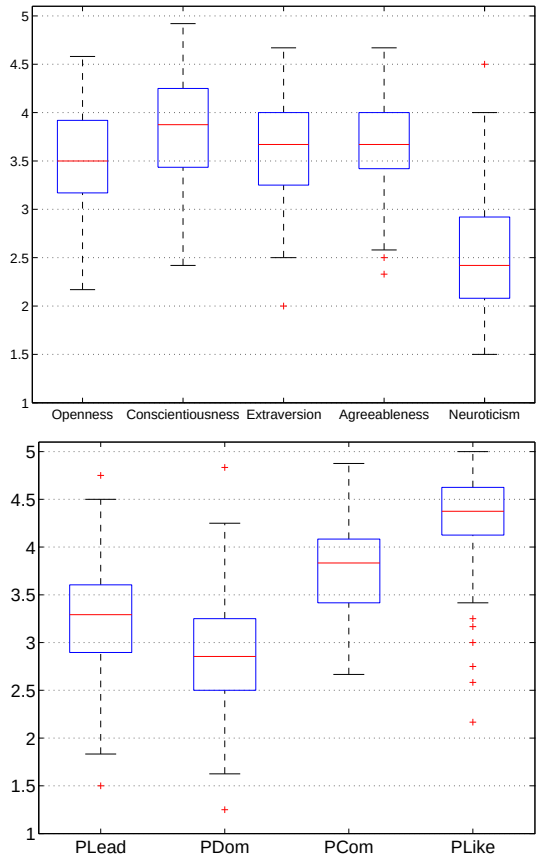
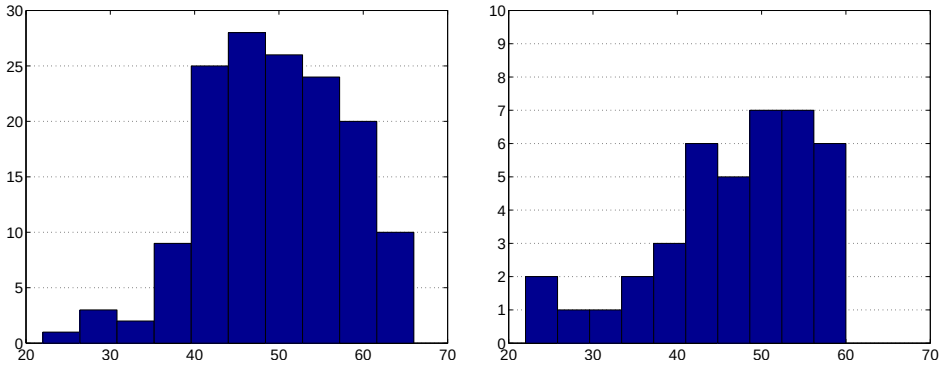


Figure 3.3: The plot shows median (central mark in the boxes), percentiles 25th and 75th (edges of the box), extreme datapoints (whiskers), and the outliers (+) for (a) personality traits (Openness to Experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism); and (b) perceived variables (PLead-Leadership, PDom-Dominance, PCom-Competence, PLike-Liking).



(a) (b)

Figure 3.4: The plot shows the distribution on the winter survival task performance. Of (a) individual performance as measured by AIS and (b) group performance as measured by AGS.

and influence. In addition, we coded two more variables considering the top-N ranked items. The variables are defined as follows:

AIS: The Absolute Individual Scores are calculated based on the absolute difference between the individual ranking list and the experts ranking list. The smaller the score, the better the answer, and according to the experts the more chances to survive. Figure 3.4(a) shows the histogram of AIS values.

AGS: The Absolute Group Score is calculated based on the absolute difference between the group ranking list and the experts ranking list. The smaller the score, the better the answer. Figure 3.4(b) shows the absolute group scores.

AII: The Absolute Individual Influence is calculated accumulating the absolute difference between the individual and the group ranking list. The smaller the score, the higher the influence of the individual in the group.

NTII: Number of Top Individual Items in the top group items, which counts the number of items in the individual list that also appear in the top items of the group list.

DTII: Absolute Distance of the Top items in the individual list with respect to the position of the same item in the top group list. If one individual item is not in the group top rank, it is assigned with the maximum distance + 1.

3.2.7 Perception from External Observers

As explained before, using the questionnaires that the participants filled based on their interaction, we extracted the perception of the participants themselves on their interaction. However, research shows that the perception of the participants themselves and external observers could differ (29). To be able to evaluate these differences, we also collected judgments from external observers for two of the variables, leadership and dominance.

We use the same questionnaire as filled by the participants, focusing only on leadership and dominance and excluding the questions related to other concepts. For each meeting, we assigned two external observers, one male and one female, who watched the first five minutes of the meeting video and answered eight questions for each of the participants in the meeting. The mean values were then calculated for the leadership and dominance variables of external observers, denoted as ELead and EDom.

3.3 Data Subsets

Among the 40 meetings in the ELEA corpus, 27 were completely recorded with the portable setup, and 10 with the static setup. In three meetings, the portable video recordings were not successfully recorded and thus discarded for experiments. Given that the ELEA corpus was collected with two video set-ups, and given some problems encountered during the recordings regarding video synchronization failures, different subsets have been defined in order to allow comparison.

3.3.1 ELEA-A

This data set refers to the full corpus, considering the 40 audio recorded meetings and the respective set of questionnaires. 12 groups were recorded with three participants and 28 with four participants. Among the 40 audio recordings, 29 groups were recorded in English, 10 groups in French and one group in Spanish. The full ELEA corpus is used in the experiments described in Chapter 4.

3.3.2 ELEA-AV

This data set contains only recordings gathered with the portable video setup, to control for variability in the video quality. It includes 27 meetings in total, from which 6

meetings were recorded with three participants and 21 with four participants. This subset is used in the experiments described in Chapter 5.

3.3.3 ELEA-AVS

This dataset is a subset of the ELEA corpus containing only recordings gathered with the portable video setup and synchronized with the audio stream from the microcone, with no frame dropping. The audio and visual streams were aligned, by manually localizing the synchronization point for each audio-visual sequence (i.e. using the clapping event that indicates the beginning of the group interaction). The set includes 22 meetings in total, from which 3 meetings were recorded with three participants and 19 with four participants. This subset is used in the experiments described in Chapter 6.

3.3.4 ELEA-EN

This dataset is a subset of the ELEA corpus containing only audio recordings in English. This subset is composed of 29 meetings in total, from which 9 meetings have three participants and 20 meetings have four participants. The ELEA-EN corpus is used in the experiments described in Chapter 7.

Table 3.1 summarizes the subsets of the ELEA corpus described in this section.

Corpus	Total Recordings	3-person	4-person
ELEA-A	40	12	28
ELEA-AV	27	6	21
ELEA-AVS	22	3	19
ELEA-EN	29	9	20

Table 3.1: Subsets from the ELEA corpus

3.4 Conclusion

In this Chapter we presented the design and collection of a new audio-visual group interaction corpus for the study of Emergence of Leadership. Overall, the corpus is relatively small, despite our efforts to collect data. This has to do with the requirement of having to engage only people who do not know each other, and shows the difficulty of collecting data even with portable sensors. The ELEA corpus contains approximately

10 hours of audio and video, and several annotations. The annotations of the corpus include personality, concepts related to leadership, and participants' performance in the survival task. In addition, several features have been extracted automatically from audio and video. Finally in subsequent chapters, we present the research questions that have been addressed using the ELEA corpus.

Chapter 4

Emergent Leader Inference with Nonverbal Audio Cues

In this chapter, we firstly address an analysis on how an emergent leader is perceived nonverbally in newly formed small-groups, and secondly, we infer automatically the emergent leader in the group using a variety of extracted nonverbal communicative cues. We hypothesize that the difference in individual nonverbal features between emergent leaders and non-emergent leaders is significant and measurable using speech activity. For the inference task we use rule-based, support vector machine and collective classification approaches with the combination of acoustic features extracted from the ELEA corpus described in Chapter 3. We show that adding relational information to the nonverbal acoustic cues improves the inference of each participant's leadership rankings in the group. Overall, our study shows that it is feasible to identify emergent leaders from only automatically extracted acoustic features.

This chapter is organized as follows. First, we present a correlation analysis of how the emergent leaders in a group are perceived based on their automatically extracted nonverbal behavior. Then we present two methods to infer emergent leaders using these nonverbal cues: a simple, person-wise, rule-based method; and a collective, group-wise classification approach. Furthermore, we analyze the temporal effect of the nonverbal cue extraction process on the accuracy of the emergent leader inference. Finally, we discuss the performance of the proposed methods and we draw our conclusions in Section 4.7

An initial study of the acoustic features performance on a partial version of the

ELEA corpus was published in (99). The study presented here was originally published in a longer journal version in (100).

4.1 Our approach

To analyze emergence of leadership two sets of data from the full ELEA corpus are used. The first set includes the ELEA-A subset, with 40 audio recordings of the groups performing the survival task. The second set includes the questionnaires filled by each group member.

From the questionnaires, we coded and averaged several variables, as described in Chapter 3. From the audio recordings, we automatically extract a number of nonverbal cues to characterize individual participants. We then analyze the correlation between variables coded from questionnaires and acoustic features. We also present two methods to automatically infer the emergent leader using acoustic nonverbal cues. Figure 4.1 shows our approach.

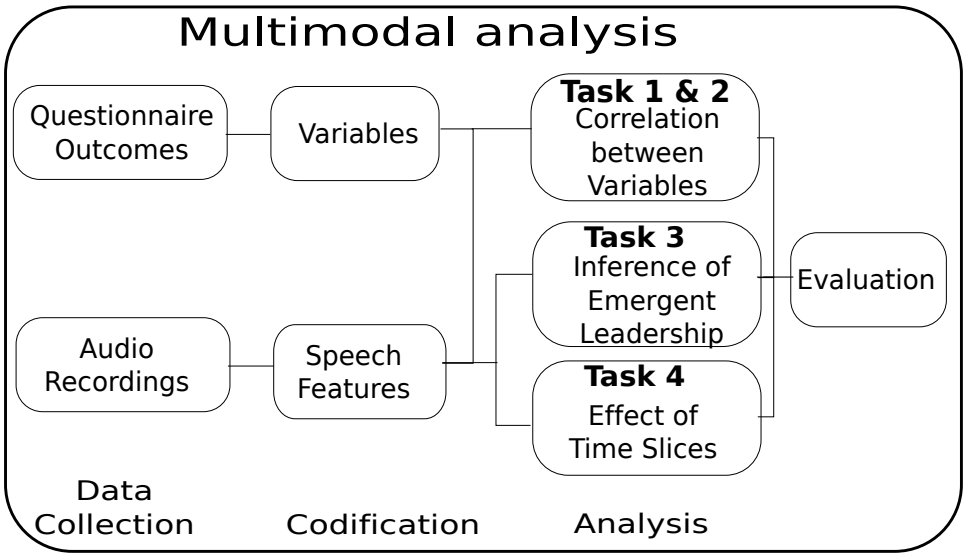


Figure 4.1: Our approach.

4.2 Data

As described in Chapter 3 this is ELEA-A corpus, consisting of 40 audio recorded meetings and the respective set of questionnaires. 12 groups were recorded with three

participants and 28 with four participants. Participants in ELEA meetings were asked to participate in the Winter Survival Task with no roles assigned. Participants were asked to answer 17 statements that capture how they perceive each participant (including themselves) after the recorded interaction. Variables included in these statements are: perceived leadership (PLead), perceived dominance (PDom), and perceived competence (PComp). One other statement asked for the ranking of group dominance (RDom) for all participants in the group.

4.3 Nonverbal Feature Extraction from Audio

In this section, we present a description of the extracted audio nonverbal features. The audio features include speaking turn and prosodic cues.

4.3.1 Speaking Turn Features

The Microcone described in Chapter CORPUS! automatically generates a speaker segmentation (2), using as a basic principle a filter-sum beamformer followed by a post-filtering stage, for each of the six spatial segments of the microphone array. The segmentation is stored in a file containing relative time in seconds (start and end), the subject label, and the Microcone sector. Similar techniques (e.g. (72)) have shown that the performance in terms of speech quality is relatively close to the performance using headset microphones, and better than lapel microphones. We did not evaluate objectively the accuracy of the speaker segmentation in a systematic manner, but had close interaction with the device's manufacturer as beta tester. Furthermore, we inspected many files and observed that the speaker turns (even if they are short) are detected correctly by the device, and that the device can recover turns' beginning and endings well. Note that as our study aims at aggregating features over longer periods of time, the features tolerate minor errors in the estimation of exact boundaries of speaker turns.

The speaker segmentation results in a binary segmentation for each participant, where status 1 represents speech and status 0 represents non-speech. From the binary segmentation, we compute the following features for each participant:

Total Speaking Length (TSL_i): The total time that participant i speaks according to the binary speaking status.

Total Speaking Turns (TST_i): The number of turns accumulated over the entire

meeting for each participant i, where each turn is a segment defined by a series of active speaking status frames. We added a variant ($TSTf_i$) which only accumulates turns longer than two seconds.

Average Speaking Turn Duration (AST_i): The average turn duration per participant i over the entire meeting.

Total Successful Interruptions (TSI_i): We use two definitions to calculate this feature:

TSI_i^1: Participant i interrupts participant j if i starts talking when j is speaking, and j finishes his/her turn before i does.

TSI_i^2: Participant i interrupts participant j if i starts talking when j is speaking; when i finishes his/her turn j is not speaking anymore.

For each of the two cases, we added a variant ($TSIf_i^1$ and $TSIf_i^2$) which only accumulates interruptions in turns longer than two seconds.

Speaking Turn Matrix (STM): The matrix which counts, as events, who speaks after whom over the entire meeting.

4.3.2 Prosodic nonverbal cues

Using the speaker segmentation, we obtain the speech signal for each participant and discard overlapped segments. We then compute two well known prosodic speech features, energy and pitch, i.e., the perceived fundamental frequency (F_0) of voice, and it is the rate of vibration of vocal cords. To extract energy, we used Wavesurf, an open source software package. For pitch extraction we used a robust method proposed in (108). The following variables were computed from energy and pitch:

Energy Spectral flatness (ESF): This is a measure often used to discriminate between voiced and unvoiced speech (37) and it is calculated as:

$$ESF = 10 * log \frac{(\prod_{i=1}^{n} a_i)^{\frac{1}{n}}}{\frac{1}{n}\sum_{i=1}^{n} a_i}, \tag{4.1}$$

where a_i denotes the magnitude of each of the bins of i (an empty bin yields a flatness of 0), and n is the number of spectral lines, in the power spectrum.

Energy variation (EVT): This feature measures the variation in energy, i.e., the loudness perceived by the ear. It is computed dividing the energy standard deviation by the mean.

Other Energy Statistics: We also estimated some statistics from the energy extracted from single speaking turns, like minimum, maximum, median and variance (denoted by EMIN, EMAX, EMED, and EVAR).

Pitch variation (PVT): This feature measures the pitch variability. It is calculated dividing the pitch standard deviation by the mean.

Other Energy Statistics: We also calculated some statistics from the F_0 from each participant's speech denoted by PMIN, PMAX, PMED, and PVAR.

4.4 Inferring the Emergent Leader

It has been shown in social psychology research that the speaking time has a strong association with individual dominance, such that people who talk more have more chances to contribute in group interaction between strangers (71). Similarly to individual dominance, emergent leaders often contribute more than nonleaders in a group discussion. If the participation in the group is quantified as single nonverbal behavior variables (like head agreement, postural shift, or rate of verbal participation) each variable alone has been shown to be a significant predictor of leadership (12, 105). Considering that there is evidence that the emergent leader can be assessed from single nonverbal features, we first present a unsupervised method that consider single nonverbal feature methods, we then present a supervised method with combination of features, and a collective classification approach.

4.4.1 Rule-Based approach

For the task of inferring the emergent leader, our hypothesis is that the emergent leader in a group is the one who has the highest value of a single nonverbal feature (i.e., the participant with the longest total speaking time). We define a rule-based inference that selects the participant with the maximum feature value in the group as the emergent leader, i.e., we infer the leader EL_m^f for group m according to feature f as

$$EL_m^f = \arg\max_p(f_p^m), p \in \{1, 2 \dots P\}, \tag{4.2}$$

where p is the participant number, f_p^m is the value of feature f for participant p in group m, and P is the number of participants (3 or 4 in our case).

4.4.2 Rank–Level Fusion approach

To investigate whether the combination of features has an advantage over using single features, we fuse rule-based estimators defined on different individual features, and used the ranked feature values of each inference as recently proposed in (5). Instead of selecting the participant with the maximum feature value, the participants are ranked and the rank information is used to fuse inferences based on different features. For group m, using feature combination $\mathcal{C}$, we sum up the ranks for each participant and select the participant with the highest total rank as the inferred leader:

$$EL_m^{\mathcal{C}} = \arg\max_p \left(\sum_{f \in \mathcal{C}} r_{f_p}^m\right), \quad \mathcal{C} \subseteq \mathcal{F}, \tag{4.3}$$

where $r_{f_p}^m$ is the rank of participant p using feature f in group m, and $\mathcal{F}$ is the set of all features. In case of ties, we select the leader based on the z-normalized scores (5).

4.4.3 Support Vector Machine

As a supervised alternative we used a support vector machine (SVM), a supervised learning method that constructs an hyperplane by mapping the nonverbal input vector in higher dimensions.

$$\sum_j \alpha_j K(x_j, x) = C \tag{4.4}$$

Where K represents the kernel function, in this case a linear kernel, α is a parameter that represents a linear combination, C is a constant value, and x_j is the input vector composed of nonverbal features. As implemented in (50), we use the SVM score to rank each participant in the group. The rankings are then used to determine which participant is assigned the Emergent-Leader person label, by considering the point which is the furthest from the class boundary. This procedure generates exactly one Emergent-Leader person in the group. For training and testing, we applied the leave-one-meeting-out approach, and the test accuracy is calculated based on the average performance.

4.4.4 Collective Classification approach

We also investigated a novel approach based on statistical relational learning. The relation among instances in network data has been exploited in several ways, ranging from classifying scientific papers with related topics to finding ways to understand centrality in online communities, and the propagation of ideas or opinions (34, 68, 77).

In a network of data, the data instances are related in some ways, and this relation can be learned to infer several instances simultaneously. This is the aim of collective classification (53, 74). The label inference of a data point can be influenced by inferences of its neighboring labels.

Taking into account that our data is not independent and possibly not identically distributed, we propose to investigate collective classification in our problem. A collective approach improves probabilistic inference when the data is relational and correlated. In the context of web data analysis, it has been proved that adding relational information when instances are not independent improves inference (53). As we mentioned in Chapter 2, there are nonverbal speaking features highly correlated with dominance, and dominance is also correlated with emergent leadership, as described in Section 2.2. Our hypothesis is that by considering the relational information and given that the data is correlated, collective inference might improve the leader estimations.

The data is modeled as follows: We have a graph $G = (V, E, X, Y, C)$ where V is the set of participants $v_i \in V$, E is a set of directed edges, coded from the speaking turn matrix (STM), each $x_i \in X$ is an attribute vector composed of nonverbal features for participant v_i, each $y_i \in Y$ is a label variable for v_i, and C is the set of possible labels (i.e., 1 for Emergent Leader or 0 for Non-Emergent Leader). Figure 4.3 shows the model.

To perform collective classification, the Iterative Classification Algorithm (ICA) has been defined in (74), the execution of the algorithm is summarized in Figure 4.2. The algorithm makes an initial label inference y_i for each v_i, using only the individual nonverbal features in x_i. In the second step, the algorithm computes the relational information f_i considering the labels from the previous inferences. The iteration step (third step), the labels y_i are re-estimated with a local classifier using the features x_i and the relational information f_i computed in the previous step, in addition the confidence

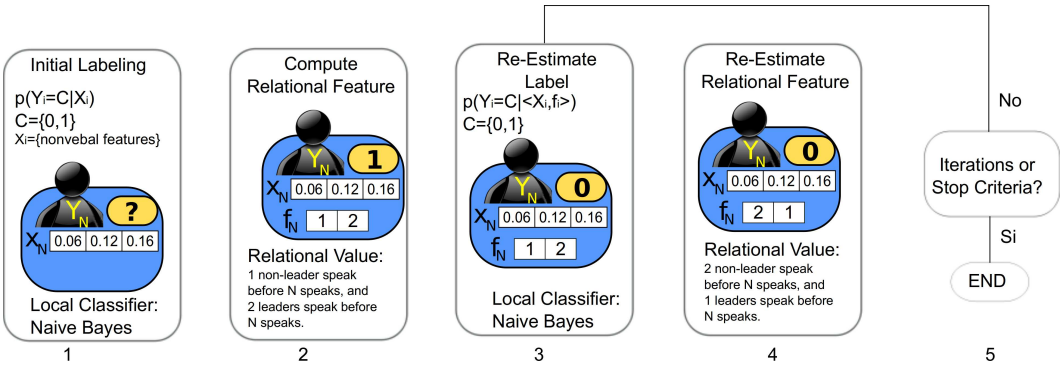


Figure 4.2: Iterative Collective Algorithm (ICA). The algorithm makes an initial label inference on y_N using a Naive Bayes classification using x_N. In the second step, the algorithm computes the relational information f_N, in this example we use *count* as the relational information, with two values, the first value counts the number of participants labeled NonEmergentLeader that speaks after participant N does, and the second value counts the number of participants labeled EmergentLeader that speaks after participant N. In the third step, the label y_N is re-estimated with a Naive Bayes classifier using the features x_N and the estimated f_N. Then, the relational information f_N is re-estimated using the new label y_N. Finally, the algorithm verify the stop criteria to iterate (back to step 3 and so on) or end.

of the inferred labels is stored. In the next step, the relational information f_i is re-estimated using the new labels. Then, the algorithm verifies for the stop criterion and iteratively re-estimates labels (i.e., third step to end) or ends the execution.

There are two tasks that can be performed using the ICA algorithm, named out-of-sample and in-sample (74). For the in-sample task, we are given a set of known labels Y^K for a subset of participants $V^K \subset V$, so that $Y^K = \{y_i | v_i \in V^K\}$. Then, the task is to infer Y^U, the values of y_i for the remaining participants with unknown labels ($V^U = V - V^K$), or a probability distribution over those values. We implemented the three variants for the ICA algorithm described in (74), denoted by ICA, ICA$_{kn}$ and ICA$_c$. All three algorithms are based on iterations, in out case up to 5, ICA considers all the estimations from the previous iteration, ICA$_{kn}$ uses only known labels V^K in the first iteration, and from the second to the last iteration it works like ICA. Finally, ICA$_c$ uses the known and the most confident estimated labels, and increases gradually the number of estimated labels in each iteration.

For the out-of-sample task, no labels are known, thus V^K is empty, and there are only two variants to the algorithm, namely ICA and ICA$_c$. For both tasks we

follow a similar procedure as proposed in (50), for dominance in small groups, i.e., the algorithm infers exactly one Emergent-Leader in the last iteration, which corresponds to the participant with the highest posterior probability for the Emergent-Leader class.

Several relational features can be used in our problem. The simplest one is coded as a *count*, which represents the number of participants who take turns after participant i and that belong to a particular class. For instance, $f_i(0) = 2$ indicates that two participants labeled as non-emergent leaders take turns after participant i. A second relational feature, called *proportion*, is coded as the proportion of participants taking turns after participant i and that have a particular label. For instance $f_i(0) = 2/3$ indicates that three participants take turns after participant i, from which two are labeled as non-emergent leaders. Finally, the relational feature *multiset* produces a single numerical value for each possible label for the participants who take turns after participant i. This value can be compared against the mean value from the training set (missing labels are not used). For instance, $f_i = \{1, 1, 1\}$ means that for participant i there is one participant labeled as non-leader that takes a turn after him, there is one participant labeled as leader that takes a turn after, and one more participant with an unknown label takes a turn after him.

To our knowledge, weighted links have not been explored as a potential relational feature. Given that we have the weights that represent the amount of turns that participants take after each other during the 15-minute interaction, we defined a new relational feature named *weighted proportion*. This relational feature considers weights, direction, and number of participants taking turns after participant i does. For instance, from Figure 4.3 $f_K(0).IN = (0.2759)/3$, where $f_K(0).IN$ represents the fact that participant K takes turns 27.59% of the time after participant N who is in turn (labeled as class 0) does, and the value is then divided by the number of neighbors, i.e., the number of participants who take turns before K takes a turn.

The ICA algorithm requires a local classifier for training and for the initial labeling. The variant ICA_c needs as well the confidence values for the labels. For confidence estimation, we use the posterior probability for the most likely label for participant v_i, calculated with a naive Bayes classifier. The local classification is performed as well using a naive Bayes classifier. For training and testing, we applied the leave-one-meeting-out approach, and the test accuracy is calculated based on the average performance.

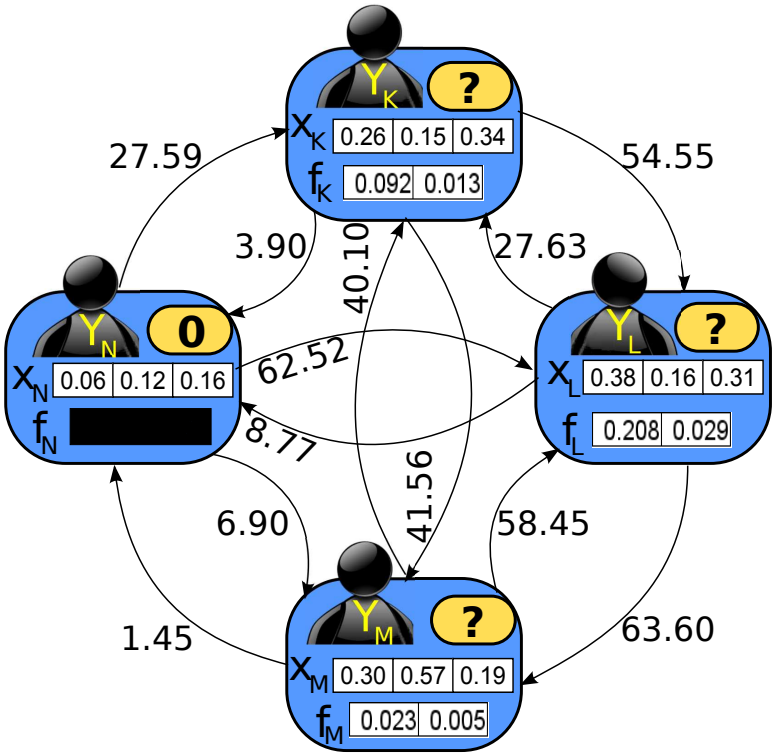


Figure 4.3: Data modeled for collective classification algorithm. The weighted links between participants represents percentage of turns taken after one another (the direction indicates who takes the turn). x_i shows values from three audio features: TSL_i, AST_i and TSI_i^1. In this case we have a known label, participant N is non-emergent leader ($y_N = 0$). In this example, the y_i labels for the other 3 participants is unknown. The relational feature called weighted proportion f_i is estimated considering the known label y_N, and the number of participants that have a turn before participant i does.

4.5 Experiments and Results

In this section, we first present a correlation analysis between self-reported questionnaires and nonverbal features, and then present results on leadership estimation.

4.5.1 Correlation Analysis

We estimate correlations between perceived variables and audio features, calculating the Pearson correlation per group, then applying a Fisher transformation. Finally we test if the correlations are statistically significant with a T-test, at 5% significance level (i.e., $p < 0.05$).

Table 4.1: Pearson correlation values between variables from questionnaires outcomes ($* : p << 0.005$, $\dagger : p < 0.05$).

	PLead	**PDom**	**PCom**	**PLike**	**RDom**
PLead		0.77*	0.30*	-0.30†	0.79*
PDom			0.25†	-0.33*	0.69*
PCom				0.26	0.31*
PLike					-0.34*
RDom					

4.5.1.1 Questionnaire output analysis.

First, we analyze the correlation of the questionnaire outputs filled by participants after the interaction. Each perceived variable is averaged over all participants per group, and the group ranking is normalized according to the number of participants per group. Table 4.1 shows the Pearson correlation values. PLead shows significant correlation with PDom and RDom (0.77 and 0.79, respectively). These results suggest that the emergent leader is perceived as a dominant person by the other participants. Interestingly, the correlation between perceived leadership and competence (PCom) is significant but less strong, and lower between perceived or ranked dominance and competence, which suggests that participants might not have used often the latter construct as part of their judgments.

4.5.1.2 Survival task top ranking analysis.

Given that the task in the groups is to come up with a group rank list (composed of twelve items), we compute the correlations with the aim of discovering the influence of individuals in the group final decision. We analyze the correlation between the number of individual items in the top group list against the perceived variables from questionnaires. We use two approaches: In the first one, we count the number of individual items in the top group rank (see Table 4.2); In the second approach, we consider the absolute difference of each individual item rank with respect to the top group rank, and normalize this value with respect to the number of items in the top group rank (Table 4.3). If one item is not in the top rank, it is assigned the maximum distance (+ 1). From Table 4.2 we can see that the emergent leader (PLead) did not necessarily convince the group to select his/her top 1-2 individual items in the group rank, in contrast with the participants that were ranked as the most dominant (RDom).

Table 4.2: Correlation values between questionnaires variables and the number of individual items in the top rank list from the winter survival task ($* : p << 0.005$, $\dagger : p < 0.05$).

	PLead	PDom	PCom	PLike	RDom
TOP1	0.17	0.16	0.03	-0.03	$0.24^{\dagger}$
TOP2	0.16	0.17	0.03	-0.02	0.17
TOP3	$0.29^{\dagger}$	0.39^{*}	0.14	-0.01	$0.29^{\dagger}$
TOP4	$0.29^{\dagger}$	0.37^{*}	0.15	-0.04	$0.30^{\dagger}$
TOP5	$0.20^{\dagger}$	0.17	0.15	0.06	$0.15^{\dagger}$
TOP6	$0.24^{\dagger}$	$0.20^{\dagger}$	0.19	-0.05	$0.24^{\dagger}$
TOP7	0.26^{*}	$0.25^{\dagger}$	$0.18^{\dagger}$	-0.09	$0.22^{\dagger}$
TOP8	$0.25^{\dagger}$	0.19	0.39^{*}	0.16	0.17
TOP9	0.13	0.15	0.20	0.20	0.19
TOP10	-0.003	0.01	-0.05	0.08	0.18

On the other hand, stronger effects are observed both for leadership and dominance when more items are allowed in the top group rank (TOP 3 - TOP 8).

From Table 4.3 we can see another facet of the influence that the emergent leader has with respect to the final group ranking. In particular, the most dominant people (PDom and RDom) might try to make the final group rank as similar as possible to their individual ranking list (TOP 1). In this case, negative correlations are due to the absolute distance: the closest the individual list with respect to the group list, the smallest the difference. As shown in (60), dominant people tend to get their way in small group tasks related to ranking preferences.

Finally, we explored as well the individual performance in the survival task (AIS). The main effects are r=−0.22 (p=0.04) between AIS and PCom, and r=−0.23 (p=0.009) between AIS and PDom. The negative correlations a due to the AIS measure, such that the smaller the score, the better the answer, and according to the experts the more chances to survive. This might suggest that the actual individual performance in the ranking task has a relation with the perception of competence and dominance from the group.

4.5.1.3 Nonverbal speaking behavior and perception from participants

Table 4.4 shows Pearson correlation values between questionnaire outputs and individual audio nonverbal features. As we can see, there is a correlation between several features and PLead, suggesting that emergent leadership perception has a connection to the person who talks the most, has more turns, and interrupts the most. Furthermore,

Table 4.3: Correlation values between questionnaires variables and absolute distance in top rank items from the winter survival task ($* : p << 0.005$, $\dagger : p < 0.05$).

	PLead	**PDom**	**PCom**	**PLike**	**RDom**
TOP1	-0.17	-0.18†	-0.04	0.02	-0.23†
TOP2	-0.15	-0.16	-0.01	-0.003	-0.15
TOP3	-0.25†	-0.34*	-0.09	0.01	-0.28†
TOP4	-0.33†	-0.37*	-0.13	0.02	-0.33*
TOP5	-0.29†	-0.30†	-0.14	-0.06	-0.23*
TOP6	-0.34*	-0.33*	-0.20	-0.01	-0.28*
TOP7	-0.29*	-0.31*	-0.18	0.02	-0.24†
TOP8	-0.29*	-0.27*	-0.38*	-0.16	-0.24*
TOP9	-0.23*	-0.25†	-0.24†	-0.18	-0.28†
TOP10	-0.23†	-0.27*	-0.07	-0.04	-0.32*

Table 4.4: Correlation values between questionnaires variables from and nonverbal acoustic features on the ELEA-A corpus ($* : p << 0.005$, $\dagger : p < 0.05$). For Energy and Pitch features, only those that have significant correlations with at least one of the questionnaire variables are shown.

	PLead	**PDom**	**PCom**	**PLike**	**RDom**
TSL	0.52*	0.40*	0.17	-0.32*	0.51*
TST	0.32†	0.31†	0.19	0.00	0.26*
TSTf	0.50*	0.47*	0.14	-0.28*	0.44*
AST	0.48*	0.36*	0.17	-0.29†	0.46*
TSI^{1}	0.51*	0.41*	0.16	-0.21†	0.47*
TSIf^{1}	0.49*	0.38*	0.21†	-0.24	0.44*
TSI^{2}	0.33†	0.35*	0.14	-0.14	0.35*
TSIf^{2}	0.53*	0.48*	0.25†	-0.23†	0.52*
EMIN	-0.33†	-0.23†	-0.22†	0.14	-0.28†
EMED	0.23†	0.14	0.18	-0.10	0.20
PVAR	-0.14	-0.21†	-0.13	0.05	-0.27†
PVT	-0.14	-0.19†	-0.01	0.04	-0.22†

several nonverbal cues have also correlation (although with lower values) with perceived or ranked dominance. This confirms previous work showing that these features are reasonably correlated with dominance in groups (33, 71). Finally, the interruptions (TSIf^{2}) have correlation with judgment of competence (4). As shown in (112), emergent leaders do not necessarily have to be the most active participants when they are perceived as competent in a task, which could be interpreted from the absence of significant correlations between PCom and the speaking time and turn features.

4.5.2 Leadership Inference using Audio Nonverbal Cues

In this section, we present the results for each of the inference methods and the audio nonverbal cues. For the evaluation of our approach, we use the variables from the questionnaires as ground truth. Random performance in this case is 27.5% given that the ELEA-A corpus has 40 meetings, from which 28 meetings have four participants, and 12 meetings have three participants.

4.5.2.1 Rule–based approach

We calculate the accuracy of the rule-based inference by comparing the ground truth emergent leader with the participant who has the highest value for each of the nonverbal cues (Equation 4.2). Figure 4.4 shows the accuracy using single speaking turn features, where the best accuracy for variable PLead is achieved using $TSIf^2$ (63.5%), followed by TSL (60%). The best accuracy for PDom is achieved using $TSIf^2$ (55%) followed by TSTf (50%). For the case of RDom, similarly to Plead, the best performance is using $TSIf^2$ (62.5%) followed by TSL (57.5%). As we can observe, PCom and PLike are more difficult to infer, for PCom the highest inference performance is reached using $TSIf^2$ (45%), and for PLike only 20% using TST and TSI^2. It could be interpreted as follows, for PLike, having information derived only from the speaking turns, does not provide sufficient information in order to have accurate inferences.

We also explored the performance of the prosodic features using the rule-based estimator. For the PLead we obtain the highest inference using PSF (47%), i.e., the energy spectral flatness. For PDom and RDom, also the highest accuracy is reached by using PSF (42%). For PCom, the best accuracy is achieved using EMED (32.5%); and for PLike the highest performance is achieved using EMIN (40%). Figure 4.5 shows accuracy for energy and pitch, from which we can observe that all prosodic features performed better than speaking turn features for the variable PLike (e.g. EMIN, with 40.0%) and PCom (EMED, with 32.5%). Although the accuracy does not improve the performance obtained with the top speaking turn features for variables PLead, PDom, PCom, and RDom, they do provide some discriminant information.

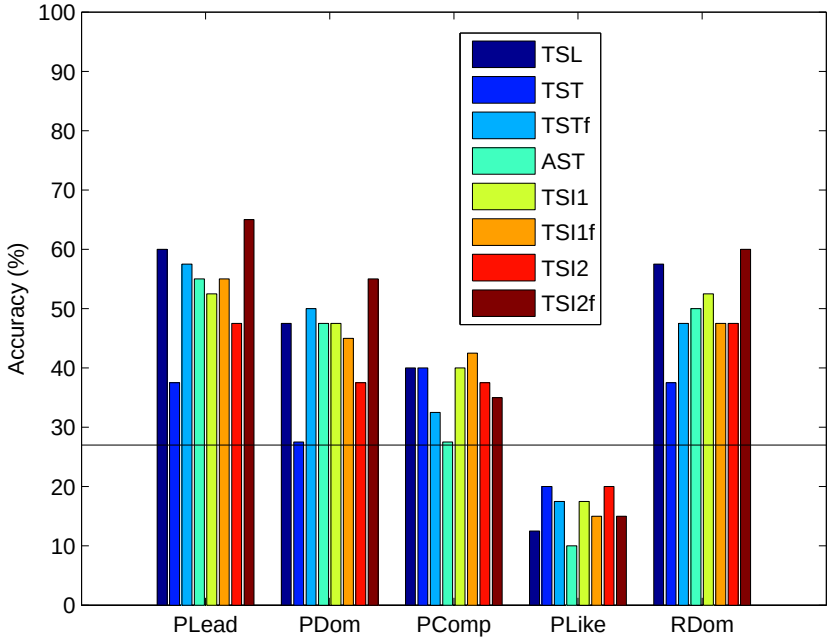


Figure 4.4: The accuracy of speaking turn features on the ELEA full corpus and rule-based estimation. The black horizontal line shows the random baseline.

4.5.2.2 Rank–level fusion approach

Table 4.5 shows the combinations of relevant nonverbal audio features to estimate emergent leadership and other related concepts. For the rank–level fusion method, the highest accuracy for PLead is 72.5% combining AST, TSI[1], TSIf[2], EMED, and EVAR, i.e., a combination of speaking activity and energy. As we can observe in the Table 4.5, for all the other variables, the combination of features consistently results in higher performance.

4.5.2.3 Support vector machine

Table 4.6 shows accuracy results on SVM. The highest accuracy for PLead is reached using AST, TSI2f, EVAR and EVT (67.9%), and for PCom the highest accuracy is 48.8%; although both results are higher than random performance, they are lower compared with rank-level fusion inferences. As we can observe, the use of SVM only improves the accuracy obtained for PLike (55.4%) as compared with the inference using

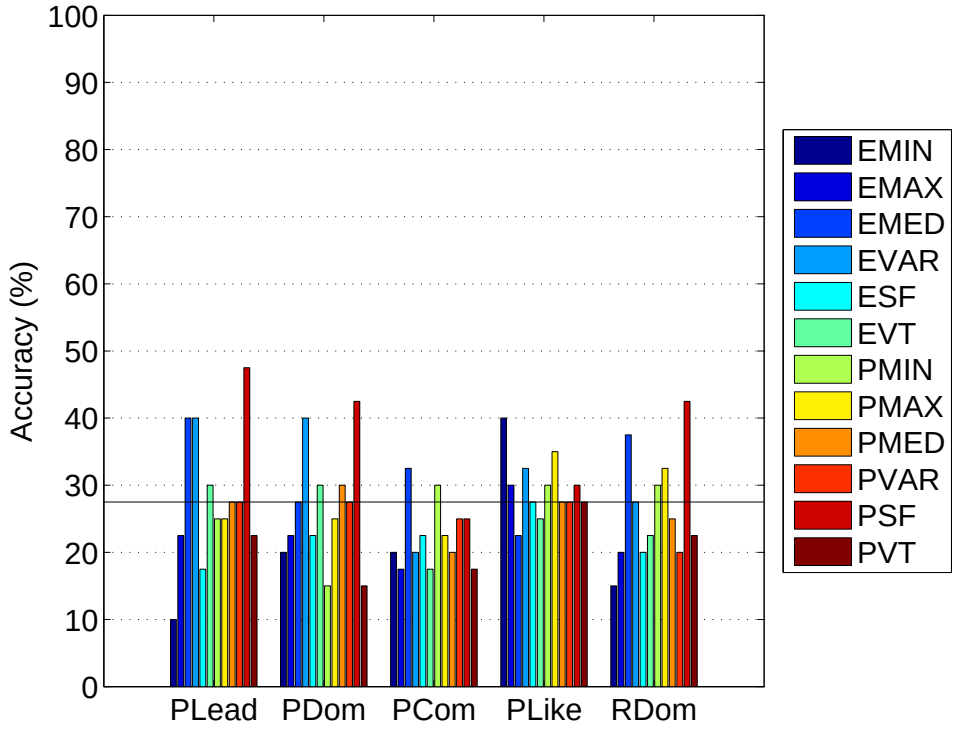


Figure 4.5: The accuracy of energy and pitch on the ELEA full corpus and rule-based estimation. The black horizontal line shows the random baseline.

rank-level fusion (40%). Although our SVM-results for PDom and RDom (64.3% and 66.7% respectively) are lower than the ones presented in (91) with up to 75% accuracy, and in (50) with up to 91.2% when there is Full agreement for the most dominant person from annotators, and up to 75.4% when there is majority agreement, is worth to mention that the scenarios differ, such that in our scenario no roles are assigned in the recordings.

4.5.2.4 Collective classification approach

Collective classification, that uses relational information improved the accuracy to infer the emergent leader and related concepts with respect to single features. The nonverbal features are selected based on the highest correlation values mentioned in section 4.5.1. We applied both the out-of-sample (two variants) and in-sample (the three variants) approaches described in section 4.4.4.

For the out-of-sample task, the accuracy is calculated on the label assigned to the

Table 4.5: Results of rank-level fusion on the ELEA-A corpus. The features combined are listed in the last column.

	Acc(%)	Fused variables
PLead	72.5	AST, TSI^2, $TSIf^2$, EMED, EVAR
PDom	65	AST, TSI^1, $TSIf^2$, EVAR, PMED
PCom	55	TST, TSI^1, $TSIf^1$, $TSIf^2$, EMIN, EVAR, PMIN, PMED
PLike	40	EMIN
RDom	72.5	TSL, AST, $TSIf^2$, EMED

Table 4.6: Best results of SVM on the ELEA-A corpus.

	Acc(%)	features
PLead	67.9	AST, $TSIf^2$, EVAR, EVT
PDom	64.3	AST, TSI^1, $TSIf^2$, EMIN, EMED, EVAR, EVT
PCom	48.8	AST, $TSIf^2$, EVAR, EVT
PLike	55.4	TSI^1, $TSIf^2$
RDom	66.7	AST, TSL, TSTf, $TSIf^2$

emergent leader or the related concepts compared with the ground truth. Table 4.7 shows the performance in terms of accuracy.

For the in-sample variant, we provide a known label per group. Since we notice from the rule based-estimator and the rank–level fusion method that participants with the lowest feature values are often perceived neither as leaders nor as most dominant, we labeled these participants as Non-Emergent Leader/Non-Most Dominant (same for competence and liking). The test is then performed by using this known label and inferring the leader or the related concepts on the remaining two or three participants per group, respectively. For this task, the baseline accuracy is 38.3%. Table 4.8 shows the accuracy results using the in-sample variant.

We can observe that the variant ICA_c (which uses the known and most confident estimated labels in each iteration) has the best performance for most of the cases. The best accuracy for emergent leadership inference is 70.2% when we provide a known

Table 4.7: Best results of collective classification on ELEA-A corpus. Out-of-sample task.

	Acc(%)	features	ICA variant
PLead	72.0	AST, $TSIf^2$, EVAR, EVT	ICA_c
PDom	60.1	AST, TSI, $TSIf^2$, EMIN, EMED, EVAR, EVT	ICA_c
PCom	46.4	TSL, TSTf, AST, $TSIf^2$, EMIN, EMED, EVAR	ICA_c
PLike	55.4	All Speaking Turn Features	ICA_c
RDom	61.9	PVAR, PSF, PVT	ICA_c

Table 4.8: Best results of collective classification on the ELEA-A corpus. In-sample task.

	Acc(%)	features	ICA variant
PLead	70.2	TSL, TSTf, TSIf^2, EMED	ICA$_c$
PDom	58.3	AST, TSIf^2, PMIN	ICA
PCom	57.7	AST, TSIf^2, EVAR, EVT	ICA$_c$
PLike	53.6	PVAR, PSF, PVT	ICA$_c$
RDom	76.2	AST, TSIf^2	ICA$_c$

label (i.e., when we assume we know a participant that is non-leader in this task, see Table 4.8), and 72% when the group does not have any known label (Table 4.7). For PDom, the highest accuracy is reached with the out-of-sample ICA$_c$ variant (60.1%). The best accuracy for RDom is obtained when we assume that we know a non-dominant participant in the interaction, i.e., using the in-sample variant (76.2%). For Pcom, the highest performance is reached using the in-sample variant (57.7%). Finally, for the case of PLike, the best performance is provided by the out-of-sample variant (55.4%).

4.5.2.5 Observation Window Analysis

As a final point, we performed an analysis to explore the temporal support needed by our approach. We computed the same audio features described in section 4.3.1, originally computed for the whole interaction, but for smaller observation windows (or slices), and then inferred the emergent leader and related concepts with the rule-based estimator per slice.

We explored three type of slices:

- Accumulated Slices: The duration of the slices is defined as multiples of 1/8 of the original duration, where each slice starts from the beginning of the interaction.
- Non-Accumulated Slices: Each slice is exactly 1/8 of the total duration, with no overlaps.
- Non-Accumulated Slices with Overlaps: The slice size is 5 minutes with two-minute overlaps, the first slice starts from the beginning of the interaction.

Figure 4.6 shows the accuracy obtained for the three types of slices with respect to the variables PLead, PDom, and RDom. We can observe that for accumulated slices (Figures 4.6 (a), (d), (g)), after the first half of the recording (7.3 minutes on average), the inferences tend to follow a trend and change slightly.

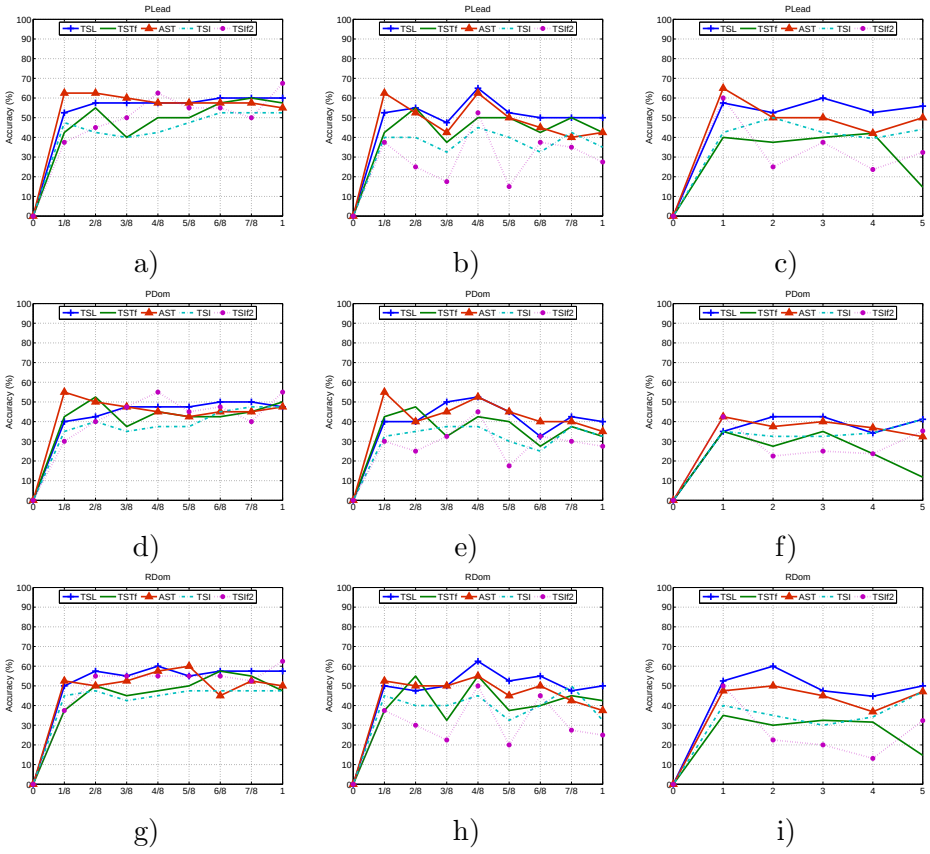
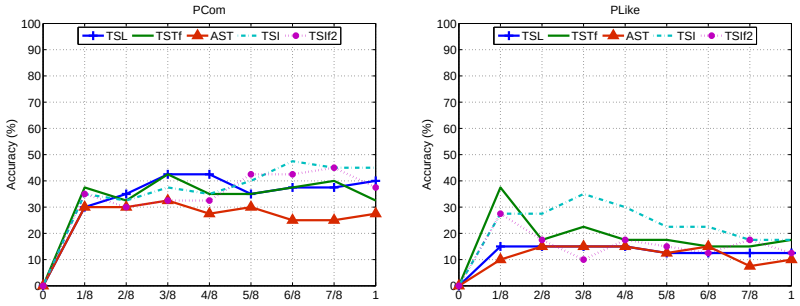


Figure 4.6: Observation window analysis for speaking turn features on the ELEA-A. The first column shows accumulated slices (a, d, g); the second column shows non-accumulated slices (b, e, h); and the last column shows non-accumulated slices with overlaps (c, f, i). Results shown per each row are: PLead, PDom and RDom.

Figures 4.6 (b), (e), and (h) show the accuracy for the non-accumulated slice with rule-based estimation with respect to PLead, PDom and RDom. It is interesting to observe that the method can often produce the best performance by just looking at the slices in the middle (slice from 3/8 to 4/8), in which the person that speaks the most, takes more turns, and interrupts more is perceived as the emergent leader and as well the most dominant. This could be due to the making-decision task, i.e., the emergent or the perceived dominant, are conscious of the time elapsed, and in order to finish on time, they take the floor to propose their solutions and make the group be aware of the time.

From Figure 4.6 (c), we also infer that on average the emergent leader talks more



a) b)

Figure 4.7: Observation window analysis for speaking turn features on the ELEA-A corpus. Results for accumulated slices for a) PCom and b) PLike.

during the first five minutes, then the other participants take turns, and during the middle of the meeting the leader again has the highest speaking time and turns. Based on the specific task performed, we might interpret these results as follows: the leader organizes the group (first five minutes), listens to opinions from the group (minutes 3-8), and then leads the discussion. Finally, considering longer slices with overlaps, in this case five minute slices, we can observe from Figure 4.6 (f) that dominant behavior is more likely to be observed approaching the middle of the meeting (minutes 3-8), and the emergent leadership is more noticeable in the middle (Figure 4.6 (c) window 3: minutes 6-11).

Figure 4.7 shows the accuracy obtained for the accumulated slices with respect to the variables PCom and PLike. We can observe that for PCom, accumulated slices (Figure 4.7 (a)), after the first half of the recording (7.3 minutes on average), the inferences follow a trend and change only slightly. For the case of PLike (Figure 4.7 (d)), after time slice 2/8 (3.75 minutes on average) the inferences does not show improvement.

4.6 Discussion

In this Section, we discuss the accuracy of all the methods including single and combined acoustic features from Section 4.5.2. Furthermore, to validate the performance of one method over another, we perform a two-tailed standard binomial significance test with z=N(0,1), i.e., mean=0 and standard deviation=1 (67). In the comparison among methods we report the significance p-value, which captures the confidence level (1-p).

As we can observe in Table 4.9, the highest accuracy for PLead is 72.5% and for PDom is 65% using Rank-level Fusion. The significance test at 95% confidence revealed that the highest accuracy for PLead is significantly better only compared with random baseline performance ($p < 1.0e-6$). The 8.8% performance improvement of Rank-level fusion over the Rule-based estimator was not found significant ($p = 0.18$). For the methods we used, the emergent leader in a group can be inferred between 63 and 72% accuracy, using unsupervised or supervised methods with only acoustic features.

For PDom, the highest performance is 65% achieved with Rank-level Fusion. However, the improvement of Rank-level fusion over Rule-based estimator was not found significant ($p = 0.13$). There is no evidence that supports the significance of performance difference of Rank-level fusion over SVM, CC-Out-of-Sample and CC-In-Sample. All methods (i.e., Rule-based Estimator, Rank-level fusion, SVM, and CC Out-of-Sample and In-Sample), were found to be statistically better than random performance ($p < 0.005$).

Similarly for RDom, based on the statistical test, there is no evidence to validate that one of the methods is significantly better than the rest. Nevertheless, the performance of all methods is significantly higher compared with random performance ($p < 0.0003$).

For PCom, the highest reached accuracy is 57.7% using CC-In-Sample, which is marginally significant compared with Rule-based estimator ($p = 0.09$). However, there is no evidence of improvement of CC-In-Sample with respect to the other three methods. Overall, Rank-level fusion, SVM and CC were found to be statistically better than random performance ($p < 0.03$).

Finally, for PLike the highest accuracy (55.4%) was obtained with SVM and CC-Out-of-Sample. After applying the significance test, there was marginal evidence of performance improvement of this method, over Rank-level fusion and Rule-based Estimator ($p = 0.09$). However, no evidence of improvement was found between CC-Out-of-Sample and CC-In-Sample. Three of the methods (SVM and CC In-Sample and Out-of-Sample) are statistically better compared with random performance ($p < 0.003$)

Table 4.9: Best accuracy (%) of all methods on the ELEA-A corpus with only audio features

	PLead	**PDom**	**PCom**	**PLike**	**RDom**
Baseline	27.5	27.5	27.5	27.5	27.5
Rule-based Estimator	63.7	53.7	42.5	40	61.3
Rank-level Fusion	**72.5**	**65**	55	40	72.5
SVM	67.9	64.3	48.8	**55.4**	66.7
CC-Out-of-Sample	72	60.1	46.4	**55.4**	61.9
CC-In-Sample*	70.2	58.3	**57.7**	53.6	**76.2**

4.7 Conclusions

In this Chapter, we proposed a computational framework to infer emergent leadership in newly formed groups, by combining speaking turns and prosodic features. We evaluated the effectiveness of individual and combined audio features in identifying the emergent leader and related constructs using various approaches. Based on the results of a correlation analysis, it was found that the emergent leader was perceived by his/her peers as an active and dominant person, who talks the most, has more turns and interruptions, and has a longer variation in the tone and energy of voice and energy. On the other hand, the individual performance in the survival ranking task has a slight effect on the perception of competence and dominance from the group.

For the emergent leader, the perceived most dominant and the ranked most dominant person in the group, we did not find out statistical evidence of better performance among Rule-based Estimator, Rank-level Fusion, SVM and CC. The emergent leader in the group can be inferred within 63.7 and 72.5% accuracy. The perceived most dominant person can be inferred within 53.7 and 65% accuracy. Similarly, the ranked most dominant person can be inferred within 61.3 and 76.2% accuracy, with either speaking turn cues, or combinations of speaking and prosodic cues (using unsupervised and supervised methods). All methods performed significantly better than random performance.

The perceived competent person can be inferred within 48.8 and 57% accuracy using Rank-level Fusion, SVM or CC. We found evidence of better performance of these methods over Rule-based Estimator and random performance. In contrast, the most agreeable (or likable) person can be inferred with supervised methods (SVM and CC) within 53.6 and 55.4% accuracy. We found evidence that SVM and CC perform

significantly better than Rule-based Estimator and Rank-level Fusion using acoustic cues.

We observed that the results for perceived leadership and perceived dominance (and rank dominance) were sometimes similar and sometimes different for the same extracted features and inference method. Given that the variables were highly correlated, it is not surprising that the results are similar. However, we opted not to combine the three questionnaire measures because they capture somewhat different aspects of verticality. Finally, through an analysis of observation windows, we found that although the entire interaction is needed to perform the task, to computationally estimate the emergent leader with the highest accuracy only the first half (approximately seven minutes) or a slice of the interaction around the middle was required. This finding could be explored in more detail in the future, given the potential value for applications that could provide reasonably accurate estimations with less data.

Some limitations of the work include the number of samples in our data, and generalizations to other populations: although the data was collected with participants from different cultural backgrounds, a larger sample would be necessary in future work to further validate our findings. Additionally, it is important to investigate whether using information beyond the acoustic channel can predict the emergent leader in the group. This is the goal of the work presented in the next chapter.

Chapter 5

Inferring Emergent Ledership from Audio-Visual Nonverbal Activity Cues

In this Chapter we address the problem of automatically inferring emergent leadership from audio-visual nonverbal cues. The nonverbal cues are automatically extracted from a the ELEA-AV corpus (described in Chapter 3) using portable audio and video sensors. The data consist of approximately 7 hours of audio/video recordings, as well as variables extracted from questionnaires filled by each group member immediately after the recordings.

We assume that features derived from the visual activity could provide more information in order to discriminate the emergent leader (and related concepts) in a group. The visual cues include head and body activity features, as well as motion-based features extracted at the individual level. We explore the performance of the combination of audio and visual features using unsupervised and supervised methods. We found that the aggregation of visual activity and acoustic information improves the inference of emergent leaders. Although, for the most agreeable person, using a supervised method and audio-only features provides better performance than the combination of audio-visual features.

In this chapter we present an analysis of nonverbal cues derived from audio and visual portable sensors. The Chapter is organized as follows. Section 5.2 describes the ELEA-AV corpus. Section 5.1 summarizes our approach. Section 4.3 introduces the

nonverbal visual cues used in the experiments. Section 5.5 presents first a correlation analysis of how the emergent leaders in a group are perceived based on their visual nonverbal behavior, followed by our experimental results using automatically extracted audio and visual nonverbal cues. Finally, we present discussion and conclusions in Sections 5.6 and 5.7 respectively.

The work reported here have been originally published in journal form in (100).

5.1 Our approach

To analyze the emergence of leadership in small groups, we used two sets of data from the ELEA-AV corpus. The first set includes audio-visual recordings from the survival task. The second set includes questionnaires filled by each group member, to capture how other participants are perceived by each other. From the recordings, we automatically extracted a number of nonverbal cues to characterize individual participants. We then analyze the correlation between variables derived from questionnaires and audio and visual features. After this, we develop methods to automatically infer the emergent leader using acoustic and visual nonverbal cues.

5.2 Data

In this chapter, we use the portable video corpus further referred as ELEA-AV and described on Chapter 3. The ELEA-AV corpus contains 27 meetings from which six meetings are three participants and 21 meetings are composed of four participants. We chose to only use the portable video corpus to control for variability in the video quality.

5.3 Visual Nonverbal Features

In this section, we present a description of the extracted visual nonverbal features. The visual features include tracking-based features and motion template-based features. These features where extracted in a collaboration with Dr. Oya Aran (Idiap Research Institute). The audio features (speaking turn and prosodic cues) can be consulted in Chapter 4.

We defined the following feature groups, based on the respective type of features.

- HA: Head activity features (Section 5.3.1).
- BA: Body activity features (Section 5.3.1).
- MT: wMEI based features (Section 5.3.2).
- ST: Speaking turn features (Section 4.3.1).
- EN: Energy features (Section 4.3.2).
- PI: Pitch features (Section 4.3.2).

5.3.1 Tracking-based features

Head activity (HA). Figure 5.1 summarizes the feature extraction process for the head activity. To measure the head activity of each participant, we first tracked the face with a Particle Filter (PF), using an ellipse face model (45). The dynamic model of the PF uses a damped velocity model for the position and velocity, and a random walk model for the shape parameters (i.e., the size of the ellipse). As observations, we use a skin color probability image, which has a positive probability for skin color pixels and zero probability for other colors. Skin color models are learned on additional data to calculate the likelihood. We make two measurements based on the ellipse that is defined by the state vector of the particle: The ratio of the skin colored pixels to the total number of pixels (i) inside the ellipse, and (ii) at the boundary of the ellipse. High likelihood is assigned to cases where the first measurement is high and the latter is low. We additionally apply the mean shift algorithm to move the particle centers to the areas with high skin color probability. This allows to use particles more effectively, and requires fewer particles than a standard PF. More details can be found in (6).

Once the face area is estimated by the PF, the optical flow vectors within the face area of two successive frames are calculated to have a fine-grained analysis of head movements. We use the hierarchical Lucas-Kanade optical flow algorithm, using points selected from the face area that indicate strong corners. The OpenCV library is used for the implementation of the optical flow algorithm (18).

Using the optical flow vectors, we calculate the average motion vector to get the average head motion on the x and y dimensions. For each participant, we obtain two

real-valued vectors, hR_x and hR_y with elements $hR_{x,t}$, $hR_{y,t}$, one for each dimension, describing the head activity of that participant during the whole meeting.

Furthermore, to identify significant head activity, we first binarized these vectors via automatic thresholding, obtaining the binary vectors hB_x, hB_y with elements $hB_{x,t}$, $hB_{y,t}$. The automatic threshold for the x dimension eliminates small movements, and it is calculated as $\mu_x + \sigma_x$, where μ_x and σ_x are the mean and standard deviation of hR_x respectively. Computed for each participant in each meeting, the values above the threshold are set to 1, indicating a significant head activity, and rest to 0. This calculation is repeated for the y dimension as well. The final binary head activity vector, hB, is then calculated by an OR operation:

$$hB = hB_x \vee hB_y. \tag{5.1}$$

For each participant, the following features are calculated using hR_x, hR_y, and hB, which represent the participant's head activity during the group interaction.

Head activity length (THL_i): The total time that participant i moves his/her head, calculated from hB.

Head activity turns (THT_i): Number of turns for each participant i, where each turn is considered as a continuous head activity, calculated from hB.

Head activity average turn duration (AHT_i): The average turn duration for participant i, calculated from hB.

Standard deviation of head activity ($stdHx_i$, $stdHy_i$): Standard deviation of head activity in x and y dimensions, calculated from hR_x and hR_y.

Body activity (BA). Figure 5.2 summarizes the process for body activity feature extraction. It is measured by simple motion differencing as the background is stationary. Hence, all the moving pixels outside the tracked head area are considered as belonging to the body area. Each frame is converted to a gray scale image, F_t, and the difference image, $\Delta_t = F_t - F_{t-1}$ is calculated.

The difference image is thresholded to identify the moving pixels, and then the total number of moving pixels in each frame, normalized by the frame size S, is recorded. We use a manually selected threshold ($Th_g = 30$) for this purpose, which means that if the difference between the gray scale values of two pixels is greater than this threshold,

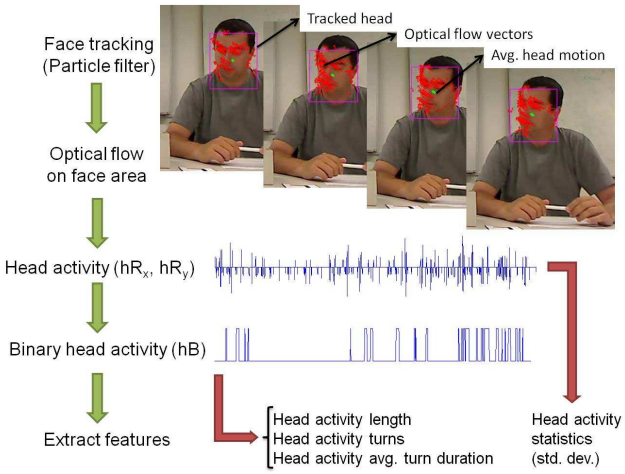


Figure 5.1: Head activity feature extraction.

it is considered as a moving pixel. For each participant, this results in a real-valued vector bR with elements bR_t describing the body activity of that participant during the whole meeting:

$$bR_t = \frac{1}{S}\sum(\Delta_t > Th_g). \tag{5.2}$$

Furthermore, to identify significant body activity, we binarized this vector with a threshold $Th_f = 0.05$, (i.e., if at least 5% of the pixels are moving in that frame, it is considered as a significant body activity), obtaining the binary vector bB. This threshold value is set such that it captures the global body movements, while filtering out the local ones.

$$bB_t = \begin{cases} 1, \ if \ bR_t > Th_f \\ 0, \ otherwise. \end{cases} \tag{5.3}$$

It is important to note that the values of the thresholds are chosen with respect to the video recordings in the ELEA-AV corpus. For different video recordings, different threshold values would be needed.

For each participant, using bR and bB, the following features, which represent the participant's body activity during the meeting, are calculated.

Body activity length (TBL_i): The total time that participant i moves his/her body, calculated from bB.

Body activity turns (TBT_i): The number of turns for each participant i, where each turn is considered as continuous body activity, calculated from bB.

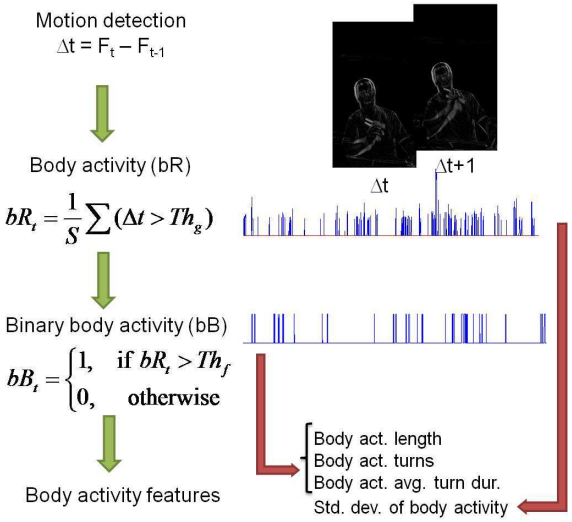


Figure 5.2: Body activity feature extraction

Body activity average turn duration (ABT_i): The average turn duration for participant i, calculated from bB.

Standard deviation of body activity ($stdB_i$): Standard deviation of body activity, calculated from bR.

5.3.2 Motion template based features (MT)

As an alternative approach to characterize visual activity, we use motion templates to extract the full body activity features of each participant throughout the meeting. Bobick and Davis proposed the Motion Energy Image (MEI) and the Motion History Image (MHI) as ways to summarize the spatio-temporal content in a single image (17). MEI is a binary image showing the location of the motion, whereas MHI is a grayscale image showing both the location and the direction of the motion. Both MEI and MHI are proposed as motion templates to describe short motion, mainly for human action recognition. We propose a modified version of MEI, what we call Weighted Motion Energy Image (wMEI) illustrated in Figure 5.3. wMEI is proposed to represent the dominant motion regions, and is suitable as a template for videos of long duration. It is a gray scale image describing the location along with the intensity of motion throughout the video in that region.

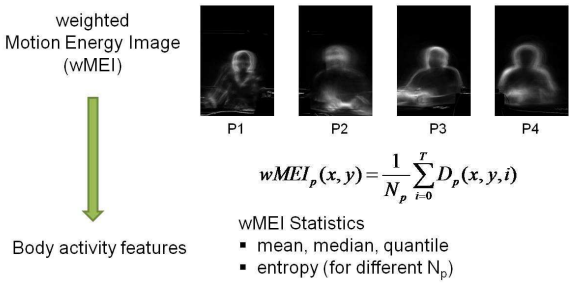


Figure 5.3: Weighted motion energy image based body activity feature extraction

A wMEI contains the accumulated motion information and is calculated as:

$$wMEI_p(x,y) = \frac{1}{N_p}\sum_{t=1}^{T}(D_p^t(x,y,t)), \tag{5.4}$$

where $D_p^t(x,y,t)$ is a binary image that shows the moving regions for participant p at time t, N_p is the normalization factor, and T is the total number of frames. Unlike motion energy images, wMEI is not a binary image. In wMEI, the brighter pixels correspond to regions where there is more motion. wMEI can be normalized by dividing all the pixel values by the maximum pixel value. Alternatively, the length of the video can be used as a normalization factor.

For each participant, we calculate the wMEI and extract several statistics as body activity features. These include the maximum (*wMEImx*), mean (*wMEImn*), median (*wMEImd*), and 75% quantile (*wMEIqn*) of the intensity value of wMEI. For mean, median and quantile calculation, we omit zero values in the wMEI and only use the non-zero intensities. In addition to these statistics, we also calculate the entropy. For entropy, we follow three different approaches to obtain the normalized wMEIs on which the entropy is calculated:

1. *wMEIeP*: $N_p = max(\sum_{t=1}^{T}(D_p^t))$.
2. *wMEIeA*: $N_p = max(N_{p,1}, N_{p,2}, ..., N_{p,P})$.
3. *wMEIeT*: $N_p = T$.

N_p is the normalization factor used in Eq. 5.4, and P is the number of participants in a meeting. The first approach, *wMEIeP*, uses the maximum value in the wMEI of each participant as the normalization factor. This value is unique for each participant

in each group. The second and third approaches use a single normalization factor for all participants in the group: in *wMEIeA* the normalization factor is calculated as the maximum intensity in all the wMEIs of participants in the meeting, and in *wMEIeT* the normalization factor is set as the length of the video.

5.4 Inferring the Emergent Leader

As in the previous chapter, we use four approaches to infer the emergent leader in each group: (i) A rule-based approach, in which the participant with the highest nonverbal feature value in the group is selected as the leader; (ii) rank–level fusion which is an extension of the rule-based approach to handle fusion of multiple features; (iii) support-vector machine, a supervised learning method using a leave-one-meeting-out cross-validation and; (iv) a collective classification approach, which uses relational information in addition to the nonverbal feature vector, using a leave-one-meeting-out cross-validation.

The description of the approaches can be found in Section 4.4. For the fusion of audio-visual features, there is no need to synchronize the two streams at frame level.

5.5 Experiments and Results

In this Chapter, we first present a correlation analysis between questionnaires on perception and nonverbal features, we then present results on leadership estimation.

5.5.1 Correlation Analysis

For this analysis, we validated correlations, by calculating the Pearson correlations per group, then applying a Fisher transformation, and finally testing whether the correlations were statistically significant with a t-test, at 5% significance level (i.e., $p < 0.05$).

Nonverbal speaking behavior and perception from participants. Table 5.1 shows Pearson correlation values between questionnaire outputs and individual audio nonverbal features. As we can see, correlations between PLead and several speaking turn features show significant correlations as similarly described in Section 4.5.1.3. Note that these results are shown for completeness purposes, as ELEA-AV is a subset of ELEA-A, and so similar results would be expected from the audio cues. As stated

Table 5.1: Correlation values between variables from questionnaires and nonverbal acoustic features on the ELEA-AV corpus ($* : p << 0.005$, $\dagger : p < 0.05$). For Energy and Pitch features, only significant correlations with at least one of the concepts are shown.

	PLead	**PDom**	**PCom**	**PLike**	**RDom**
TSL	0.72*	0.45†	0.24	-0.52*	0.70*
TST	0.56†	0.39	0.22	-0.05	0.42†
TSTf	0.68*	0.52†	0.23	-0.45†	0.60*
AST	0.74*	0.43†	0.32	-0.35†	0.71*
TSI[1]	0.72*	0.45†	0.28	-0.50†	0.59*
TSIf[1]	0.60*	0.38†	0.27	-0.43†	0.65*
TSI[2]	0.51†	0.53†	0.24	-0.23	0.53*
TSIf[2]	0.67*	0.48†	0.48†	-0.42*	0.60*
EMIN	-0.49†	-0.36†	-0.12	0.35†	-0.52†
EMED	0.23†	0.14	0.18	-0.10	0.20
PVAR	-0.18	-0.19	-0.21	0.10	-0.31†

in Chapter 4, this suggest that emergent leadership perception has a connection to the person who talks the most, has more turns, and interrupts the most.

Nonverbal visual behavior and perception from participants. We use the 27 meetings recorded with the portable setup from the ELEA corpus that include both audio and video recordings, which we call ELEA Audio-Visual (AV) corpus. Pearson correlation values between individual visual nonverbal features and questionnaire outputs are shown in Table 5.2. Significant correlations can be observed between PLead and body activity (TBL, TBT, ABT, and stdB) up to r=0.58(p=0.02), and PLead and motion statistics (wMEIeA, wMEImx, wMEImn,wMEImd, and wMEIqn) up to r=0.61(p=0.001). These results seem to support Baird's (12), that found that gesticulation of arms and shoulders is an important contributor in the perception of emergent leadership. PDom and RDom have as well significant correlations with body activity (TBL and ABT) and motion statistics, of up to r=0.44(p=0.02). As exposed in (29), dominant individuals are highly noticeable by their body movements and gestures, in association with their vocal cues. For the case of PCom, there are only significant correlations with the head activity (THL, THT, AHT, and stdHy), with up to r=0.35(p=0.004). Finally, for PLike significant negative correlations are found with motion activity (wMEImd, wMEIeA, and wMEIeP), with up to r=-0.46(p=0.03). This finding suggest that the person who is perceived most agreeable, moves the least during the interaction.

Table 5.2: Correlation values between variables from questionnaires and nonverbal visual features on ELEA-AV corpus ($* : p << 0.005$, $\dagger : p < 0.05$).

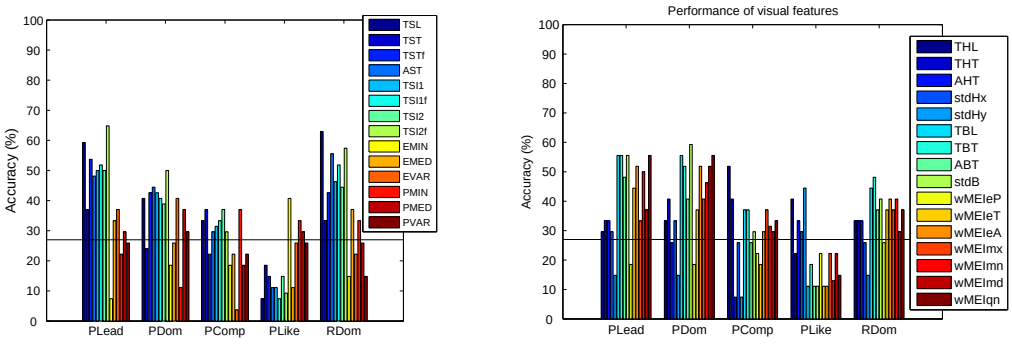
	PLead	**PDom**	**PCom**	**PLike**	**RDom**
THL	0.19	0.16	$0.29^{\dagger}$	0.19	0.29
THT	$0.41^{\dagger}$	$0.40^{\dagger}$	0.35^{*}	0.02	0.29
AHT	-0.26	$-0.24^{\dagger}$	$-0.47^{\dagger}$	0.28	-0.17
stdHx	-0.15	-0.16	-0.17	0.05	-0.07
stdHy	-0.23	-0.33	$-0.38^{\dagger}$	0.21	-0.18
TBL	$0.53^{\dagger}$	$0.40^{\dagger}$	0.17	-0.39	$0.44^{\dagger}$
TBT	$0.57^{\dagger}$	$0.37^{\dagger}$	0.15	-0.34	0.40
ABT	$0.40^{\dagger}$	$0.34^{\dagger}$	-0.02	-0.45	$0.34^{\dagger}$
stdB	$0.58^{\dagger}$	$0.43^{\dagger}$	0.07	-0.27	0.31
wMEIeP	$0.39^{\dagger}$	0.21	-0.02	$-0.46^{\dagger}$	0.33
wMEIeT	0.02	-0.12	0.06	0.34	0.01
wMEIeA	0.61^{*}	$0.37^{\dagger}$	-0.07	$-0.43^{\dagger}$	$0.48^{\dagger}$
wMEImx	$0.42^{\dagger}$	$0.34^{\dagger}$	0.07	-0.14	0.31
wMEImn	$0.31^{\dagger}$	$0.30^{\dagger}$	-0.01	-0.08	0.14
wMEImd	$0.56^{\dagger}$	0.28	-0.16	$-0.40^{\dagger}$	$0.43^{\dagger}$
wMEIqn	$0.49^{\dagger}$	$0.38^{\dagger}$	0.26	-0.27	$0.41^{\dagger}$

5.5.2 Leadership Inference using Audio-Visual Nonverbal Cues

In this section, we present the results for each of the four estimation methods and the audio, visual and audio-visual cases. Note again that the audio-only results are presented for completeness purposes, as they are in essence similar to the ones reported in Chapter 4 (the differences coming from being trained/tested on the smaller ELEA-AV corpus). For the 27-meeting ELEA-AV corpus, the random baseline performance is 26.8% for the inference of the emergent leader (or the other variables).

5.5.2.1 Single Cues Rule-Based approach

Figure 5.4 shows the accuracy of the audio features and the visual features respectively, for the five tasks on the ELEA-AV corpus. The results on the visual features show that for PLead, PDom, and RDom the body activity features and wMEI based features perform better than the head activity ones. The better performance on body activity and motion, over head activity, could be due to the natural movements while speaking, and due to the emergence of movements and gestures that complement the speech. On the contrary, for PComp and PLike, head activity features perform better, possibly due to differences in nodding and agreement/disagreement gestures, in addition the head activity serves to display interest/boreedom (i.e. looking at others while they



a) b)

Figure 5.4: Accuracy of the nonverbal features on the ELEA-AV corpus: a) audio and b) visual. The black horizontal line shows the random baseline.

speak). Some visual nonverbal features perform quite poorly, for example the standard deviation of the vertical head activity (stdHy), giving accuracies below the baseline for all variables except for PLike. As stated before, the difference could be due to nodding and agreement; mostlikely the most agreeable person will nod or agree with the group. The highest performance for emergent leadership is 55.6% and is achieved by TBL, TBT, stdB and wMEIqn features, compared with 70.4% obtained for audio. On the contrary, for PComp and PLike, head activity features perform better (THL and stdHy, with 51.85% and 44.4% respectively), compared to 37% and 40.7% obtained with audio.

5.5.2.2 Rank–Level fusion approach

We performed an exhaustive search for all feature combinations up to six features on the ELEA-AV corpus. Figure 5.5 shows the accuracies of best single audio nonverbal feature, best single video nonverbal feature, audio-visual fusion, audio-only fusion, and video-only fusion on the ELEA-AV corpus. We also show the confidence intervals of the best accuracy, with 95% confidence, with respect to the number of examples in the dataset. The results show that, for PLead and RDom, the best audio feature provides higher accuracy than the best visual feature. This fact is reversed for PDom, PComp, and PLike. For each of the variables, audio-visual fusion provides the highest accuracy, better than audio-only or visual-only fusion. Table 5.3 shows the fused variables, that produced the highest accuracy for each of the tasks. The highest achieved accuracy

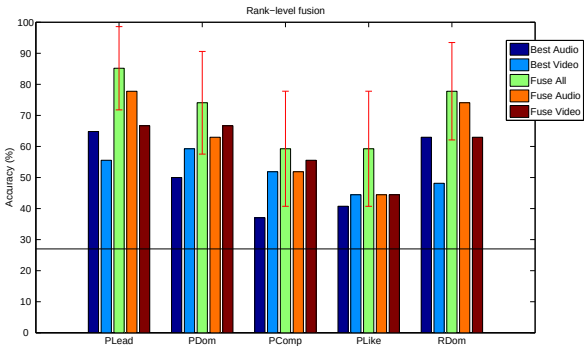


Figure 5.5: Audio-visual, audio-only, and visual-only score-level fusion results on the ELEA AV corpus. The accuracies of best single audio nonverbal feature and best single video nonverbal feature are also shown. The black horizontal line shows the random baseline.

Table 5.3: Results of rank-level fusion on the ELEA AV corpus. The last column of the table summarizes the fused features with respect to the feature groups (ST: speaking turn, HA: head activity, BA: body activity, MT: wMEI based features, EN: energy, PI: pitch)

	Acc(%)	Fused variables	Feature Groups
PLead	85.2	TSL, TSI^1, $TSIf^2$, THT, TBT, EMED	ST, HA, BA, EN
PDom	74.1	TSI^1, THT, wMEIqn, EVAR	ST, HA, MT, EN
PCom	59.3	THL, PMIN	HA, PI
PLike	59.3	THL, AHT, PMIN, EMIN	HA, PI, EN
RDom	77.8	TSL, AST, TSI^2, wMEImx, EMED, EMIN	ST, MT, EN

for leadership is 85.2% and corresponds to a variety of the extracted features. As a reference, the best achievable performance for dominance is lower than the one reported in (5, 50) (85.3% and 88.06%) which investigated a subset of the AMI meeting corpus (that is based on a different group task).

For a more detailed look into fused variables, we analyzed the pairwise feature selection frequency in the best combinations of rank-level fusion, as there are multiple combinations giving the best accuracy. For simplicity, instead of reporting the actual frequencies of features, we grouped the features into six feature groups, and report the pairwise frequencies of the feature groups in Figure 5.6. In each matrix of Figure 5.6, the diagonal corresponds to the selection frequency of that feature group, whereas off-diagonals indicate the pairwise frequencies: the brighter the pixel, the higher the frequency. Several observations can be made from this figures:

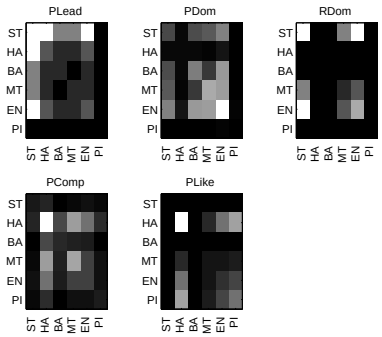


Figure 5.6: Pairwise frequency of feature groups in best combinations

- For all the variables, audio-visual fusion is essential.
- Head activity is more important for PLead, whereas it is not used in PDom or in RDom. Instead, PDom and RDom use body activity or wMEI based features as visual information.
- Pitch information is not used in PLead, PDom, and RDom. However it is informative for PLike and to a lesser degree PCom.
- For PLike and PCom, head activity, energy and pitch are the most informative features. Speaking turn features have a very little effect for these two variables.

5.5.2.3 Collective classification approach

Table 5.4 (right) shows the accuracies for emergent leader and related concepts for the collective out-of-sample task on the ELEA AV corpus. We observe that adding visual information increases accuracy inference using the ICA algorithm: PLead, PDom and PCom increased accuracy with respect to audio-only performance. The best accuracy obtained for PLead is 81.0%.

Table 5.4 (left) shows the averaged accuracy results for the in-sample task (i.e. a known label per group). Again, since participants with the lowest feature values are not perceived often as leaders nor most dominant, we labeled these participants as Non-EmergentLeader/Non-MostDominant. The test is performed using this known label and the emergent leader and related concepts are inferred from the remaining two or three participants in the group. For this task the baseline, random accuracy is 37.0%.

Table 5.4: Best accuracy results (%) of collective classification using audio and visual features on the ELEA AV corpus from Out-of-sample and In-sample tasks. Feature groups: ST-Speaking Turn, HA-Head Activity, BA-Body Activity, MT-Motion (wMEI based), EN-Energy, PI-Pitch.

		Out-of-sample	feature group	**In-sample**	feature group
Audio	PLead	59.5	ST	63.7	ST
	PDom	58.3	ST, EN	61.9	ST
	PCom	41.7	EN	57.1	ST, EN
	PLike	63.1	ST, EN	75.0	ST
	RDom	67.9	ST	82.1	ST
Visual	PLead	70.2	HA, BA	78.6	HA, BA
	PDom	67.9	BA	67.9	BA
	PCom	35.7	HA	51.2	HA
	PLike	50.0	MT, BA	42.9	BA
	RDom	53.6	BA	72.6	MT, BA
AV	PLead	81.0	ST, BA	85.7	ST, EN, BA
	PDom	70.2	ST, BA	70.2	ST, EN, BA
	PCom	46.4	HA, EN, PI	57.1	ST, EN
	PLike	63.1	ST, EN	75.0	ST
	RDom	67.9	ST	82.1	ST

In general terms, with the in-sample task less features are needed to discriminate between emergent leaders and non-emergent leaders. Additionally, the performance for the emergent leader with respect to PLead (85.7%) and RDom (82.1%) is higher than the out-of-sample task. This confirms the statement of McDowell et al. (74), which affirms that having known labels for the test phase can provide better accuracy in realistic scenarios. From Table 5.4 we can observe that for the variables PLead and PDom, the combination of audio and visual information performed better, in contrast with PLike and RDom for which audio features performed better than the combination of features. Finally, for PCom the combination of audio and visual features performed better than only audio or only visual information; on the other hand, if we provide a labeled example, only audio features performed better than the combination.

5.6 Discussion

We now discuss and compare the highest accuracy of all the methods including single and combined feature modalities. These are presented in Table 5.5. In order to compare the methods, we used a standard normal significance test as mentioned in Section 4.6. For PLead, the highest accuracy using CC-In-Sample is 85.7%. However, there is not

statistically significant evidence that this method outperforms Rule-based Estimator, Rank-level Fusion or CC-Out-of-Sample. The performance of all four methods was found statistically higher compared to random performance ($p < 0.0002$). The emergent leader with our methods can be inferred between 70.4 and 85.7% accuracy.

For PDom, the highest accuracy using Rank-level Fusion is 74.1%, the outperformance over Rule-based Estimator is statistically significant ($p = 0.05$). However, the relative improvement over CC (In-Sample and Out-of-Sample) is not statistically significant. Our methods can infer the perceived dominant person within 70 and 74.1% accuracy, significantly higher than random performance ($p < 0.0002$). Regarding RDom, the highest accuracy (82.1%) is reached using CC-In-sample. The improvement in performance is marginally significant compared with Rule-based Estimator ($p = 0.07$). However, there is no evidence that the performance of CC-In-Sample is better than Rank-level Fusion nor CC-Out-of-Sample. The methods Rank-level Fusion and CC (In-Sample and Out-of-Sample) were found statistically significantly better than random performance ($p < 0.002$). The most dominant person can be inferred with our methods within 67.9 and 82.1% accuracy.

For PCom, the Rank-level Fusion method performed up to 59.3% accuracy, which is statistically better than Rule-based Estimator ($p = 0.04$). However, there is no evidence that it performs better than CC-In-Sample and CC-Out-of-Sample. Furthermore, Rank-level Fusion and CC (In-Sample and Out-of-Sample) were found statistically better than random ($p < 0.0004$). Our methods can infer the perceived competent person within 46.4 and 59.3% accuracy.

For PLike, the highest accuracy is 75.0% using CC-In-sample, which is statistically better compared with Rule-based Estimator ($p = 0.01$). In comparison with Rank-level Fusion and CC-Out-of-Sample, there is no evidence of better performance. The methods Rank-level Fusion and CC (In-Sample and Out-of-Sample) were found statistically better than random ($p < 0.004$). Thus, our methods can infer the most agreeable person in the group within 59.3 and 75.0% accuracy.

5.7 Conclusions

In this work, we presented a computational framework to infer emergent leadership in newly formed groups from nonverbal behavior, by combining speaking turns, prosodic

Table 5.5: Best accuracy (%) of all methods on the ELEA AV corpus with audio and visual features

	PLead	**PDom**	**PCom**	**PLike**	**RDom**
Baseline	26.8	26.8	26.8	26.8	26.8
Rule-based Estimator	70.4	51.9	37	40.7	63.0
Rank-level Fusion	85.2	**74.1**	**59.3**	59.3	77.8
CC-Out-of-Sample	81.0	70.0	46.4	63.1	67.9
CC-In-Sample*	**85.7**	70.2	57.1	**75.0**	**82.1**

features, visual activity, and motion. We evaluated the effectiveness of individual and combined audio and visual features in identifying the emergent leader and related constructs using four inference approaches. Based on the results, we noticed that the nonverbal acoustic information could be augmented by using the head or body activity information. The augmentation with body activity is explained by the nature of the interaction, since there is a natural emergence of movements and gestures that complement the speech. The head activity, aside from the effect of movements due to the speaking activity, might also be due to agreement/disagreement gestures while listening. In order to infer the emergent leader, although the combination of acoustic and visual information resulted in higher values of performance than single modalities, there was not enough statistical evidence to fully validate these findings. The emergent leadership inference ranges between 70.4 and 85.7% accuracy using our investigated methods.

The perceived most dominant person in the group can be inferred within 70 and 74.1% accuracy, with combination of audio and visual cues. In contrast, the ranked most dominant person can be inferred most of the time with only audio cues, rather than with combinations of audio and visual cues, within 67.9 and 82.1% accuracy (these differences of performance were not statistically significant). Regarding other concepts related to leadership, we found that for the perception of competence, informative nonverbal cues came from head activity, energy and pitch; as we observed from the correlations. The perceived most competent person can be correctly inferred using either audio or combined audio-visual cues within 46.4 and 59.3% accuracy. For the case of perceived liking, although the most informative cues were extracted from the audio channel when using supervised approaches, the most agreeable person can be inferred within 59.3 and 75.0% accuracy, with either audio-only or combined audio-visual cues. Given the nature of the in-sample collective approach, the most informative features

differ from the ones using rule-based and rank-level fusion approaches. Note also that the results for perceived leadership, perceived dominance and ranked dominance were sometimes similar and sometimes different for the same features and inference method. As already stated in Chapter 4, we opted not to combine the three measures because they capture somewhat different aspects of verticality. Note also that perceived liking and perceived competence, which we assessed as aspects of socio-emotional and task-oriented leadership, respectively, showed results different than those obtained for perceived dominance.

Other visual cues have been discussed in the literature regarding social verticality, namely visual attention. Extracting these features is a more challenging task. As the degree of complexity in extracting visual features increases, one might expect performance improvements. We study this issue in the next chapter.

5. INFERRING EMERGENT LEDERSHIP FROM AUDIO-VISUAL NONVERBAL ACTIVITY CUES

Chapter 6

Inferring Emergent Leadership from Visual Attention Cues

In this Chapter we focus on the study of features that characterize visual attention and speaking activity of group members for inference of emergent leadership. Some of these features are derived from classic studies in psychology (13, 30) but not yet studied in the context of computational inference. As with the previous two chapters, we first present a correlation analysis between the automatically extracted features and the concepts related to emergent leadership. The nonverbal features are measures of visual attention and speaking activity including synchronized features that are multimodal in nature, such as measures of looking at participants while speaking and the visual dominance ratio. Then, we study the performance of the nonverbal features in estimating the emergent leader in the group. Finally, we present effects of possible misalignments in the multimodal features on the estimation performance. We found that emergent leadership as measured by these features is related, but not equivalent, to dominance, and while multimodal features are relatively effective in inferring the leader, much simpler features extracted from the audio channel are found to perform better.

This paper is organized as follows: we first introduce the ELEA-AVS corpus in Section 6.1, we then present the nonverbal features in Section 6.2. The method to infer emergent leadership and related concepts is presented in Section 6.3. Experimental results are shown in Section 6.3. Finally, we present our conclusions in Section 6.6.

The work reported in this Chapter was published as a journal paper in (98).

6.1 Data

The ELEA-AVS corpus is a subset of selected recordings from the ELEA corpus described in Chapter 3. This subset corresponds to audio-video synchronized data, allowing multimodal synchonous analysis of emergence of leadership. The corpus consists of 22 meetings (19 meetings with four participants and 3 meetings with three participants).

6.2 Visual Attention Features

In addition to manual coding, our corpus includes a number of automatically extracted features. Table 6.1 summarizes the list of features extracted from the corpus, described in this section. We first describe speaking activity features, then visual attention features, and finally audio-visual features that combine speaking activity and attention.

Table 6.1: Feature groups: AT-Visual Attention, SA-Speaking Activity, AV-Audio-visual features.

Feature type	Acronym	Definition
Visual Attention (AT)	ATR	Attention Received
	ATG	Attention Given
	ATQ	Attention Quotient (ATR/ATG)
	ATC	Attention Center
Speaking Activity (SA)	TSL	Total Speaking Length
	TSTf	Total Speaking Turns (longer than 2 seconds)
	TSI	Total Speaking Interruptions
	TSTD	Average Speaking Turn Duration
Audio-Visual (AV)	LWS	Looking While Speaking
	LWL	Looking While Listening
	BLWS	Being Looked While Speaking
	CAWS	Center of Attention While Speaking
	VDR	Visual Dominance Ratio (LWS/LWL)

6.2.1 Visual Attention Features

The extracted visual features are based on attention (denoted VFOA for Visual Focus of Attention), specifically 'who is looking at whom or what'. First, we extract the VFOA and then construct features that could characterize an individual's behavior in group interactions. Gaze cues, along with conversational cues are known to be informative to characterize small group interactions (63). Apart from facilitating the turn-taking

patterns, they also signal socially relevant information, for example dominance or status (40, 41). Features were extracted in a collaboration with Dr Dinesh Babu Jayagopi (Idiap Research Institute).

As tracking eye gaze requires high-resolution videos, and head direction captures eye gaze direction relatively accurately in conversational settings (106), we first estimate the head pose automatically. The head pose is characterized by three angles: pan, tilt, and roll. Then, we assign the head pose to a discrete VFOA label in every frame. We use the method proposed in (90), that employs a dynamic, probabilistic framework to estimate the head location and pose jointly based on a standard state-space formulation. The states correspond to the location and scale of the head as well as the discretized head pose. The observation model uses both color features and texture features (based on Histograms of Oriented Gradients (HOG)). The inference is done using particle filters, which represent the distribution of states at each frame by a finite set of samples (or particles). The left image in Fig. 6.1 shows the tracker output location, which is computed as the mean (in green color) and median (in red color) of the state distribution. The right part of Fig. 6.1 shows the estimated pan and tilt head pose angles represented by the green line over a semi-circumference spanning $\pm 90°$.

Figure 6.1: Tracking, head-pose estimation, and VFOA estimation for an individual in a group interaction in the ELEA AVS corpus. See main text for details.

Considering only pan and tilt angle, the VFOA is later estimated by Maximum a Posteriori (MAP) rule. The MAP rule assumes a Gaussian distribution with mean and standard deviation pre-specified manually (in pan-tilt space), for each of five visual

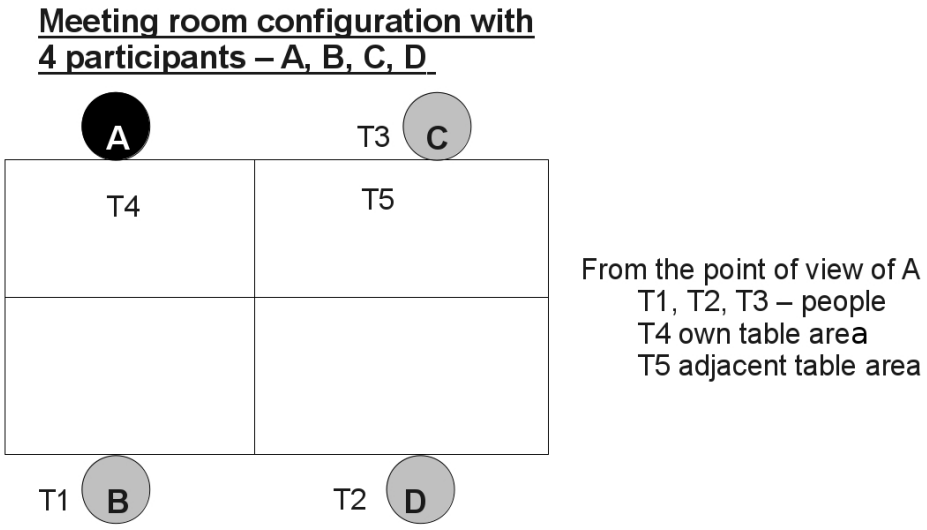


Figure 6.2: The configuration of the meeting room (where the group interaction took place).

targets T1 to T5. Fig. 6.2 shows the position of these visual targets with respect to the configuration of the room. T1, T2 are the participants sitting opposite to the participants shown in Figure 6.1. T3 is the participant sitting next to the tracked participant. T4 and T5 represent the table area close to the tracked participant and participant T3, respectively. Finally, a final UN class is added, where UN stands for unfocused (i.e. any other possible VFOA). The bottom right part of Fig. 6.1 shows the estimated VFOA target (T1 for this particular frame).

In order to assess the VFOA recognition accuracy, we carried out manual annotations of the VFOA of every participant, for one randomly chosen discussion in the ELEA-AVS corpus. Every 15 seconds, the VFOA of every participant was annotated using one annotator. Using this ground truth, the automatic method had an accuracy of 42% (frame-level) when compared to the manual annotation. The cases where the method failed belonged to two categories. The first one was due to tracking failures, which were typically due to background color effects or illumination issues. The second source of error are inaccuracies in head-pose estimation. Errors in tilt estimation sometimes resulted in the wrong assignment of automatic VFOA targets. Our method used a fixed mapping from head-pose angles to VFOA. As mentioned in the previous paragraph, this mapping was pre-specified for every participant. Importantly, typi-

cal VFOA accuracies obtained with similar methods in other group interaction data (e.g. the AMI corpus) are roughly in this order (see for instance (10)). Also note that more sophisticated methods, which for instance model the joint VFOA of multiple people (11), could probably result in higher recognition performance but have not been studied here.

From the recognized VFOA labels, i.e. the visual target of each participant, the following features that capture socially relevant information are extracted:

Attention Received (ATR): ATR is the number of frames in which the participant i is looked by the other participants.

Given Attention (ATG): ATG is the number of frames in which a participant i looks at other participants.

Attention Quotient (ATQ): is the ratio between the amount of attention that participant i received from the other participants (ATR) and the amount of attention that participant i gives to the other participants in the group (ATG).

Attention Center (ATC): ATC is the total number of frames in which participant i received attention from all the participants in the group at the same time.

Similar features were originally used by Hung et al., (43) to characterize dominance in small groups in the AMI corpus. Furthermore, other related features have been used to capture connections between attention and personality (107), and to investigate interpersonal influence (79). Furthermore, attention features have been discussed in some of the classic works in social psychology on dominance and nonverbal behavior (25, 30).

6.2.2 Speaking Activity Features

The speaking activity features used in this chapter, are described in detail in Chapter 4. The acronyms and short definitions are listed in Table 6.1.

6.2.3 Multimodal Features

The fusion of features obtained from different channels can provide a better understanding of the group interactions (78). As described by Dovidio, the proportions of look-speak and look-listen in a conversation provide information about dominance and power (28). This finding has been verified with automatic features by Hung et al. (43). We extracted the following variables.

Looking while Speaking (LWS): Amount of attention (in frames) that participant i gives to the participants in the group while i is speaking.

Looking while Listening (LWL): Amount of attention (in frames) that participant i gives to the participants in the group while i is not speaking. Note that we cannot infer that a person is listening, so we approximate this by non-speaking.

Being Looked at while Speaking (BLWS): Amount of attention that participant i receives from the other participants while i is speaking.

Center of Attention while Speaking (CAWS): Number of frames that participant i is the center of attention (i.e. all the participants are looking at her/him at the same time) while i is speaking.

Visual Dominance Ratio (VDR): Ratio of Looking while Speaking and Looking while Listening (LWS/LWL).

To compute these features, audio-visual synchronization is needed and thus the ELEA-AVS corpus is used.

6.3 Inferring Emergent Leaders

To infer the emergent leader in a group, we use the same rule-based estimator defined in Chapters 4 and 5.

6.4 Experiments and Results

In this section we first present a correlation analysis between the visual attention, audio and multimodal features with the perceived variables. We then present the results on the inference of emergent leaders and related concepts.

6.4.1 Visual attention cues and perception from participants

In this section we present correlations between the visual attention and the perceived variables. Table 6.2 shows Pearson correlations between the features extracted from attention and the perceived variables. The Pearson correlations are calculated per group, followed by Fisher transformation and a t-test at 5% significance level. The mean value of the Fisher transformation is calculated and then passed through a inverse Fisher transformation function. As we can observe, there are significant correlations

Table 6.2: Pearson correlation from attention features and speaking activity ($^{+}$: $p < 0.05$, $*$: $p < 0.01$). ATR-Attention Received, ATG-Attention Given, ATQ-Attention Quotient and ATC-Attention Center, TSL-Speaking Time, TSTf-Turns, TSI-Interruptions and TSTD-Average Speaking Turn Duration.

	ATR	ATG	ATQ	ATC	TSL	TSTf	TSI	TSTD
PLead	0.46*	0.02	0.27	0.37*	0.69*	0.70*	0.68*	0.65*
PDom	0.54*	-0.17	0.49^{+}	0.45^{+}	0.42	0.55*	0.52^{+}	0.32
PCom	-0.16	0.19	-0.12	0.06	0.21	0.07	0.40	0.25
PLike	-0.60*	0.34	-0.71^{+}	-0.60*	-0.49^{+}	-0.35^{+}	-0.37^{+}	-0.34
RDom	0.41*	0.14	0.1	0.22	0.67*	0.65*	0.66*	0.69*

Table 6.3: Pearson correlation between attention features and multimodal features ($^{+}$: $p < 0.05$, $*$: $p < 0.01$).

	TSL	TSTf	TSI	TSTD	LWS	LWL	BLWS	CAWS	VDR
ATR	0.33^{+}	0.32^{+}	0.53*	0.20	0.13	-0.35*	0.75*	0.81*	0.33*
ATG	0.05	0.01	-0.262	0.06	0.45*	0.56*	-0.12	-0.22	0.011
ATQ	0.25^{+}	0.30^{+}	0.45*	0.14	-0.19	-0.65*	0.56*	0.67*	0.24^{+}
ATC	0.12*	0.15*	0.23	0.06	-0.037	-0.37*	0.58*	0.85*	0.19

between ATR, and the variables PLead, PDom and RDom. For PLike there is negative correlation with ATR and ATC, suggesting that the person with high score in PLike received the less amount of attention from the group. Table 6.2 also shows correlations between the speaking activity features and the perceived variables.

Further, we reviewed the correlations between the visual attention and the acoustic nonverbal features. In Table 6.3 we can observe significant correlations between the attention received ATR and TSL, TSTf and TSI. Also the correlations between ATQ and, TSL, TSTf and TSI are significant. Finally, low (but significant) correlations can be observed between ATC and, TSL and TSTf.

The correlations between TSL and ATR, although lower compared with the ones reported in (107) using a winter survival task scenario, show that the attention received in small groups is correlated to the total amount of speaking activity and, in our case it also correlates with the successful interruptions to grab the floor. We also performed correlations between multimodal (i.e. audio-visual) and visual attention features, shown in Table 6.2. We can observe that there are significant correlations between CAWS and ATR, CAWS and ATC, and, CAWS and ATQ. The strong correlations suggest that being the center of group attention while speaking is connected to the amount

Table 6.4: Accuracy (%) performance from visual attention and speaking activity features on the ELEA-AVS corpus. Random performance is 26.1%

	ATR	ATG	ATQ	ATC	TSL	TSTf	TSI	TSTD
PLead	**59.1**	22.7	40.9	40.9	54.5	45.5	**72.7**	45.5
PDom	**68.2**	22.7	59.1	54.6	31.8	40.9	**45.5**	40.9
PCom	31.8	22.7	18.2	**36.4**	**31.8**	13.6	**31.8**	**31.8**
PLike	4.5	**27.3**	22.7	13.6	9.1	**18.2**	4.5	9.1
RDom	**45.5**	22.7	22.7	27.3	54.5	36.4	**63.6**	50

of attention received as much as being the visual attention center during the meeting. Similarly, significant correlations can be observed between BLWS and ATR, BLWS and ATQ, and BLWS and ATC. Finally, there are significant negative correlations between LWL and ATR, ATQ and ATC, which indicates that the participants that look the most at others while not speaking, capture less amount of attention from the group.

6.4.2 Leadership Inference with Visual Attention Cues

In this section we present the results on the emergent leadership inference (and related concepts) using the visual attention features and the rule-based method. In Table 6.4 we observe that the amount of attention received (ATR) from participants is the most informative cue for emergent leadership (59.1%), followed by the amount of attention received from the group (ATC) with 40.9%. For the case of PDom, the best performance is 68.2% as well with the feature ATR, which suggest that the most dominant participant receives a large amount of visual attention in the group. For PLike, the best performance is 27.3% with the feature ATG and is about random. For the case of PCom, the best accuracy is 36.4% with ATC, suggesting that the perceived most competent person gets a significant amount of attention from the group. In Table 6.4 we can also observe accuracy performance of single nonverbal speaking cues extracted and the rule-based method in Chapter 4. As we can observe, the highest accuracy with visual attention features (ATR with 59.9%) is lower compared with the most informative speaking activity feature (72.7% with TSI).

6.4.3 Multimodal features

In this section we present the results of identification of the emergent leader and related concepts using multimodal features. Considering that nonverbal behavior extracted

Table 6.5: Accuracy (%) performance from frame based multimodal features on the ELEA AVS corpus. Random performance is 26.14%.

	LWS	LWL	BLWS	CAWS	VDR
PLead	50.0	4.5	**63.6**	**63.6**	50.0
PDom	31.8	27.3	59.1	**63.6**	36.4
PCom	27.3	22.7	22.7	**36.4**	31.8
PLike	18.2	**36.4**	9.1	4.6	13.6
RDom	50.0	13.6	45.5	45.5	**54.5**

from audio and visual single channel can be used to identify the emergent leaders (100), multimodal features extracted from synchronized audio and video might provide better information about the nonverbal behavior of the emergent leader. Table 6.5 shows performance using the unsupervised method and the multimodal features, where the best performance to identify the leader is using either BLWS or CAWS with up to 63.6%. For PDom the best accuracy is 63.6% with CAWS, for the case of RDom the best accuracy is 54.4% with VDR, this feature has been previously shown as informative nonverbal feature of dominance (43). For Pcom highest accuracy is 36.4% using the information being the center of attention while speaking (i.e., CAWS). Finally, for PLike highest accuracy is 36.4% using LWL.

With the aim of having a better understanding on how multimodal features can perform for PLead, PDom and RDom, we also considered an event-based evaluation strategy. To do this, we count only the times that an event (i.e. segment of consecutive frames with the same multimodal feature) occurs during the meeting instead of counting the exact number of frames in which this event occurs. Considering this option, we can observe in Table 6.6 that the event-based accuracy to infer the emergent leader in the group increases up to 68.2%, on the other hand the inference of the perceived dominant participant in the group decreases from 63.6% to 59.1% for the best multimodal feature (CAWS).

6.4.4 Effect of Stream Asynchrony in Multimodal Features

Frame dropping could occur during video recordings, given to several reasons including applications running in background. As a final experiment, to test the effects of possible

Table 6.6: Accuracy (%) performance from event based multimodal features on the ELEA AVS corpus. Random performance is 26.14%.

	LWS	LWL	BLWS	CAWS	VDR
PLead	50.0	54.5	54.5	**68.2**	50.0
PDom	40.9	45.5	45.5	**59.1**	36.4
RDom	50.0	45.5	**54.5**	50.0	**54.5**

misalignment between the audio and the video channels, we define a alignment-match from the video frame i to a window from i to $i + \delta$ with the respective audio stream, where δ denotes the width of the temporal window in frames (see Figure 6.3).

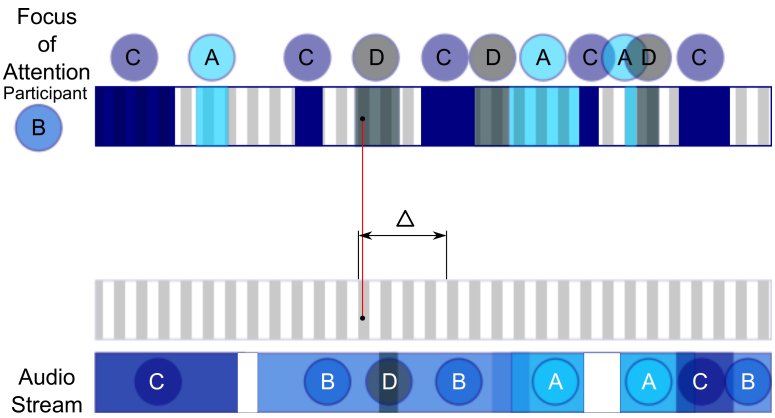


Figure 6.3: Frame alignment window between visual attention and speaking activity streams. Frame i in the attention stream is "aligned" with a window $(i,i + \delta)$ in the speaking activity stream, by allowing the event of interest in the audio stream $(i,i + \delta)$ occur anywhere in the window rather than exactly at frame i, thus relaxing the synchrony assumption.

A video could be susceptible to frame dropping. If it is not well synchronized, we could notice a delay between the visual activity (while speaking) and the audio sound. Considering that our corpus was collected using separate audio and video recording devices, we explored the impact of possible asynchrony in the multimodal extracted features. In our experience, as it is most likely that the frame dropping occurs in the video stream, we considered the effect of slight dropping frame in the video channel on the extraction of multimodal features. More explicitly, when defining looking while speaking at time t, we compute instead "looking at t while speaking anywhere in the

window $(t,t+\delta)$".

Figure 6.4 shows the accuracy considering the variables PLead, PDom and RDom where this effect is model. The X axis represents the amount of frames considered (δ from 1 to 60), i.e., 2 seconds. The Y axis represents the accuracy performance, using the rule-based method in Section6.3.

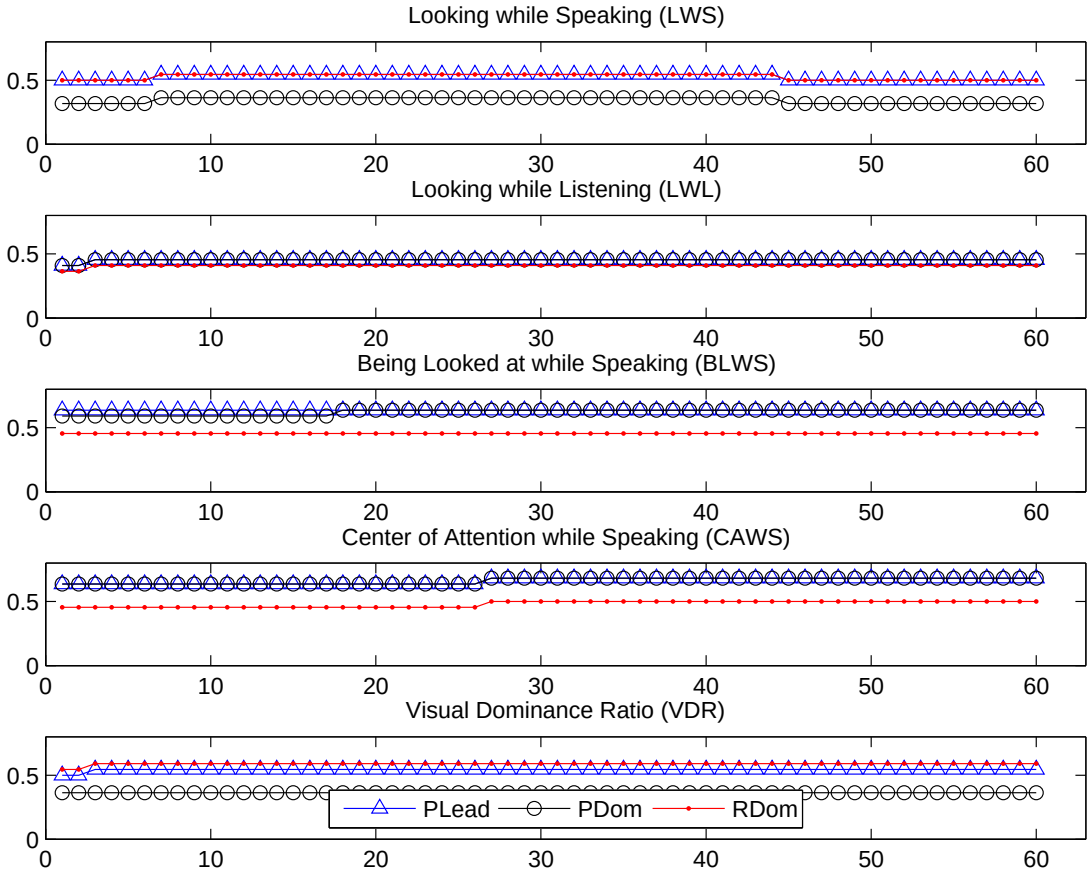


Figure 6.4: Accuracy performance (%) from multimodal features using a time delay alignment window with the audio stream. The X axis represents the amount of frames considered (δ from 1 to 60), i.e., 2 seconds and the Y axis represents the accuracy performance. The extraction of the coordinated visual and speaking activity features is stable, even assuming video frame dropping and using a sliding window $(i,i+\delta)$, as we can see from the stability on the inferences.

As we can observe, in Figure 6.4 the extraction of multimodal features can be robust in frame dropping situations, if an alignment window is considered with respect to the audio stream. The multimodal features can be captured and still provide accurate

inferences if the dropping frame is not to severe. For the case of looking while speaking (LWS), is we use a window of 7 frames (δ=6) and up to 44 frames, we observe a slight improvement in the accuracy. This suggest that the LWS feature can infer leaders better if we use small audio-window for the feature extraction i.e., an alignment of width 0.2 seconds and a maximum width of one and a half seconds (45 frames). For the case of being the center of attention while speaking (CAWS), the performance improves slightly if use δ=26, which suggest that having long continuous speaking activity will allow to recover more accurately the group of attention, even in the video missed up to 26 frames (i.e. a synchronization delay of almost one second). Similarly for BLWS, accurate inferences of leadership and perceived dominance can be done, assuming an asynchrony of approximately half a second (i.e., δ=17).

6.5 Discussion

In this section, we discuss the overall performance of single and multimodal features presented in Section 6.4. To allow comparison among modalities, we performed a binomial significance test, similarly as in Section 4.6. In Table 6.7, we can observe that the highest predictor of emergent leader is TSI with 72.7%. However, the performance of this cue over attention prediction (ATR) and multimodal cues (CAWS) is not statistically significant. On the other hand, speaking, attention, and multimodal cues were found statistically better than random performance ($p < 0.01$). Thus, the emergent leader in the group can be inferred within 59.1 and 72.7% accuracy. The findings suggest that although the focus of attention tracker does not perform accurately all the time, it could perform reasonably well in predicting the emergent leader in cases where the audio channel was not available.

For perceived dominance (PDom), visual attention cues (68.2%) performed marginally better than only speaking cues ($p = 0.07$). However, the performance with visual attention cues is not statistically significant compared with multimodal cues. Moreover, visual attention and multimodal cues were found statistically better than random performance ($p < 0.003$). The methods can infer the perceived dominant person within 63.6 and 68.2% accuracy.

For RDom, the highest accuracy is reached using speaking cues with up to 63.6%, which is marginally higher compared with visual attention cues ($p = 0.07$). In other

Table 6.7: Best accuracy performance (%) from the single and multimodal features on the ELEA AVS corpus. Random performance is 26.1%. TSI-Speaking Interruptions, ATR-Attention Received, CAWS-Center of Attention while Speaking, VDR-Visual Dominance Ratio.

	Variable	Accuracy (%)	**feature**
SA	PLead	**72.7**	TSI
	PDom	45.5	TSI
	PCom	31.8	TSL, TSI, TSTD
	PLike	18.2	TSTf
	RDom	**63.6**	TSI
AT	PLead	59.1	ATR
	PDom	**68.2**	ATR
	PCom	**36.4**	ATC
	PLike	27.3	ATG
	RDom	45.5	ATR
AV	PLead	63.6	CAWS
	PDom	63.6	CAWS
	PCom	**36.4**	CAWS
	PLike	**36.4**	LWL
	RDom	54.5	VDR

words, the opposite result to the one obtained for PDom. On the other hand, the improvement is not statistically significant compared with multimodal cues. Speaking and multimodal cues were found statistically higher compared with random performance ($p < 0.02$). Thus, the ranked most dominant person can be inferred within 54.5 and 63.6% accuracy. Our findings revealed similarity with previous research in ranked dominance using the AMI corpus, which reported better performance using speaking nonverbal cues, as compared with visual attention cues (43).

For PCom, the highest performance of attention features is 36.4%, which is not statistically significantly better than random performance (26.1%). Similarly for PLike, the highest accuracy is obtained with multimodal features (36.4%), not statistically better compared with random performance.

6.6 Conclusions

In this chapter we presented a framework for inference of emergent leadership using visual attention features. Our findings in the ELEA-AVS corpus revealed that for the emergent leadership inference, speaking activity cues performance is neither statistically better than visual attention nor than multimodal cues when using unsupervised rule-

based estimators. The emergent leader can be inferred within 59.1 and 72.7% accuracy using either multimodal cues or single speaking and attention cues. On the other hand, the amount of visual attention received was slightly more informative for the perception of dominance with respect to speaking activity ($p = 0.07$), but the opposite result was found for ranked dominance. Additionally, the multimodal features provided some information about perceived leadership, such that being the center of attention while speaking correlates with being perceived as the leader.

From the multimodal analysis, we conclude that there is a connection between the visual dominance ratio (Looking while Speaking/Looking while Listening) and the most dominant person, in concordance with previous findings in social psychology and social computing (40, 41, 43). Furthermore, the multimodal feature extraction could be robust to slight frame dropping, if we use a sliding window to compute the synchronized features.

For the cases of the perceived competent and agreeable person in the group, we cannot predict better than chance using an unsupervised Rule-based Estimator approach. That said, from the correlation analysis we saw significant but negative correlations between PLike and the amount of attention received. The use of machine learning techniques, as the ones used in Chapter 4 could perhaps capture these connections and improve the inferences.

One practical question that emerges from here is whether the much more complex attention features are justified for applications. Our experiment with the ELEA corpus suggest that this might not be the case. However, addressing this question in more depth would involve replicating these experiments in other settings, and employing better methods to estimate the visual focus of attention.

As limitations, of our work the main one is the size of the ELEA-AVS corpus, which is partly the result of sensing failures during the data collection process, and that limits the observations of statistically significant differences. Second, the automatic extraction of attention features has a relatively low performance on a frame-based accuracy in our corpus (reported to be of up to 42% in (52)). We did not conduct studies using clean manual VFOA labels. This is clearly an important issue to address as part of future work. Third, clearly other better inference methods could have been used, but in this work we made the explicit decision of using simple inference methods and focus on the analysis of the visual attention features.

Until now, we have concentrated in analyzing emergent leadership using only non-verbal behavior, which has been shown to provide information for several vertical traits (33, 50, 71, 83, 100). In the next chapter, we turn to the study of the impact of the verbal content in the perception of emergent leadership.

6. INFERRING EMERGENT LEADERSHIP FROM VISUAL ATTENTION CUES

Chapter 7

Language Style

Many aspects of our identity and relationships are embedded in the words we say and write (24, 35, 54, 81). Existing findings in psychology reveal a strong connection between personality traits and the language embedded in written or spoken forms (69, 81). Language cues also provide information in the prediction of successful relationships (44). Language has also been used to analyze the presidentiality of candidates by using manual transcriptions from publicly available interviews, speeches, and debates (103).

The use of high quality audio recordings in scenarios in which privacy is not an issue allows for verbal content analysis. In this Chapter, we present a framework to identify emergent leaders from automatically transcribed spoken words in face-to-face group interactions. We are not aware of any work that has attempted to use automatic transcription of spoken words in interactions for predicting emergent leaders. We study two novel research questions in the context of predicting emergent leadership in small groups. First, is there any correlation between how an emergent leader is perceived and his/her verbal language style (as opposed to nonverbal cues)? And secondly, can emergent leadership be inferred from only partial verbal samples of the full conversation? The language style is extracted using a psychologically validated content analysis module (LIWC), and investigated both under ideal conditions (clean manual speech transcriptions) and realistic automated conditions where a highly accurate keyword spotter is used in the audio channel. Our findings first reveal a significant correlation between language styles and the perceived emergent leader in a group. Second, a simple word counting approach can also provide an accurate inference of perceived dominance, a variable related (although not identical) to leadership. Third, by using fully auto-

matic extraction of verbal content, we can correctly identify the emergent leaders with an accuracy of up to 82%.

In Section 7.1, we describe our approach. In Section 7.2, we explain the verbal feature extraction, as well as the leadership inference methods. In Section 7.3, we present the analysis of our results. Our final discussion and conclusions are presented in Sections 7.4 and 7.5.

This chapter was published as a long conference paper in (102).

7.1 Overview of our Approach

The analysis for the verbal content, is based on a subset of the ELEA corpus, described in Chapter 3 containing approximately 7 hours of audio. The ELEA-EN corpus contains only English spoken meetings. The ELEA-EN corpus contains 29 groups of three and four participants that are asked to solve the winter survival task as a group while being recorded. There are 20 four-person meetings and 9 three-person meetings. Our approach is summarized in Figure 7.1.

From the audio recordings, we first generate a manual word transcription from the whole interaction. The scripted conversations are trasncribed by a professional company that required only the audio files. In addition, we run an automatic keyword spotting detection system. Potential (i.e., most probable) keywords are automatically detected by the system and assigned a confidence value.

From both the full manual transcriptions and the detected keywords, we proceed to extract word categories using the Linguistic Inquiry and Word Count software (LIWC) (1). LIWC is a text content software that allows language analysis (from text, transcribed speech, blogs etc.) based on high level categories defined by psychologists.

For the emergent leader analysis, we generate individual files per participant accumulating their respective dialog segments and keywords, and then extract the language categories per participant using LIWC. As we work with transcriptions and ASR, we do not compute the categories that involve punctuations like periods, dots, exclamation, etc.

With the resulting 64 word categories, we perform a feature selection applying support vector machine-recursive feature elimination (SVM-RFE) (38). Finally, using the

top relevant features from the categories, we use a supervised method to automatically infer the emergent leader using the top relavant features.

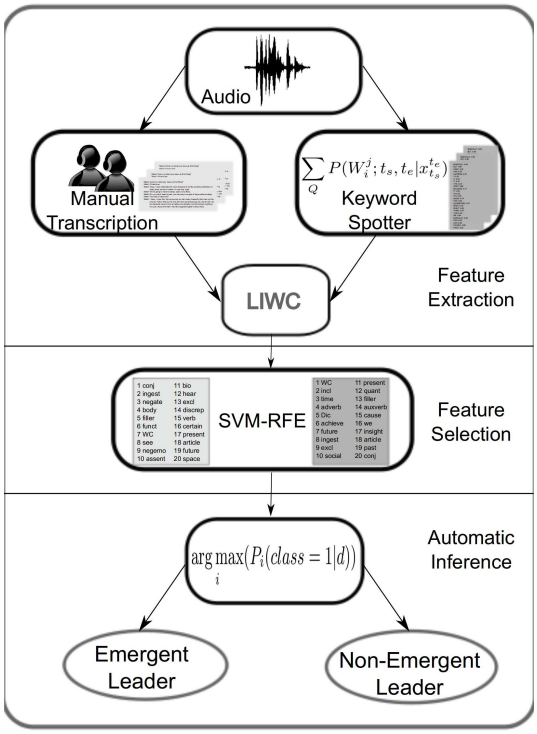


Figure 7.1: Our approach.

7.2 Automatic Analysis of Emergent Leadership

This section details the procedure followed in order to obtain the word categories from the conversations and the estimation techniques that we use. Automatically extracted word categories are needed in order to infer the emergent leader and the perceived dominant person in the group using a supervised model.

7.2.1 Full Transcription

For each audio recording, we generate the full transcription of the conversations. The transcriptions were generated manually, performed by a professional company from only the audio recordings. The level of the transcriptions includes time stamps, gender, natural utterances, and description of events like crosstalk or laughter. Additionally, whenever there were long pause segments (silence), they are interpreted as sequences of

dots (e.g., a 5 second pause of speech is captured as ".....") and unintelligible speech or inaudible words are described as "xxxx". Few manual transcriptions failed in tracking speakers due to similarities in tone of voice. For the incorrectly transcribed recordings, the corresponding video files were used to correct the transcriptions. Each group transcription is segmented to have individual speech transcription files per participant.

7.2.2 Keyword Spotter

As a keyword spotting system (KWS), we use a relatively complex system based on Large Vocabulary Continuous Speech Recognition (LVCSR) in collaboration with Dr Petr Motlicek (Idiap Research Institute). Our LVCSR is built on HMM/GMMs and is trained on 16 kHz multi-distant microphone recordings from several standard meeting corpora (ICSI, NIST, AMI) (39). The LVCSR system decodes the input speech in several passes. The acoustic models are trained discriminatively and in speaker adaptive manner. We also use state-of-the-art, discriminatively trained posterior-based speech features trained using Neural Networks.

During the decoding, a 50k dictionary is used together with a 3-gram Language Model. Such system reaches a Word Error Rate (WER) of about 3% on the well-known Wall Street Journal task (reading Wall Street Journal Sentences (80), in this case independent-head microphone recordings provided by same datasets are used for training of acoustic models)[1].

To perform the search of selected keywords in meeting recordings, the recordings are first pre-processed using the LVCSR system that produces word recognition lattices (representing the set of most probable hypotheses generated by the LVCSR). The word lattices are then converted into a candidate term index accompanied with times and detection scores. The detection scores are represented by the word posterior probabilities, estimated from the lattices using the forward-backward reestimation algorithm (31), and defined as:

$$P(W_i; t_s, t_e) = \sum_{Q} P(W_i^j; t_s, t_e | x_{t_s}^{t_e}), \tag{7.1}$$

where W_i is the hypothesized word identity spanning the time interval $t \in (t_s, t_e)$; t_s and t_e denote the start and end time interval, respectively; j denotes the occurrence of word W_i in the lattice; $x_{t_s}^{t_e}$ denotes the corresponding partition of the input speech (the

[1] www.amiproject.org/ami-scientific-portal/documentation/annual-reports/pdf/D4_2.pdf/view

observation feature sequence) and Q represents a set of all word hypotheses sequences in the lattice that contain the hypothesized word W_i in $t \in (t_s, t_e)$.

Keyword spotting system is performed on our full corpus to extract verbal cues from all the recordings. Obtained manual transcriptions are then used to evaluate the keyword spotter on target data.

Our recordings are acoustically very challenging due to several reasons:

- the corpus is recorded using multi-distant microphones,
- there are many cross-talk segments created by multiple speakers,
- the interaction represent an open vocabulary scenario.

Since the obtained manual transcriptions are not segmented (according to time into speech/silence), we evaluate the keyword spotting system only in terms of ASR using WER. The achieved WER is about 60%. Although this is a quite high number, it also includes errors due to many cross-talks and mainly due to lack of speech/non-speech segmentation (many insertions in non-speech parts of recordings appear and are used for scoring). By applying automatic segmentation, keyword spotter could perform reasonably well as it has been shown on real lecture recording scenarios (76).

7.2.3 Automatic Analysis of Verbal Content

In a spontaneous conversation, people do not often control how or when to use pronouns, articles or auxiliary verbs, but their unconscious use has an impact in the listener and reflects the individual linguistic style. For instance, the preferences in the use of function words, (i.e., use of personal and impersonal pronouns, and articles) have been proved to be related to individuals with skills to socialize (24).

In order to analyze the linguistic style by using high level word categories, we use LIWC to process the manual transcriptions per participant. Similarly, we extract the word categories from the verbal content captured with the keyword spotter, using only the words with high score negative log posterior probability. In order to keep the most accurate words, we use a high threshold score from -1 to 0, where 0 is the highest score for a given word posterior probability equal to 1, i.e., $P(W_i; t_s, t_e) = 1$.

The LIWC dictionary is composed of almost 4,500 words (82). Each word belongs to one or more word categories. For example, the word “agree” is part of three word

categories: *affect*, *posemo* and *assent*. So, whenever the word "agree" is found, the scores in the categories *posemo* and *assent* will be incremented. Additionally, the positive emotion category (*posemo*) belongs to a higher category that considers all the emotion words (*affect*), so this general category will be also incremented. More details on the categories and the dictionary can be found in (1).

For the extraction of categories, we consider 64 categories. Since the scenario is a conversation, punctuations are not considered as relevant as the verbal content per se, thus discarded from the analysis. The results from the categories are interpreted as percentages of full content (speech transcription or keywords per participant), except for the *WC* category (word count) which only accumulates the total number of words per participant.

Considering the resulting 64 categories from the full transcription, 12 categories were discarded from the analysis due to low to null occurrences (e.g., money, religion, etc.). This leaves 52 categories for the analysis.

As a first exploratory step, we performed a correlation analysis in order to see the potential of applying a classification method. As done in previous chapters, a Pearson correlation was computed per group, followed by a Fisher transformation, then a T-test was applied to validate the significance, and the mean correlation values were computed using the Fisher inverse function.

7.2.4 Feature Selection

To analyze the relevance of the features coded from the word categories described in Section 7.2.3, we use support vector machines recursive feature elimination (SVM-RFE) based on a 5-fold cross-validation approach. SVM-RFE is an algorithm that performs feature selection, using SVM to estimate relevance or "weights" of the features. The algorithm eliminates in each step non-discriminative features (i.e., features with the smallest weights), and continues the elimination of features until the remaining features can accurately discriminate between the different classes (38).

Considering the 52 categories obtained from LIWC, we performed SVM-RFE in order to obtain the top most informative features. For the process, we used two SVM-RFE respectively for groups with three and four participants, considering as Emergent-Leader (class 1) and Non-Emergent-Leader (class 0), and similarly for Perceived-Dominant (class 1) and Non-Dominant (class 0), and the other leadership related variables. The

labels are assigned considering the outputs from the questionnaires described in Section 3.2, such that the person with the highest score in perceived leadership is assigned with the label Emergent-Leader per group, and so on for the other variables.

7.2.5 Inferring the Emergent Leader

To infer the emergent leader in the group and the perceived dominant person, we use a supervised method in which the input is a feature vector composed of the top 20 categories selected from the SVM-RFE described in the previous section.

We employ a linear kernel SVM ($k(x, x') = \langle x, x' \rangle$) by using a leave-one-meeting-out approach. As SVM outputs, we obtain posterior probabilities instead of class labels (87), resulting from fitting a sigmoid function

$$P(class = 1|d) = \frac{1}{1 + e^{Ad+B}}, \tag{7.2}$$

where d are the decision values of the SVM, and A and B are estimated by minimizing the negative log-likelihood function. The outputs are then interpreted such that we assign only one Emergent-Leader label (class 1) per group, computed by:

$$Emergent_Leader^f = \arg\max_i (P_i(class = 1|d)), \tag{7.3}$$

where f represents the input feature vector (composed of the top word categories), $nPart$ is the number of participants in the group, and $i = 1, 2, .., nPart$. Finally, the emergent leader in the group is the one with the highest posterior probability belonging to class 1. We follow similar procedure for the perceived dominant person in the group.

Considering the amount of participants and meetings in the corpus, the random accuracy in this case is 27.6%, given that we have 20 meetings with four participants and 9 meetings with three participants.

7.3 Results

In this section we present the results for the inference of the emergent leader in the group using automatically extracted features. First, we present the correlation analysis results using the manual transcriptions. Then, we present the resulting top categories using SVM-RFE and their accuracy in the inference of the emergent leader.

Category	PLead	PDom
WC	0.819*	0.680*
assent	-0.747*	-0.549*
funct	0.645*	0.568*
WPS	0.503*	0.301
article	0.415	0.564*
time	0.400^{+}	0.466
conj	0.391^{+}	0.562*
work	0.349^{+}	0.379^{+}

Table 7.1: Correlation values between word categories from the manual transcription and perceived variables PLead and PDom. Significance values $+ : p < 0.05$, $* : p < 0.01$.

7.3.1 Correlation Analysis

In this section, we present correlations of the top word categories and the perceived leadership variables.

Table 7.1 shows the top high correlated values for the variables PLead and PDom, ranked with respect to PLead. As we can observe, for the variable PLead the highest correlation corresponds to the category *WC* (word count), followed by the *assent* category. Assent is a spoken category related to opinions that express agreement or consent. The negative correlation between PLead and *assent* suggests that the perceived leaders use less words from this category (for instance "agree", "mm*", "ok", "yeah", "yes"). Further, the *WPS* category (words per sentence) shows a high correlation with PLead, suggesting that the emergent leader produces longer sentences. Similarly for the PDom case, the highest correlation is again with the category *WC*, followed by the *funct* category (total function words, this category includes pronouns and articles).

Table 7.2 shows the most significant correlations between PCom, PLike and the word categories. As we can observe, the highest correlation for PCom is with the word category *you*, followed by the word category *social*. The positive correlation with the category *social*, suggests that the perceived competent person uses words like: "anyone", "deal","help","let's","share","us". For the case of PLike the strongest correlation is with the word category *percept*, which includes words like: "chocolate","cold","dry","feel","fire","warm", etc.; this suggest that the person who uses the most words that involve any of the five senses or well being, is perceived as the most

Category	PCom	PLike
you	0.494^{+}	-0.121
social	0.479^{+}	0.098
time	-0.439^{+}	-0.170
WPS	0.177	-0.489^{+}
achieve	0.133	0.391^{+}
home	0.090	-0.539^{+}
percept	-0.004	0.575^{+}

Table 7.2: Correlation values between word categories from the manual transcription and perceived variables PCom and PLike. Significance values $+ : p < 0.05$, $* : p < 0.01$.

agreeable. The negative correlation between PLike and *WPS* (words per sentence), suggests that the most likable person in each group tends to produce short sentences.

7.3.2 Feature Selection

The features described in this section correspond to the resulting outputs from the SVM-RFE based on a 5-fold cross-validation approach. The features correspond to the top word categories from the manual speech transcription and keywords. For the analysis, we used separate SVM-RFE training and test sets for groups with three and four participants.

7.3.2.1 Full manual transcription

In the full transcription case, for groups with four participants, the top 20 ranked features that discriminate between Emergent-Leaders and Non-Emergent-Leaders are shown in Table 7.3 (left). Similarly, the top 20 ranked features according to the SVM-RFE that can discriminate between the classes Perceived-Dominant and Non-Dominant are listed in Table 7.3 (right). As we can observe, the ranking of relevance of the categories differs between the two sets (i.e., for groups with three and four participants), partly due to the size of the data set.

Although the order in the ranking of the features for the variable PLead between the group sizes differ, there are 9 categories that are relevant in the two sets. The similar categories in PLead capture linguistic processes (*WC, article, verb* and *conj*), cognitive processes (*certain*), perceptual processes (*see* and *hear*), biological processes

	PLead		PDom	
Top	**Gp 3**	**Gp 4**	**Gp 3**	**Gp 4**
1	conj	WC	conj	WC
2	ingest	see	incl	motion
3	negate	hear	article	tentat
4	body	you	WC	negate
5	filler	achieve	hear	excl
6	funct	article	negate	filler
7	WC	adverb	assent	we
8	see	quant	verb	achieve
9	negemo	certain	social	certain
10	assent	bio	discrep	incl
11	bio	assent	ipron	space
12	hear	we	ingest	article
13	excl	pronoun	funct	Sixltr
14	discrep	social	i	you
15	verb	conj	cause	ingest
16	certain	death	negemo	hear
17	present	verb	work	feel
18	article	affect	motion	negemo
19	future	cause	nonfl	quant
20	space	WPS	tentat	future

Table 7.3: Top 20 word categories from the SVM-RFE for PLead and PDom, resulting from categories extracted from the manual transcriptions.

(*bio*) and eloquence (*assent*). Similarly for PDom, from the 20 top categories, there are 9 common relevant categories for the two groups although in different ranking. The categories include linguistic processes (*WC, article* and *negate*), affective processes (*negemo*), cognitive processes (*incl* and *tentav*), perceptual processes (*hear*), biological processes (*ingest*) and relativity (*motion*).

Note also that PLead and PDom have categories in common. Although this could suggest that both concepts share a language style, there are also clear differences in the top categories that captures PDom. For example, the categories *motion*, *negate (negations), tentat (tentative), ipron (impersonal pronouns)* and *negemo (negative emotion)*, that capture PDom are not relevant for PLead.

As we can observe from Table 7.4, for PCom there are 7 categories that appear to be relevant for groups with three and four participants. The relevant categories for PCom capture linguistic processes (*WC, article* and *conj*), cognitive processes (*cogmech*), a perceptual process (*feel*), a biological process (*ingest*) and negations (*negate*). PCom has come similarities with the top categories from PLead, specifically *WC*, *negate* and *conj*, in addition the category *certain*, relevant for PLead, belongs to the category cognitive processes (*cogmech*), which is relevant for PCom.

For the case of PLike, there are 8 similar top categories, relevant for the groups. The similar categories capture exclusive (*excl*, which is sub category of cognitive processes), function words (*WC*, *funct* and *pronoun*), social processes (*social*), ingestion (*ingest*), negative emotion (*negemo*), and auxiliary verbs (*auxverb*). Finally, PLike has only one category in common with PLead in a one-to-one base, the category *WC*.

7.3.2.2 Keywords

The top ranked categories derived from the most accurate keywords are shown in Table 7.5. As we can observe, the category word count (*WC*) is the most relevant category for the Emergent-Leader class in the two sets of groups. In contrast for the variable PDom, only for groups with four participants the category *WC* is relevant, and for groups with three participants *WC* is not in the top 20 ranked features.

Considering the top categories for the variable PLead, only 4 categories are similar among the two set of groups (see Table 7.5, left). For the case of PDom, 8 categories are commonly relevant among the groups in the top 20 (see Table 7.5, right).

	PCom		PLike	
Top	**Gp 3**	**Gp 4**	**Gp 3**	**Gp 4**
1	quant	negate	percept	preps
2	article	conj	excl	past
3	tentat	time	negemo	achieve
4	feel	motion	quant	excl
5	negate	ipron	ingest	ingest
6	percept	cogmech	funct	pronoun
7	ppron	WC	conj	social
8	assent	feel	pronoun	motion
9	ingest	space	social	certain
10	work	death	space	death
11	WPS	nonfl	work	negemo
12	cogmech	certain	posemo	nonfl
13	preps	past	verb	WC
14	conj	ingest	auxverb	funct
15	social	inhib	leisure	Dic
16	discrep	article	future	assent
17	WC	auxverb	WPS	cause
18	Dic	leisure	incl	time
19	affect	excl	Sixltr	auxverb
20	future	hear	WC	relativ

Table 7.4: Top 20 word categories from the SVM-RFE for PCom and PLike, resulting from categories extracted from the manual transcription.

	PLead		PDom	
Top	**Gp 3**	**Gp 4**	**Gp 3**	**Gp 4**
1	WC	WC	affect	assent
2	incl	relativ	you	WC
3	time	conj	adverb	cogmech
4	adverb	posemo	posemo	time
5	Dic	space	quant	funct
6	achieve	negemo	filler	bio
7	future	hear	work	past
8	ingest	we	we	negate
9	excl	pronoun	future	inhib
10	social	work	time	leisure
11	present	achieve	verb	work
12	quant	i	tentat	filler
13	filler	cogmech	relativ	adverb
14	auxverb	funct	Sixltr	motion
15	cause	feel	leisure	ipron
16	we	discrep	certain	relativ
17	insight	tentat	social	social
18	article	bio	negate	percept
19	past	negate	Dic	insight
20	conj	social	assent	verb

Table 7.5: Top 20 word categories from the SVM-RFE for PLead and PDom, resulting from categories extracted from the keyword spotter.

Although the automatic keyword spotter can not recover the full conversation (due to overlaps, words that are not in the dictionary, and other reasons) it captures the interaction by recovering some top categories from the full transcription.

For groups with three participants, the categories extracted from the keywords for the variable PLead are consistent with the categories considering the full transcription, such that there is an overlap of 8 categories among the top 20. Additionally, some other categories are captured, if the hierarchical category is considered. For example the *verb* category (top 15 from the full transcription in Table 7.3 left), includes the subcategories *adverb* and *auxverb*, that are captured in the top 20 categories from the keywords (see Table 7.5 left, top 4 and 14). For the case of four participants, 7 out of the top 20 categories are recovered, and it also captures subcategories that appear in top categories from the full transcription, e.g., the *posemo* and *negemo* categories (top 4 and 6 from Table 7.5) are subcategories of the *affect* category (top 18 from Table 7.3). This indicates that the keyword spotter does a reasonably job at extracting relevant feature language related to emergent leadership.

Similarly, for the variable PDom, 6 categories overlap considering the top ranked categories for groups with three participants. For groups with four participants, although only 4 categories overlap one-to-one, higher categories captured from the keywords appear as relevant categories in the full transcription, e.g., the *cogmech* category (top 3 from Table 7.5 right) includes the subcategories *tentat, excl, certain,* and *incl* (top 3, 5, 9 and 10, respectively from Table 7.3 right).

As we can observe from Tables 7.3 and 7.5, the top ranked categories capture the discussion, such that the interaction involves decisions about motion, time and the needs in order to survive an accident. Although some top categories are linked to the context, other categories more related to the language style (free of the context) are relevant in order to discriminate between Emergent-Leaders and Non-Emergent-Leaders (e.g., *WC, conj, excl and assent*). While the WC does not exactly reflect language style, it provides relevant information in the perception of leadership and dominance. Despite the fact that the keyword spotter can not recover exactly the top 20 relevant categories for PLead (from the manual transcription), it captures a significant number of top categories in a one-to-one-base (7 categories out of 20, for both cases, groups with three and four participants).

	PCom		PLike	
Top	**Gp 3**	**Gp 4**	**Gp 3**	**Gp 4**
1	space	filler	space	auxverb
2	tentat	you	quant	Dic
3	filler	i	discrep	funct
4	certain	quant	we	tentat
5	feel	present	nonfl	certain
6	work	achieve	time	motion
7	WC	space	WC	insight
8	pronoun	excl	conj	ipron
9	cogmech	cogmech	future	article
10	see	nonfl	cogmech	filler
11	discrep	relativ	ipron	WC
12	past	Dic	verb	feel
13	leisure	inhib	filler	discrep
14	inhib	time	incl	verb
15	percept	verb	leisure	present
16	you	certain	present	inhib
17	achieve	ingest	past	negate
18	negate	future	work	cause
19	affect	preps	posemo	adverb
20	incl	leisure	hear	relativ

Table 7.6: Top 20 word categories from the SVM-RFE for PCom and PLike, resulting from categories extracted from the keyword spotter.

Table 7.6 shows the top 20 categories for PCom and PLike, for groups with three and four participants. As we can observe, for PCom there are 7 similar categories for both, groups with three and four participants. The similar categories include *space, filler, certain, cogmech, leisure, inhib, you,* and *achieve*. For the case of PLike there are 6 similar top categories for groups with three and four participants. The similar categories capture *WC*, discrepancy (*discrep*), impersonal pronouns (*ipron*), common verbs (*verb* and *present*), fillers (*filler*).

For the case of PCom, the keyword spotter recovers 9 top word categories (relevant as per the manual transcription) from all groups. For PLike, the keyword spotter recovers 10 and 8 top word categories from the manual transcriptions, from groups with three and four participants respectively.

7.3.3 Inference with Manual Transcriptions

The results in this section show the performance on the top 20 features obtained from the SVM-RFE and the manual transcriptions described in the previous section.

Figure 7.2 (a) shows the performance of the perceived emergent leader inference method, using categories derived from the full transcription. As we can observe, the best overall performance is 86.2% reached by using the top 12 and 13 ranked features, having 25 correctly inferred emergent leaders, out of the total of 29 leaders in the corpus. For groups with three participants, the best performance is already reached starting with the top 8 features. For the case of four participants, more features are needed to achieve the highest accuracy (12 features).

For the perceived dominant person in the group (Fig. 7.2 (b)), the top 10 and 11 ranked features provide the best achieved performance, namely 69.0%, i.e., 20 dominant participants are correctly inferred. As we can observe for groups with four participants, one single feature provides 50% accuracy (the category *WC*), and the best performance is reached with the top 10 and 11 ranked features. For the case of groups with three participants, more features are needed, starting from top 14 on, to get the best performance, inferring accurately all the 9 perceived dominant participants in the three person groups.

For the perceived most competent person, the best accuracy is 48.3% with the top 16 features, and we can observe the correct inferences are mostly done on the performance with groups with four participants. For the most likeable person we reach only about

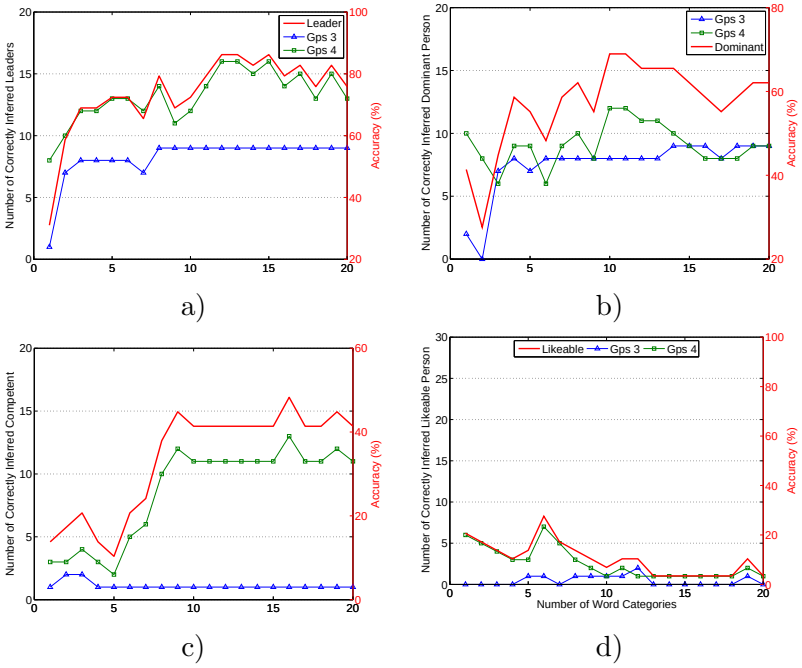


Figure 7.2: Accuracy of Perceived variables in the ELEA-EN corpus, using categories extracted from the manual transcriptions. a) PLead, b) PDom, c) PCom and d) PLike.

27.6% accuracy performance with top 6 word categories, which is just above chance. For the inferences in groups with three participants, given that maximum two most competent and one most agreeable person are accurately inferred, we hypothesize that the algorithm could provide a higher posterior probability to the emergent leader, and estimate as second high posterior probability at the actual perceived most competent and most agreeable person respectively.

7.3.4 Inference with Keywords

The results in this section show performance for the resulting top 20 features from the SVM-RFE and the automatically spotted keywords.

Figure 7.3 (a) shows the performance of the perceived emergent leader inference method, using the most accurate detected keywords. As we can observe, the best performance is 82.8%, reached by using the top 13 ranked features, having 24 correctly inferred emergent leaders, out of 29 in the corpus. Interestingly, this performance is only slightly lower than the one obtained with manual transcriptions. For groups with

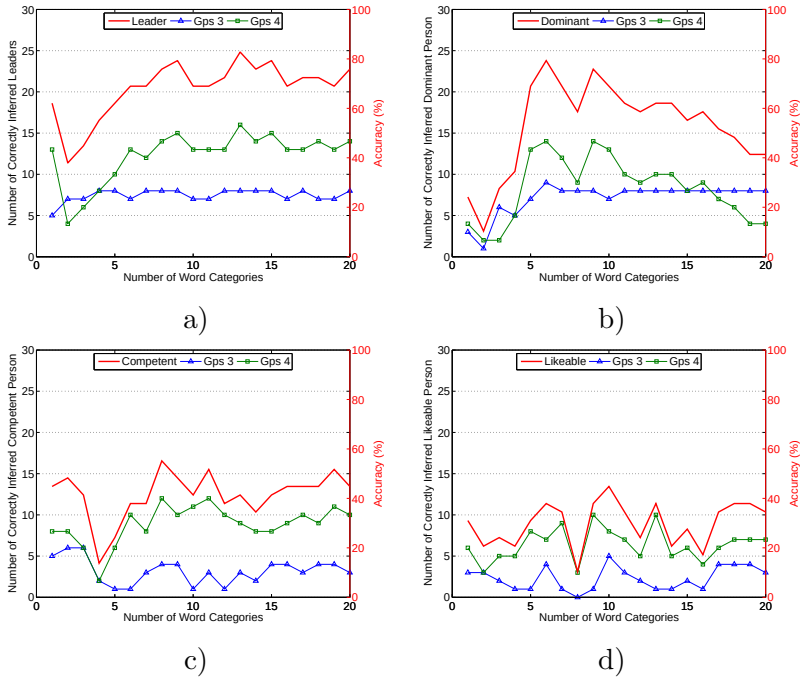


Figure 7.3: Accuracy of Perceived variables in the ELEA-EN corpus, using categories extracted from the keyword spotter. a) PLead, b) PDom, c) PCom and d) PLike.

three participants, the top 4-5 features provide the same performance as using the top 12 to 15. For groups with four participants, although the performance is relatively high with only one feature (*WC*, 13 out of 20 perceived leaders), the top 13 features are needed to reach the highest performance (16 out of 20).

For the perceived dominant person in the group, the best performance is 79.3% by considering the top 6 categories, which results in 23 correctly inferred most dominant participants out of the 29. Somewhat surprisingly, the performance using keywords is higher than the one obtained with manual transcriptions. This result could be due to the fact that the keyword spotter might be overestimating words from categories that are highly correlated with dominance, e.g. *negemo* (negative emotions). We plan to explore the hypothesis that the keyword spotter could overestimate words from specific word categories as part of future work.

For PCom, the best accuracy is higher than random performance, up to 55.2% with the top 8 features. For groups with three participants the highest performance is reached with top 2 and 3 features. For the case of PLike, the best performance

is reached with top 10 word categories with up to 44.3% accuracy, also better than random performance.

7.4 Discussion

Figure 7.4 shows the highest accuracy using word categories from both manual transcriptions and automatic keyword spotter, the category WC, and one nonverbal feature (TSL: total speaking time). The amount of speaking time (TSL) is computed in the ELEA-EN corpus as described in Chapter 4. We use the unsupervised rule-based estimator defined in previous chapters (see Chapter 4), to infer the emergent leader and related concepts using the individual measures of speaking TSL, and the category *WC* extracted from manual transcriptions and the keyword spotter.

We again performed a standard binomial significance test to validate the differences in performance among methods. As we can observe for PLead, using only the category *WC* from manual transcriptions results in lower performance as compared to using TSL (marginally significant, $p = 0.07$). Although similar performance could be expected between *WC* from manual transcriptions and TSL, there are differences in the nature of the feature extraction, TSL captures continuous speaking status and *WC* only counts instances of words i.e., while WC count as one more instance each of the words "ok" and "misunderstanding", TSL accumulates each word as the amount of time used to express each word.

Although the highest accuracy is reached with the top word categories extracted from the manual transcription (86.2%), in comparison with inferences using the keyword spotter (82.8%) there is no statistically significant improvement. In addition, even though manual transcription or keywords appear slightly better than TSL, we did not find significant evidence to validate a better performance ($p = 0.12$ and $p = 0.19$ respectively). Moreover, the performance of top word categories from the manual transcriptions over *WC* from manual transcriptions and keywords, was found statistically significantly better ($p = 0.0003$ and $p = 0.01$ respectively). Overall, all methods were found statistically significantly better than random performance ($p < 0.009$). Our methods can predict the emergent leader within 72.4 and 86.2% accuracy.

For PDom, there is no statistically significant difference between *WC* from the manual transcription, *WC* from the keyword spotter and TSL. Nevertheless, we can

observe in Figure 7.4 that keyword spotter shows the highest accuracy (79.3%), which is statistically better than *WC* from the keyword spotter ($p = 0.0005$) and than *WC* from manual transcriptions ($p = 0.03$). All methods were found statistically significantly better than random performance ($p < 0.02$). Although is has been shown that TSL is a good predictor of dominance (50, 71), our findings suggest that some top categories could also provide relevant information, by using only partial content of the speech.

For PCom, TSL has the higher accuracy (68.9%), statistically better compared with top word categories from manual transcriptions ($p = 0.04$), *WC* from manual transcriptions and *WC* from keywords ($p = 0.01$). In addition, the performance observed for PCom is higher compared with the performance reported in previous chapters. However, TSL is not significantly higher compared with categories from the keyword spotter ($p = 0.13$). TSL, word categories from the keyword spotter, and *WC* (manual transcriptions and keywords) were found statistically higher compared with random performance ($p < 0.04$). Thus, the most competent person can be inferred within 48.3 and 68.9% accuracy. Moreover, TSL accuracy shows consistency with previous literature in competence (4), that states that in order to appear competent, people speak the most in order to get the control over the group decisions. The findings suggest that the nonverbal feature captured with TSL could be augmented using features derived from the verbal channel, specifically the language style. Furthermore, the low performance of the verbal content is due to the low performance for the groups with three participants, which counts for 30% of the accuracy.

Finally, for PLike, the highest performance is reached using top categories extracted with the keyword spotter (44.3%). There is a statistically marginal improvement in performance of keyword spotter over TSL ($p = 0.07$). Only categories from the keyword spotter were found marginally better than random performance ($p = 0.07$).

7.5 Conclusions

In this Chapter, we conducted a study on the language style that characterizes people being perceived as emergent leaders and dominant individuals. For this purpose, word categories were extracted from manual transcriptions and an automatic keyword spotter using 7 hours of recorded interactions from the ELEA-EN corpus. By using the top

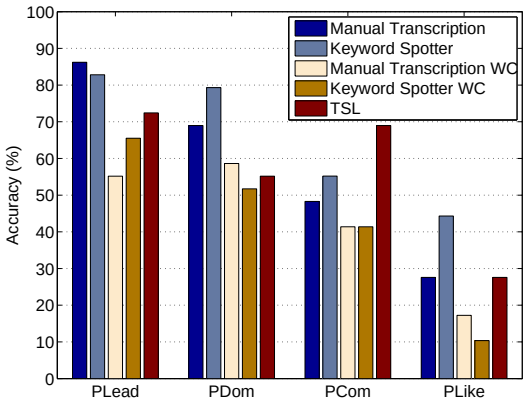


Figure 7.4: Best accuracy from perceived variables in the ELEA-EN corpus. Inferences using manual transcriptions, keyword spotter, the category *WC* (from manual transcriptions and the keyword spotter) and amount of speaking time (TSL).

relevant word categories and a supervised method, we inferred the emergent leader and related concepts in the group.

Our analysis on the relevance of categories shows that perceived leadership uses words related to *motion (to walk or to stay), space (map, far, close, etc)* and *basic needs (eat, sleep, etc)*. These categories are clearly scenario-dependent, and suggest that the emergent leader gets involved in the scenario, in this case the decision making process of how to survive an airplane crash in winter. Although these categories are linked to the context, other relevant categories like *WC, conj, excl, and assent* (i.e., context-free) are also informative in order to discriminate the emergent leader.

While the *WC* feature (word count) does not exactly reflect language style, it provides relevant information in the perception of leadership and dominance. The *WC* category, is as a feature, correlated with the amount of speaking time. In that sense, our findings are consistent with previous work by others, and our previous chapters: the amount of speaking time is highly correlated with dominance (50, 71) and the perception of a leader (99, 100, 104).

Despite the fact that the keyword spotter can not recover exactly the top 20 relevant categories for leadership extracted from the manual transcription, it captures a significant number of the top categories on a one-by-one-base (7 categories out of 20). Although the performance of the keyword spotter regarding inference of the emergent leader seems promising (82.8%), we can overall infer the perceived leader within 72.4 and 86.2% accuracy, using either speaking cues or word categories (derived from man-

ual transcriptions or keywords). For the case of perceived dominance, our framework can achieve within 69 and 79.3% accuracy using language style, i.e., word categories derived from manual transcriptions or automatically detected keywords. It is worth to mention that the performance of the keyword spotter based-inference is specially interesting, considering that the audio track is challenging due to an open vocabulary scenario and many cross-talk segments. Finally, a keyword spotting system can be implemented in a realistic scenario (i.e., as an automated and fast process).

We conclude this chapter by pointing out several future research lines. First, the verbal content has been analyzed considering only English spoken meetings, and therefore we can not generalize our findings in the perception of emergent leadership to other spoken languages. Other datasets would be needed to confirm our findings. Second, nonverbal features could be used as additional source of information, more specifically to augment information derived from categories that capture positive or negative emotions. We did not conduct experiments that integrate verbal and nonverbal features, but this is clearly a direction of future work. Finally, we need to investigate in depth the seemingly counter-intuitive result for the most agreeable person, in which automatic keywords might produce better performance than manual annotations.

Chapter 8

Conclusion

In this Chapter we present the conclusions of this dissertation, centered on computational methods to analyze emergence of leadership in small groups using audio-visual cues. We further point out the limitations of our work, followed by a proposal for future research lines.

In this thesis we present a behavioral framework approach to infer emergent leaders in face-to-face, small group interactions. For our study, we first proposed a portable recording setup to collect a new audio-visual corpus for analysis of emergent leadership in teams. We automatically extracted audio-visual behavioral cues, and we proposed the use of a variety computational approaches to infer the emergent leader in a group of previously unacquainted people.

In Chapter 2, we discussed the related work in the emerging leadership context, from the points of view of social psychology and social computing. We also described existing corpora recorded with the aim of analyzing small group interactions and concluded that our work would fill an existing research gap.

In Chapter 3 we described the design and collection of a new audio-visual group interaction corpus for the study of emergence of leadership. The collected corpus, called ELEA, contains 40 interactions and approximately 10 hours of audio and video streaming. Groups of unacquainted people engaged in a decision-making task that allows an open discussion interaction. The annotations of the ELEA corpus include self-reported personality, concepts related to perceived leadership from the participants in the groups and from external observers, and the participants' actual performance in the survival task.

In Chapter 4, we proposed a computational framework to infer emergent leadership in newly formed groups by combining automatically extracted speaking turn and prosodic features. We investigated the effectiveness of individual and combined audio features in order to infer the emergent leader using three machine learning approaches. Our results showed that the emergent leader is perceived as a person who talks and interrupts the most, and varies the tone of voice. The most informative turn-taking cue is the amount of interruptions to grab the floor, which captures the decision-making scenario and it is perhaps influenced for the limited time to come up with a group decision. In addition, the individual performance in the survival task has an slight effect in the perception of competence and dominance. We found connections between being perceived as leader and being perceived as dominant, by both the participants in the group and external observers. The emergent leader in the group can be inferred with supervised and unsupervised approaches, within 63.7 and 72.5% accuracy. The Emergent leadership inference using acoustic nonverbal cues is statistically better compared with 27.5% random performance ($p < 0.0001$). However, no statistically significant difference was found among Rule-based Estimator, Rank-level Fusion, SVM and CC. Although, it is important to investigate whether using information beyond the acoustic channel to predict emergent leadership, a fully automated inference of the emergent leader could be trusted without compromising privacy sensitive content, i.e. using acoustic nonverbal cues.

In Chapter 5, we investigated the effect of audio and visual nonverbal cues in the identification of the emergent leader in the group. In addition to the audio features from Chapter 4, we automatically extracted visual activity features to characterize individual participants in terms of statistical measures on head and body activity, and motion energy. Our analysis showed that the emergent leader was perceived as a visually active and talkative person. The augmentation with body activity is explained by the nature of the interaction, since there is a natural emergence of movements and gestures that complement the speech. The head activity, aside from the effect of movements due to the speaking activity, might also be due to agreement/disagreement gestures while listening. Although in some cases adding visual information to the existing acoustic cues increased the performance, we did not find enough evidence to confirm that indeed visual activity improves the inference of the emergent leader in the group. Overall, we can infer the emergent leader within 70.4 and 85.7% accuracy, using

either single acoustic cues or combined audio and visual cues. The proposed methods Rule-based Estimator, Rank-level Fusion and CC, were found to be statistically better than random performance ($p < 0.0002$). However, we found no statistically significant difference in performance among them.

In Chapter 6, we presented a synchronous multimodal approach to infer emergent leadership. For the multimodal analysis we automatically extracted visual attention features, and audio-visual features that combine speaking activity and attention. The visual attention features characterized attention received and given during the interaction. The combined speaking activity and attention features include classic measures in social psychology like the visual dominance ratio. Our correlation analysis revealed a connection between the amount of received attention and being perceived as a leader, for the perception of the most dominant person. Similarly, the amount of attention received from the team while speaking was found to be an informative cue for emergent leadership inference. On average we can infer the emergent leader within 59.1 and 72.7% accuracy using either multimodal cues or single speaking and attention cues that are statistically better than random performance ($p < 0.01$). Nevertheless, we found no statistically significant difference in accuracy performance among speaking time, attention and multimodal cues. Although social attention features capture relevant information, they did not outperform estimations from simple audio features using unsupervised methods. This raises a practical question about the justification for the additional computational cost involved in extracting visual attention cues.

Finally, in Chapter 7 we investigated the connection between language style and emergent leadership. We used a framework for language style analysis, which involves manual transcriptions, an automatic keyword spotter and a state-of-the-art content analysis module. Our analysis showed that the emergent leader gets involved in the interactions, and uses a significant amount of word categories related to the scenario and task. The analysis also suggests that, context-free and privacy-preserving word categories like conjunctions (e.g., also, although, then, etc.) can be used to accurately infer the emergent leader in the group. Although the performance of the keyword spotter is not very high, relatively accurate inferences of leadership (82.8%) could be achieved. Overall, we can infer the perceived leader in group within 72.4 and 86.2% accuracy, using either speaking cues or word categories (derived from manual transcriptions or keywords). Although the performance is statistically significantly better than random

performance ($p < 0.009$), we found no statistical difference among speaking time and word categories derived from both manual transcriptions and keywords.

8.1 Limitations

There are some limitations in our work. First, and foremost the corpus is small (N=40), despite our best efforts to collect data. This has to do with the requirement of having to engage only people who do not know each other, and shows the difficulty of collecting face-to-face data even with portable sensors that can be easily deployed for recordings. We made the decision to consider all the groups, assuming that the emergent leader in the group is the person with the highest score, even though there were three teams in which there was not a clear emergent leader (in terms of relatively similar scores). Furthermore, we were unable to use all data points for all modalities, given some technical failures of the recording system and some human choices (like recording groups in two languages). The size of the corpus puts limits on the statistical confidence of some of the results, as discussed at length in the previous chapters.

Second, regarding the automatic extraction of features, the visual attention features were found to have a performance that is not very high. This is clearly an important issue, as they involve significant computation. An alternative for the study would have been the use of manually annotated visual attention but this involves significant amount of time to produce annotations per video-frame, per participant, and per group.

Third, clearly other well known machine learning methods could have been used for the group and the individual based analysis, but that is considered as part of future work and expanded in the next section.

Fourth, although the performance of the framework proposed using a keyword spotter is relatively high, the open vocabulary scenario and missing words due to many cross-talk segments, would require further testing on a different problem-solving task to validate our findings.

A fifth limitation of the work includes generalization; although the data was collected from people with a variety of backgrounds, larger and more diverse populations could be used in future work to validate our findings in further cases.

8.2 Future Work

The analysis on emergent leadership on the ELEA corpus can be extended in several directions. From the individual point of view, other nonverbal features could be extracted and coded to complement the visual cues presented in this thesis. It would be interesting to explore the impact of emotional states of the participants on the perception of leadership. The emotional states of the participants could also be extracted from several channels. For example, on the visual channel, by using the facial action coding system, facial action units could serve as source of information of disgust or contempt. The amount of stress per participant, could also be extracted by using voice stress analysis on the acoustic channel.

The labeling of functional roles within the interaction could also be coded, in order to identify the producer of ideas, the seeker, the giver, etc, as proposed in (27). It would be interesting to identify which specific roles, either related to the task or focused in the socio-emotional aspects, are predictive of emergent leadership, competence or dominance in zero-acquaintance groups. In addition, several emerging social interactions in small groups, such as involvement, trust or control, which are known to be informative for emergent leadership (60, 61, 114), could also be studied. Furthermore, it would be interesting to track an emergent leader in a long term interaction, i.e., once the leader has emerged in a first encounter, follow the interactions of a group on a week or month based, and analyze the changes in behavior, as well as in the perception from the participants.

As pointed in the limitations section, the language style presented uses only English conversations. The framework proposed could be used in different language scenarios, to analyze the relevance on non-contextual language style. By analyzing interactions on different language scenarios scenarios, we could investigate if the emergent leadership' relevant context-free word categories reported in this dissertation, could be also found relevant for leadership in new scenarios. This would reveal if there exist a consistent pattern of verbal context-free emergent leadership behavior in zero-acquaintance groups.

We also assume that an approach considering the integration of verbal and nonverbal information could improve the accuracy performance reported in this thesis.

Taking into account that in spontaneous face-to-face interactions, we use all the available information (e.g., verbal, visual and acoustic) to create good or bad impressions of people. It would be interesting to integrate all the automatic extracted nonverbal cues presented in this thesis (i.e., speaking turns, prosody, visual activity, motion, attention and verbal content), and use a supervised learning algorithm to predict emergent leadership and related concepts.

Another dimension of future work would be to study the personality of the participants as an influence factor during the interaction and its influence on the perception of the leader in the group. First, by considering apriori participants self-reported personality and the extracted nonverbal features, a machine learning method could predict the emergent leader and related concepts, either Hidden Markov Models or Bayesian Networks could be used as a starting point. A Markov model approach would use the personality as apriori information, the nonverbal behavior as observable data, and the emergent leader and related concepts as the hidden states.

Finally, other machine learning techniques could be explored to model the small group interactions and emergent leadership communicative behaviors. For example, the nonverbal behavior information could feed an influence model (9, 15), such that the model would learn the interacting patterns within a group in order to form an estimation network that collectively infer emergent leadership and related concepts. Also, probabilistic topic models that have been used to model small group interactions (49), could be used to generate a bag of audio-visual nonverbal features and in addition a bag of word categories, that model individual verbal and nonverbal emergent leadership behavior.

References

[1] Liwc inc. http://www.liwc.net/index.php. 92, 96

[2] The Microcone website, 2011. 22, 35

[3] N. Ambady, M. Hallahan, and R. Rosenthal. On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, **69**:518–529, 1995. 1

[4] C. Anderson and G. J. Kilduff. Why do dominant personalities attain influence in face-to-face groups? the competence- signaling effects of trait dominance. *Journal of Personality and Social Psychology*, **96**(2):491–503, 2009. 1, 12, 45, 110

[5] O. Aran and D. Gatica-Perez. Fusing audio visual nonverbal cues to detect dominance in small group conversations. In *ICPR*, Aug 2010. 15, 38, 68

[6] Oya Aran. *Vision Based Sign Language Recognition: Modeling and Recognizing Isolated Signs With Manual and Non-manual Components*. PhD thesis, Bogazici University, Istanbul, Turkey, 2008. 59

[7] Oya Aran and Daniel Gatica-Perez. Analysis of group conversations: Modeling social verticality. In A. A. Salah and T. Gevers, editors, *Computer Analysis of Human Behavior*, Lecture Notes in Computer Science. Springer, 2011. 16

[8] Oya Aran, Hayley Hung, and Daniel Gatica-Perez. A multimodal corpus for studying dominance in small group conversations. In *Workshop International Conference on Language Resources and Evaluation*, LREC, 2010. 17, 18

[9] C. Asavathiratham, S. Roy, B. Lesieutre, and G. Verghese. The influence model. *Control Systems, IEEE*, **21**(6):52–64, 2001. 16, 118

[10] S.O. Ba and J.M. Odobez. Recognizing visual focus of attention from head pose in natural meetings. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **39**(1):16–33, 2009. 79

[11] S.O. Ba and J.M. Odobez. Multi-person visual focus of attention from head pose and meeting contextual cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **33**(1):101–116, 2011. 79

[12] J. E. Baird. Some non-verbal elements of leadership emergence. *Southern Speech Communication Journal*, **42**(4):352–361, 1977. 1, 12, 37, 65

[13] R.F. Bales and F.L. Strodtbeck. Phases in group problem-solving. *Journal of Abnormal and Social Psychology*, **46**:485–495, 1951. 1, 75

[14] B. M. Bass. *Bass and Stogdill's handbook of leadership. Theory, research, and managerial applications*. Free Press, 1990. 1

[15] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Towards measuring human interactions in conversational settings. In *IEEE CVPR-CUES*, Dec 2001. 2, 16, 118

[16] Kenneth D. Benne and Paul Sheats. Functional roles of group members. *Journal of Social Issues*, **4**(2):41–49, 1948. 13

[17] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**:257–267, 2001. 62

[18] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 59

[19] S. Burger, V. MacLaren, and H. Yu. The isl meeting corpus: the impact of meeting type on speech style. In *International Conference on Spoken Language Processing*, Interspeech-ICSLP, 2002. 17, 18

[20] N. Campbell, T. Sadanobu, M. Imura, N. Iwahashi, S. Noriko, and D. Douxchamps. A multimedia database of meeting and informal interactions for tracking participant involvement and discourse flow. In *Workshop International Conference on Language Resources and Evaluation*, LREC, 2006. 17, 18

[21] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, V. Karaiskos J. Kadlec, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus: A pre-announcement. In *Workshop on Machine Learning and Multimodal Interaction*, ICMI-MLMI, 2005. 17

[22] T. A. Carte, L. Chidambaram, and A. Becker. Emergent leadership in self-managed virtual teams. *Group decision and negociation*, **15**:323–343, 2006. 14, 15

[23] L. Chen, T. R. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, and T. Huang. Vace multimodal meeting corpus. In *Workshop on Machine Learning and Multimodal Interaction*, ICMI-MLMI, 2005. 17, 18

[24] Cindy Chung and James Pennebaker. The psychological functions of function words. In *Social communication*, pages 343–359. Psychology Press, 2007. 91, 95

[25] Mark Cook and Jacqueline M. C. Smith. The role of gaze in impression formation. *British Journal of Social and Clinical Psychology*, **14**(1):19–25, 1975. 79

[26] P. Costa and R. McCrae. *NEO PI-R profesional manual*. 1992. 24

[27] Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. Using the influence model to recognize functional roles in meetings. In *International Conference on Multimodal Interfaces*, ICMI '07, pages 271–278, 2007. 16, 117

[28] J.F. Dovidio and S.L. Ellyson. 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, 1982. 13, 79

[29] N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, **22**(2):207–233, 2005. 1, 29, 65

[30] J.S. Efran. Looking for approval: effects of visual behavior of approbation from persons differing in importance. *Journal of Personality and Social Psychology*, **10**(1):21–25, 1968. 1, 75, 79

[31] G. Evermann and P.C. Woodland. Large vocabulary decoding and confidence estimation using word phoneme accuracy posterior probabilities. In *International Conference on Acoustics, Speech and Signal Processing*, **3** of *ICASSP*, pages 2366–2369, Jun 2000. 94

[32] I. Garofolo, M. Michel, C. Laprun, V. Stanford, and E. Tabassi. The nist meeting room pilot. In *International Conference on Language Resources and Evaluation*, LREC, 2004. 17

[33] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: a review. *Image and Vision Computing*, **1**(12), Dec 2009. 16, 45, 89

[34] Peter A. Gloor and Yan Zhao. Analyzing actors and their discussion topics by semantic social network analysis. *Information Visualisation, International Conference on*, **0**:130–135, 2006. 39

[35] Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, **333**(6051):1878–1881, 2011. 91

[36] L. D. Goodstein and R. I. Lanyon. Applications of personality assessment to the workplace: a review. *Journal of Business and Psychology*, **13**(3), 1999. 1

[37] Jr. A. Gray and J. Markel. A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, **22**(3):207–217, Jun 1974. 36

[38] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**(1):389–422, mar 2002. 92, 96

[39] Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Vincent Wan, Martin Karafiat, Jithendra Vepa, and Mike Lincoln. The ami system for the transcription of speech in meetings. In *International Conference on Acoustics, Speech and Signal Processing*, **4** of *ICASSP*, pages 357–360, Apr 2007. 94

[40] J. A. Hall, E. J. Coats, and L. Smith. Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological bulletin*, **131**(6):898–924, 2005. 1, 13, 15, 77, 88

[41] J.A. Harrigan. Proxemics, kinesics, and gaze. *The new handbook of methods in nonverbal behavior research*, pages 137–198, 2005. 13, 77, 88

[42] Hayley Hung, Yan Huang, Gerald Friedland, and Daniel Gatica-Perez. Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**(4):847–860, may 2011. 15

[43] Hayley Hung, Dinesh Babu Jayagopi, Sileye Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *International Conference on Multimodal Interfaces*, ICMI, pages 233–236, 2008. 79, 83, 87, 88

[44] Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. Language style matching predicts relationship initiation and stability. *Psychological Science*, **22**(1):39–44, Jan 2011. 14, 91

[45] Michael Isard and Andrew Blake. Contour tracking by stochastic propagation of conditional density. In Bernard Buxton and Roberto Cipolla, editors, *Computer Vision ECCV '96*, **1064** of *Lecture Notes in Computer Science*, pages 343–356. Springer Berlin / Heidelberg, 1996. 59

[46] D. N. Jackson. *Personality research form manual*. Research Psychologists Press, 1967. 25

[47] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The icsi

meeting corpus. In *International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, 2003. 17, 18

[48] D. Jayagopi and D. Gatica-Perez. Discovering group nonverbal conversational patterns with topics. In *International Conference on Multimodal Interfaces (ICMI)*, Nov 2009. 15

[49] D. Jayagopi and D. Gatica-Perez. Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia*, **12**(8):790–802, Dec 2010. 15, 118

[50] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on audio, speech and language processing*, **17**(3), Mar 2009. 15, 25, 38, 41, 48, 68, 89, 110, 111

[51] D. Jayagopi, B. Raducanu, and D. Gatica-Perez. Characterizing conversational group dynamics using nonverbal behavior. In *ICME*, Jun 2009. 2

[52] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez. Linking speaking and looking behavior patterns with group composition, perception, and performance. In *International Conference on Multimodal Interfaces (ICMI)*, Oct 2012. 88

[53] David Jensen, Jennifer Neville, and Brian Gallagher. Why collective inference improves relational classification. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, August 2004. 39

[54] Matthew Jensen, Thomas Meservy, Judee Burgoon, and Jay Nunamaker. Automatic, multimodal evaluation of human interaction. *Group Decision and Negotiation*, **19**:367–389, 2010. 91

[55] Oliver P. John and Lawrence A. Pervin. The big five factor taxonomy: Dimensions of personality in the natural language and in questionnaires. *Handbook of personality: Theory and research*, pages 66–100, 1990. 1

[56] D. W. Johnson and F. P. Johnson. *Joining together: Group theory and group skills*. Boston: Allyn and Bacon, 1994. 24

[57] N. Jovanovic, R. op den Akke, and A. Nijholt. A corpus for studying addressing behavior in multi-party dialogues. In *The sixth SigDial conference on Discourse and Dialogue*, 2005. 17, 18

[58] Kyriaki Kalimeri, Bruno Lepri, Oya Aran, Dinesh Babu Jayagopi, Daniel Gatica-Perez, and Fabio Pianesi. Modeling dominance effects on nonverbal behaviors using granger causality. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 23–26, 2012. 16

[59] Kyriaki Kalimeri, Bruno Lepri, Taemie Kim, Fabio Pianesi, and AlexSandy Pentland. Automatic modeling of dominance effects using granger causality. In *Human Behavior Understanding*, **7065** of *Lecture Notes in Computer Science*, pages 124–133. Springer Berlin Heidelberg, 2011. 16

[60] A. K. Kalma, L. Visser, and A. Peeters. Sociable and aggressive dominance: Personality differences in leadership style? *Leadership Quarterly*, **4**(1):45–64, 1993. 1, 13, 44, 117

[61] J. Kickul and G. Neuman. Emergent leadership behaviours: The function of personality and cognitive ability in determining teamwork performance and ksas. *Journal of Business and Psychology*, **15**(1), 2000. 1, 13, 24, 117

[62] T. Kim, A. Chang, L. Holland, and A. Pentland. Meeting mediator: enhancing group collaboration with sociometric feedback. In *Conference on CSCW*, pages 457–466, 2008. 17, 18

[63] M. L. Knapp and J. A. Hall. *Nonverbal Communication in Human Interaction*. Wadsworth, Cengage Learning, 2008. 10, 76

[64] J. M. Kouzes and B. Z. Posner. *The leadership challenge: How to get extraordinary things done in organizations*. Jossey-Bass, 1987. 1

[65] A. Leffer, D. L. Gillespie, and J. C. Conaty. The effects of status differentiation on nonverbal behavior. *Social Psychology Quaterly*, **45**(3):153–161, Sep 1982. 13

[66] R. G. Lord, J. S. Phillips, and M. C. Rush. Effects of sex and personality on perceptions of emergent leadership, influence, and social power. *Journal of applied psychology*, **65**(2):176–182, 1980. 1

[67] Richard Lowry. Concepts and applications of inferential statistics, 1998. 52

[68] A. Madan, K. Farrahi, D. Gatica-Perez, and A. Pentland. Pervasive sensing to model political opinions in face-to-face networks. In *Pervasive*, 2011. 39

[69] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, **30**:457–501, 2007. 14, 91

[70] Nadia Mana, Bruno Lepri, Paul Chippendale, Alessandro Cappelletti, Fabio Pianesi, Piergiorgio Svaizer, and Massimo Zancanaro. Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. In *Workshop on Tagging, mining and retrieval of human related activity information*, TMR, 2007. 17, 19

[71] M. Schmid Mast. Dominance as expressed and inferred through speaking time: A meta-analysis. *Human Communication research*, **28**(3):420–450, 2002. 1, 13, 15, 37, 45, 89, 110, 111

[72] I. McCowan, M.H. Krishna, D. Gatica-Perez, D. Moore, and S. Ba. Speech acquisition in meetings with an audio-visual sensor array. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1382 – 1385, july 2005. 35

[73] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(3):305–317, 2005. 15, 17

[74] L. K. McDowell, K. M. Gupta, and D. W. Aha. Cautious collective classification. *Journal of Machine Learning Research*, **10**:2777–2836, 2009. 39, 40, 70

[75] Darren C. Moore. The idiap smart meeting room. Technical report, Idiap Research Institute, 2002. 22

[76] Petr Motlicek, Fabio Valente, and Philip N. Garner. English spoken term detection in multilingual recordings. In *Interspeech*, Sep 2010. 95

[77] Jennifer Neville, David Jensen, and Brian Gallagher. Simple estimators for relational bayesian classifiers. *Data Mining, IEEE International Conference on*, **0**:609, 2003. 39

[78] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase. Probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns and head directions and utterances. In *International Conference on Multimodal Interfaces*, ICMI, 2005. 16, 17, 19, 79

[79] Kazuhiro Otsuka, Junji Yamato, Yoshinao Takemae, and Hiroshi Murase. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *CHI '06 extended abstracts on Human factors in computing systems*, CHI EA '06, pages 1175–1180, New York, NY, USA, 2006. ACM. 79

[80] D. Paul and J. Baker. The design for the wall street journal-based csr corpus. In *Proc. of the DARPA SLS Workshop*, February 1992. 94

[81] James Pennebaker and Laura King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**(6):1296–1312, 1999. 14, 91

[82] J.W. Pennebaker, M.E. Francis, and R.J. Booth. *Linguistic Inquiry and Word Count: LIWC2001*. Mahwah, NJ: Erlbaum Publishers, 2001. 95

[83] A. Pentland. *Honest signals: how they shape our world*. MIT Press, 2008. 89

[84] F. Pianesi, N. Mana, and A. Cappelletti. Multimodal recognition of personality traits in social interactions. In *International Conference on Multimodal Interfaces (ICMI)*, Oct 2008. 2, 15

[85] Fabio Pianesi, Massimo Zancanaro, Bruno Lepri, and Alessandro Cappelletti. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, **41**:409–429, 2007. 15, 17, 19

[86] Fabio Pianesi, Massimo Zancanaro, Elena Not, Chiara Leonardi, Vera Falcon, and Bruno Lepri. Multimodal support to group dynamics. *Personal Ubiquitous Computing*, **12**(3):181–195, 2008. 15

[87] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999. 97

[88] M. S. Poole, Andrea B. Holligshead, Joseph E. McGrath, Richard L. Moreland, and John Rohrbaugh. Interdisciplinary perspectives on small groups. *Small Group Research*, **35**(1):3–16, 2004. 1

[89] B. Raducanu, Jordi Vitria, and D. Gatica-Perez. You are fired! nonverbal role analysis in competitive meetings. In *ICASSP*, Apr 2009. 2, 15

[90] E. Ricci and J.M. Odobez. Learning large margin likelihoods for realtime head pose tracking. In *International Conference on Image Processing*, ICIP, 2009. 77

[91] R. Rienks and D. Heylen. Automatic dominance detection in meetings using easily detectable features. In *Workshop on Machine Learning for Multimodal Interaction (MLMI)*, 2005. 48

[92] Rutger Rienks and Dirk Heylen. Dominance detection in meetings using easily obtainable features. In *In Bourlard, H., and Renals, S. (Eds.), Revised Selected Papers of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, pages 76–86. Springer Verlag, 2005. 15

[93] Rutger Rienks, Dong Zhang, Daniel Gatica-Perez, and Wifried Post. Detection and application of influence rankings in small group meetings. In *International Conference on Multimodal Interfaces*, ICMI, 2006. 17, 18

[94] H. Salamin, S. Favre, and A. Vinciarelli. Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, **11**(7):1373–1380, 2009. 15

[95] Hugues Salamin and Alessandro Vinciarelli. Automatic role recognition in multiparty conversations: An approach based on turn organization, prosody, and conditional random fields. *IEEE Transactions on Multimedia*, **14**(2):338–345, 2012. 15

[96] E. Salas, D. E. Sims, and C. S. Burke. Is there a big five in teamwork. **36**(5):555–599, Oct 2005. 1

[97] D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez. An audio visual corpus for emergent leader analysis. In *Workshop on Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future*, ICM-MLMI, Nov 2011. 21

[98] D. Sanchez-Cortes, O. Aran, D. Jayagopi, M. Schmid Mast, and D. Gatica-Perez. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 2012. 3, 4, 9, 21, 75

[99] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez. Identifying emergent leadership in small groups using nonverbal communicative cues. In *International Conference on Multimodal Interfaces (ICMI)*, Nov 2010. 4, 34, 111

[100] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, **14**(3):816–832, 2012. 3, 4, 9, 21, 34, 58, 83, 89, 111

[101] D. Sanchez-Cortes, D. Jayagopi, and D. Gatica-Perez. Predicting remote versus collocated group interactions using nonverbal cues. In *ICMI-MLMI'09: Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*, Nov 2009. 15

[102] D. Sanchez-Cortes, P. Motlicek, and D. Gatica-Perez. Assessing the impact of language style on emergent leadership perception from ubiquitous audio. In *Conference on Mobile and Ubiquitous Multimedia*, MUM, Dec 2012. 4, 92

[103] Richard B. Slatcher, Cindy K. Chung, James W. Pennebaker, and Lori D. Stone. Winning words: Individual differences in linguistic style among u.s. presidential and vice presidential candidates. *Journal of Research in Personality*, **41**(1):63 – 75, 2007. 14, 91

[104] R. T. Stein. Identifying emergent leaders from verbal and nonverbal communications. *Personality and Social Psychology*, **32**(1):125–135, 1975. 1, 12, 21, 111

[105] R. Timothy Stein and Tamar Heller. An empirical analysis of the correlations between leadership status and participation rates reported in the literature. *Journal of Personality and Social Psychology*, **37**(11):1993–2002, 1979. 1, 11, 37

[106] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *CHI'02 Extended abstracts on Human factors in computing systems*, CHI EA '02, 2002. 77

[107] R. Subramanian, J. Staiano, K. Kalimeri, N. Sebe, and F. Pianesi. Putting the pieces together: Multimodal analysis in social attention in meetings. In *ACM Multimedia*, MM, 2010. 79, 81

[108] David Talkin. A robust algorithm for pitch tracking (rapt). In *In Speech Coding and Synthesis*, pages 495–518. Elsevier Science, 1995. 36

[109] Punnarumol Temdee, Bundit Thipakorn, Booncharoen Sirinaovakul, and Heidi Schelhowe. Of collaborative learning team: An approach for emergent leadership roles identification by using social network analysis. In *Technologies for E-Learning and Digital Entertainment*, **3942** of *Lecture Notes in Computer Science*, pages 745–754. 2006. 14, 15

[110] Fabio Valente, Samuel Kim, and Petr Motlicek. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *Proceedings of Interspeech 2012*, 2012. 15

[111] G. Varni, G. Volpe, and A. Camurri. A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. *IEEE Transactions on Multimedia*, **12**(6):576–590, 2010. 15

[112] D. K. Wentworth and L. R. Anderson. Emergent leadership as a function of sex and task type. *Sex Roles*, **11**(5/6):513–524, 1984. 13, 45

[113] M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *International Conference on Multimodal Interfaces (ICMI)*, Nov 2006. 15

[114] Xiaomeng Zhang and Kathryn M. Bartol. Linking empowering leadership and employee creativity: The influence of psychological empowerment, intrinsic motivation, and creative process engagement. *Academy of Management Journal*, **53**(1):107–128, 2010. 117