Automated analysis of cohesion in small groups interactions

Lucien Maman

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAT030

Thèse de doctorat

Automated Analysis of Cohesion in Small Groups Interactions

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Palaiseau, le 15 Septembre 2022, par

LUCIEN MAMAN

Composition du Jury :

Catherine Pelachaud CNRS Research Director, Sorbonne University (ISIR)	Présidente
Albert Ali Salah Professor, Utrecht University (ICS)	Rapporteur
Dominique Vaufreydaz Associate Professor, University Grenoble Alpes (LIG)	Rapporteur
Slim Essid Professor, Télécom Paris (LTCI)	Examinateur
Laurence Likforman-Sulem Associate Professor, Télécom Paris (LTCI)	Directrice de thèse
Giovanna Varni Associate Professor, Télécom Paris (LTCI)	Co-encadrante de thèse
Mohamed Chetouani Professor, CNRS/Sorbonne University (ISIR)	Invité (co-encadrant de thèse)

Cette Thèse est dédiée à ma famille, sans qui, toute cette aventure n'aurait été possible.

Résumé

Au cours de la dernière décennie, un nouveau domaine de recherche multidisciplinaire appelé traitement des signaux sociaux (SSP) a émergé. Il vise à permettre aux machines de détecter, reconnaître et afficher les signaux sociaux humains. L'analyse automatisée des interactions de groupe est l'une des tâches les plus complexes abordée par ce domaine de recherche. Récemment, une attention particulière s'est portée sur l'étude automatisée des états émergents. En effet, ceux-ci jouent un rôle important dans les dynamiques de groupe car ils résultent des interactions entre les membres d'un groupe.

Dans cette Thèse, nous abordons l'analyse automatique de la cohésion dans les interactions de petits groupes. La cohésion est un état émergent affectif multidimensionnel qui peut être défini comme un processus dynamique, reflété par la tendance d'un groupe à rester ensemble pour poursuivre des objectifs et/ou des besoins affectifs. Malgré la riche littérature disponible sur la cohésion du point de vue des Sciences Sociales, l'analyse automatique de la cohésion en est encore à ses débuts.

En s'inspirant de connaissances tirées des Sciences Sociales, cette Thèse vise à développer des modèles informatiques de cohésion suivant quatre axes de recherche, en s'appuyant sur des techniques d'apprentissage automatique et d'apprentissage profond. Ces modèles doivent en effet tenir compte de la nature temporelle de la cohésion, de sa multidimensionnalité, de la façon de modéliser la cohésion du point de vue des individus et du groupe, d'intégrer les relations entre ses dimensions et leur évolution dans le temps, ainsi que de tenir compte des relations entre la cohésion et d'autres processus de groupe. De plus, face à un manque de données disponibles publiquement, cette Thèse a contribué à la collecte d'une base de données multimodales spécifiquement conçue pour étudier la cohésion, et pour contrôler explicitement ses variations dans le temps. Une telle base de données permet, entre autres, de développer des modèles informatiques intégrant la cohésion perçue par les membres du groupe et/ou par des points de vue externes.

Nos résultats montrent la pertinence de s'inspirer des théories tirées des Sciences Sociales pour développer de nouveaux modèles computationnels de cohésion et confirment les avantages d'explorer chacun des quatre axes de recherche.

Abstract

Over the last decade, a new multidisciplinary research domain named Social Signal Processing (SSP) emerged. It is aimed at enabling machines to sense, recognize, and display human social signals. One of the challenging tasks addressed by SSP is the automated group interaction analysis. Recently, a particular emphasis is given to the automated study of emergent states as they play an important role in group dynamics. These are social processes that develop throughout group members' interactions.

In this Thesis, we address the automated analysis of cohesion in small groups interactions. Cohesion is a multidimensional affective emergent state that can be defined as a dynamic process reflected by the tendency of a group to stick together to pursue goals and/or affective needs. Despite the rich literature available on cohesion from a Social Sciences perspective, its automated analysis is still in its infancy.

Grounding on Social Sciences' insights, this Thesis aims to develop computational models of cohesion following four research axes, leveraging Machine Learning and Deep Learning techniques. Computational models of cohesion, indeed, should account for the temporal nature of cohesion, the multidimensionality of this group process, take into account how to model cohesion from both individuals and group perspectives, integrate the relationships between its dimensions and their development over time, and take heed of the relationships between cohesion and other group processes. In addition, facing a lack of publicly available data, this Thesis contributed to the collection of a multimodal dataset specifically designed for studying group cohesion and for explicitly controlling its variations over time. Such a dataset enables, among other perspectives, further development of computational models integrating the perceived cohesion from group members and/or external points of view.

Our results show the relevance of leveraging Social Sciences' insights to develop new computational models of cohesion and confirm the benefits of exploring each of the four research axes.

Acknowledgments

First of all, I would like to warmly thank Professors Laurence Likforman-Sulem, Mohamed Chetouani, and Giovanna Varni for being my Thesis supervisors. I am grateful for their constructive feedback and guidance all along the Thesis, with particular attention to Giovanna Varni for the numerous efforts and time she invested in my work.

I also thank Professors Albert Ali Salah and Dominique Vaufreydaz for accepting to review this Thesis and Professors Catherine Pelachaud and Slim Essid for accepting to be part of my committee.

I would like to express my gratitude towards my co-authors Eleonora Ceccaldi and Professor Nale Lehman-Willembrock for helping me with the choice of the questionnaires for the data collection, and their statistical analysis, as well as Professor Gualtiero Volpe, his team, and Giovanna Varni, for taking care of the technical setup of the GAME-ON dataset. I learned a lot from you and I am proud having worked with you. I would also like to thank Fabian Walocha for co-designing and implementing the motion capture features, our labeling strategy and a first version of our baseline with the computation of its Shapley values, Soumaya Sabry for co-designing and implementing different approaches to integrate the link between leadership and cohesion, and Cédric Siboyabasore for his hard work and the fun moments we shared during his internship at Télécom Paris.

I also thank André-Marie Pez for having developed a VR application specifically for my Thesis, and all the people involved in the 2 data collections. Thank you also to Professors Jan Gugenheimer, Matthieu Labeau, Enzo Tartaglione, Nicolas Rollet, Attilio Fiandrotti and Luca Oneto, for their interesting discussions and feedback over the years.

Thank you to Pierre, Emile, Guillaume, Tanvi, Leo, Gaël, Cyril, Kamélia, Marc, Emilia, Tamim, Dimitri, the two Anas, Arturo, Junjie , Brice, Amaury, Nathan as well as other members of Télécom Paris, for making my time at the lab particularly enjoyable. *Grazie* to Gualtiero, Simone and Roberto for the warm welcome at Casa Paganini and, in particular, to Eleonora for showing me Genova at its best.

I also owe a lot to all the people who shared my daily routine and I would like to give a particular thanks to my friends Pierre and Emile, all the Epikoa family, and all my teammates from the P.U.C. rugby club with whom I spent countless hours on the pitch.

I would also like to give special credit to Marine, who supported me from the very beginning of this adventure.

Last but not least, I am forever grateful to my family. They have contributed more than they will ever imagine to this journey.

This work has been partially supported by the French National Research Agency (ANR) in the framework of its JCJC program (GRACE, project ANR-18-CE33-0003-01, funded under the Artificial Intelligence Plan).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Contents

THIS Chapter presents the context in which this Thesis is placed and describes its goal. It also introduces the four research axes from which stemmed the two research questions here investigated. The motivations behind these research questions and the main steps we realized to answer them are also explained. A list of the contributions of this Thesis and the publications that resulted from them is also given. The Chapter ends with the organization of the different Chapters of the Thesis.

1.1 Context of the Thesis

Throughout the human evolution, group interactions have been key to our species' success and have played a central role in the development of today's society (Van Vugt and Schaller, 2008; Tomasello et al., 2012). From an evolutionary perspective, humans are *ultra* social animals (Tomasello, 2014) that gather and cooperate to deal with specific threats and opportunities. Such a behavior can be viewed as an adaptive strategy that increased the survival and reproductive success of ancestral humans (Van Vugt and Schaller, 2008). Belonging to a group is, indeed, one of the most important human needs (Baumeister and Leary, 1995), and being accepted by it (and by society) is located at the top priority in psychological needs according to Maslow (1943)'s hierarchy of needs. Thus, humans multiplied their interactions for survival. As of today, interactions happen in everyday life

with a broad range of people (e.g., family) and in diverse contexts (e.g., in a workplace), requiring people to constantly adapt to appropriately behave. These social skills define the social intelligence which refers to the ability to *read* other people, understand their intentions and motivations, and act accordingly to manage human relations (Thorndike, 1920; Ambady and Rosenthal, 1992; Albrecht, 2006). The skills of social intelligence have been argued to be indispensable and perhaps the most important for success in life (Goleman, 2006). These are, however, not inherited and are learned from very early ages throughout face-to-face as well as group interactions.

Group interactions and group dynamics, in particular, have a long history in Social Sciences disciplines (e.g., Lewin, 1951; Bennis and Shepard, 1956; Tuckman, 1965; Wheelan, 1994; Birmingham and McCord, 2002; Cronin et al., 2011; Kozlowski, 2015). Scholars in Social Sciences shifted from a static to a dynamic view of the processes and showed that group-level processes (e.g., cohesion) are fundamentally different from individual-level processes (Abrams and der Pütten, 2020). On one hand, according to what Lewin (1951) called "*interactionism*", individual behavior results from personal and environmental factors as well as from the interaction of both. Thus, a group influences the individuals' behavior since they interact in a social setting. On the other hand, group processes can only be understood trough the group perspective and cannot be fully understood by observing its individuals without considering the group influences or social settings (Forsyth, 2012).

Recently, Kozlowski and Chao (2018) highlighted that, despite the agreement on the dynamic nature of group processes, they have primarily been assessed as static constructs. Also, they stated that more research is needed to appropriately conceptualize group processes that are "*multilevel phenomena that emerge, bottom-up from the interactions among group members over time*". Still according to Kozlowski and Chao (2018), one of the promising streams of scientific inquiry is the computer study of group processes.

With the advent of new technologies, new research domains such as Social Signal Processing (SSP) and Affective Computing (AC), emerged with the aim of developing machines that are socially and emotionally aware. In fact, being able to automatically analyze, detect and reproduce social and affective skills and enhance group processes, would provide machines that are, a priory, socially ignorant (Pentland, 2005), the power to dynamically adapt to humans and support them, opening new opportunities in a broad range of domains (e.g., virtual agents, robotics). At first, these research domains focused on analyzing individuals (e.g., emotion recognition, Cowie et al., 2001). A recent shift towards the study of groups emerged, with a particular emphasis on group affect. This is primarily due to its central role in group dynamics (Waller et al., 2016) and the fact that it is a potential driver of emergent states such as cohesion (Allen et al., 2021). Emergent states are social processes that result from the micro-level affective, behavioral, and cognitive interactions among group members (Marks et al., 2001). This new focus entails both technological and social difficulties due to the high diversity of groups and the complexity of modeling human interactions and their evolution over time.

The aim of the Thesis is to develop automated methods to study cohesion, a multidimensional group affective process that develops over time, in small groups interactions. According to Carron and Brawley (2000), cohesion is "*a dynamic process that is reflected*

in the tendency for a group to stick together and remain united in the pursuit of its instrumental objectives and/or for the satisfaction of member affective needs". Despite disarray concerning the number of dimensions of cohesion and their functions, scholars in Social Sciences agree on its Social and Task dimensions. Social cohesion refers to the social bonds between group members while Task cohesion corresponds to the degree of commitment to the group's tasks and goals. This Thesis encompasses the exploration of the cohesion dynamics as well as the interplay between its Social and Task dimensions over time. It also includes the modeling and integration of a group and the relationships between cohesion's dimensions and with other group processes such as emergent leadership and group emotion. These directions of research are grounded on Social Sciences insights (e.g., López-Zafra et al., 2008; Kozlowski and Chao, 2012; Severt and Estrada, 2015; Grossman et al., 2015; Salas et al., 2015; Vanhove and Herian, 2015) and highlight some of the multiple challenges that this Thesis aims to address. The research questions (RQs) described in this Chapter are first motivated and the main steps we undertook to answer them are presented.

All the work presented in this Thesis has been partially supported by the French National Research Agency (ANR) in the framework of its JCJC program (GRACE, project ANR-18-CE33-0003-01, funded under the Artificial Intelligence Plan).

In addition to my Thesis supervisors, I received help from Professor Gualtiero Volpe, University of Genoa, and his team, as well as from Professor Nale Lehmann-Willenbrock, University of Hamburg (see work in Chapter 3). At Télécom Paris, Professor Nicolas Rollet provided some feedback on the theoretical aspect of interactions, and André-Marie Pez developed the VR platform mentioned in Chapter 7. Moreover, I co-supervised, with Giovanna Varni, Fabian Walocha and Soumaya Sabry that partially contributed to the work presented in Chapter 4, Chapter 5 and Chapter 6.

1.2 Research Questions

The motivation of the work presented in this Thesis was to design and implement computational models of cohesion. Inspired by Social Sciences insights and existing computational methods, we first identified and targeted specific problems that are related to the automated study of cohesion and, sometimes, that are shared with other emergent states and group processes. Thus, in Chapter 2, we organized the literature into four research axes (RAs) in which we clustered the various existing approaches taken for the automated analysis of cohesion and we also introduce our own approaches according to them. These research axes are described in Chapter 2 and are the following:

RA1: Temporal nature of cohesion

RA2: Group modeling

RA3: Interplay between its dimensions

RA4: Relationships with other group processes

The two research questions that we address in this Thesis, stem from these research axes.

1.2.1 RQ1: What computational architectures can be implemented to automatically predict cohesion and its dynamics?

The first step in answering this RQ was to identify and implement the features characterizing the interaction that could be relevant to predict cohesion dynamics and its Social and Task dimensions, in particular. While verbal communication plays an important role in social interactions, it is known that a valuable amount of information is delivered non-verbally (Knapp et al., 2013). Thus, most of the studies interested in the automated analysis of cohesion and related group processes focused on extracting nonverbal features, showing that nonverbal communication is a more powerful predictor of group-level cohesion than verbal behavior (e.g., Kubasova et al., 2019; Alsulami, 2021). For these reasons, the Thesis concentrates on extracting nonverbal features.

Taking inspiration from Social Sciences' insights and computational studies on cohesion, we developed nonverbal features, using audio and motion capture data. Furthermore, to start investigating RA2 from the input, we extracted features computed from individuals and from the group as a whole.

Building on the first three research axes, we implemented our computational architectures to answer the following subresearch questions (SRQs):

(SRQ1) How to integrate the temporal nature inherent to cohesion?

(SRQ2) How to take into account both individuals and group behaviors that result from, and are influenced by, the group members' interactions?

(SRQ3) How to model the interplay between the Social and Task dimensions of cohesion over time?

Most of the computational studies related to cohesion rely on different definitions, making it difficult to compare findings across studies (e.g., Hung and Gatica-Perez, 2010; Kantharaju et al., 2020; Ghosh et al., 2022). Moreover, these studies use Machine Learning models built to predict cohesion over a short period of time (e.g., 2 minutes, Hung and Gatica-Perez, 2010), without taking previous elements of interaction into account. Also, these models are designed to predict the presence or the absence of cohesion as a whole (i.e., without distinguishing between dimensions) or for the Social and Task dimensions, separately, without investigating the relationships between cohesion's dimensions over time. Finally, group processes such as cohesion should be studied from both individual and group perspectives to fully capture such an affective emergent state, by integrating the complex relationships between the group members and their group behavior that results from their interactions.

Concretely, in Chapter 5, we describe various computational models of cohesion that range from a simple but consolidated state-of-the-art approach to more sophisticated approaches that increasingly address each SRQ. The development of these architectures was part of an iterative process that led to a large number of experiments. Only the relevant architectures are presented and discussed in this Thesis.

1.2.2 RQ2: How other group processes can inform the modeling of cohesion?

We grounded our work on Social Sciences' knowledge. Severt and Estrada (2015), indeed, state that various links between cohesion and other group processes may be observed depending on the function (e.g., instrumental vs affective), the dimension (e.g., Social vs Task), or the level of analysis of cohesion (e.g., horizontal vs vertical) that is being investigated. Furthermore, cohesion is an affective emergent state (e.g., Kozlowski and Chao, 2012; Maynard et al., 2015; Rapp et al., 2021) that develops over time. With this in mind, the focus is on integrating the links between cohesion and emergent leadership and between cohesion and group emotion into the computational models' architectures.

The cohesion-leadership link has already been proved (e.g., Light Shields et al., 1997) and could help to understand the group dynamics while the link between emotions and cohesion is straightforward (cohesion is, indeed, an affective emergent state) and could provide contextual information regarding the group members' affective states. In addition, Barsade and Gibson (1998) demonstrated a relationship between Social cohesion and emotion as well as between Task cohesion and emotion, corroborating the relevance of integrating these relationships into our computational models.

Leveraging the relationships between two (or more) group processes to computationally investigate them has already been explored, and showed promising results (e.g., Parthasarathy and Busso, 2017). Links with cohesion, however, had only been used in studies to improve computational models of other group processes (e.g., emotions, Ghosh et al., 2022). There is, to the best of our knowledge, no automated study of cohesion taking advantage of its links with other group processes to improve cohesion predictions and, in particular, for its Social and Task dimensions. Thus, the challenge of **RQ2** is to fill this gap by integrating the links between cohesion and other group processes into our computational models of cohesion, with the aim of improving their performances.

In Chapter 6, we computationally explore how to integrate the links between (1) cohesion and group emotion, and (2) cohesion and emergent leadership, respectively. As for (1), we leveraged Social Sciences' insights to design two DNN architectures to integrate such links, following the Top-down and Bottom-up approaches (Barsade and Gibson, 1998). About (2), we introduced two different families of approaches focusing on integrating leadership information into our models (i.e., by intervening in the features or in the architecture of the models).

1.3 Contributions of the Thesis

First contribution: *A structured survey on cohesion for supporting the automated analysis of cohesion in small groups interactions.*

Multiple definitions, theoretical models, and frameworks of cohesion exist with major differences (e.g., the number of dimensions). Such disarrays in the theoretical conceptualization of cohesion ultimately slow the emergence of robust and reliable computational studies of cohesion. Moreover, automatically analyzing cohesion is a complex task. Cohesion is, indeed, a group affective multidimensional emergent state. Thus, its temporal

nature, its various dimensions (and their interplay), and its links with other group processes are all challenges to be met, with a plethora of approaches that could be employed. In Chapter 2, we defined four research axes to structure our work and ease the development of new approaches for addressing the automated analysis of cohesion.

Based on these axes, we reviewed and clustered the approaches employed in existing work on the automated analysis of cohesion according to their research axis and goal. In fact, an approach could be employed with the goal of impacting the computational model at its input (e.g., segmentation of the features), model's architecture (e.g., type of layer in a DNN) or output (e.g., labeling strategy). Such an "*Input-Model-Output*" categorization is inspired by the "*Input–Process–Output*" (IPO) theoretical framework (Hackman and Morris, 1975) for conceptualizing teams in Social Sciences and its subsequent enhancements (e.g., Kozlowski et al., 1999; Ilgen et al., 2005). In the IPO framework, the input refers to any antecedent that may influence the group, directly or indirectly, the process (here, cohesion), is an activity that mediates the relationships between the inputs and the group's outcomes, while the outputs are the consequences of the group's actions (Forsyth, 2012). The parallel with our categorization is straightforward: inputs correspond to the features, model refers to the computational model architecture, and outputs are related to the purpose of the model (e.g., predicting cohesion for a specific dimension). Such a categorization highlights the state of the literature on the automated analysis of cohesion regarding each research axis. In addition, we also suggest new approaches to computationally address such a complex emergent state (cf. Chapter 5 and Chapter 6).

With this contribution, we aim to provide a structured way to comprehend the existing literature, highlighting the open challenges of automatically analyzing cohesion. It also helps appreciating the novelties introduced by our computational models of cohesion.

Second contribution: *Multimodal dataset for the automated cohesion analysis.*

Facing a lack of publicly available data specifically designed for the automated analysis of cohesion, we collected GAME-ON, a multimodal dataset that contains more than 11 hours of audio, video, and motion capture data of 17 groups of friends interacting in the context of an escape game. The escape game was thought to elicit variations of the Social and Task dimensions of cohesion across five tasks that require different skills. Before and after each task, a cohesion questionnaire (i.e., the GEQ, Carron et al., 1985) as well as questionnaires related to other group processes (e.g., leadership, emotion) were administered to each group member. In addition, annotations of cohesion were collected afterward by external annotators watching the recordings of a group. The dataset is presented in Chapter 3.

In this Thesis, features were extracted from the GAME-ON dataset (cf. Chapter 4), and the labels used to train our computational models of cohesion were computed from the self-assessments of cohesion (cf. Chapter 5) as well as from the self-assessments of emotion and leadership (cf. Chapter 6). The designs of the computational models were also directly impacted by the set-up of the escape game. Architectures are, indeed, designed for predicting, within the same model, the Social and/or Task cohesion dynamics, for each of the five tasks of the escape game. The motivation of GAME-ON is to provide the scientific community with an asset for studying cohesion and its relationships with other group processes.

Third contribution: *Design and implementation of computational models of cohesion.*

Previous work on the automated analysis of cohesion (e.g., Hung and Gatica-Perez, 2010; Nanninga et al., 2017) either focused on predicting cohesion as a whole, without distinguishing between its dimensions (e.g., Hung and Gatica-Perez, 2010; Ghosh et al., 2022), or on designing models for predicting only a specific dimension (e.g., Nanninga et al., 2017). In both cases, they did not explore the interplay between its dimensions over time, nor how other group processes such as leadership could impact each dimension. Furthermore, as cohesion develops over time, more effort is needed to capture the dynamics of such an affective emergent state. Finally, existing computational studies rely on external assessments of cohesion to train their models. While differences in the perception of cohesion exist between both self- and external assessments (Vinciarelli and Mohammadi, 2014), it remains to be seen how these could impact computational models of cohesion.

Following the four research axes mentioned in Section 1.2, we took inspiration from Social Sciences' theories and insights on small groups interactions and cohesion to design various computational models of cohesion to answer **RQ1** and **RQ2**. In Chapter 5 and Chapter 6, we introduce a collection of computational models of cohesion. Each model addresses at least one of the research axes and implement approaches with various degree of complexity. Performances of the models are compared and discussed, providing us elements of answer for each SQR.

The following references are the work accepted and published during the Thesis.

Journals

- Maman, Ceccaldi, Lehmann-Willenbrock, Likforman-Sulem, Chetouani, Volpe, and Varni, 2020, GAME-ON: A Multimodal Dataset for Cohesion and Group Analysis. **IEEE Access**.

Conferences

- Maman, Volpe, and Varni, 2022, Training Computational Models of Group Processes without Groundtruth: the Self- vs External Assessment's Dilemma. **International Conference on Multimodal Interaction (ICMI)** (Late-Breaking Results) - Accepted.

- Maman, Likforman-Sulem, Chetouani, and Varni, 2021b, Exploiting the Interplay between Social and Task Dimensions of Cohesion to Predict its Dynamics Leveraging Social Sciences. **ICMI** - Best Paper award.

- Maman, Chetouani, Likforman-Sulem, and Varni, 2021a, Using Valence Emotion to Predict Group Cohesion's Dynamics: Top-down and Bottom-up Approaches. **International Conference on Affective Computing & Intelligent Interaction (ACII)**.

- Maman, 2020, Multimodal Groups' Analysis for Automated Cohesion Estimation. **ICMI** (Doctoral Consortium).

Workshops

- Sabry, Maman, and Varni, 2021, An Exploratory Computational Study on the Effect of Emergent Leadership on Social and Task Cohesion. **Insights on Group & Team Dynamics (IGTD)** at ICMI.
- Walocha, Maman, Chetouani, and Varni, 2020, Modeling Dynamics of Task and Social Cohesion from the Group Perspective Using Nonverbal Motion Capture-based Features. **IGTD** at ICMI.
- Maman and Varni, 2020, GRACE : Un projet portant sur l'étude automatique de la cohésion dans les petits groupes d'humains. **Workshop sur les Affects, Compagnons artificiels et Interactions (WACAI)**.

1.4 Organization of the Thesis

This Thesis is organized in seven Chapters, including this Introduction. In Chapter 2, we introduce the theoretical background on emergent states and cohesion. Based on the theoretical foundations of cohesion, we defined four research axes for the automated analysis of cohesion. These axes helped us organizing the related work as well as highlighting the contributions of this Thesis.

In Chapter 3, we review the existing datasets used in Social Signal Processing and Affective Computing research domains to study small groups interaction and we present GAME-ON, a multimodal dataset specifically designed for the automated study of cohesion. Details about the data collection process are given followed by an analysis of the cohesion questionnaires.

Chapter 4 is devoted to the description of the features that we extracted and used in every computational model of cohesion presented in this Thesis. First, we briefly review the important automatically extracted features in the automated group interaction analysis. Then, we justify and give computational details of each feature extracted.

Chapter 5 presents our models' evaluation and comparison procedures as well as all the computational models developed following the research axes presented in our structured survey. Models are compared between each other according to their characteristics and novelties introduced.

In Chapter 6, we detail the various approaches that we implemented to integrate the links between cohesion and group emotion and leadership, respectively.

Finally, Chapter 7 summarizes the contributions of our work as well as its limitations and suggests some perspectives to improve it in short and long-term perspectives.

Chapter 2

Background and Related Work

Contents

COHESION is a group process that has been extensively studied since the 1920s and that is still being investigated through the lens of new group development theories. It is one of the most studied emergent state (Rosh et al., 2012) - i.e., a social process that results from the micro-level affective, behavioral, and cognitive interactions among group members (e.g., Marks et al., 2001) - due to its influence on desirable group outcomes such as group effectiveness and performance.

In this Chapter, we introduce the research domains in which this Thesis is placed and we describe the main theories in Social Sciences about emergent states and cohesion in particular. Based on such a review, we identified four research axes for the development of computational models of cohesion. Then, we present the computational studies that

are related to the automated analysis of cohesion from nonverbal features and we cluster them according to the research axes previously defined and the level at which it operates (i.e., input, model, or output of the computational model). Such an organization helps identify what are the axes that are under-investigated from a computational point of view, hence, guiding the design of new approaches.

In this Chapter, I did all the work, including the choice of the four research axes and the design of the new approaches introduced in the structured survey.

2.1 Introduction

Understanding humans is a complex task that fascinates and engages scholars in disciplines ranging from Social Sciences (e.g., Psychology) to Computer Sciences (e.g., Human-Centered Computing). In the last 20 years, the computing community shifted from a computer-centered to a more human-centered vision of computing (Pantic et al., 2007, 2008) with the aim of providing new methods and tools to endow machines with social and affective intelligence (Picard, 1999; Pentland, 2007; Vinciarelli et al., 2008). Nowadays, two growing and active interdisciplinary research domains embrace such a vision. On one hand, *Affective Computing* (AC) aims to develop systems and devices that can recognize, interpret, process, and simulate human emotions, affect, and moods as humans would, by relying on their senses to assess each other's communicative and affective states (Picard, 2000). AC is, indeed, more focused on providing affective intelligence to machines (see Picard, 1999, 2000, 2003; Tao and Tan, 2005; Zeng et al., 2008; Calvo et al., 2015; Cambria et al., 2017, for various reviews). On the other hand, Social Signal Processing (SSP) aims to provide machines the ability to integrate and support human-human interactions and embody natural modes of human communication for interacting with their users (Vinciarelli et al., 2011). To achieve such a goal, machines should ultimately model, analyze and synthesize behavior in social interactions. Thus, SSP is more focus on the social intelligence (see Pentland, 2007; Vinciarelli et al., 2008, 2009a,b, 2011; Salah et al., 2011; Pantic et al., 2011; Pantic and Vinciarelli, 2014; Gunes and Hung, 2015; Burgoon et al., 2017, for the various challenges and applications of SSP).

Initially, these research domains focused on individuals. Recently, a particular emphasis is given to the study of groups. In fact, group affect plays an important role in group dynamics (Waller et al., 2016) as it potentially leads to the emergence of group processes such as cohesion (Allen et al., 2021). Such a shift towards groups unlocks a broad new range of applications that encompass, but are not limited to, smart surveillance, ambient intelligence, social robotics, human-computer interfaces, virtual agents, entertainment, education, social skills training and so on (Pantic et al., 2011; Salah et al., 2011).

Automatically studying groups and their processes, however, entails both technological and social difficulties due to the high diversity of groups and the complexity of modeling human interactions and their evolution over time. For example, Salah et al. (2011) identified the fact that numerous definitions, theoretical models, and frameworks exist for a same group process, as one of the major issues for the development of more robust computational models. Thus, collaborations with other disciplines (e.g., Psychology) should be considered at each step of the study and, in particular, for automatically studying emergent states such as cohesion. Emergent states are social processes that result from the

micro-level affective, behavioral, and cognitive interactions among group members (e.g., Marks et al., 2001). Some of these states are multidimensional making their automated analysis even more complex, hence, potentially explaining why emergent states remain under-investigated, despite their important role in group dynamics.

The focus of this Thesis is on the automated analysis of cohesion, a multidimensional group affective emergent state, in small groups. Thus, our research lies at the intersection of the AC and SSP research domains and requires a multidisciplinary approach. In this regard, it is necessary, at first, to accurately define what are emergent states and cohesion in particular.

2.2 Background

2.2.1 Emergent States

The term *emergent state* was first coined by Marks et al. (2001). In their seminal article, they first differentiated group processes from group emergent states. Group processes describe interdependent group activities that lead group members to pursue their goals, while emergent states express cognitive, motivational, and affective states of groups. Thus, emergent states differ from group processes as they do not describe the nature of the group members' interactions. Such a characterization of emergent states builds upon Klein and Kozlowski (2000)'s work stating that a "*phenomenon is emergent when it originates in the cognitive, affect, behaviors, or other characteristics of individuals, is amplified by their interactions and manifests as a higher level, collective phenomenon*". Furthermore, Marks et al. (2001) add that emergent states are "*dynamic in nature and vary as function of team context, inputs, processes, and outcomes*". Following these definitions, emergent states were first clustered into three families (Kozlowski and Ilgen, 2006; Grossman et al., 2017). *Cognitive* emergent states are related to the management of the group's collective knowledge; *behavioral* emergent states concern the activities and the interactions among group members; *affective* emergent states deal with the relationships among group members and their emotional responses. Recently, Rapp et al. (2021) provided a review of the literature on emergent states over the past 20 years. Building upon Marks et al. (2001) categories of emergent states, they introduce a new category, i.e., the *Motivational* emergent states, and suggest that some emergent states can blend into two or more categories.

In detail, the cognitive emergent states concern group members beliefs or thoughts regarding a specific factor. It includes, for example, constructs related to team cognition (i.e., the manner in which knowledge for group functioning is mentally organized, represented, and distributed within a team, DeChurch and Mesmer-Magnus, 2010), and team climates (i.e., the group members' perceptions of norms, attitudes, and expectations perceived to operate within a specific context, Schneider, 1990). Shared mental models (i.e., a shared understanding of the task that is to be performed and of the involved teamwork required, Converse et al., 1993) and transactive memory systems (i.e., a group-level knowledge sharing and memory system in which group members share responsibility for encoding, storing, and retrieving of information from different knowledge areas,

and have a shared awareness about each member's knowledge responsibilities, Wegner, 1987), are examples of team cognition group processes.

Behavioral emergent states involve what team members do, that is, the activities and interactions primarily focused on accomplishing task objectives (Kozlowski and Ilgen, 2006). According to Grossman et al. (2017), transition, action, and interpersonal processes are categorized as behavioral emergent states. Marks et al. (2001) define transition processes as "*periods of time when teams focus primarily on evaluation and/or planning activities to guide their accomplishment of a team goal or objective*", whereas action processes involve "*periods of time when teams conduct activities leading directly to goal accomplishment*". Finally, interpersonal processes are described as "*processes teams use to manage interpersonal relationships*". Team reflexivity (i.e., the extent to which group members overtly reflect upon the group's objectives, strategies, and processes, and adapt them to current or anticipated endogenous or environmental circumstance, West, 1996), coordination processes (i.e., processes that involve orchestrating the sequence and timing of interdependent actions, Marks et al., 2001) and conflict management are examples of transition, action and interpersonal processes, respectively.

Affective emergent states concern group members' feelings, attitudes, and emotions. Among the many group processes classified as affective emergent states (see Rapp et al., 2021, for a review), psychological safety and cohesion are the two most investigated. Psychological safety concerns beliefs that the group is safe for interpersonal risk-taking (Edmondson and Lei, 2014). Cohesion refers to the tendency of a group to stick together to pursue goals and/or affective needs (Carron et al., 1985). Various definitions and theoretical frameworks, however, exist (e.g., Festinger et al., 1950; Dion, 2000; Severt and Estrada, 2015) defining cohesion with a various number of dimensions (i.e., from two to five), making it difficult to compare results and insights between studies. Cohesion remains one of the most studied emergent states due to its relationships with other group processes (e.g., group emotion and leadership). More importantly, scholars in Social Sciences provided evidence linking cohesion to a broad range of positive outcomes such as team performance (Gully et al., 2012; Levi, 2001), effectiveness (Tekleab et al., 2009), and creativity (Zhang, 2016), stimulating researchers to investigate such an affective emergent state through its multiple dimensions.

Motivational emergent states concern group members' intensity, direction, and regulation of effort toward task accomplishment. It includes processes such as team potency (i.e., beliefs regarding general team ability) and team efficacy (i.e., beliefs about task-specific team ability).

One of the most dominant theoretical frameworks for conceptualizing teams and group processes is the "*Input–Process–Output*" (IPO) framework, developed by Hackman and Morris (1975). In this framework, the *input* refers to any antecedent that may influence the group, directly or indirectly, the *process*, is an activity that mediates the relationships between the inputs and the group's outcomes, while the *outputs* are the consequences of the group's actions (Forsyth, 2012). Kozlowski et al. (1999) and Ilgen et al. (2005) also enhanced the IPO framework by incorporating an iterative feedback loop to take into account the influences of the actions taken by the group members. In fact, the outputs of a group's actions can provide the input for their next action. IPO remains, to the best of our knowledge, the theoretical framework of reference for studying emergent states and

is also applied to address various types of groups such as virtual groups (Webster and Staples, 2006; Hoch and Kozlowski, 2014; Dulebohn and Hoch, 2017).

2.2.2 What Is Cohesion?

Cohesion, derived from the Latin word *cohaesus* meaning "*staying together*", is one of the most studied emergent states in Social Sciences (LePine et al., 2008; Rosh et al., 2012). Since the early 1920s, it has been extensively studied through the lens of various group dynamics theories (e.g., Schneider and Mcdougall, 1921; Moreno, 1934) and because of its relationship with group performance (e.g., Bird et al., 1980; Spink, 1990; Mullen and Copper, 1994). Multiple definitions, conceptual models, and theoretical frameworks, however, exist, reflecting the complexity of such a group process (see Buton et al., 2006, for a review of the early works on cohesion). Initially defined as a unidimensional process, scholars in Social Sciences rapidly adopted a multidimensional conceptualization of cohesion without necessarily agreeing on the number of dimensions and their functions. The plethora of definitions, conceptual models, and theoretical frameworks of cohesion that emerged, ultimately slowed the comprehension of such a complex process and limited the generalization of the results. Scholars in Social Sciences, however, agreed on two distinct but interrelated dimensions of cohesion, i.e., the Social and the Task, as they play an important role in group interactions. Depending on many factors (e.g., relationships among group members, size of the group), one dimension might be predominant and impact the development of the other over time. Multiple group development theories accounted for the emergence and the interplay of these dimensions but, again, provided opposite findings.

2.2.2.1 From Unidimensional to Multidimensional Definitions of Cohesion

In the early 1940s, Lewin (1939) laid the foundation for the concept of group cohesion, under the framework of the field theory. Such a theory examines patterns of interaction between the individuals and their total field, or environment. He considered cohesion (or the willingness to stick together) as an essential property of groups that depends on the group size, organization, and intimacy. He defined it as the set of forces keeping members together, including both the positive forces of attraction and the negative forces of repulsion. Building on their work, Festinger et al. (1950) define groups as a set of connections (i.e., friendship bonds) between group members and extended Lewin's definition of cohesion, defining it as the "*total field of forces causing members to remain in the group*". These forces are related to (1) the individual attraction to the other group members, relying on the need to belong to a group, (2) the operating forces, corresponding to the ones related to the group activities, and (3) the group prestige, referring to the group members pride of being part of the group. Such a categorization of forces highlights that cohesion was already understood as a multidimensional construct. Due to the difficulties to control and measure the impact of each force, scholars in Social Sciences, however, continued to consider cohesion as a unidimensional construct. Later, researchers started to focus either on the forces related to the social aspects of cohesion (e.g., Schachter et al., 1951; Lott and Lott, 1965) or on those related to the task aspects of cohesion (e.g., Back, 1951; Van Bergen and Koekebakker, 1959). These studies had a relevant impact on the devel-

opment of multidimensional models and frameworks of cohesion. In fact, these forces became two distinct dimensions, namely Social cohesion, and Task cohesion. Figure 2.1 shows a timeline of the main studies defining cohesion as unidimensional.

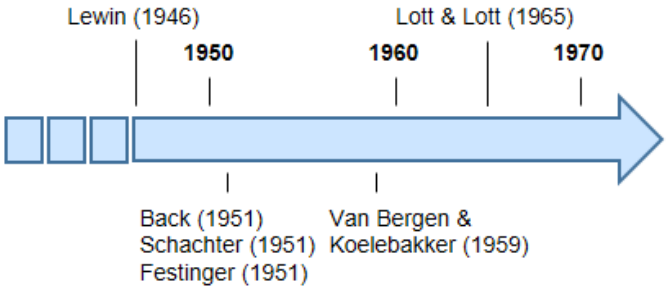


Figure 2.1: Timeline of the main studies defining cohesion as unidimensional, from the 1940s to the 1970s.

2.2.2.2 Multidimensional Models and Frameworks of Cohesion

Since the 1980s, the idea that cohesion is a multidimensional construct is well accepted and many scholars in Social Sciences introduced multidimensional definitions (e.g., Bollen and Hoyle, 1990; Dion, 2000) and models (e.g., Carron et al., 1985; Hogg and Hardie, 1991; Cota et al., 1995; Bliese and Halverson, 1996) of cohesion, designed for studying such an affective emergent state in different contexts (e.g., military, sport). Most of these definitions and models are bi-dimensional and suggest multiple ways to define cohesion. For example, Bollen and Hoyle (1990)'s definition as well as the models introduced by Hogg and Hardie (1991) and Bliese and Halverson (1996), focus on the social aspects of cohesion and categorize various types of attraction (e.g., feelings towards other group members, sense of belonging). Based on previous works indicating the need to incorporate a Task dimension (e.g., Festinger et al., 1950; Hersey and Blanchard, 1969) and to distinguish individual and group levels at which cohesion can emerge (e.g., Van Bergen and Koekebakker, 1959; Hagstrom and Selvin, 1965), Carron et al. (1985) introduced a model to study groups in sport teams (see Figure 2.2). This model grounds on a definition of cohesion that considers such an emergent state as a dynamic process that can be reflected by the tendency of a group to stick together to pursue goals and/or affective needs. This model comprises two major dimensions: *Individual attraction to the group* and *Group integration*. Individual attraction to the group represents all the reasons that would motivate a group member to remain in the group, while Group integration represents the degree of unification of the group. Each one of these dimensions can manifest through the Task and Social dimensions. The Task dimension relates to the degree of commitment to group tasks and goals while the Social dimension relates to the relationships and friendships between group members. This model was adopted by many scholars in Social Sciences as the reference model. Other studies, however, questioned the ability of Carron's model to generalize to interactions that are outside of a sport context (e.g., Cota et al., 1995; Dion, 2000) and advised designing multidimensional frameworks that consider *Primary* dimensions that are applicable to most groups and *Secondary* dimensions that are able to adapt to specific contexts and groups. For example, Dion (2000),

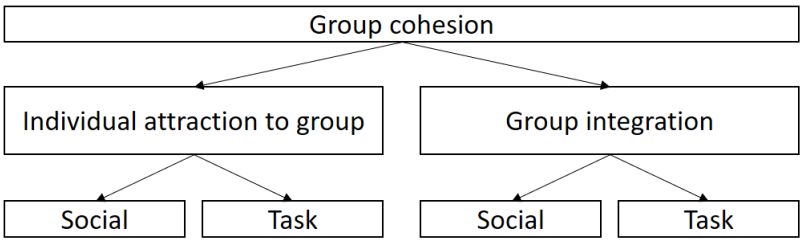


Figure 2.2: Model of cohesion developed by Carron et al. (1985). It is composed of two dimensions (*Individual attraction to the group* and *Group integration*) that, in turn, are expressed in the Social and Task dimensions.

suggested that the Social and Task dimensions, the sense of belonging, and vertical cohesion (i.e., in the context of hierarchical relationships, it refers to the subordinates' perceptions of their leaders' competence and considerateness) were the Primary dimensions while valued roles, as identified by Yukelson et al. (1984), and risk-taking were examples of Secondary dimensions. Despite the disarray in the number of dimensions and the functions of cohesion, scholars in Social Sciences always agreed on the Social and Task dimensions of cohesion.

More recently, Severt and Estrada (2015) proposed an integrative framework taking into account Carron's model and other researchers' ideas and improvements (i.e., Griffith, 1988; Bollen and Hoyle, 1990; Dion, 2000; Beal et al., 2003). This framework posits that cohesion can be categorized by two main functions, an *Affective* function and an *Instrumental* function. Figure 2.3 summarizes their framework.

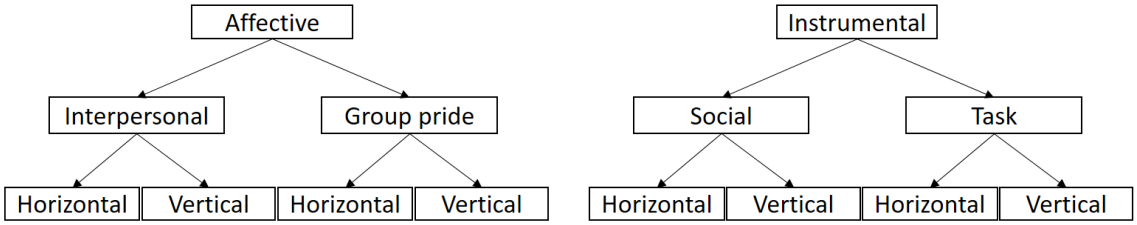


Figure 2.3: Multidimensional framework of cohesion developed by Severt and Estrada (2015). It is composed of two functional properties (i.e., *Affective* and *Instrumental*). Each property has two dimensions, which are also divided into two levels (i.e., *horizontal* and *vertical*) depending on the hierarchical relationships that may exist in the group.

The Affective function of cohesion refers to all the aspects that highlight the emotional impact on a group member and, by extension, the group as a whole (e.g., behaviors or elements of interaction such as cooperation or exchange). Severt and Estrada divided it into two dimensions that they refer to as facets. First, the *Interpersonal* dimension lies in how much one likes, dislikes, or hates the other group members. It can be viewed as a force acting between people that tends to draw them together and to resist their separation. The second one, the *Group pride* dimension, results from a deep sense of belonging to a group as a whole. It creates a sense of community which strengthens the bonds of unity. A group member may be attracted to the group because being part of it is viewed as an honor (Back, 1951). This dimension emphasizes the importance that

members place on identifying themselves to the group and being part of it (Beal et al., 2003). Friendship bonds and the desire to identify with a group are often signals of the emergence of cohesion through its affective dimensions. A group of coworkers going out for an event outside of work hours is an example of the emergence of interpersonal cohesion whilst observing group members wearing group t-shirts is an example of group pride cohesion.

The Instrumental function of cohesion refers to "*those aspects that highlight the goal- and task-based activities of the group*" (Severt and Estrada, 2015). Following Katz (1960)'s statement about the instrumental function of cohesion, Severt and Estrada (2015) suggest that it is the instrumental function of cohesion that "*keeps the group intact so that it can achieve the set goals of the group, all the while maximizing the rewards gained from achieving those goals, and minimizing penalties or losses in the process*". Within the Instrumental function of cohesion, they distinguish between *Social* and *Task* cohesion.

The Social dimension refers to the social bonds between group members that are bound by the group's *working* relationship. It might be counterintuitive to categorize Social cohesion as an instrumental function, but social bonds can indeed serve the group's goal. The higher Social cohesion will be in a group, the more its members will value the relationships and friendships that the group provides (Lott and Lott, 1965), resulting in a positive climate where group members engage in high-quality social working relationships. An example of Social cohesion is when group members play board games together during their break.

Task cohesion relates to the degree of commitment to group tasks and goals. It is implied that group members need to share a sufficient level of confidence in the task(s) realization. An example of task cohesion is when a leader supports another group member by creating conditions that will ease the resolution of the task.

For each dimension of the two functional properties of cohesion (i.e., Affective and Instrumental), two levels can be distinguished according to hierarchy differences among members: *Horizontal* and *Vertical*. Horizontal cohesion concerns relations among group members of the same authority level, whereas Vertical cohesion implies hierarchy and refers to the relations between a member of authority and a subordinate within the group context. It is important to differentiate these levels as cohesion can emerge from relationships among various types of groups and group members and across the entirety of the group's hierarchy. Cohesion also manifests differently according to the dimension and level of measurement. Figure 2.4 shows a timeline of the main studies defining cohesion as a multidimensional construct.

Nowadays, more and more scholars in Social Sciences include and consider autonomous systems (e.g., robots) as a group member. This is one of the reasons that could explain the recent development of new theoretical frameworks designed for studying cohesion in hybrid groups composed of humans and robots (e.g., Abrams and der Pütten, 2020; Lakhmani et al., 2022). In particular, Abrams and der Pütten (2020) introduce the Ingroup identification (I), Cohesion (C), and Entitativity (E) conceptual framework as a theoretical foundation for group dynamics research in Human-Robot Interaction (HRI). Such a framework is based on the Social and Task dimensions following Carron et al. (1985)'s model and has the particularity to consider cohesion at individual and at group levels while integrating the perception of group unity with other members, including the robots (i.e., "*Entitativity*" from an outside observer perspective and "*Ingroup identifica-*

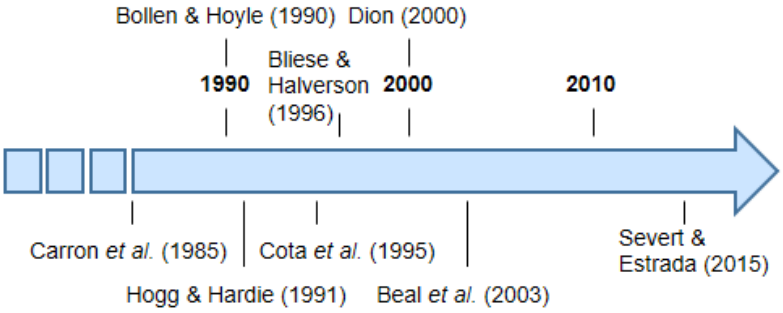


Figure 2.4: Timeline of the main studies defining cohesion as multidimensional construct, from the 1980s to nowadays.

tion" from an inside member perspective). While the ICE framework focuses on existing dimensions (i.e., Social and Task cohesion), Lakhmani et al. (2022) completely redefine a set of dimensions, subdimensions, and factors of cohesion based on the existing Social Science literature. They also introduce a new dimension (i.e., team resilience) and factor (i.e., complementarity) that are specific for studying hybrid teams composed of both humans and robots or virtual agents. Figure 2.5 shows an overview of this framework. It is composed of three major dimensions: *Functions of cohesion*, *Directions of cohesion* and *Team resilience*. Each dimension contains sub-dimensions (e.g., *Interpersonal*) and relevant factors such as *Morale* (Berg et al., 2021).

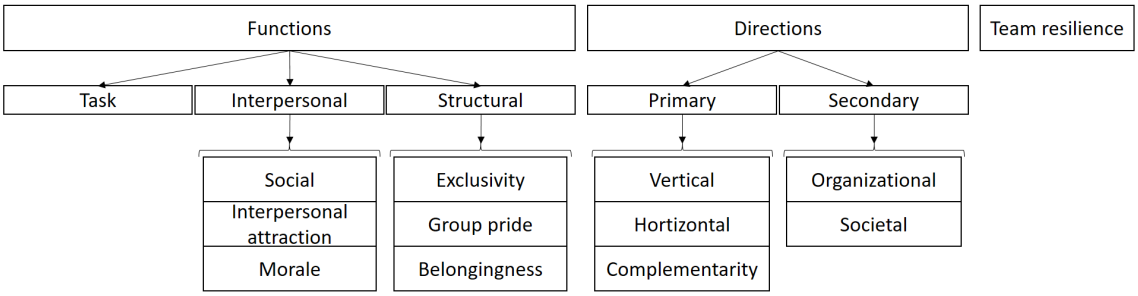


Figure 2.5: Framework of cohesion developed by Lakhmani et al. (2022). Cohesion is here defined into three dimensions (i.e., *Functions of cohesion*, *Directions of cohesion* and *Team resilience*. Some of these dimensions are composed of sub-dimensions that are associated with relevant factors (e.g., the *Interpersonnal* dimension and its *Morale* relevant factor).

Even if these frameworks are very specific to HRI research (i.e., they were specifically designed to study mixed teams composed of robots and/or virtual agents), they highlight the complexity of defining a universal framework of cohesion that adapts to various types of groups and contexts.

2.2.3 The Temporal Nature of Cohesion

Cohesion is an affective emergent state. Therefore, by definition, it develops over time through the group members' interactions. While scholars in Social Sciences agree on the temporal nature of cohesion, there is a lack of study explicitly addressing it (Coultas et al., 2014; Salas et al., 2015; Grossman et al., 2015). Groups are, indeed, dynamic entities, and cohesion is likely to change and operate differently as various group processes and situational variables unfold over time (Grossman et al., 2015). Many factors could, in fact, impact the emergence and development of cohesion (and its dimensions). One of the main factors is the developmental phase of the group. For example, in newly formed groups, cohesion may rapidly emerge but is highly volatile (Mullen and Copper, 1994), especially for its Social dimension (Siebold, 2006). Such an observation led to the notion of *swift* cohesion (Coultas et al., 2014; Salas et al., 2015). On the contrary, in groups with longer tenure, cohesion is more stable. In this case, variations of cohesion might occur depending on the task that is being performed. These variations might also be different for each dimension. For example, Bartone and Adler (1999) looked at cohesion in deployed military units and showed that cohesion emerged in an inverted-U function, with low levels of cohesion pre-deployment, high levels mid-deployment, and a leveling off toward the end of the deployment cycle. Later, Siebold (2006) corroborated these findings and suggested that this U-shape pattern of cohesion was more prevalent for Social cohesion while an inverted U-shape pattern occurred more prominently for Task cohesion.

That being said, it appears crucial to incorporate time into computational models of cohesion. It is, indeed, highly recommended in various studies highlighting the difficulties of measuring such a process (e.g., Coultas et al., 2014; Salas et al., 2015; Grossman et al., 2015). Furthermore, modeling dynamics of emergent states was also identified as one of the key challenges of SSP according to Brunet et al. (2012). Thus, addressing the temporal nature of cohesion corresponds to our first research axis (RA1).

2.2.4 Cohesion, a Group Emergent State

Automatically studying groups and their dynamics is a complex task that has received little attention compared to individual processes. This is primarily due to the many non-linear interactions that occur between group members (Gilbert, 2004), that dynamically affect individuals' behaviors. Also, it comes from the fact that group processes are not necessarily well defined from a Social Sciences point of view (i.e., multiple definitions exist). This is the case for cohesion. It is, indeed, a group emergent state that unfolds through its members' interactions.

The two main streams for considering groups processes, and cohesion by extension, are the *Top-down* and *Bottom-up* approaches (Barsade and Gibson, 1998). The first one focuses on the group as a whole and is close to the group conception developed by Schneider and Mcdougall (1921). This approach is based on the assumption that the whole is greater than the sum of its parts, hence, it considers the group processes such as cohesion, as responsible for the influences on the members' feelings and behavior. Following this approach, scholars in Social Sciences characterized group processes as (1) forces which shape individual emotional response (e.g., Le Bon, 1897), (2) social norms (e.g., Gibson, 1997), (3) the interpersonal glue that keeps groups together (e.g., Festinger et al., 1950)

and (4) a display of group's maturity and development (e.g., Bales and Strodtbeck, 1951). Oppositely, the Bottom-up approach views group processes as the sum of its individuals. This approach is derived from the "*individualist*" view of group processes exposed by Moreno (1934). Such a view led researchers to examine the group through a variety of compositional perspectives such as the mean of the group's members, the degree of variance of the observed process within the group, and the influence of the most extreme members on the others. There is, however, an open debate on defining the best approach since both bring different characterizations of group processes.

As cohesion is a group emergent state, we believe it is worth investigating it through the group as a whole as well as from its individuals. Hence, group modeling constitutes one of our research axes (i.e., RA2) for developing computational models of cohesion.

2.2.5 The Social and Task Cohesion Interplay

As previously mentioned, traditional definitions, theoretical models, and frameworks of cohesion agree on the importance of the Social and Task dimensions of cohesion in group interactions. Although scholars in Social Sciences clearly state that cohesion's dimensions interplay somehow and somewhere over time, some of them argue that Social cohesion emerges first and impacts Task cohesion (e.g., Tuckman, 1965; Grossman et al., 2015). Other ones affirm that, especially at an early stage of group formation, Task cohesion might emerge before Social cohesion, and it could be seen as a shared experience auspicious to group bonding (e.g., Kozlowski et al., 1999). These two opposite points of view might hold depending on many factors (e.g., the nature of the group members and the group's goals). In their work, Severt and Estrada (2015), indeed, highlight that not every group exploits each dimension of cohesion. Moreover, Grossman et al. (2015) state that once Social cohesion appeared, followed by Task cohesion, after a while, a dynamical reciprocal adjustment between the two dimensions occurs, at the expense of Social cohesion. Bartone and Adler (1999) and Siebold (2006) also provide evidences in that direction (i.e., Task cohesion increases while Social cohesion decreases). They, however, add that U- and inverted U-shape patterns are observed for the Social and Task dimensions of cohesion, respectively, over the lifespan of a military group, implying dynamical adjustments over time. These results highlight that these dimensions have a reciprocal impact on each other, opening a third way to study the interplay between these dimensions.

2.2.5.1 From Social Cohesion to Task Cohesion

Early work by Tuckman (1965) on small groups development suggests that cohesion is part of the life cycle of a group and that the social aspects of cohesion develop first. Empirical work confirmed and extended Tuckman's hypothesis (e.g., Zurcher Jr, 1969; Runkel et al., 1971) stating that groups go through the stages of "*forming*", "*storming*", "*norming*", "*performing*", and, finally, "*adjourning*" (Tuckman and Jensen, 1977). During the forming, group members develop social bonds and get to know each other, while, in the storming, they start learning about each others' strengths and weaknesses, leading to the definition of their roles. Such a categorization of the different stages of a group encourages to consider the Social dimension as a potential driver for the Task dimension. Moreover, Carron and Brawley (2000) state that all dimensions are not equally present

across groups and that some dimensions might be more salient depending on the developmental phase of the group (e.g., a newly formed group vs. a group of friends), and the specific interaction settings such as a meeting. In addition, the influence of a dimension is likely to change gradually over time. In their study, they also conclude that, in particular contexts (e.g., in social groups), Social cohesion would be more salient. Grossman et al. (2015) support the predominance of Social cohesion in social groups and argue that Social cohesion emerges first in a group, and sets the stage for Task cohesion, which develops later. Lending further support to the notion that Social cohesion breeds Task cohesion, Severt and Estrada (2015) advanced that Social cohesion facilitates flexible and constructive relationships in groups and teams, hence, promoting Task cohesion.

2.2.5.2 From Task Cohesion to Social Cohesion

While the path from Social cohesion to Task cohesion may be more intuitive from a developmental point of view, the other direction (i.e., Task cohesion influencing Social cohesion) may also occur in group interactions. Prior theorizing has hinted at the possibility that Task cohesion might emerge earlier in a group's developmental trajectory, before group bonding and relationship formation come into play and create shared experiences of Social cohesion (Kozlowski et al., 1999). In earlier stages of team development, task aspects can be more salient than social aspects of a team, which may require an extended period of interaction (Carron and Brawley, 2000). Empirical work indicates support for the notion of task aspects promoting subsequent Social cohesion. A study of youth athletes showed that members of task-focused teams report personal enjoyment and friendship development (Balaguer et al., 2003). Similarly, a study of teams of male college athletes showed that a task-involving team climate predicts aspects of Social cohesion (Boyd et al., 2014). The authors discuss that a task-involving climate can help reduce social barriers, foster interdependence, and trigger positive social interactions, which paves the way for Social cohesion. While it remains to be seen whether these findings extend to other types of groups with, for example, a more heterogeneous gender distribution, they highlight a possible direction of influence from Task cohesion to Social cohesion, as opposed to previous findings derived from Tuckman's hypothesis on small groups development.

While opposite points of view emanate from the Social Sciences literature regarding the way the Social and Task dimensions of cohesion interplay, there is no doubt that such relationships impact cohesion dynamics. We believe that computational models would benefit from integrating such an interplay. Thus, this constitutes another research axis (i.e., RA3).

2.2.6 Relationships between Cohesion and Other Group Processes

Cohesion is one of the most studied emergent state in Social Sciences (Rosh et al., 2012) as many studies focused on investigating the links between cohesion dimensions and other group processes such as collective efficacy (e.g., Spink, 1990; Zaccaro et al., 1995; Paskevich et al., 1999; Estabrooks and Carron, 2000b; Kozub and McDonnell, 2000), emotions (e.g., Lawler and Yoon, 1996; Barsade and Gibson, 1998; Lawler et al., 2000; Thye et al., 2002; Zheng et al., 2015) or leadership (e.g., transformational leadership,

Light Shields et al., 1997; López-Zafra et al., 2008; Callow et al., 2009; Vincer and Loughead, 2010; Smith et al., 2013). A takeaway from these studies is that, during an interaction, a group process linked to cohesion could lead to the emergence and development of cohesion or, oppositely, could result from the emergence and dynamics of cohesion.

Integrating the links between cohesion and group processes related to cohesion into its computational models appears necessary given their impact on cohesion. Thus, our fourth research axis (RA4) consists of addressing such an open challenge.

2.3 A Structured Survey for the Automated Analysis of Cohesion

In this Section, we first introduce the studies addressing the automated detection of cohesion. Then, we present how the survey is organized and how the studies fit into it by presenting the different approaches they explored.

2.3.1 Automated Studies on Cohesion Using Nonverbal Features

Hung and Gatica-Perez (2010) were the first to include both audio and video nonverbal descriptors to computationally investigate cohesion in a meeting context using the AMI dataset (Carletta et al., 2006). They also collected annotations of cohesion provided by external observers to establish a reference for evaluating automated methods. Even if the questionnaires used to assess cohesion were based on a multidimensional conceptualization of cohesion, the models presented in this study focused on predicting cohesion as a whole, without distinguishing between its Social and Task dimensions. They extracted features from audio and video data and, all the features but the ones related to turn-taking were computed from individuals. Some of the individual features were amalgamated to reflect their distribution over the group as a whole. Their results showed that using an SVM classifier, the best performing features to estimate high and low levels of group cohesion during meetings were the following: the total pause time between each individual's turns during a meeting segment (extracted from the audio), the total visual activity for each person in the meeting (extracted from the video), and the visual activity during periods of overlapped speech (extracted from both the audio and video). With these features, their SVMs reached 90%, 83%, and 82% classification accuracy, respectively. Nanninga et al. (2017) recently extended this work, integrating pairwise and group descriptors related to the alignment of para-linguistic speech behavior (e.g., the Mel Frequency Cepstral Coefficients, speech rate) to study if these could improve estimations of Social and Task dimensions of cohesion in a meeting setting. They found that such kind of descriptors outperform the traditional turn-taking-based descriptors (i.e., the ones used in Hung and Gatica-Perez, 2010), for Task cohesion. With the same experimental setting, they reached a mean area under the ROC curve (AUC) of 0.64 by combining both types of features as opposed to an AUC of 0.53 when using turn-taking related features only. The authors also evaluated the performances of two supervised classification methods (a Gaussian Mixture Model and a Kernel Density Estimation) fed with nonverbal features combining mimicry-, synchrony- and turn-taking-related. Results show that these models, in a meeting setting,

performed well for classifying the Social dimension of cohesion (low or high), for which they achieved a performance of 0.71 Area under the ROC Curve (AUC). Concerning the Task dimension of cohesion, they managed to reach a performance of 0.64 AUC. These results confirm that quantifying mimicry is useful for automatically assessing cohesion, especially for its Social dimension and suggest that Social cohesion is more clearly expressed by behaviors in general than Task cohesion. In this study, however, they did not focus on how the Task and Social dimensions are related to each other over time.

Kantharaju et al. (2020) also investigated cohesion in a meeting context. As in the previously mentioned studies, they present a multimodal analysis of cohesion using 16 two-minute segments from the AMI dataset. They, however, explored how both verbal (e.g., dialogue acts) and nonverbal (e.g., laughter duration) features are related to high and low cohesive segments. Their results indicate that the occurrence of some nonverbal features (i.e., laughter and interruption) are higher in high cohesive segments and that verbal features did not have an impact on the level of cohesion by itself. This is in line with previous works in other group contexts showing that nonverbal communication is a more powerful predictor of cohesion than verbal behavior (e.g., Kubasova et al., 2019; Alsulami, 2021).

Lately, Dhall (2019) provided a bench-marking platform to investigate methods on affect labeled data through the EmotiW challenge. In this challenge, researchers implemented different Deep Neural Networks to predict group cohesion from images (e.g., Zhu et al., 2019; Wang et al., 2020) taken from the GAF 3.0 dataset (Dhall et al., 2017). For all the EmotiW-related studies, the aim is to predict a group cohesion score comprised between zero and three included. Zhu et al. (2019) proposed a hybrid network including regression models which are separately trained on face features, skeleton features, and scene features. Then, they fused each regression value into a final layer predicting the group cohesion score. They reached a Mean Square Error (MSE) of 0.44, outperforming the baseline MSE of 0.50. In their study, Wang et al. (2020) developed a deep neural network (DNN) architecture that takes both an image and its textual description. Such an approach aims at picking-up extra information contained in the textual description to improve cohesion prediction. They reach an MSE of 0.47, hence, showing that their approach is outperforming their baseline (i.e., the same DNN without the image description).

All of the previously mentioned studies, however, implemented their models without considering the temporal aspect of cohesion. They, indeed, considered each sample (e.g., images, segments of meeting videos) independently. In their longitudinal study (i.e., a study where the same individuals or groups are repeatedly examined to detect any changes that might occur over a period of time), Zhang et al. (2018) addressed the temporal nature of cohesion by studying small groups collaborations during long-duration missions in confined spaces with the use of sociometric badges. These can be anything placed on a person or on its phone, that is able to track the person's movement and activity. In order to recognize group members' affect states and group cohesion (i.e., through its Social and Task dimensions), they collected and analyzed data from a group of six members involved in a 4-months simulation of a space exploration mission. They defined cohesion detection as a binary classification problem (negative or positive) and they used features in their models both from individual members and the group as a whole. Their results show that Task cohesion can be correctly classified with a high performance of over 0.80 AUC.

The fact that Task cohesion is well classified (as opposed to previous studies) could be explained by the task-driven nature of the group. Also, an interesting conclusion from this study is that quantifying behavior patterns including dyadic interactions and face-to-face communications is important in assessing the group process. Results are promising and show the benefits of integrating the temporal nature of cohesion in computational models. These results, however, concern a quite specific scenario. It remains to be seen whether they apply to other types of groups (e.g., social groups).

While the aforementioned studies focused on analyzing and predicting cohesion only, other studies also introduced computational architectures to jointly predict cohesion with other related processes in order to take advantage of their interplay. Wang et al. (2012), for example, conducted a study about the joint prediction of leadership and cohesion based on verbal features in multiparty dialogues and broadcast conversations in English and Mandarin. Using AdaBoost algorithm (Freund and Schapire, 1997), they achieved F1-scores ranging from 0.73 to 0.95 for leader detection and around 0.80 F1-score for group cohesion detection across multiple datasets. Except for this pioneering study, it is only very recently that researchers in Computer Sciences started again exploring the relationships between cohesion and other group processes.

Fang and Achard (2018) attempted to link specific audio and video nonverbal features as well as personality traits to cohesion using segments of meeting interactions from the ELEA dataset (Sanchez-Cortes et al., 2011b). They treated the problem of cohesion prediction as a binary classification problem and clustered their 2-minute long videos into high or low levels of cohesion. They used a Ridge Regression (Hastie et al., 2009) to classify each video and showed that speech turn and variation of speech energy are related to cohesion and that the Big Five Personality Trait (John, 1990) "*Agreeableness*", is highly correlated to cohesion compared to other personality traits. This study, however, does not explicitly integrate the relationships between cohesion and these personality traits in their models. Furthermore, interactions from this dataset are scripted, hence, conclusions should be interpreted with care as they might differ in a natural unscripted interaction setting.

As part of the EmotiW challenge (Dhall, 2019), researchers implemented various methods to jointly predict emotion and cohesion's level in images (e.g., Guo et al., 2019; Xuan Dang et al., 2019; Gavrikov and Savchenko, 2020; Ghosh et al., 2022; Zou et al., 2020; Tien et al., 2021) and videos (e.g., Sharma et al., 2019), using DNNs and the same experimental settings (see Dhall, 2019, for more details). Thus, results are comparable across the following studies. In their work, Ghosh et al. (2022) introduced a DNN to jointly predict cohesion and emotion that uses the whole image as input and that is composed of a pre-trained Inception V3 (Szegedy et al., 2016) model. It classifies cohesion as a regression task (between zero and three) and emotion as a 3-classes classification (positive, negative, neutral). Results show that, when using group cohesion as a secondary task, it helps increase the performance for group emotion prediction. This study sets the baseline for the other works that attempted to jointly predict both group emotion and cohesion. In fact, Guo et al. (2019) jointly trained the group cohesion prediction task with the group emotion recognition task using a multi-task learning approach. They tested their models with different visual features (i.e., extracting only faces or only bodies from images or the whole image, respectively). They also designed two different losses (i.e., a rank loss and an hourglass loss) and achieved an MSE of 0.44 using the whole images. Similarly,

Zou et al. (2020), presented a hybrid deep learning network for the prediction of group emotion and level of cohesion from images. They first used a model to classify emotions according to their valence (positive, neutral, negative) and used the model's output into a regression layer to predict the cohesion level (between zero and three). They also implemented a multitask loss to merge the regression task (i.e., the prediction of the level of cohesion) with the classification task (i.e., emotion prediction). They reached a classification accuracy of 74.80% for the prediction of the valence of emotion, and an MSE of 0.70 for their cohesion regression task. Xuan Dang et al. (2019) also designed a custom loss for predicting group cohesion from images. In order to integrate the influence of emotion on cohesion, they designed a custom weighted loss. Tien et al. (2021) extended this study and explained in depth their DNN model. They exploited four types of visual features: the scene, skeletons, UV coordinates (also known as texture coordinates which define a map of a 2D image onto a surface in 3D space, Hughes et al., 2014) and faces from the images, along with convolutional neural networks (CNNs). DNN's architecture is composed of one independent branch for each of the visual features and one additional branch that results from the concatenation of all of the branches except the one extracting faces. Then, the five branches are concatenated into a final layer to predict the group cohesion score. With their architecture, they managed to reach an MSE of 0.42. Finally, there is, to the best of our knowledge, only one study that attempts to jointly predict group emotion and cohesion from videos. Sharma et al. (2019), indeed, designed a multimodal DNN based on the inception V3 pre-trained model (Szegedy et al., 2016). The videos used in this study largely differed, for example, in scenarios and poses, making it difficult for the model to capture the dynamics of group emotion since it was trained on images. Their model predicts the valence of emotion (i.e., positive, neutral, negative) with 47.50% accuracy and predicts cohesion with an MSE of 0.80.

While these approaches computationally confirm Social Sciences' insights regarding the relationships between cohesion and other persons' characteristics (e.g., personality traits), cohesion and leadership as well as cohesion and emotion, they all defined cohesion without distinguishing between its dimensions. A multidimensional approach would help advance further our understanding of the way these group processes interplay over time.

2.3.2 Organization of the Survey

Given the relatively small literature on the automated analysis of cohesion, we provide a structured survey that highlights the various approaches employed to develop computational models of cohesion. Some of them could also be applied to the automated analysis of other group processes and emergent states. The survey is organized following the four research axes identified in the previous Section (i.e., Temporal nature of cohesion, Interplay between dimensions, Group modeling, and Relationships with other group processes). In addition, inspired by the IPO theoretical framework (Hackman and Morris, 1975) and its subsequent enhancements that incorporate iterative feedback loops (e.g., Kozlowski et al., 1999; Ilgen et al., 2005), we clustered the approaches according to three levels: "*Input*", "*Model*" and "*Output*". In the IPO theoretical framework, input refers to the antecedents that could influence the group, directly or indirectly while, in our categorization, the inputs of a computational model are either the raw data (in the case of an end-to-end deep neural network) or the extracted features. Process refers to the ac-

tivity that mediates the relationships between the inputs and the group's outcomes (i.e., cohesion). In our categorization, we replaced it with the Model level, which corresponds to the architecture of the computational model. Finally, in IPO, the outputs are considered as the "*consequences of the group's actions*" (Forsyth, 2012) which, in the context of a computational model, refer to its purpose (e.g., predicting cohesion for a specific dimension).

The approaches that are presented in this survey could be applied independently of the environment in which the interactions take place (e.g., real, virtual, or mixed) and of the technology from which the signals are extracted (e.g., video, audio, motion capture). In fact, cohesion has been traditionally automatically investigated in a real-world setting. With the advent of new technologies and the actual world context (e.g., health crisis), more and more tools are, however, developed to encourage people to meet and gather virtually (e.g. virtual and hybrid conferences). Thus, researcher in Computer Sciences, started investigating emergent states (e.g., Curşeu, 2006; Moustafa and Steed, 2018) and cohesion in particular (e.g., Torro et al., 2022) in virtual environments as well as collaborative systems in mixed environments (see Ens et al., 2019, for a review). While the social interactions of virtual and mixed groups usually follow the conventions of the real world (e.g., keeping distances from each other, turning to face the conversation partners), multiple studies (e.g., Axelsson, 2002; Salinäs, 2002) highlight the differences in the group members' behavior (e.g., people use fewer gestures, Schroeder, 2002). In addition, new technologies extended the range of human signals available to capture, with increasing precision. While the most common data consist of video and audio, human signals are also captured through devices such as electroencephalogram (EEG) (Soroush et al., 2017), electrocardiogram (ECG) (Santamaria-Granados et al., 2018), motion capture (Kapur et al., 2005) and so on. Thus, our survey only provides approaches that are independent of the environment (i.e., they can be implemented in real, virtual, and mixed environments) and of the signals extracted. Finally, it is worth mentioning that such approaches could be complementary. In fact, for each level, multiple approaches to investigate cohesion through more than one research axis could be applied. Figure 2.6 shows the structure of our survey according to the four research axes and the three levels at which approaches could be applied (Input, Model, Output). Existing approaches are classified according to their complexity of integration into a computational model (i.e., the darker the color is, the more complex it is). All the approaches in bold are the ones we implemented in Chapter 5 and Chapter 6.

In the following, each approach employed by the previously mentioned studies is presented, at each level, and for every research axis. New approaches introduced in this Thesis are also motivated and detailed.

2.3.3 Approaches Employed at the Input Level

2.3.3.1 Addressing the Temporal Nature of Cohesion (RA1)

At the Input level, time is addressed by choosing the most appropriate segmentation of the data. This choice depends on the goal of the study, the eventual technical constraints, or the specific research questions (Lehmann-Willenbrock and Allen, 2018). Such a task remains, however, complex but crucial to catch the multiple facets of cohesion (Ceccaldi

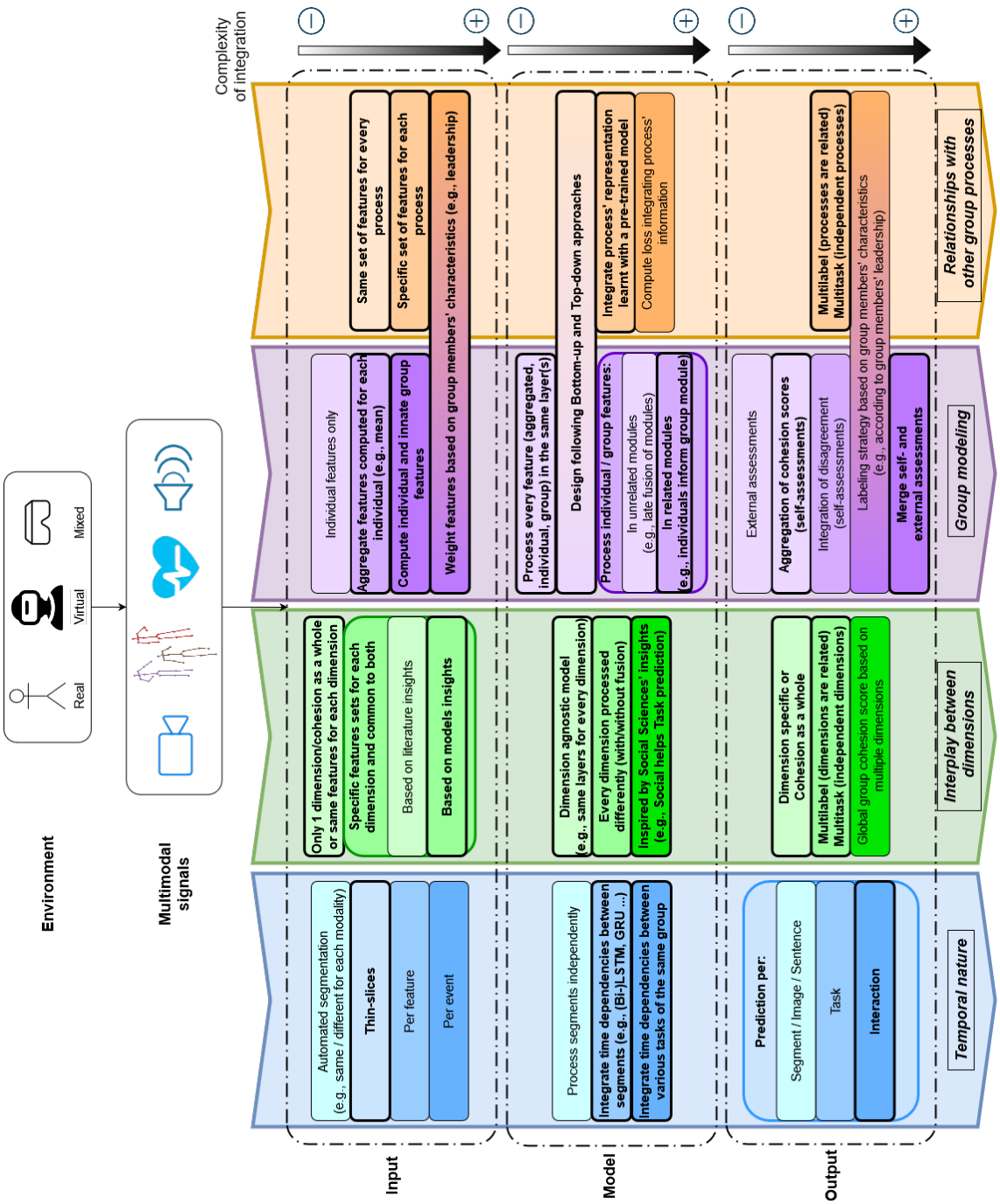


Figure 2.6: Structured survey of approaches for the automated analysis of cohesion. We suggest investigating cohesion at *Input*, *Model* and *Output* levels of the computational models, following four research axes: *Temporal nature* (in blue), *Interplay between dimensions* (in green), *Group modeling* (in purple) and *Relationships with other group processes* (in orange). For each axis, multiple approaches are presented according to their complexity of integration. Approaches framed in bold are the ones implemented in Chapter 5 and Chapter 6.

et al., 2019). In fact, many approaches exist to segment the data and we only present the ones that had been used in the previously mentioned studies or that are promising for the automated analysis of cohesion.

One of the simplest, but time-efficient approach, consists of extracting the features on consecutive or overlapped fixed-length windows (i.e., automated segmentation). The length of the window as well as the duration of the overlap can be similar across all the modalities or specific to the modalities' characteristics. Another approach exploits a *thin-slice* strategy (Ambady and Rosenthal, 1992). This refers to the process of making very quick inferences about the individual and/or group processes with a minimal amount of information. According to Ambady and Rosenthal (1992), fixed-length time windows of behavior from 2 seconds to 5 minutes are deemed to provide an efficient assessment of personality, affect, and interpersonal relations. This approach is extensively used in Social Psychology and is leveraged in some computational studies (e.g., Hung and Gatica-Perez, 2010; Nanninga et al., 2017). These strategies have the advantage of being easily automatized and fast to compute and, in the case of automated segmentation, do not require prior knowledge of the content of the interaction. They, however, might break the interaction in the middle of a social signal (e.g., during a turn), leading to a potential loss of meaning. Overlapped windows help reduce such a risk but increase the amount of redundant information and make the process of integrating time dependencies at Model level harder due to the fact that the natural time dependencies between two segments no longer exist.

Another approach is to segment the data per relevant features. It implies that every feature of interest is annotated throughout the whole interaction and that only the segments of interest are selected. This methodology usually requires a coding scheme. The most popular one that has been applied for the automated study of cohesion is ACT4Teams (Kauffeld et al., 2018). It was initially designed for measuring problem-solving dynamics in groups but has been applied to annotate Social and Task cohesion from audio content (Nanninga et al., 2017). Despite the convenience of being tailored for a specific objective (i.e., automatically analyzing cohesion), this approach is time-consuming and often requires annotators to be trained on the methodology. In a similar vein, another approach consists of segmenting the interaction per *event* which is, according to Zacks and Tversky (2001), "*a segment of time at a given location, that is conceived by an observer to have a beginning and an end*". Such an approach leverages the Event Segmentation Theory (EST) that exploits the innate ability of human beings to parse an ongoing interaction into meaningful units (Zacks and Swallow, 2007). Such an approach has been used by Ceccaldi et al. (2019) to explore how it affects external observers' annotation of Social and Task cohesion. Their results reflect more variability in cohesion in different interactions as compared to traditional automatic and continuous types of segmentation and provide hints to automate this segmentation.

2.3.3.2 Addressing the Group Modeling (RA2)

Concerning RA2, three approaches had been identified at the Input level. The first one, which is the approach used in most of the computational studies of cohesion using multiple modalities, consists of aggregating features computed from individuals to produce group features (e.g., Fang and Achard, 2018; Zhu et al., 2019). In these studies, all the features extracted from the video were first computed from individuals and then, amal-

gamated to approximate the distribution of the feature for the group. A more complex approach consists of computing innate group features (i.e., features that are computed over the whole group). This is done, for example, for all the turn-taking-related features in the studies presented by Hung and Gatica-Perez (2010) and Nanninga et al. (2017). Extracting innate group features from different modalities would help capture social signals that are induced by individuals' interaction (e.g., F-formations, Kendon, 1990). Such innate group features are extracted (see Chapter 4) and used in the computational models (see Chapter 5 and Chapter 6). Lastly, an approach that addresses both RA2 and RA4 consists of taking into account group members' characteristics at the Input level by weighting individual or group features depending on the process studied. For example, the features of a leader could be amplified to accentuate the differences with its followers. The last approach is implemented in Chapter 6.

2.3.3.3 Addressing the Interplay between the Social and Task Dimensions (RA3)

To the best of our knowledge, there is no automated study on cohesion interested in the interplay between cohesion's dimensions and, in particular, between its Social and Task dimensions. Hence, when cohesion is considered as a multidimensional process, the same set of features is used to infer Social and Task cohesion (e.g., Nanninga et al., 2017; Zhang et al., 2018). While features might contain important information for predicting both dimensions (e.g., features related to touch can help communicate task-related information as well as convey social status and emotions, Saarinen et al., 2021), the Social Sciences literature suggests that particular behaviors are particularly relevant for studying either Social or Task cohesion. For example, big overall posture expansion is positively correlated to Social cohesion (Weisfeld and Beresford, 1982) while overlapping of speeches is usually a sign of engagement in the task (Hilton, 2016). Thus, an approach consisting of building multiple features sets that either characterize both dimensions or each dimension, specifically, could help improve predictions for both dimensions. These features sets, indeed, could be processed differently at the Model level. How to compose such features sets could be done by leveraging Social Sciences' insights or by running post-hoc analysis on existing trained computational models of cohesion to select the most important features. The latter approach is addressed in Chapter 5 and requires different techniques depending on the nature of the model. For example, decision tree algorithms offer importance scores based on the reduction in the criterion used to select split points (e.g., Gini or entropy) while techniques based on Shapley values could be used to explain Deep Learning models (e.g., SHapley Additive exPlanations values, Lundberg and Lee, 2017).

2.3.3.4 Addressing the Relationships with Other Group Processes (RA4)

The approaches following RA4 share some similarities with RA3. Cohesion (and its dimensions), indeed, have relationships with other group processes (e.g., group emotion). Among the existing studies exploiting information from other group processes to predict cohesion (e.g., Fang and Achard, 2018; Xuan Dang et al., 2019; Ghosh et al., 2022), they all use the same set of features for both processes, with the exception of Wang et al. (2012) that distinguished between leadership features and cohesion features. As for studying the interplay between cohesion's dimensions, one approach could consist of defining vari-

ous features sets that are common to all the processes predicted by the model or that are specific to each process. In that way, this will enable the model to process features differently depending on its architecture. All of the approaches addressing RA4 at Input level presented in the structured survey are explored in Chapter 6.

2.3.3.5 Summary

Many approaches could be implemented at Input level. As summarized in Figure 2.7, the current literature on the automated analysis of cohesion usually employs approaches that are simpler to integrate into computational models of cohesion. Since automatically analyzing cohesion is a complex task, empirical analysis will provide more insights into the most relevant approaches to apply.

At Input level, each research axis, has room for improvement, especially for studying the interplay between Social and Task cohesion where there are only a few computational studies that make a distinction between its Social and Task dimensions (e.g., Nanninga et al., 2017).

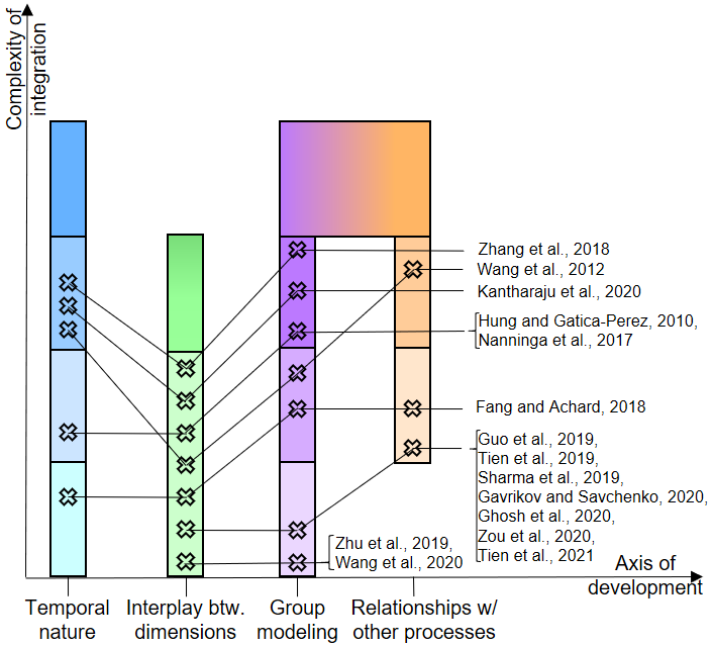


Figure 2.7: Overview of the approaches employed at the Input level, according to their complexity of integration. At this level, most studies apply simple strategies, especially with respect to the second RA. Only a few, indeed, make a distinction between the Social and Task dimensions of cohesion. For more details about the approaches, refer to the corresponding color in Figure 2.6.

2.3.4 Approaches Employed at the Model Level

2.3.4.1 Addressing the Temporal Nature of Cohesion (RA1)

At the Model level, RA1 has only been addressed by employing a simple approach consisting of processing each segment (e.g., Hung and Gatica-Perez, 2010; Nanninga et al., 2017) or images (e.g., Zhu et al., 2019; Xuan Dang et al., 2019; Gavrikov and Savchenko, 2020) independently, without integrating the potential time dependencies that exist from samples of the same interaction. Thus, a first approach consists of integrating these short-time dependencies between consecutive windows in the model. For example, this can be addressed through DNNs architectures. Multiple layers such as Gated Recurrent Unit (GRU), Long short-term memory (LSTM), and Bidirectional Long short-term memory (Bi-LSTM), indeed, allow to process entire sequences of time series data. In the case a group performs multiple interactions (e.g., accomplishing multiple tasks), a more advanced approach involves capturing the long-time dependencies that may exist between the different interactions by integrating elements from the past (e.g., from the previous interaction) to give context to the current interaction. The two last approaches are implemented in Chapter 5 and Chapter 6.

2.3.4.2 Addressing the Group Modeling (RA2)

We identified three main approaches at the Model level. The first one, which is currently applied in all the computational studies of cohesion, consists of processing every feature together, whether it is computed from the individuals or the group. A more advanced approach consists of processing individual and group features differently to learn both individual and group contributions to the interaction. This could be done in two ways. Firstly, individual and group features can be processed independently and fused only before the model's predictions. Secondly, to account for the relationships between individual and group manifestations of cohesion, individual features could be first processed to inform the representation learned from group features. Such approaches are explored in Chapter 5. Another approach that also concerns RA4, consists of considering cohesion (and other group processes) from a *Top-down* or *Bottom-up* approach. Top-down focuses on the group as a whole. This means that group dynamics influence the feelings and behaviors of members of the group. Bottom-up approximates the group as the sum of its parts. This approach led researchers to examine the group through a variety of compositional perspectives such as the mean of the group's members. To the best of our knowledge, only Ghosh et al. (2022)'s study acknowledges these Top-down and Bottom-up approaches to define both group emotion and cohesion. There is, however, an open debate on defining the best approach. As both the Top-down and the Bottom-up approaches bring different characterizations of group processes such as group emotions, Barsade and Gibson (1998) recommend exploring methods following both these approaches to have a complete picture of the process. Literature on cohesion, and group emotion in particular, highlight the importance to consider them from both individual and group perspectives (Braun et al., 2021). Both Top-down and Bottom-up approaches are investigated in Chapter 6.

2.3.4.3 Addressing the Interplay between the Social and Task Dimensions (RA3)

Concerning RA3, the same architecture could be used for predicting each dimension, independently. This approach is employed in Nanninga et al. (2017) to predict the Social and Task dimensions of cohesion. In this study, both dimensions are predicted separately, hence, implying that the relationships between the Social and Task dimensions are not taken into account. Since only a few computational studies are differentiating between the Social and the Task dimensions (e.g., Nanninga et al., 2017; Zhang et al., 2018), the other approaches remain to be explored. The first one consists of processing each dimension differently (e.g., by using different layers) according to the dimensions' specifics (e.g., Social cohesion might require different model parameters). Finally, the design of the model could be inspired by Social Sciences' insights on the dynamics of cohesion in small groups. Various theories on small groups development, indeed, exist (see Section 2.2.5.1), hence, opening different ways to integrate the Social and Task interplay (e.g., Social cohesion helps predict Task cohesion or inversely). All of these approaches are investigated in Chapter 5 and Chapter 6.

2.3.4.4 Addressing the Relationships with Other Group Processes (RA4)

Relationships with other group processes could be studied through two extra approaches at the Model level. The first one requires to initially train a model for a specific group process (e.g., leadership) and to use it within the cohesion model. In that way, the pre-trained model could be used directly within the computational model architecture to help learning contextual information. This approach implies, however, having computational resources to pre-train a model on a specific process of interest. This approach is implemented in Chapter 6. The second approach follows Tien et al. (2021)'s study in which they computed a multitask loss based on both emotion and cohesion losses. Similarly, other loss functions could be developed to integrate other group processes.

2.3.4.5 Summary

To summarize, many paths could be undertaken at Model level. Figure 2.8 shows the current state of the literature on automated analysis of cohesion. As for the Input level, studies first focused on the simpler approaches.

At Model level, most of the approaches remain to be explored. This is primarily due to the fact that DNN architectures, that offer more flexibility in the architecture design, are, to the best of our knowledge, under-investigated for the automated analysis of cohesion.

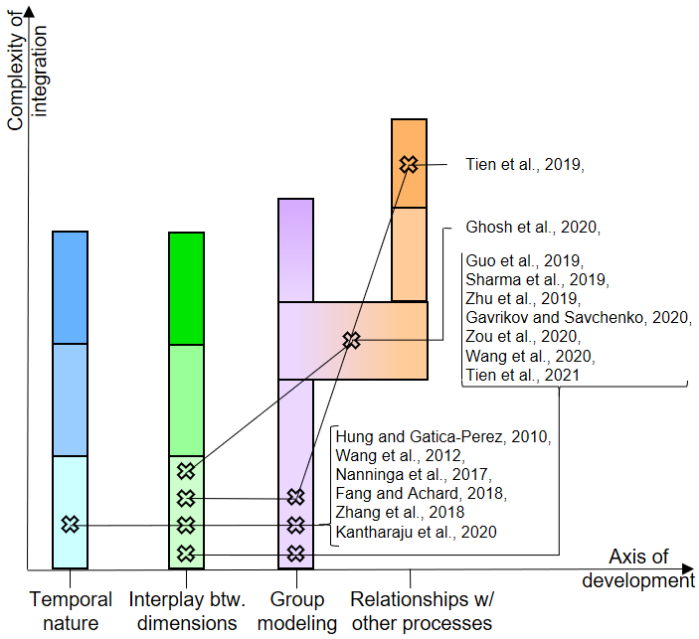


Figure 2.8: Summary of the approaches implemented at the Model level, according to their complexity of integration. At this level, most studies applied simple strategies, especially regarding the first three RAs (i.e., time dependencies between segments, the interplay between Social and Task cohesion and the group modeling were not investigated). For more details about the approaches, refer to the corresponding color in Figure 2.6.

2.3.5 Approaches Employed at the Output Level

2.3.5.1 Addressing the Temporal Nature of Cohesion (RA1)

When designing computational models of cohesion, its aim should be clearly defined as it would impact its design and the approaches employed at every level of the structured survey. It is, indeed, crucial to define how the model should handle the temporal nature of cohesion. In fact, it can predict cohesion for each segment (e.g., time window, image), as in all of the previously mentioned studies on the automated analysis of cohesion, for a specific task or interaction (i.e., a part of the overall group interaction), or for the whole interaction. In all of the models presented in Chapter 5 and Chapter 6, we predict the cohesion dynamics for multiple tasks, taking into account the whole interaction.

2.3.5.2 Addressing the Group Modeling (RA2)

Modeling the group at the Output level is a complex task. As of today, most of the computational studies on cohesion rely on external assessments of cohesion to provide labels for each interaction (e.g., Hung and Gatica-Perez, 2010; Zhu et al., 2019; Zou et al., 2020). Collecting these external assessments is time-consuming and prone to errors and subjectivity. In studies where there are two or more raters, statistics such as Intraclass Correlation Coefficient (ICC), are applied to make sure ratings are reliable, for each annotated interaction. While such a labeling strategy based on external assessments is well defined and broadly used, the use of either internal or external observations of a group in

terms of cohesion remains an open debate (Casey-Campbell and Martens, 2009). There are, however, to the best of our knowledge, only a few computational models based on self-assessments of cohesion (i.e., Wang et al., 2012; Zhang et al., 2018). This is probably due to the fact that no dataset designed specifically for the automated analysis of cohesion exists, hence, making it impossible to collect self-assessments during the interaction.

The structured survey presents four labeling strategies based on self-assessments. The easiest to implement consists of aggregating the cohesion scores collected for each individual (e.g., by taking the mean of individual scores) to produce a group cohesion score. Such an approach is used in our models (see Chapter 5 and Chapter 6). Most of the questionnaires used to assess cohesion, however, contain questions that concern the individual towards the group as well as the group as a whole (e.g., "I was unhappy with my team's level of desire to win" and "Our team did not work well together", respectively, from the GEQ questionnaire). Thus, disagreements in the perception of cohesion can occur between the group members. These could be taken into account in the labeling strategy to compute more nuanced or robust labels.

Another approach would consist, if available, of merging both self- and external assessments of cohesion to produce a "*true*" label. Both types of assessments, indeed, have pros and cons (Vinciarelli and Mohammadi, 2014). Self-assessments of cohesion might be over-optimistic since group members tend to provide ratings towards socially desirable characteristics. External assessments reflect the behavior that people adopt toward others, without necessarily corresponding to their true internal state (Uleman et al., 2008). Implementing a labeling strategy able to handle both types of assessment could help develop more robust computational models. Such an approach is explored in Chapter 5.

Finally, similarly to an approach described at the Input level, we could build labels for group modeling that integrate the relationships with other group processes. Labels could, indeed, be based on group members' characteristics such as leadership. For example, more weight could be given to the cohesion ratings provided by the leader of the group. Such a strategy would help compute more robust labels since it would integrate, at the Output level, extra information that is relevant to cohesion.

2.3.5.3 Addressing the Interplay between the Social and Task Dimensions (RA3)

For this research axis, it must be decided at which granularity cohesion is defined. Cohesion can, indeed, be defined as a whole, without any distinction between its dimensions (as in most of the computational studies on cohesion, e.g., Hung and Gatica-Perez, 2010; Fang and Achard, 2018; Wang et al., 2020; Kantharaju et al., 2020) or, oppositely, as a multidimensional construct (as in Nanninga et al., 2017). In the latter case, the output of the model can (1) only predict a specific dimension (cf. Nanninga et al., 2017), (2) predict both dimensions, and (3) predict cohesion as a whole, building on the combination of multiple dimensions (e.g., Social and Task cohesion). The first approach is the simplest to implement. The second one depends on how we consider the relationships between the dimensions. In fact, a multilabel setting implies that both dimensions are strongly related as both dimensions would be predicted from the same model or layer in the case of a DNN architecture. These two approaches are implemented in Chapter 5 and Chapter 6. A multitask approach, however, suggests that both dimensions are less related since they are predicted independently and could potentially share almost no layers. The last approach

would allow the model to predict a cohesion score that already takes into account the interplay between both dimensions, hence, easing the implementation of more complex computational models. This approach remains to be addressed.

2.3.5.4 Addressing the Relationships with Other Group Processes (RA4)

In addition to the last approach presented in RA3, another one consists of jointly predicting cohesion alongside another related group process. Thus, as for RA2 at the Output level, this can be done in a multilabel or a multitask setting. In these ways, cohesion would benefit from the knowledge learned from another process. That is the approach exploited by Guo et al. (2019), Xuan Dang et al. (2019), Ghosh et al. (2022) and Zou et al. (2020). In Chapter 6, we also implement such an approach to jointly predict cohesion and group emotion.

2.3.5.5 Summary

At Output level, the choice of the appropriate approach may vary a lot depending on the aim of the model. In fact, depending on the application, some approaches from the structured survey may not be applicable (e.g., if self-assessments are not available). Figure 2.9 highlights what are the approaches employed at Output level in the current literature on automated analysis of cohesion.

As at previous levels, most of the time, approaches that are easier to integrate into the models are employed. Thus, it remains to be seen to what extent each approach described in this structured survey contributes to improving computational models of cohesion.

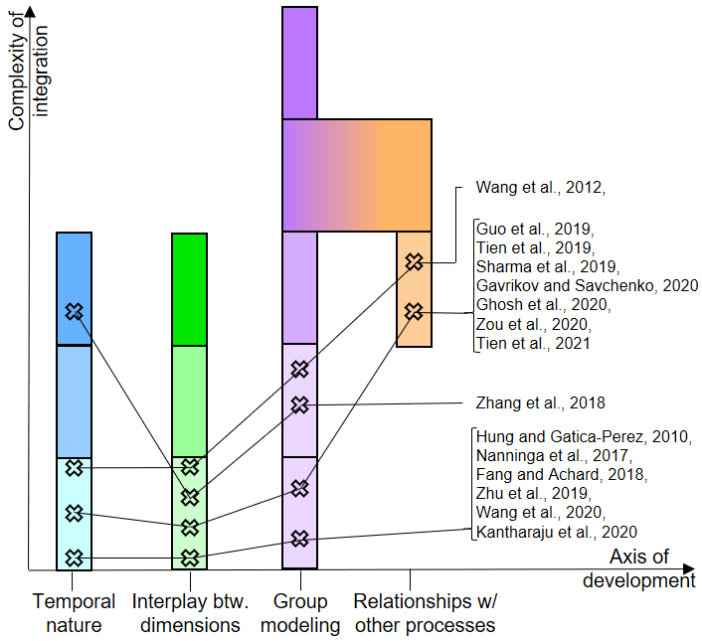


Figure 2.9: Overview of the approaches applied at the Output level, sorted according to their complexity of integration. For RA2 and RA4, approaches employed at this level are the easier to integrate, due to the characterization of the group processes of study. For more details about the approaches, refer to the corresponding color in Figure 2.6.

2.4 Conclusion

COHESION is a complex group affective emergent state that evolves over time. Multiple definitions, models, and theoretical frameworks exist, with a various number of dimensions. Scholars in Social Sciences, however, all acknowledge the existence of its Social and Task dimensions. In the remaining of the Thesis, the work is based on the Severt and Estrada (2015)'s framework. Such a framework considers cohesion as a multidimensionnal emergent state that has both Affective and Instrumental functions and that can be studied at different levels (i.e., horizontal and vertical), hence, adapting to various types of groups and contexts. This framework also integrates that cohesion is interrelated with other group processes such as group emotion.

Based on the review of the Social Sciences literature, we identified four research axes to develop computational models of cohesion. These, aim to address the Temporal nature of cohesion (RA1), the Group modeling (RA2), the Interplay between its dimensions (RA3), and the Relationships with other group processes (RA4).

In this Chapter, we also presented the existing work on the automated analysis of cohesion from nonverbal behaviors. We also organized and explained the various approaches implemented in these studies into a structured survey. Each approach is clustered according to a research axis and the level at which it can be applied (i.e., Input, Model, or Output). Such a categorization was inspired by the IPO theoretical framework for studying emergent states.

Finally, given the relatively small literature on the automated analysis of cohesion, we also described, at each level, novel approaches to investigate cohesion through each of the research axes. These are implemented and discussed in Chapter 5 and Chapter 6.

Chapter 3

Data Collection

Contents

IN this Chapter, we first review the existing datasets that are used for automatically studying small groups within the Social Signal Processing community. Then, we introduce *GAME-ON* (*Group Analysis of Multimodal Expression of cohesiON*), a multimodal dataset specifically designed for studying cohesion and for explicitly controlling its variations over time. GAME-ON is composed of more than 11.5h of audio, video, and motion capture data, as well as assessments of various processes. These assessments include self- and external repeated assessments of cohesion, self- repeated assessments of warmth and competence, emotions, leadership, competitivity, and motivation. Then, we perform a data analysis to assess the design of the data collection.

In particular, I co-designed and co-participated in the data collection. I also post-processed motion capture and audio data and ran a part of the statistical analysis of the cohesion questionnaires.

3.1 Available Datasets

The rise of interest in the automatic analysis of human-human interactions, coupled with a lack of publicly available data, led researchers to collect datasets to capture various group

processes (see Čereković, 2014, for a review). Collecting data, especially in a multimodal fashion is, however, a long and costly process that becomes increasingly complex in the context of small groups' interactions, depending on the number of persons and devices required. Most of the publicly available datasets that involve social interactions among at least three persons have been designed either to record social interactions in different environments to improve group and crowd recognition algorithms (see Borja et al., 2017, for a review), or in a specific context such as meetings (e.g., AMI, VACE, ELEA - Carletta et al., 2006; Chen et al., 2006; Sanchez-Cortes et al., 2011b, respectively), conversational groups (e.g., SALSA, MatchNMingle - Alameda-Pineda et al., 2017; Cabrera-Quiros et al., 2021, respectively), working on a task (e.g., MULTISIMO, AMIGOS, WoNoWa - Koutsombogera and Vogel, 2018; Miranda Correa et al., 2018; Biancardi et al., 2020, respectively) or playing games (e.g., Idiap Wolf, Panoptic, MUMBAI - Hung and Chittaranjan, 2010; Joo et al., 2019; Doyran et al., 2021, respectively).

These datasets also differ in terms of (a) the amount of data available, ranging from a few interactions corresponding to 1h of data (i.e., SALSA Alameda-Pineda et al., 2017) to 167 interactions corresponding to 100h of data (i.e., AMI Carletta et al., 2006), (b) the number of persons (e.g., fixed to three in MULTISIMO vs varying between eight and twelve in the Idiap Wolf), and (c) the technology used to capture data. Regarding the last point, most of the datasets consist of video and/or audio recordings of the interactions (e.g., AMI, ELEA, Idiap Wolf, MUMBAI). To capture more diverse and precise data, other sensors have been used such as optical motion capture in VACE, accelerometers in SALSA, wearable devices and identifiers in MatchNMingle and WoNoWa, 360°cameras in MULTISIMO, or electroencephalogram (EEG), electrocardiogram (ECG) and Galvanic Skin Response (GSR) in AMIGOS. In addition, authors of the Panoptic dataset built a complex setup composed of various sensors (i.e., the Massivey Multiview System[1]), dedicated to the recording of social interaction.

Finally, among all of the previously mentioned datasets, ELEA was one of the first to collect data for specifically investigating a group process (i.e., emergent leadership). It, indeed, addresses emergent leadership in groups by using a well-known meeting situation called the "*Winter Survival Task*", a game where two participants have to identify objects (out of a predefined list) that would increase their chances of survival in a polar environment. ELEA, however, did not refer to emergent group states but rather focused on the emergence of individual leaders in group interactions. Nevertheless, it includes self- and/or external annotations (i.e., personality traits, Big Five, leadership, dominance, competence, likeness) that give the opportunity to use such a dataset for other purposes. Similarly, AMIGOS was also designed to study specific processes. It, indeed, focuses on the affect, mood, and personality of individuals and groups and provides a large variety of self- and external annotations (e.g., emotions, valence, arousal, dominance, liking). MUMBAI also provides a significant amount of data (i.e., more than 46 hours available) for automatically studying emotions and expressions. It is extensively annotated with emotional moments and self-assessments of the personality of each group member are also provided. While ELEA, AMIGOS, and MUMBAI explore individual processes in multi-person settings, the WoNoWa dataset is designed to study a group emergent state over time. It provides around 6 hours of multimodal data to study Transactive Memory

[1] http://domedb.perception.cs.cmu.edu/

System (TMS), a group emergent state characterizing the group's meta-knowledge about "*who knows what*". In addition to the video and audio data available, this dataset provides self-assessments of warmth and competence, TMS, and leadership.

Despite the ever-growing number of datasets for studying small groups' interactions and, lately, for studying specific group processes and emergent states, there is, to the best of our knowledge, no existing dataset that explicitly addresses cohesion. Thus, we introduce GAME-ON (Group Analysis of Multimodal Expression of cohesiON), a multimodal dataset specifically designed for studying group cohesion and for explicitly controlling its variation over time. It consists of multimodal (audio, video, and motion capture data) synchronized recordings of small groups (three persons) playing an *escape game*, that is a game where the players, in a limited amount of time, have to escape a room by collaborating and solving puzzles and other tasks. Such a context helped to engage participants in various tasks, hence, eliciting natural behaviors. This dataset is dedicated to the study of cohesion, and more specifically to its instrumental dimensions (i.e., Social and Task) according to the Severt and Estrada (2015)'s theoretical framework of cohesion. Our dataset also provides a significant amount and diversity of data with the use of a combination of two motion capture systems, in addition to HD video and audio recordings. It also contains repeated self-annotations per participant about their perception of cohesion over time as well as external assessments, giving insights into the dynamics of this emergent group state from both perspectives. We also collected assessments about participants' emotions as well as about their perception of leadership and warmth and competence of each group member including themselves.

Table 3.1 shows the main datasets reviewed in this Section, that are used for automatically studying small groups interactions within the Social Signal Processing community. Moreover, it also provides the characteristics of the GAME-ON dataset.

3.2 The GAME-ON Dataset

3.2.1 Data Collection Design

3.2.1.1 The Game

The game scenario is inspired by the rules of Cluedo[2] and is conceived as an *escape game*. Cluedo is a board game where three to six players try to figure out three main facts of a murder: the murderer, the location of the murder, and the weapon used to kill the victim. An *escape game* is a physical social game in which a small group of players is fake locked in a room set up according to a specific theme. The players have to cooperatively discover clues, solve puzzles, and so on to accomplish a specific goal (e.g., escaping, finding an object, or solving a murder) in a limited amount of time. Social games, such as escape games, are, indeed, a form of socially rich multi-party problem solving where people coordinate and like to spend time together to achieve common goals. They have been considered a viable research methodology to address the subtle nuances of human-human communication in several research domains, from Psychology (Freedman and Flanagan, 2017) and neuroscience (Redcay and Schilbach, 2019) to behavioral economics (Van Dijk

[2]See the Cluedo game at https://www.hasbro.com

Table 3.1: A selection of the main datasets used for automatically studying small groups interactions. Datasets are grouped by scenario. Information about their focus, the size of the groups, the recordings' duration, the type of annotation (self and/or external), and the different technologies used to collect the data are provided.

Dataset	Scenario	Focus	Group size	Duration	Annotations *Self (*), External (*)*	Data: Video (HD)	Data: Audio	Data: Motion capture *Inertial*	Data: Motion capture *Optical*	Other
AMI (McCowan et al., 2005)	Meeting	Individual actions, face behaviors, speech	4	167 meetings 100h	Agreements*, disagreements*, dominance*	✓	✓	×	×	×
VACE (Chen et al., 2006)	Meeting	Event interpretation, multimodal signal processing	5	N/A	Speaker segmentation*, speech transcription*, F-formations*	✓	✓	×	✓	×
ELEA (Sanchez-Cortes et al., 2011b)	Meeting	Leadership, non verbal behaviors	3-4	40 meetings ~10h	Personality traits*, Big Five*, perceived leadership*, dominance*, competence*, likeness*, ranked dominance**	✓	✓	×	×	×
SALSA (Alameda-Pineda et al., 2017)	Free Standing Conversational Group	Natural social interactions, F-formations	2-18	1h	Personality*, position*, head*, body orientation*, F-formation*	✓	✓	×	×	ID/RFID, bluetooth, Accelerometers
MatchNMingle (Cabrera-Quiros et al., 2021)	Free Standing Conversational Group, speed dates	Automatic analysis of social signals and interactions	2-8	2h	HEXACO*, Self Control Scale*, Sociosexual Orientation Inventory*, social cues*, social actions* F-formations*	✓	✓	×	×	Wearable devices recording triaxial acceleration and proximity
MULTISIMO (Koutsombogera and Vogel, 2018)	Experiment Solving a quiz	Human-human interactions, groups' multimodal behavior	3	23 sessions ~4h	Personality*, experience*, speaker segmentation*, dominance*, transcripts*, turn-taking*, emotions*	✓	✓	×	×	360°camera, 2 Kinects
AMIGOS (Miranda Correa et al., 2018)	Experiment Watching videos	Affect, personality, mood	4	~9h	Big-Five*, PANAS*, valence**, arousal**, dominance*, liking*, familiarity*, emotions*	✓	✓	×	×	EEG, ECG, GSR
WoNoWa (Biancardi et al., 2020)	Experiment Collaborative tasks	Transactive Memory System (TMS)	3	~6h	Audio/Video features*, Warmth and competences*, TMS*, leadership*	✓	✓	×	×	Wearable identifiers
The Idiap Wolf (Hung and Chittaranjan, 2010)	Game	Deceptive roles, group interaction	8-12	4 groups ~7h	Speaker segmentation*, roles identifications*	✓	✓	×	×	×
Panoptic (Joo et al., 2019)	Game	Capturing social interactions	3-8	65 sequences 5,5h	×	✓	✓	×	×	Massively Multiview System
MUMBAI (Doyran et al., 2021)	Game	Automated analysis of multimodal behavior, expression detection, emotion classification	4	~46h	Game outcome*, player affect*, personality*, game experience*	✓	×	×	×	×
GAME-ON (Maman et al., 2020)	Game	Automated analysis of multimodal behavior, cohesion, non verbal behaviors	3	17 groups ~11,5h (as a group) ~34.5h (as individual)	Cohesion**, leadership*, emotional state*, warmth and competences*	✓	✓	✓	✓	×

et al., 2020) and human-computer interaction (Bonillo et al., 2019). There exist, indeed, several datasets in which social games are exploited as an experimental tool for eliciting socio-affective behavior such as laughter (Niewiadomski et al., 2013), deceptive behavior (Hung and Chittaranjan, 2010) and members' affect and personality (Doyran et al., 2021), or for evaluating interaction capture methods (Joo et al., 2019). At the time of the data collection, there was, however, no dataset specifically designed for studying a specific emergent state, according to Social Sciences' theories and theoretical models. It is only very recently that Biancardi et al. (2020) collected the WoNoWa dataset for automatically studying Transactive Memory System, a cognitive emergent state, through its three dimensions (i.e., specialization of members' knowledge, credibility, and coordination).

In the context of GAME-ON, the game created an engaging experience for the participants, allowing us to collect measurements of the dimensions of cohesion multiple times by naturally breaking the whole interaction into distinct tasks. The game scenario was the following:

During the XIIth century, a brilliant mathematician, student of Leonardo Fibonacci[3]*, was assassinated. His ghost is trapped in a theater. Every year the ghost locks people there asking them to help him to discover* ***who*** *killed him,* ***with what*** *weapon, and* ***where****.*

The scene contains five posters of the suspects, with a short description of their personality, eight potential weapons, with a symbol attached to them, and seven different places where the murder could have occurred (see Figure 3.1). The game is divided into

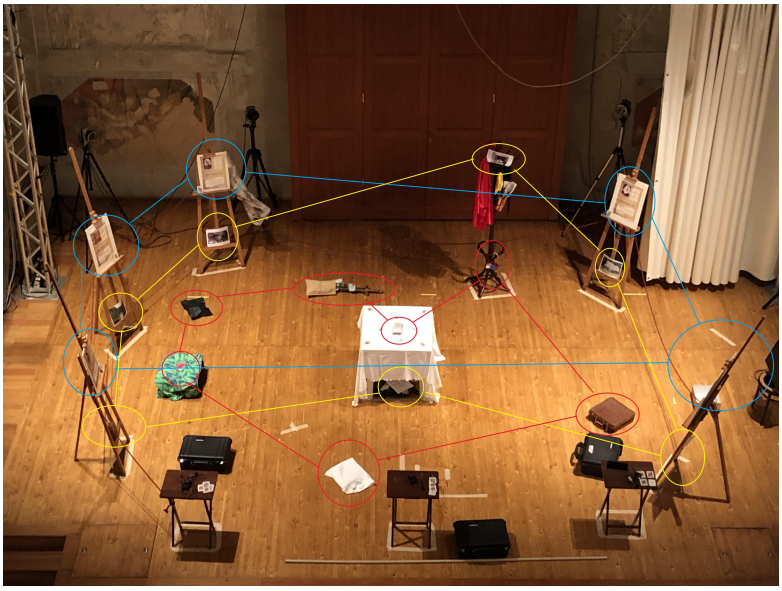


Figure 3.1: The game area and the material required to solve the murder. Blue circles correspond to the posters of the suspects, yellow circles represent the places where the murder could have occurred and the potential weapons are circled in red.

[3]Leonardo Fibonacci (c.1170 – c.1240–50) was an Italian mathematician from the Republic of Pisa. He is best known for his discovery of a particular number sequence, which has since become known as the Fibonacci Sequence

five tasks and participants were instructed to finish the game as quickly as possible. In fact, they had up to 1h to solve the murder and escape from the theater. During each task, they could find different clues, helping them to solve the murder or unlock a new task of the game. Between each task, participants were asked to fill up questionnaires that were conceived as part of the game (e.g., once completed, they received a code for unlocking the next instructions). Details about the questionnaires are provided in Section 3.2.1.4.

To create some competition between the groups and/or among the members of each group, we established a group and an individual leaderboard. This was based on the time participants took to solve the murder and on their performances on the different tasks. Leaderboards are an effective way to motivate participants through competition (Nov and Arazy, 2013; Codish and Ravid, 2014; Hamari et al., 2014).

The design of the game has been tested and incrementally adjusted until the beginning of the data collection to ensure that the game flow was coherent and that the tasks were understandable by the participants (e.g., we displayed some hints on the wall to make sure that everyone could still progress in the game).

3.2.1.2 Participants

We ran a campaign for recruiting participants through the GRACE website and social media, mailing lists, and the distribution of flyers. The protocol was approved by the Ethics Committee of the Department of Informatics, Bioengineering, Robotics and System Engineering of the University of Genoa, Italy.

To take part in the data collection, participants gave written informed consent and needed to be over 18 (legal age in Italy), to have a good understanding of written and spoken Italian (as all the rules, questionnaires and hints were in Italian) and to participate in a group of three friends without any hierarchical status among them. This last point is very important as we are only controlling the functional property of cohesion at the horizontal level. Having participants considering themselves as friends allowed us to infer that the affective property of cohesion, according to Severt and Estrada (2015)'s framework, is approximately constant over the time of the data collection, hence, providing us a baseline for studying variations of the instrumental property of cohesion (i.e., the Social and Task dimension). We also observed during the pre-tests, that having participants considering themselves friends, really impacted the spontaneity of the reactions and the dynamics of the group. Also, cohesion can take a long time to emerge in groups of strangers. For instance, previous studies show how cohesion is more volatile during the early phases of team functioning (Mullen and Copper, 1994) and sustainable Task cohesion emerges more quickly than does sustainable Social cohesion (Grossman et al., 2015).

A total of 17 groups (i.e., 51 persons) participated in the data collection. Participants ages ranged from 21y to 33y (M = 25.3y, SD = 3.1y) with 69% identified as female (i.e., 35 participants) and 31% identified as male (i.e., 16 participants). Participant's friendship duration ranged from 1 month to 22 years (M = 3.1y, SD = 2.5y). Concerning the escape game experience of the participants, 65% (i.e., 33 participants) had never participated in an escape game before, 25% (i.e., 13 participants) only tried once and 10% (i.e., five participants) participated multiple times. Only two participants had already gone to an escape game together before.

Participants received a small gift having a value inferior to 10 euros as a nominal honorarium for their participation.

3.2.1.3 Procedure

The data collection took place at Casa Paganini in Genoa, Italy[4]. This is an ancient monumental building having a space, which was formerly used as a theatre. This space is now exploited as a location for experiments on movement analysis in naturalistic settings and is endowed with a technological infrastructure for motion capture and multimodal recordings. First, we welcomed participants in a room next to the theater and we asked them to read the purpose of the data collection and sign the consent form. Before starting the game, participants also filled up a set of questionnaires to assess their level of friendship, their experience in escape games, their initial perception of cohesion within the group, participants' warmth and competence, and, finally, their attitude toward group games. More details and explanations of these questionnaires are in Section 3.2.1.4. They were filled up on an Android tablet (one for each participant) that we lent them for the time of the game. Then, the participants entered the theater. Researchers helped them to wear the motion capture suits and the radio-microphones followed by a full check of the setup to make sure that the data was streamed properly. Participants were allowed to interact freely on stage for a few minutes to get acquainted with the sensors. Then, the game started with a pre-recorded audio-video presentation explaining the context and the rules. The presentation was displayed on a wall of the game area. This was done to avoid any bias in providing participants with instructions. Similarly, we used another presentation during the game, automatically displaying additional information, clues, or reminders.

The game consisted of five tasks and was designed *ad hoc* to control the instrumental functional property of cohesion. Each task was conceived for a specific purpose to elicit a controlled variation of the Social and Task dimensions of cohesion (i.e., its increase or its decrease). In the following, we refer to those as Increase of Cohesion (I) and Decrease of Cohesion (D). The duration of each task was timed according to the feedback collected during the pre-tests and its difficulty (see Table 3.2 for the timings of the tasks). Figure 3.2 summarizes the flow of the game. In this Figure, bubbles indicate the questionnaires administered before, during, and after the game. To not break the dynamics of the game and to avoid weariness, we integrated the questionnaires into the game logic. For example,

Table 3.2: Expected variations of cohesion per task and the duration of each task. DS and IS refer to a decrease and an increase in Social dimension, whereas DT and IT refer to a decrease and an increase in Task dimension.

No. Task	Task name	Social dimension	Task dimension	Duration (min)
1	Discovery	Decrease (DS)	Decrease (DT)	10
2	Enigmas	DS	Increase (IT)	9
3	The impossible	Increase (IS)	DT	7
4	The weird object	IS	IT	7
5	The presentation	IS	IT	8

[4] http://www.infomus.org/index_eng.php

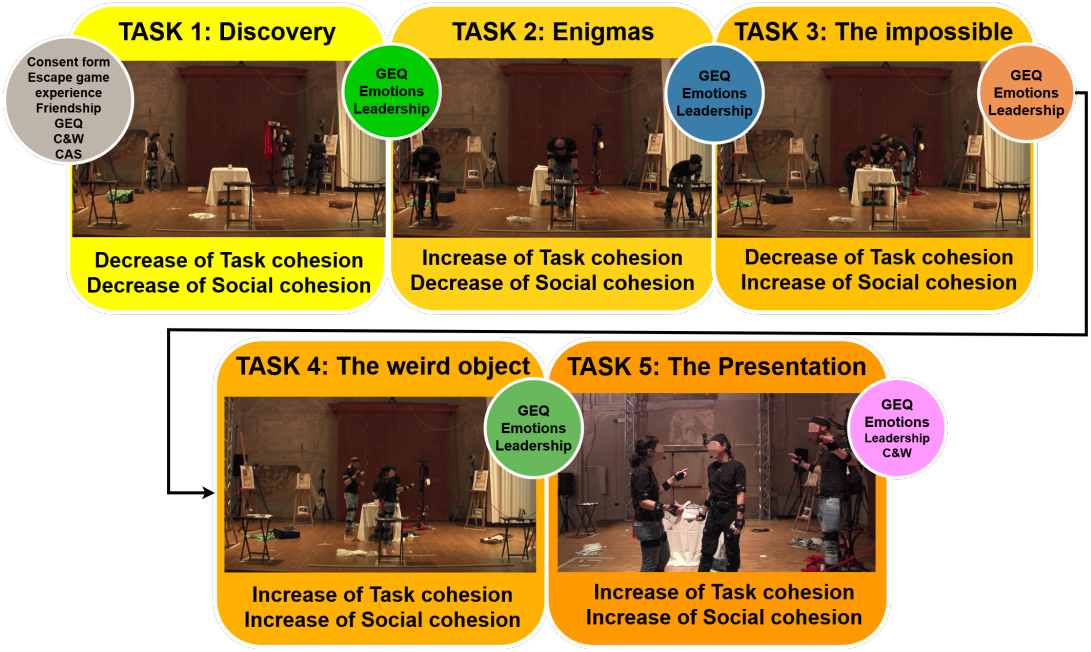


Figure 3.2: Timeline of the flow of the game. The questionnaires are displayed in chronological order before, between and after the tasks. The expected variations in cohesion are indicated at the bottom of each image taken from the dataset.

participants had to finish the questionnaires to get a code to unlock the next task. In that way, we ensured that all the participants filled up all the questionnaires at the same moment of the game.

Below, we report a detailed description of each of the five tasks:

- **Task 1: *Discovery* (DS & DT)**
  Participants were asked to find two objects, a box and its key, hidden in the game area. The box contained the instructions and materials for the next task. Participants had up to 10 minutes to complete this task. By finding objects, they get bonus points, otherwise, they lose points for their personal score on the leaderboard. This task was conceived to encourage participants to discover the game area while being in competition among them to find the objects in order to limit social interactions.

- **Task 2: *Enigmas* (DS & IT)**
  17 enigmas were divided into the following different categories: 1) *Matchsticks*: these are rearrangement puzzles in which a number of matchsticks are arranged as squares, rectangles or triangles. The aim is to move one, or a limited number, of matchsticks to create a new shape; 2) *Logic*: these enigmas describe a specific situation or context and ask the participant to find a logical explanation for it; 3) *Numbers*: these problems require calculations and ask the participant to give a mathematical solution to the problem; and 4) *Observation*: these enigmas propose visual scenes with squares or circles and participants need to link different objects together. We intentionally chose enigmas that require different skills to make sure that every participant could contribute.

Participants had 4 minutes to split all the enigmas taking into account every participant's skills. This brainstorming was expected to elicit an increase in the Task dimension (IT). Once participants split the enigmas, or if the 4 minutes were over, they had to start working on them in dedicated areas of the stage. They were not allowed to talk, otherwise, they would lose points. We established this rule to limit social interactions. Every time a participant completed an enigma, she had to put it on a box located outside of the game area. This added some stress and we could observe interesting behaviors (e.g., we noticed that successful participants were often looked at by the other group members when they moved to the box).

Participants had 5 minutes to solve a maximum of enigmas. At the end of the game, we added or subtracted points to the group regarding the number of correct and wrong answers. All groups received a 4 minutes extra time reward at the end of the last task.

- **Task 3: *The impossible* task (IS & DT)**
  This task included three different sub-tasks. Participants still needed to collaborate as two out of three puzzles gave hints about the murderer and the weapon. The group received 60 square pieces of paper of different sizes and colors with a number written on the front and a letter written on the back. One person had to reconstruct a part of the Fibonacci sequence (i.e., a sequence starting with 1 and 1, where each subsequent number is the sum of the previous two), another one had to reconstruct a palindrome spotted on a murderer poster, and the last one had to construct a Fibonacci clock indicating 3:45 pm. A Fibonacci clock is composed of five squares whose side lengths match the first Fibonacci numbers (i.e., 1, 1, 2, 3, and 5). The hours are displayed in red and the minutes in green. When a square is blue, it means that it indicates both the hours and the minutes. White squares are ignored. Hours are obtained by summing the values of the red and blue squares while minutes are five times the sum of the values of the green and blue squares. Figure 3.3 shows a Fibonacci clock indicating 3:45pm. On each weapon, a different Fibonacci clock was printed and participants had to find the clock indicating 3:45 pm to guess the weapon used for the murder.

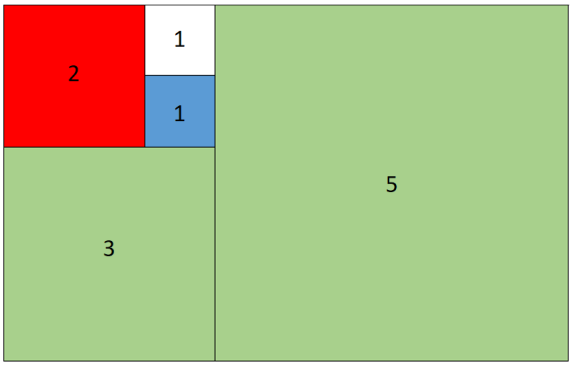


Figure 3.3: Fibonacci clock indicating 3:45pm. Hours are the sum of the red and blue squares' values (i.e., $2 + 1 = 3$) while minutes are five times the sum of the green and blue squares' values (i.e., $5 \times (5 + 3 + 1) = 45$).

We made this task impossible to achieve as each problem required the same pieces of paper. Moreover, the task had to be done within 7 minutes, adding some pressure on the participants. As each participant could not complete their part of the puzzle without negatively impacting other members of their group, we expected a decrease in the Task dimension of cohesion (DT), whereas the Social dimension was expected to increase (IS) due to the high number of interactions provoked by a stressing situation.

- **Task 4: *The weird object* task (IS & IT)**
  It consisted of guessing what an unusual object was. Participants had to link it to one of the seven potential places of the murder. Then, the group had to write their solution and explanation on a paper and put it in a box. If they guessed it right, they earned extra points at the end of the game. This task was timed to 7 minutes.

- **Task 5: *The presentation* (IS & IT)**
  The group had 4 minutes to provide a first solution to the murder in an original way (e.g., acting). At the end of the presentation, a red signal was always given by the researcher in charge of the session, indicating that the group provided a wrong solution. This was designed to observe the group's reaction after failing. We gave them an extra 4 minutes to present a second solution. At the end of it, a green signal was always given, indicating that they found the solution.

Task 4 and Task 5 required participants to be creative. We did this choice due to the fact that creativity enhances social interactions, eliciting situations with an increase of cohesion for the Social dimension (Keller, 1986). Also, the fact that the group had to reach a common decision was expected to amplify the Task dimension of cohesion. In both Task 4 and Task 5, Social and Task cohesion were expected to increase.

At the end of the data collection session, participants were briefed about the details, the aims, and the context of the study. Moreover, researchers answered all of the participants' questions. Before leaving the theater, participants were asked to fill up a last questionnaire to obtain their feedback on the game.

3.2.1.4 Questionnaires

Participants were asked to fill up several questionnaires at the beginning, at the end of the data collection, and after each task to further assess group cohesion as well as other group processes such as leadership. We chose to adopt repeated measures at regular intervals to reach a good level of granularity and to be able to detect changes in the processes. The questionnaires were presented in the same order after each task, but the order of the items of each questionnaire was randomized to keep participants' attention. All the English version of the items from the questionnaires we analyzed are in Appendix A, and Figure 3.2 shows the order in which we distributed the questionnaires.

As this data collection involved Italian speakers, we used validated Italian versions of each questionnaire, when they were available. Otherwise, we translated the items without changing the valence nor the grammatical construction of the questions, according to the guidelines provided by Carron and Brawley (2000). Original Likert scale formats were retained. In the following a description of each questionnaire is provided:

- **Cohesion**: We used the Group Environment Questionnaire (GEQ) (Carron et al., 1985), an 18-items self-report survey with a 9-point Likert scale answering format (from 1: "*Strongly disagree*" to 9: "*Strongly agree*") that is designed to assess Social and Task cohesion according to Carron's model (see Figure 2.2). Even if it was initially designed for studying cohesion in a sport environment, several studies have shown how it can be leveraged for addressing group situations in other contexts, for example in work meetings (Carless and De Paola, 2000; Michalisin et al., 2004) or in exercise classes (Estabrooks and Carron, 2000a) and even in different cultural contexts (Heuzé and Fontayne, 2002). Thus, we selected the GEQ to measure group members' self-assessment of cohesion. In particular, we used an Italian version of the GEQ (Andreaggi et al., 2000) to match the participants' first language. We administrated such a questionnaire before the data collection (i.e., to obtain a baseline of cohesion within the group) and after each task. The first time we administrated the GEQ, before Task 1, we decided to discard the two following items as we considered that they were not related to the escape game context and hardly adaptable: "*I'm not happy with the amount of playing time I get*" and "*Members of our team do not stick together outside of practice and games*". Concerning the questionnaires administered between the tasks, we used a shorter version of the GEQ as the answers to some items would not evolve during the time of the data collection. We discarded the two following items: "*For me, this team is one of the most important social groups to which I belong*" and "*Some of my best friends are on this team*". We also slightly adapted the items without changing the valence nor the grammatical construct of the questions. For example, "*Our team members have conflicting aspirations for the team's performance*" became "*Our team members had conflicting aspirations for finding the key*" after the *Discovery* task. We also decided to replace two items by ones from Michalisin et al. (2004) as we believe that they are close enough to the originals and more suited to our context. In that way, "*I enjoy other parties rather than team parties*" became "*I wish I was on a different team*" and "*I do not like the style of play on this team*" was replaced by "*Our team does not work well together*".

  The used version of the GEQ used between the tasks contains 14 items: eight related to the Task dimension, and six to the Social dimension (see Appendix A.1 for all the items).

- **Warmth and competence (W&C)** (Aragonés et al., 2015): This questionnaire is a set of eight items to measure warmth and competence, answered on a 9-points Likert scale from 1 ("*I completely disagree*") to 9 ("*I completely agree*"). We used a round-robin rating, meaning that each participant had to rate all the other participants and themselves. Half of the items are related to the warmth dimension whilst the other half focus on the competence dimension. The warmth dimension captures traits that are related to perceived intent, including friendliness, helpfulness, sincerity, trustworthiness and morality whereas the competence dimension reflects traits that are related to perceived ability, including intelligence, skill, creativity, and efficacy (Fiske et al., 2007). Participants were asked to fill up this questionnaire before and at the end of the data collection.

- **Competitivity**: The Italian version of the Competitivity Attitude Scale (CAS) questionnaire was used (Menesini et al., 2018). It consists of 10 items on participants' attitudes toward competition. This is a self-assessment questionnaire on a 5-point Likert scale from 1 ("*Never true for me*") to 5 ("*Always true for me*"). This questionnaire was administered just before the *Discovery* task with a twofold aim: to foster participants' competitiveness by having them reason about it and to gain further information on participants' attitudes towards group games.

- **Emotions**: To get some insights into participants' emotions at each task, we asked them to answer a question about their feelings by picking one among six different labels of emotion (see Appendix A.3). Moreover, participants could select the "*other*" option and provide their own label of emotion. The labels were selected by relying on Roseman (2001)'s Emotion Theory. According to this theory, emotions depend on the subjective perception of the ongoing situation (i.e., one's own appraisal), in terms of causal attribution (the situation was caused by someone else, by the self, or was due to external circumstances) or in terms of being consistent or not with one's goals and motivations. Each emotion can be identified by a specific combination of causal attribution and goal consistency (i.e., its appraisal configuration, Roseman and Smith, 2001). For instance, a player winning a game may feel *pride* as a consequence of perceiving herself as responsible for the victory (causal attribution) and because winning satisfies her goal of being a good player (consistency with personal goals and motivations). According to their appraisal configuration, emotions can be categorized as positive or negative (Roseman, 2013). Following this, we selected six emotions (i.e., three positive and three negative) that, given their specific appraisal configuration, might be elicited by the game.

- **Leadership**: We used a set of six items on a 6-point Likert ranging from 1 ("*Completely disagree*") to 6 ("*Completely agree*"), following Gerpott et al. (2019)'s study that based their work on Lanaj and Hollenbeck (2015) and McClean et al. (2018) items. For the same reasons as in the W&C questionnaire, we decided to use a round-robin rating. Hence, participants had to rate every member of the group, including themselves, resulting in answering each item three times. The five items are reported in Appendix A.2.

- **Motivation**: We used the Intrinsic Motivation Inventory (IMI) questionnaire developed by McAuley et al. (1989) to assess the participants' subjective experience with our escape game. It is on a 7-point Likert scale from 1 ("*Completely disagree*") to 7 ("*Completely agree*"). We decided to leverage this tool at the end of the data collection session as a guide for our debriefing phase. Having participants' opinions about the game and their enjoyment would be useful for further studies. With this in mind, we selected the *Interest/Enjoyment* and *Perceived Competence* subscales from the IMI.

As we modified the questionnaires, we assess their consistency via an Exploratory Factor Analysis (EFA) with oblique rotation (promax). For the GEQ and the W&C scales, EFA was performed for both dimensions (i.e., Social/Task and Warmth/Competence, respectively), each time the questionnaire was administered. The first step of the EFA

consists of applying the Kaiser criterion (Fabrigar et al., 1999) to select all the factors holding eigenvalues greater than 1. Then, we performed a Scree test to visually determine the number of factors to adopt. Results are reported here below.

Consistency results (EFA)
Scree plots analysis suggested a one-factor solution for each dimension measured by the GEQ (i.e., Social and Task cohesion) and for the W&C scale (i.e., Warmth and Competence). We obtained such a result at each time the questionnaires were administered, hence, supporting the idea that all of the items related to a specific dimension are loading into the same factor. It also indicates the consistency of our questionnaires. Regarding the Leadership questionnaire, Scree plots analysis suggested a multiple-factor solution. We observed that the items were loading into multiple factors (i.e., two factors for Task 2 and Task 4 or three factors for Tasks 1, Task 3 and Task 5). Our results can be explained by the fact that each task elicited and required different group dynamics and different aspects of leadership. This is in line with the functional leadership theory (Morgeson et al., 2010), according to which team leaders should adapt their behavior depending on the group's needs during a specific situation. Hence, we opted for a more parsimonious solution relating all the different functions to one overall leadership factor.

Finally, regarding CAS and IMI scales, scree plot analysis suggested a two factors solution which is in line with previous work on the CAS study (Menesini et al., 2018) and coherent regarding the IMI scale as we only selected two subscales from the original questionnaire (McAuley et al., 1989).

Reliability results (GLBs)
We calculated Greatest Lower Bounds (GLB) to establish the reliability of the scales (Jackson and Agunwamba, 1977). It corresponds to the lowest possible value that a scale's reliability can possess. GLB provides, indeed, a viable option in cases of a low number of items and small sample sizes (Ten Berge and Sočan, 2004; Revelle and Zinbarg, 2009; Sijtsma, 2009; Bendermacher, 2010; Peters, 2014; Trizano-Hermosilla and Alvarado, 2016; McNeish, 2018). The GLBs obtained for each questionnaire administered are reported in Table 3.3. All of the GLBs are over 0.700, hence, indicating the reliability of the used questionnaires (George and Mallery, 2016) for each task.

Table 3.3: GLBs obtained for each questionnaire. All are over 0.700, for each task, hence, indicating the reliability of the questionnaires used during the data collection.

		GLB values for each questionnaire					
		Baseline	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**
GEQ	**Social**	.882	.726	.849	.876	.922	.974
	Task	.920	.902	.822	.917	.919	.915
W&C	**Warmth**	.988	-				.996
	Competence	.996	-				.995
Leadership		-	.989	.954	.931	.990	.992
CAS		.882	-				
IMI		-					.909

3.2.2 Technical Setup

To collect GAME-ON, we built a setup that allowed us to capture, manage and visualize data from different sources. Synchronization of the data was handled via hardware and software, as explained in Section 3.2.2.2.

3.2.2.1 Equipment

We captured the behaviors of three participants interacting simultaneously. For this purpose, we adopted a hybrid motion capture approach combining together three Shadow inertial motion capture suites[5] with a Qualisys optical motion capture system[6]. This choice was made to take advantage of the strengths that each technology offers, correct the drifts that may occur in long recording sessions, and avoid occlusions. Shadow's suite is a wireless wearable system composed of 17 IMU sensors (3-axis accelerometers, gyroscopes, and magnetometers), placed on the body at some precise reference points (see Figure 3.4) plus two additional sensors, placed in the participants' shoes. Qualisys configuration included 16 infrared cameras optimally placed to cover the whole game area. In our setup, Shadow and Qualisys data were captured at 100Hz. With the aim of having a perfect coupling between the two systems, 17 infra-red reflective Qualisys markers were attached to the Shadow's IMUs with Velcro straps. Additionally, audio and video were recorded. We used three wireless headsets microphones (AKG wireless set 800MHz with C555L headsets, Mono, 48kHz, 16 bits per sample), and two static professional JVC video-cameras (1280×736, 50fps) frontally (at about 9m from the center of the scene) and laterally (at about 4.5m from the center of the scene) placed with respect to the game area. Moreover, two additional Panasonic handy cameras (1920×1080, 50fps) completed the setup. These last two video-cameras were used as backup cameras and were not synchronized.

For data acquisition and synchronization, we used four desktop PCs (I7 Intel processor, eight GB DDR3 RAM, Windows 10x64), one devoted to audio capturing, one devoted to video capturing, one for the Qualisys system, and one for the Shadow system.

3.2.2.2 Software Platform

Data recordings were handled by using an EyesWeb[7] application developed for the data collection purpose by Professor Gualtiero Volpe and his team. EyesWeb is a software platform that supports real-time capturing and processing of multimodal data streams. It handles data synchronization by time-stamping each received frame or sample. Time-stamping is based on SMPTE time codes[8], with the additional possibility to use sub-sample accuracy. When the hardware supports it, the SMPTE signal is used as a reference clock. For example, the Qualisys system can receive an SMPTE signal as input and lock to it. This mechanism is also used by the JVC video cameras. In such cases, the received samples are automatically timestamped by the capture device. Other devices are synchronized by EyesWeb, which timestamps each sample when it is received by the host

[5] https://www.motionshadow.com/

[6] https://www.qualisys.com/

[7] http://www.infomus.org/eyesweb_eng.php

[8] See standard ST 12-1:2014, which is available at the SMPTE website: https://www.smpte.org/standards/document-index/ST

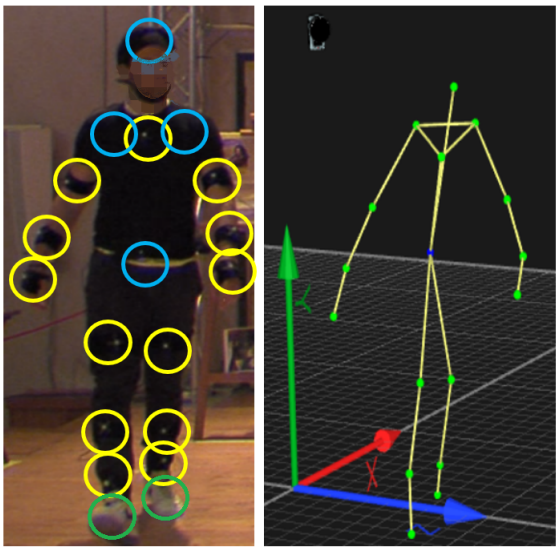


Figure 3.4: Position of the 17 IMU Shadow sensors and 17 Qualisys reflective markers (yellow and blue circles) on a participant and its associated reconstructed skeleton. Sensors circled in blue are positioned at the back of the participant (the two shoulders, the head, and the hip). Sensors in yellow are at the front of the participant. Green circles correspond to the two Shadow sensors placed in the shoe of the participant.

computer. By means of these timestamps, EyesWeb can accurately play the data back with the same timings as they were captured. That is, this process preserves each raw signal's native frame rate, when performing multimodal analysis.

In the case of the GAME-ON dataset, the frontal JVC camera was generating the SMPTE time codes, which were received by the lateral JVC camera, by the audio card of the PC for audio recordings, by the Qualisys system, and by the PC running the Shadow recorder. Thus, audio, video, and Qualisys recorders were all locked to the same SMPTE signal. The Shadow system generates its own timestamps. Shadow data, including the timestamp, was received by an *ad-hoc* C# console application connected to both the Shadow system and to EyesWeb. Shadow data was thus received by EyesWeb, and the correspondence between the SMPTE time code and the Shadow timestamp, for each Shadow sample, was recorded in a separate file, letting us manage synchronization between Shadow data and other data.

3.2.2.3 Post-processing

Post-processing included several steps. As data was recorded separately for each task, the first step was to trim the data to only keep the interesting content, discarding the moments where participants were filling out questionnaires or were waiting for the others to start a new task. We used ffmpeg[9] to trim our audio and video files and discarded the data that was not tasks-related. Then, the second step consisted of determining what data got lost for each sensor. Among all the groups (representing 11h36m16s of data) we had to discard two groups (1h16m48s), representing 11% of the data, due to connectivity

[9] https://www.ffmpeg.org/

problems between the C# application and the Shadow system, causing deep gaps in the data.

We only needed to label one point (i.e., hip or head) with the Qualisys Track Manager (QTM) software[10] to get the drift-corrected translation values for all the other points. We used the hip marker except for the frames where it was not visible. Concerning the video, we managed to save 100% of the files, whilst we lost 3% of the audio data, representing 24m16s of content. Missing audio was however available on the backup cameras.

3.2.2.4 Data Visualization

Another EyesWeb application was developed by Professor Gualtiero Volpe and his team to visually check that the motion capture data concerning the 17 points representing joints in the participants' skeletons was coherent (see Figure 3.5). As the data was recorded and stored in a specific architecture and format, this application automatically selects and plays the audio, the video, and the motion capture data files belonging to the same recording session in a synchronized way. Below is the organization of the recorded files:

```
Date of the session (e.g., 2019-10-28)
├── audio
│   └── Audio files (.aif)
├── qtm
│   └── Qualisys' Qtm files (.qtm)
├── shadow
│   ├── Shadow's CSV files
│   └── Shadow's text files (timestamps)
└── video
    ├── Video files (.avi)
    └── Video's text files (timestamps)
```

We recorded one audio file per participant and per task for a total of 15 audio files per group. We recorded one QTM file per task for a total of five QTM files per group. Concerning the Shadow data, we stored all the data in a CSV file containing all the sensors' values per participant per task and one text document per CSV file, storing the shared timestamps for a total of 30 files. We saved the frontal and lateral video recordings for each task, but also one text file per recording storing the shared timestamps, for a total of 20 files.

3.2.3 Collecting External Assessments of Cohesion

According to Vinciarelli and Mohammadi (2014), both self- and external assessments has pros and cons. When a person assesses themselves, they tend to provide ratings toward socially desirable characteristics, especially when the assessment can have negative consequences. External assessments reflect the behavior that people adopt toward others, without necessarily corresponding to their true internal state (Uleman et al., 2008). Depending on the application, researchers might favor one type of assessment over the

[10] https://www.qualisys.com/software/qualisys-track-manager

other one. Furthermore, having access to both assessments could help limit the cons they introduce, by developing strategies to combine them. Thus, we ran an external annotation campaign for assessing cohesion. This campaign has been approved by the Ethics Committee of Paris-Saclay. External assessments of cohesion were collected and stored through a PHP website hosted on Télécom Paris' servers.

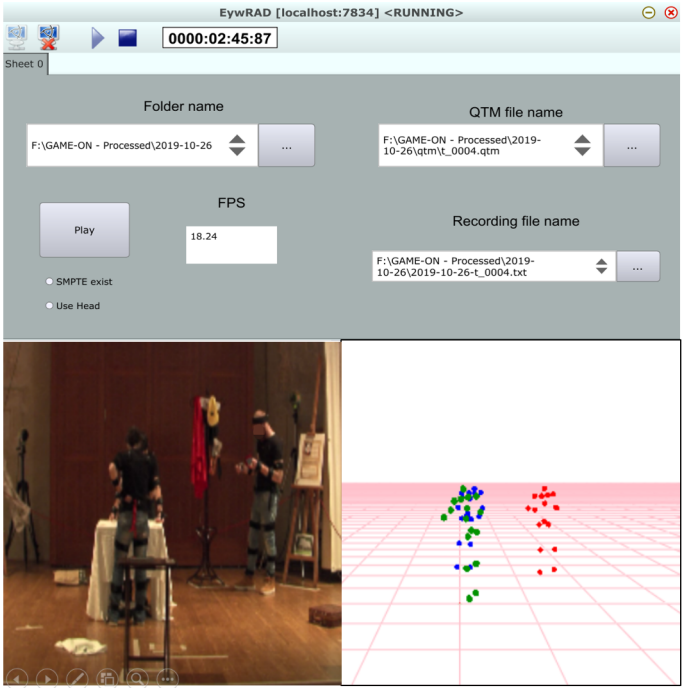


Figure 3.5: Interface of the EyesWeb application for visualizing synchronized data streams. At the top, paths to the files are provided. At the bottom, the recorded video is played (left) while the 17 body points are displayed for each person (right).

3.2.3.1 Procedure

First, an introduction to the GRACE project and the purpose of the external annotation campaign was given to the participants. Then, they had to agree to the terms and conditions of the data collection as well as to give their consent to participate. Before starting the annotation, they filled, anonymously, demographic information that include: age, gender, and if they understood Italian or not (i.e., the language spoken in the videos). Once completed, one of the 17 groups available in GAME-ON was randomly selected and its videos of the five tasks were displayed, in the correct order (i.e., from Task 1 to Task 5), to the same rater. After watching a video, the participant was asked to fill up the GEQ questionnaire to process to the next video. Here, we adapted all the items of the GEQ to allow the assessment of cohesion from an external point of view. For example, the item "*Our team did not work well together*" became "*The team did not work well together*". It is only once all the five videos were assessed that the data about their demographics and answers to the five GEQ questionnaires were stored, anonymously.

3.2.3.2 Participants

All the participants were recruited through international mailing lists of researchers and by sharing the campaign with our acquaintances. Their participation was voluntary, anonymous, and not remunerated. The only conditions that all the participants met are to understand English (i.e., the language of the instructions and the questionnaires) and to be over 18.

In total, we collected annotations of cohesion from 59 participants. Each group has been assessed by at least three different participants. Ages ranged from 18y to 42y (M = 26.8y, SD = 4.8y) with 45% identified as female (i.e., 27 participants) and 55% identified as male (i.e., 32 participants). 69% (i.e., 41 participants) did not understand Italian, the language spoken in all the interactions. This is, however, not problematic since we were interested in the assessment of cohesion from nonverbal behavior.

3.3 Data Analysis

In this Section, we present an analysis of the data collected through the GEQ questionnaire, both from self- and external assessments of cohesion. The analysis of leadership and emotion assessments is performed in Chapter 6. We used an alpha level of 0.05 for all statistical tests.

3.3.1 Analysis of Cohesion from Self-assessments

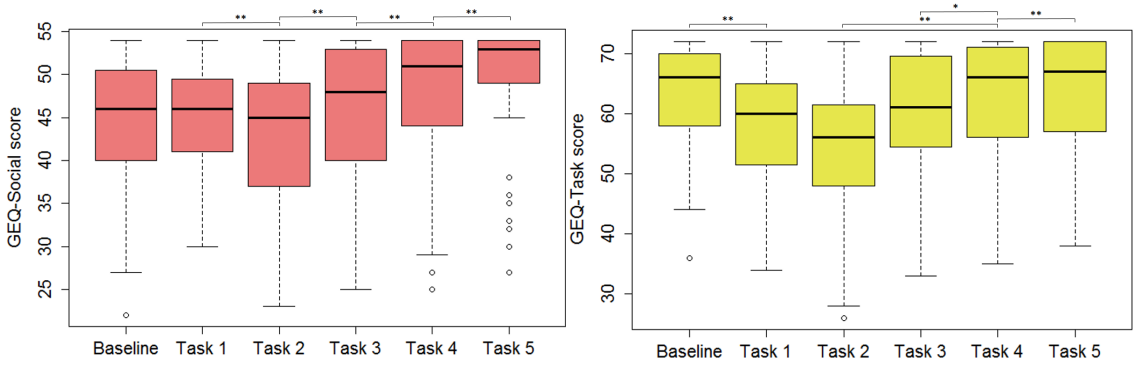
The following analysis is aimed at understanding and evaluating the dynamics of cohesion over time, regarding its Social and Task dimensions. In Task 1, we looked at the variations of cohesion (i.e., increase or decrease) with respect to the baseline obtained from the first administration of the GEQ questionnaire before starting the data collection. In each of the other tasks, we looked at the variations with respect to the previous one.

To analyze such variations, we computed two scores of cohesion from the GEQ questionnaire, for every participant, and for each task. We named these scores as *GEQ-Social* and *GEQ-Task*, respectively. The former relates to the Social dimension and it results from the sum of the items 1 to 6 reported in Appendix A. The latter one corresponds to the Task dimension and it results from the sum of the items 7 to 14 reported in Appendix A. Figure 3.6 shows the box-plots of the GEQ-Social and GEQ-Task scores, respectively.

We first examine the normality of the data by using a Shapiro-Wilk test. The test shows a significant departure from normality for both the Social dimension ($W = 0.88$, $p < .001$) and the Task dimension ($W = 0.91$, $p < .001$). Thus, non-parametric tests are used for the data analysis.

3.3.1.1 The Social dimension

A non-parametric Friedman test of differences among repeated measures shows a significant difference between the GEQ-Social scores across tasks ($X^2(5) = 57.83$, $p < .001$). Post-hoc Conover's tests with a Bonferroni-adjusted alpha level confirm that we managed to almost control the Social dimension of cohesion accordingly to the sequence in Figure 3.2. In Task 1 and Task 2, we expected to break the Social cohesion of the group,



(a) Box-plot of the GEQ-Social scores

(b) Box-plot of the GEQ-Task scores

Figure 3.6: Box-plots of the GEQ-Social and GEQ-Task scores, per task. Medians of GEQ scores are represented by the bold black lines. White dots represent mild outliers, computed using the interquartile range (IQR) criterion. Significant differences between tasks are displayed with a "*" ($.001 < p < .05$) or "**" ($p < .001$).

developed prior to the data collection as participants were friends. Hence, the expected variation was from an Increase in Social cohesion (IS) to a Decrease in Social cohesion (DS). Then, we wanted to observe an increase in Social cohesion in Task 3, Task 4, and Task 5 (from DS to IS). Post-hoc tests show a significant difference between the Baseline and all the tasks except Task 1 ($p < .001$, for all the transitions)[11], proving that the game had an impact on the Social dimension of cohesion after the first Task. Moreover, as expected, there is a significant decrease of Social cohesion between Task 1 and Task 2 ($p < .001$, $Mdn = 46$ for Task 1, and $Mdn = 45$ for Task 2). There is also a significant increase of Social cohesion between Task 2 and Task 3 ($p < .001$, $Mdn = 45$ for Task 2 and $Mdn = 48$ for Task 3), and between Task 3 and Task 4 ($p < .001$, $Mdn = 48$ for Task 3, and $Mdn = 51$ for Task 4), indicating that the expected variation of Social cohesion was indeed obtained. Post-hoc tests also show significant differences between Task 4 and Task 5 ($p < .001$). Again, the medians increased ($Mdn = 51$ for Task 4 to $Mdn = 53$ for Task 5), indicating that this last task can also be considered as IS.

In conclusion, we managed to control the direction of variations of the Social dimension of cohesion between all the tasks (i.e., from Task 1 to task 5). We, however, did not manage to break the Social cohesion between the Baseline and Task 1. This is probably due to the fact that group members shared strong bonds (i.e., they considered themselves friends).

3.3.1.2 The Task Dimension

A non-parametric Friedman test of differences among repeated measures shows a significant difference between the GEQ-Task scores across tasks ($X^2(5) = 36.14$, $p < .001$). Post-hoc Conover's tests with a Bonferroni-adjusted alpha level show, however, differences compared to the expected variations of Task cohesion (see Figure 3.2). We first

[11] All the p-values presented are already Bonferroni-adjusted.

expected Task cohesion to decrease (DT) from Baseline to Task 1 and then, to observe an increase (IT) in Task 2, followed by another decrease in Task 3. Finally, we expected Task cohesion to increase in Task 4 and Task 5.

Regarding the Task dimension, post-hoc tests show a significant difference between the Baseline and Task 1, Task 2, and Task 3, respectively ($p < .001$), proving that the game had an impact on the Task dimension. There also was not a significant difference between Task 1 and Task 2 and medians decreased instead of increasing as we expected (*Mdn* $= 60$ for Task 1, *Mdn* $= 56$ for Task 2).

Several explanations account for this result. A visual inspection of the video data showed that the participants did not fully understand the aim of Task 2. We noticed that the researcher in charge of the session had to remind the instructions more than once during the other tasks as participants were not following or understanding the guidance. Also, Task 2 was designed to allow time for participants (4 minutes) to organize the distribution of the enigmas among them. This was expected to result in an increase in Task cohesion, but most of the groups rushed to the next phase of the task and randomly assigned enigmas. As participants were not allowed to interact during the second part of the task (5 minutes), it is very likely that their answers about the Task dimension were biased by the decrease in Social cohesion. Also, whereas we were aware that eliciting multiple changes of one single dimension over a very short period of time (i.e., the Task 1 - Task 2 - Task 3 sequence) was complicated, this indeed revealed more complex than expected.

In brief, there is only a significant decrease of Task cohesion between the Baseline and Task 1 ($p < .001$, *Mdn* $= 66$ for the Baseline, *Mdn* $= 60$ for Task 1) and a significant increase of Task cohesion between Task 3 and Task 4 ($p = .003$, *Mdn* $= 61$ for Task 3, *Mdn* $= 66$ for Task 4) as well as between Task 4 and Task 5 ($p = .001$, *Mdn* $= 66$ for Task 4 and *Mdn* $= 67$ for Task 5). Indeed, according to Conover's post-hoc results, there is also a significant difference in Task cohesion between Task 2 and Task 4 ($p < .001$), and a significant difference between Task 2 and Task 5 ($p < .001$). GEQ-Task scores in Task 3 and Task 5 were significantly different ($p < .001$) too.

We can consider that GEQ-Task scores from the Baseline to Task 3 reflect a downward variation of Task cohesion as the medians significantly decreased. Conversely, there is an upward variation between Task 3, Task 4, and Task 5, so we can conclude that Task cohesion increased in Task 4 and Task 5.

In conclusion, we managed to control the direction of variations of the Task dimension of cohesion over time, from a decrease in the Baseline to the Task 3 followed by an increase until Task 5. We, however, probably miss-evaluated Task 2 as we were expecting an increase in Task 2 followed by a decrease in Task 3.

3.3.2 Analysis of Cohesion from External Assessments

3.3.2.1 Reliability of the External Annotations

We first assessed whether external assessments were reliable. Thus, we leveraged the intra-class correlation (ICC) (Fisher, 1992), one of the most commonly-used statistics for assessing inter-rater reliability for ordinal variables (Hallgren, 2012).

The design of our study is not fully crossed and every group has been assessed by a subset of at least three different raters. We first computed the GEQ-Social-ext and GEQ-Task-ext scores at each of the five tasks as we did for the self-assessments. Since we re-used the GEQ questionnaires (adapted to the external annotation), the theoretical minimum and maximum scores are identical for the Social (i.e., 6 and 54, respectively) and Task (i.e., 8 and 72, respectively) dimensions of cohesion. Figure 3.7 displays the boxplots of the GEQ-Social-ext and GEQ-Task-ext scores, respectively. Then, we summed the GEQ-Social-ext scores obtained for each of the five tasks. Similarly, we summed the GEQ-Task-ext scores, hence, producing a cohesion score for each dimension. Based on these cohesion scores, we computed $ICC(1, k)$ with a consistency definition, following Shrout and Fleiss (1979)'s convention. We chose such an ICC as groups were annotated by different sets of randomly selected raters. According to Cicchetti (1994), we obtain a *poor* inter-rater agreement of 0.24 for the Social dimension ($p = .045$) while no significant agreement is found for the Task dimension (i.e., $p > .05$). Such a result is, however, expected due to the variations of both GEQ-Social-ext and GEQ-Task-ext scores across the five tasks but also between the groups. Thus, we decided to analyze the inter-rater agreement for each group.

We choose the $ICC(2, k)$ with a consistency definition. For 12 out of 15 groups, we obtain a *good* ICC (i.e., over 0.60, with $p < .050$) for both the Social and Task dimensions of cohesion. For the remaining three groups, we reach a significantly *good* ICC for only one dimension over the two (i.e., Social cohesion for two of the groups and Task cohesion for the third group). Thus, we decided to keep them in the analysis. Such results confirm the reliability of the external annotations. For this reason, we perform a similar analysis than for the self-assessments of cohesion. External raters, however, could not evaluate cohesion before the beginning of Task 1 because they did not know the group members. Thus, we have external assessments from Task 1 to Task 5 from which we computed the GEQ-Social-ext and the GEQ-Task-ext scores for each group. These are used in the remaining of the analysis.

We first test the normality of the data by running a Shapiro-Wilk test, for both the Social and Task dimensions. It shows a significant departure from normality for the Social dimension ($W = 0.96$, $p < .001$) and the Task dimension ($W = 0.97$, $p < .001$). Thus, we perform non-parametric tests.

3.3.2.2 The Social Dimension

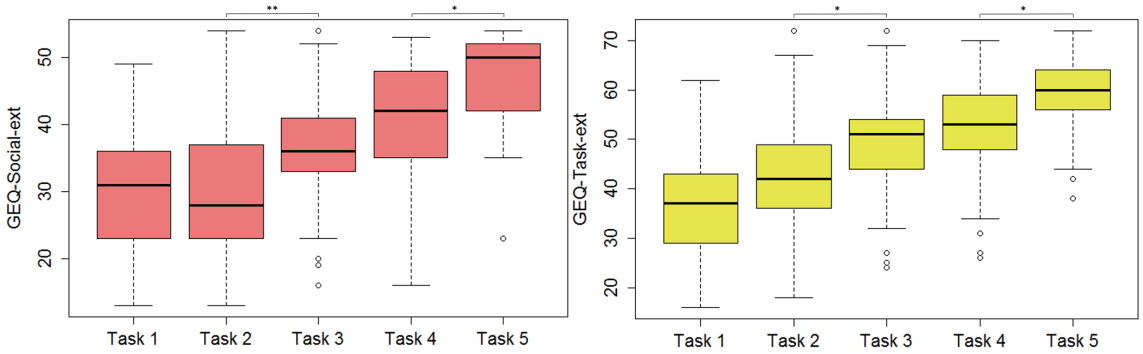
First, we run a non-parametric Friedman test of differences among repeated measures between the GEQ-Social-ext scores across the five tasks. This shows that a significant difference between these scores exists ($X^2(4) = 130.58$, $p < .001$). Then a post-hoc analysis is performed using Conover tests with a Bonferroni-adjusted alpha level, showing that we almost managed to obtain similar observations than for the self-assessments. They are, indeed, close to the expected variations of cohesion (as in Table 3.2). Results show that, for the transition between Task 1 and Task 2, the difference in scores for the Social dimension ($Mdn = 31$ for Task 1 and $Mdn = 28$ for Task 2) is not significant while expecting a significant decrease. Also, there is no significant difference in score for the transition between Task 3 and Task 4 despite the increase in Social cohesion expected. Except for these transitions, Social cohesion significantly increases between Task 2 and

Task 3 ($p < .001$, $Mdn = 28$ for Task 2 and $Mdn = 36$ for Task 3) and between Task 4 and Task 5 ($p = .008$, $Mdn = 42$ for Task 4 and $Mdn = 50$ for Task 5).

3.3.2.3 The Task Dimension

A non-parametric Friedman test of differences among repeated measures between the GEQ-Task-ext scores across the five tasks shows a significant difference ($X^2(4) = 112.01$, $p < .001$). The following post-hoc analysis using Conover tests with a Bonferroni-adjusted alpha level reveals a similar pattern than for the Social dimension: from Task 1 to Task 2, there is no significant difference between the scores while there is a significant augmentation of Task cohesion between Task 2 and Task 3 ($p = .003$, $Mdn = 42$ for Task 2 and $Mdn = 51$ for Task 3). Again, there is no significant difference in the scores between Task 3 and Task 4 and, finally, Task cohesion significantly increased, as expected, between Task 4 and Task 5 ($p = .004$, $Mdn = 53$ for Task 4 and $Mdn = 60$ for Task 5).

These results show that, for both the Social and Task dimensions of cohesion, Task 2 and Task 4 were miss-evaluated by external raters. As interactions were limited in the second half of Task 2, external raters might have focused on the beginning of the interaction (i.e., during the split of the enigmas) to assess both dimensions. Also, the fact that no significant difference is found between Task 3 and Task 4 might indicate that, from an external point of view, the contradictory goals of each group member in Task 3 were not evidently displayed, hence, making both tasks very similar (i.e., solving a problem as a group).



(a) Box-plot of the GEQ-Social-ext scores

(b) Box-plot of the GEQ-Task-ext scores

Figure 3.7: Box-plots of the GEQ-Social-ext scores and GEQ-Task-ext scores, per task. Medians of GEQ scores are represented by the bold black lines. White dots represent mild outliers, computed using the interquartile range (IQR) criterion. Significant differences between tasks are displayed with a "*" ($.001 < p < .05$) or "**" ($p < .001$).

3.4 Conclusion

THIS Chapter initially reviewed the main available datasets that are used for the automated analysis of small groups interaction. It highlighted the need for a dataset specifically designed for the automated study of cohesion in small groups of humans. Thus, we collected GAME-ON, a dataset in which 17 groups of three friends interact in the context of an escape game divided in five tasks explicitly designed for observing variations of the Social and Task dimensions of cohesion according to Severt and Estrada (2015)'s theoretical framework of cohesion.

This Chapter presented the data collection design as well as the various questionnaires used to assess group members' cohesion, warmth and competence, competitivity, emotions, leadership, and motivation. Since we modified some items of the questionnaires, we also confirm the questionnaires' psychometric properties. Then, the technical setup is explained in depth to understand the equipment used to collect the data (e.g., hybrid motion capture setup) and the various software developed to capture, manage and visualize data from different sources. The external annotation campaign ran after processing and analyzing all the data collected is also presented. Finally, this Chapter provides a first analysis of the answers from both the self- and external assessment of cohesion to validate the experimental design for the Social and Task dimensions of cohesion.

Chapter 4

Feature Extraction

Contents

THIS Chapter introduces what are the multimodal nonverbal features used for the automated analysis of cohesion. Then, it presents the motion capture-based and audio-based features we extracted. For each feature, motivations and computational details are provided.

I co-designed the motion capture-based features and partially implemented them. I also designed and implemented the turn-taking-related features.

4.1 What Matter for Automatic Analysis of Cohesion?

During everyday interactions, people coordinate their vocal and visual behavior to convey messages to others. When people interact, they, consciously or not, adopt various strategies to dynamically adapt to the audience's reactions (e.g., locating their bodies, assuming various postures, directing their eyes, moving their hands). Both verbal and nonverbal behaviors are co-occurring and are interrelated (Jones and LeBaron, 2002). While verbal communication plays an important role in social interactions, it is known that a valuable amount of information is delivered nonverbally (Knapp et al., 2013). Previous studies on the automated analysis of affective group processes explored verbal and nonverbal features for predicting, for example, group affect and satisfaction (e.g., Lai and Murray, 2018) or group performance (e.g., Kubasova et al., 2019). They showed that nonverbal

communication is a more powerful predictor than verbal behavior for the automated analysis of such group processes.

As of today, the strategy that is the most often employed to extract features that account for the group behaviors, consists of computing group features by deriving key statistics of their individual features set by, for example, calculating the average, the median, the maximum, and the minimum values, as well as the variations from the mean for all individuals or dyads in the group.

In this Section, the features extracted that are generally used for the automated analysis of cohesion are presented. Since most of the existing studies (e.g., Hung and Gatica-Perez, 2010; Nanninga et al., 2017; Fang and Achard, 2018; Kantharaju et al., 2020) use datasets that contains video and audio data only (e.g., Carletta et al., 2006; Sanchez-Cortes et al., 2011b), we present these features according to their media (i.e., video and audio).

4.1.1 Video-based Features

From the video data, various types of features can be extracted. First of all, as in Hung and Gatica-Perez (2010)'s study, the amount of motion found in a video can be extracted. Motion features are, indeed, very salient in portraying interpersonal relationships as a lot of information can be inferred based on the way people move. Based on the motion information, features such as the total distance traveled by an individual or by the group (e.g., Okada et al., 2015), the average velocity of hands and their synchronization among group members (e.g., Müller et al., 2018) and head-nodding (e.g., Feese et al., 2012; Kantharaju et al., 2020; Kantharaju and Pelachaud, 2021) had been extracted.

Also, a large number of features were extracted from the gaze of group members. For example, Kantharaju et al. (2020) computed the overlapping gaze between any two participants at a given point in time and the total amount of time spent by each group member looking at the others. Such features are, indeed, particularly helpful at predicting turn-taking activities (Jokinen et al., 2013).

In addition, some studies also extracted information about facial expressions (Müller et al., 2018; Kantharaju et al., 2020; Kantharaju and Pelachaud, 2021). They focus on various Facial Action Units (FAUs) and extracted both the duration and intensity of the activated FAUs. These features convey a plethora of affect-related information (the activation of cheek raising or lip corner puller units are often associated with happiness and smile Ekman et al., 1990), hence, are relevant for the automated analysis of cohesion (Kantharaju and Pelachaud, 2021)

Another feature that has been extracted for studying cohesion across studies is the amount of self- and inter-member synchrony. According to Delaherche et al. (2012), synchrony is "*the dynamic and reciprocal adaptation of the temporal structure of behaviors*" between group members. It can be studied using visual (e.g., by analyzing gestures) and auditory (e.g., by analyzing communication patterns) information. Group members that work well together and are closer to each other, indeed, gradually adopt each other's behavioral patterns, such as the way they talk, their body pose, and the way they communicate (Lakin and Chartrand, 2003; Campbell, 2008). The concept of synchrony is, however, complex. Measuring synchrony, hence, can be computed over the whole interaction or on smaller units of interactions. Multiple approaches had been used to extract such a feature by, for example, using Pearson correlation to quantify the amount of linear correlation

among two time series of a similar length from two group members (Zhang et al., 2018), as well as their amount of mutual information (Hung and Gatica-Perez, 2010). Another possibility consists in quantifying the distance via dynamic time warping (Müller et al., 2018).

Features extracted from videos are, however, often found to perform poorly in comparison to the ones extracted from audio (e.g., audio, Jayagopi et al., 2009; Hung and Gatica-Perez, 2010). One possible reason for this is that extracting nonverbal behavior from videos or images might be challenging (e.g., due to occlusions). They, indeed, contain a lot of information and often require further processing to extract relevant information on group activity. Furthermore, video cameras might not be available for privacy and/or convenience reasons. For example, to capture group behavior without video cameras, Zhang et al. (2018) used sociometric badges to track the number of interactions, their frequency, and the energy and consistency of each member's movements as well as their synchrony. Their results show that these features could be useful for predicting cohesion and its Task dimension in particular. Such a device might be a good alternative in longitudinal studies.

4.1.2 Audio-based Features

Prosody has been extensively studied to automatically study cohesion and related processes in group settings. The most common features extracted to describe prosody are the loudness of the voice, the F0 envelope and contour, the voicing probability and a set of features detailing voice quality, the differential and pitch jitter and shimmer of the voice (Nanninga et al., 2017; Lai and Murray, 2018; Murray and Oertel, 2018; Kubasova et al., 2019). The latter yields information on the laxness or the tenseness of the vocal tract. Lastly, some studies include the Pulse-code modulation (PCM) into their features set (Kubasova et al., 2019) to detail the encoding of the digital audio as well as the Line spectral frequency pairs (LSF/LSP), which are expressed by coefficients representing the channel transmission (Lai and Murray, 2018). Most of the prosody-related features can be extracted using the openSMILE software (Eyben et al., 2010) and its various features sets (e.g., the Computational Paralinguistics ChallengE or the Geneva Minimalistic Acoustic Parameter Set, Schuller et al., 2013; Eyben et al., 2015, respectively).

Aside from prosody, the most studied features are the ones related to turn-taking. Turn-taking can best be analyzed if individual speakers can be identified, either by using individual microphones to detect separate speech signals or by using automatic source separation techniques. Turn-taking is often quantified over the participation rate of all group members by looking at the median pause-to-speech ratio (Lai and Murray, 2018), the probability to take a turn after any other individual (Müller et al., 2018), the higher-level participation equality, and turn-taking freedom (Lai and Murray, 2018) with the expectation that groups with higher cohesion allow for higher equal and unconditional participation among members. Also, the speaking rate is tracked over the total amount of pauses versus speech (Hung and Gatica-Perez, 2010) and by counting the number of syllables per second (Hung and Gatica-Perez, 2010; Nanninga et al., 2017) and the rate of speaker changes (Jayagopi et al., 2009; Hung and Gatica-Perez, 2010; Lai and Murray, 2018; Müller et al., 2018). Features related to the time and frequency of overlapping speech are also computed in various studies (e.g., Jayagopi et al., 2009; Hung and Gatica-

Perez, 2010; Lai and Murray, 2018) since these are indicative of conflict or backchannels/feedback if the utterances are short. Overlapping speech can be tracked over the total amount of time (Hung and Gatica-Perez, 2010; Lai and Murray, 2018), the rate and amount of successful and unsuccessful interruptions (Jayagopi et al., 2009; Hung and Gatica-Perez, 2010) and the average duration of uninterrupted speech (Lai and Murray, 2018).

As for the video-based features, synchrony can also be extracted from audio. In their work, Nanninga et al. (2017) modeled the prosodic behavior of the group members as a mixture of Gaussian and non-parametric distributions over time. In that way, they could relate the findings over multiple time segments and extract the synchrony over the whole interaction.

4.2 Multimodal Nonverbal Feature Engineering

In line with previous studies showing the relevance of extracting nonverbal behaviors for the automated analysis of group processes (e.g., Müller et al., 2018; Kubasova et al., 2019) and, in particular for cohesion (e.g., Hung and Gatica-Perez, 2010; Nanninga et al., 2017; Alsulami, 2021), we focus on extracting a set of nonverbal multimodal features characterizing social interaction. The design of each feature is either inspired by Social Sciences' insights (e.g., Tannen, 1994; Wallbott, 1998) or by the most relevant features extracted in previously mentioned studies on the automated analysis of cohesion (e.g., Hung and Gatica-Perez, 2010; Nanninga et al., 2017) and other affective processes such as group emotion (e.g., Chao et al., 2015) and stress (e.g., Aigrain et al., 2018). In the following, we describe how each feature was computed. Features are either extracted from the motion capture data or from the audio data of the GAME-ON dataset and are computed either from individuals (I) or for the group as a whole (G). They were extracted on fixed-length consecutive time windows. The duration of these time windows is determined in Chapter 5. For some of the features, we also applied functionals (i.e., mean, standard deviation, minimum, maximum, and skewness) to their values over the whole time windows. Thus, in total, we provide the computational models with an input vector of size 91. Table 4.1 recapitulates all of the features extracted.

4.2.1 Motion Capture Features

As previously mentioned in Chapter 3, motion capture data was collected at 100Hz. To compute the features, we down-sampled it to 50Hz. The motion capture-based features are related to proxemics and kinesics as they both play an important role in nonverbal communication and social interaction (Hans and Hans, 2015).

4.2.1.1 Proxemics Features

Proxemics is the study of how humans use and structure space around them (Hall, 1966). As empirically demonstrated by Ashton et al. (1980), we expect groups that are standing closer together to not interpret the presence of others as invading, meaning they have

Table 4.1: Summary of all the features extracted. They are computed from motion capture or audio data and are either extracted from the individuals or the group. A "⋆" indicates that the mean, standard deviation, minimum, maximum, and skewness were applied.

		Individual	**Group**
Motion capture	**Proxemics**	Distance from group barycenter ⋆ Total distance traveled	Histogram of the interpersonal distances ⋆ Maximum of the interpersonal distances ⋆ Time in F-formation ⋆
	Kinesics	Longitudinal posture expansion ⋆ Lateral posture expansion ⋆ Occupied volume ⋆ Kinetic energy ⋆	Average amount of motion ⋆ Difference ratio of motion ⋆ Touches' duration ⋆ Synchrony among kinetic energies Average amount of hands movements while not moving ⋆ Difference ratio of hands movements while not moving ⋆
Auditory	**Turn-taking**	Laughter duration Total speaking time	Time of overlapping speech Average turn duration
	GeMAPS	**Individual**	
		Pitch Jitter Shimmer Loudness HNR F1, F2, F3 frequencies and F1 bandwidth	F1, F2, F3 relative energies H1-H2 H1-A3 Spectral slope (0-500Hz and 500-1500Hz) Alpha ratio (50-1000Hz and 1-5kHz) Hammarberg Index (0-2kHz and 2-5kHz)

stronger social bonds, and to trigger positive affective reactions. We computed distance-based features using Equation 4.1:

$$d(a,b) = \sqrt{\sum_{n \in N} (b_n - a_n)^2} \tag{4.1}$$

For the distances that are computed over the transverse plane (here the XZ-plane, see Figure 3.4), $N = \{x, z\}$ and point a have Cartesian coordinates (a_x, a_z) while point b have coordinates (b_x, b_z). For the 3D distances, then $N = \{x, y, z\}$ and point a have Cartesian coordinates (a_x, a_y, a_z) while point b have coordinates (b_x, b_y, b_z).
The following distance-based features are then extracted:

- *Histogram of the interpersonal distances (G)*: The 2D Euclidean distances between each pair of the hips of the participants on the transverse plane were computed frame by frame. The instantaneous position of the hips of the participants, at a time t, is indicated as $hip\_p_i(t)$. Then, the distances are clustered into three bins to reflect the different categorizations of interpersonal distances introduced by Hall (1966): Public space ($>$ 3.6m), Social space (in 3.6m and 1.2m) or Personal space ($<$ 1.2m), respectively (see Figure 4.1 for an example).

- *Maximum of the interpersonal distances (G):* Based on the interpersonal distances computed previously, the maximum distance among the three pairs of hips at each frame is selected. The mean, standard deviation, minimum, maximum, and skewness statistics are then applied over the values obtained over the whole time window.

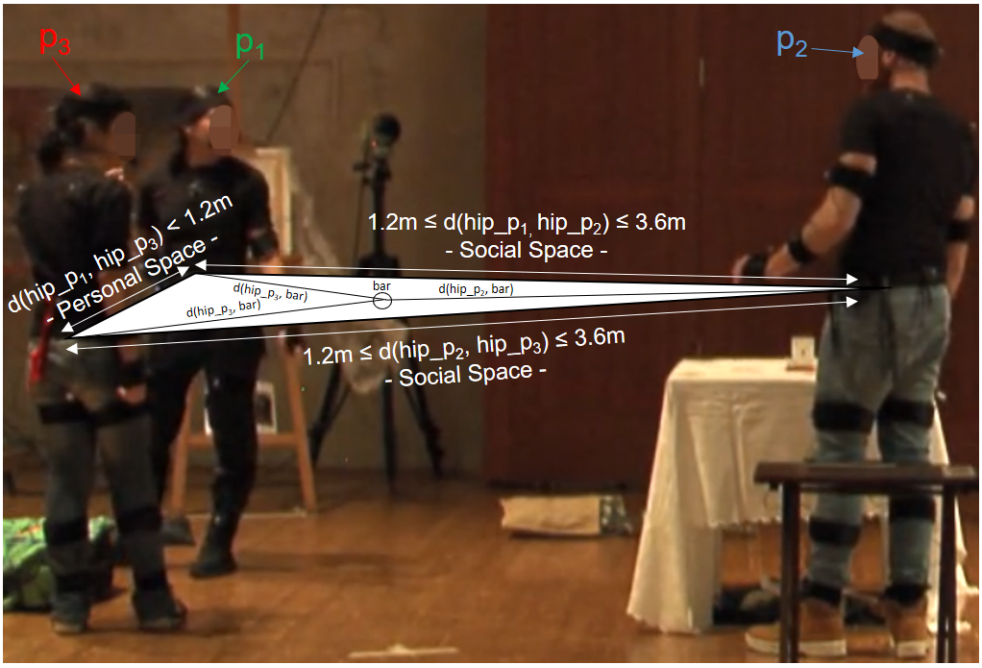


Figure 4.1: Example of the different distances computed: (1) interpersonal distances, here clustered into Social and Personal distances, according to Hall (1966) and (2) distances from the hip of each group member and the group barycenter.

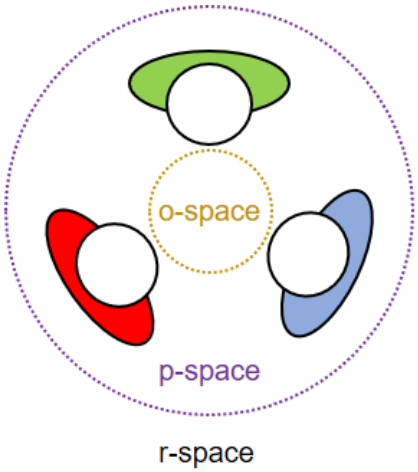
- *Distances from the group barycenter (I):* We compute the Euclidean distances over the transverse plane between the hip of each group member (i.e., hip_p_i, hip_p_j, hip_p_k, respectively) and the group barycenter (i.e., bar). Such a group barycenter corresponds to the barycenter of the triangle shaped by the hips of the three group members. Distances are computed at each frame, following Equation 4.2 (see a visual example in Figure 4.1).

$$d(hip\_p_i, bar) = \frac{d(hip\_p_i, hip\_p_j) + d(hip\_p_i, hip\_p_k)}{3} \tag{4.2}$$

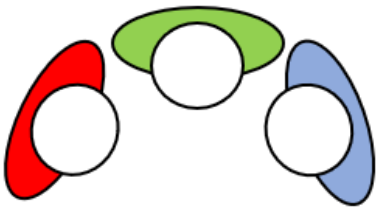
  With $hip\_p_i \neq hip\_p_j \neq hip\_p_k$, corresponding to the hips of the three persons composing the group and bar, the barycenter of the group. The mean, standard deviation, minimum, maximum, and skewness statistics are then applied to all the distances computed over the time window, for each group member.

- *Total distance traveled (I)*: This feature corresponds to the total length of the trajectory covered by the hip of each group member on the transverse plane during the whole duration of the time window. Equation 4.1 is used to compute the distance traveled by a hip between two consecutive frames. This feature is computed for each group member.

In an effort to capture how a group structures itself in the space, we focus on detecting specific spatial formations by detecting the *facing formation* (or F-formation) of a group. Concretely, an F-formation occurs when two or more group members are engaged in a joint activity (Kendon, 1990), and denotes a shared-interested in the interaction (Kendon,



(a) Spaces defined during a F-formation



(b) Semi-circular F-formation

Figure 4.2: Figure 4.2a displays the three regions formed by the persons during a F-formation (i.e., o-space, p-space and r-space), according to Kendon (1990). In particular, Figure 4.2a and Figure 4.2b show a circular and a semi-circular F-formation, respectively.

2010). According to Kendon (2010), when a group involving at least three persons arranges itself in an F-formation, the members' bodies define three regions: the inner o-space, the ring of p-space, and the surrounding r-space (see Figure 4.2a). Here, we focus on the 3-person F-Formations. These include circular and semi-circular arrangements (see Figure 4.2). Below are the features extracted based on the F-Formations detection:

- *Time in F-formation (G)*: It is the amount of time during which a group makes a circular or a semi-circular F-formation. To automatically detect these F-formations, a cone is computed from the chest of each member, in the direction in which she is facing. This is done to approximate the area where the group members' attention is directed. An F-Formation is detected when the cones of every group member intersect (i.e., an o-space exists).

  Multiple steps are needed to define each group member's cone. First, the direction in which each group member is facing is obtained to compute the cone of attention in the correct direction. This "facing" vector is obtained, for each group member, by rotating the vector obtained from the angular displacement of its hip over 1s by 90°. The length of the facing vector is then normalized and multiplied by 3600. In that way, it does not go beyond the Public space (i.e., 3.60m). Then, starting from the chest, a cone is constructed around the facing vector, using the *Shapely* Python package (Gillies et al., 2007), by rotating it by $\frac{1}{3}\pi$ (i.e., +60°) and by $\frac{2}{3}\pi$ (i.e., 120°), given an inner angle of $\frac{1}{3}\pi$, or 60° in the middle.

  We check if the cones of the three group members all intersect and if they share an overlapped area for at least 1s. If both conditions are met, we consider that an F-Formation exists. In particular, if the group barycenter is in the overlapped area, we approximate such an arrangement as a circular F-Formation. Otherwise, we

consider it as a semi-circular one. We, however, did not focus on detecting specific types of F-Formations as the distinction between them might require a more robust and precise F-Formation detection method.

Finally, we apply the mean, standard deviation, minimum, maximum, and skewness statistics to the time in F-formation over the whole time window.

4.2.1.2 Kinesics Features

Kinesics concerns the study of how humans communicate using posture and gesture (Birdwhistell, 2010). They may indicate active engagement in the task and thus are expected to have a positive impact on predicting cohesion (Goldin-Meadow and Alibali, 2013). Features related to the posture are expected to be particularly associated with dominance and hierarchy since small differences and big overall expansion are positively correlated to Social cohesion (Weisfeld and Beresford, 1982) and emotion (Tracy and Robins, 2004). Thus, we extracted features related to the variations of the posture of each group member over the transverse and frontal planes as well as through the volume variations of the bounding boxes computed around each group member. This bounding box method is inspired by Piana et al. (2013). Here, are the posture-related features we extracted:

- Longitudinal posture expansion (I): At each second, we first compute the maximum 3D distance possible between one foot to the head (named md_lon) to approximate each group member's maximum height size. In detail, md_lon is obtained by computing the distances between the sensors located in the feet, thighs, upper legs, hip, chest, and head, using Equation 4.1. Then, we normalize the head position (i.e., $head\_p_i$) over the ordinate axis by md_lon. The minimum between the result obtained by this normalization operation and one is selected as in Equation 4.3:
  Let the head point $head\_p_i$ of group member i have 3D Cartesian coordinates $(head\_p_{i,x}, head\_p_{i,y}, head\_p_{i,z})$. The longitudinal expansion for each group member $i \in \{1, 2, 3\}$ is given by:

$$lon\_exp_i = min\left(\frac{head\_p_{i,y}}{md\_lon_i}, 1\right) \tag{4.3}$$

  In this way, we get a value between zero and one, indicative of the longitudinal expansion of each group member proportional to its size. The mean, standard deviation, minimum, maximum, and skewness statistics are then applied over the entire time window.

- Lateral posture expansion (I): Similarly to the longitudinal expansion features, we first compute, at each second, the maximum 3D distance possible between the sensors located from the left hand to the right hand (including sensors on the forearms, arms, and chest) to approximate each person maximum lateral size (named md_lat). Then, based on the coordinates on the transverse plane of every body joint (i.e., 17 in total), the smallest enclosing circle is computed using Nayuki's Python implementation[1] of a variant of the Welzl (1991)'s algorithm. The radius

[1] https://github.com/nayuki/Nayuki-web-published-code/blob/master/smallest-enclosing-circle/smallestenclosingcircle.py

of this circle corresponds to the lateral expansion over the transverse plane. It is then normalized by md_lat. The lateral posture expansion is obtained following the same procedure as in Equation 4.3 by replacing $head\_p_{i,y}$ by the radius of the smallest enclosing circle of group member i and by swapping md_lon_i by md_lat_i. As previously, the mean, standard deviation, minimum, maximum, and skewness statistics are then applied over the entire time window.

- Occupied volume (I): We first approximate the maximum volume that each group member could possibly occupy (i.e., max_vol) based on its initial position at the beginning of a task, following Equation 4.4. Each member, indeed, started with the "T-pose" which consists of standing still and opening up both arms to mimic the letter "T". Figure 4.3 shows a person from the GAME-ON dataset doing a T-pose. Then, at each second, the occupied volume ($OccVol$) is computed as in Equation 4.4:

$$OccVol = \frac{1}{max\_vol} \prod_{n \in \{N\}} (max(\{body\_joints\})_n - min(\{body\_joints\})_n) \tag{4.4}$$

With $\{body\_joints\}_n$ the list of the coordinates of all of the 17 body joints from axis $n \in N = \{x, y, z\}$. Figure 3.4 shows the location of each body joint. As for the other expansion-related features, such a volume is computed at each second and the mean, standard deviation, minimum, maximum, and skewness statistics are then computed over the entire time window.

Figure 4.3: An example of "T-pose". This pose enables the calibration of the motion capture system. Moreover, it was used as a marker for the beginning of a task and to compute the Occupied Volume feature.

Gestures might be intentional (e.g., touching someone else to give comfort, Sahi et al., 2021) or unconscious (e.g., moving hands to unconsciously mimicry another person,

Van Baaren et al., 2009). In both cases, they could be indicative of a large range of meanings (Calbris, 2011). Since we are interested in features that are specifically relevant for cohesion, we focus on extracting features related to body movements, with a particular emphasis on the hands. Hands movements are, indeed, a vector for specific emotions communication (Wallbott, 1998). The amount of hand movement within the group might also be indicative of the group engagement in the task. We extract the kinetic energy of each group member as well as the amount of synchrony between these kinetic energies. Synchrony refers to the ability of a group to coordinate collective action efficiently and it has been proved to be positively related to cohesion and cooperation (e.g., Wiltermuth and Heath, 2009; Gordon et al., 2020). Finally, we are interested in touch-related features among the group members. Signaling by touch can, indeed, work both at communicating task-related information (e.g., a tap on the shoulder to require attention) as well as conveying social status (Saarinen et al., 2021) and emotions (e.g., hugging another person, Teyssier et al., 2020). The sensitivity of motion capture allowed us to approximate haptic communications without the use of tactile sensors. The following gesture-based features are then extracted:

- *Group amount of motion (G)*: We compute the average change in the position of each group member's chest coordinates in the transverse plane over 1s. Then, based on such values, we compute the average amount of motion among the three group members by averaging them over 1s and we also compute the difference ratio (i.e. the difference between the highest and the lowest amount of motion in the group) over 1s too. Finally, the mean, standard deviation, minimum, maximum, and skewness statistics are then applied over the entire time window on the average amount of motion and on the difference ratio.

- *Group amount of hands movement while not moving (G)*: Similarly to the group amount of motion feature, we compute the group amount of hand movement while not moving as follows: first, we compute the average change in the position of each group member's left and right hands coordinates, in the transverse plane over 1s. These values are only computed while the person is not walking. Here, we approximate that a person is not walking if the distance traveled of the chest joint, on the transverse plane, is less than 50cm over 1s. This choice was made to only account for the hands movements that are not provoked by group members' displacement (i.e., a lot of arm swinging happens during a walk, Collins et al., 2009). At each second, the mean between both hands movement of the three group members is applied.

  We also calculate the amount of 3D hands rotation by computing, for each person and for each of its hands, the distance (accounting for the sign ambiguity) between its quaternions coordinates at time t and $t+1$ using the *Pyquaternion* Python package[2]. The mean of the amount of rotation between both hands is obtained for each group member. Again, these values are only computed when the person is not walking and the mean of the amount of rotation of each group member's hands is obtained. Statistics (i.e., mean, standard deviation, minimum, maximum, and skewness) are applied to the group amount of hands motion and rotation.

[2] https://pypi.org/project/pyquaternion/

- *Kinetic energy (I)*: We compute the total kinetic energy of the whole body (K_{tot}), for each group member by summing the kinetic energies (i.e., translational and rotational) of each of the 17 body joints, as follows in Equation 4.5:

$$\begin{aligned} K_{tot} &= \sum_{j \in \{J\}} K_{translational}(j) + K_{rotational}(j) \\ &= \frac{1}{2} \sum_{j \in \{J\}} \left(m_j v_j^2 + I_j \omega_j^2 \right) \end{aligned} \tag{4.5}$$

  Where $\{J\}$ is the list of the 17 body joints, m_j is the mass of body joint j (for the sake of simplicity, m_j is, here, set to one), v_j is the velocity of body joint j, I_j, the moment of inertia (here, also set to one) and ω_j, the angular velocity of body joint j, using the Euler angles derived from the quaternion coordinates obtained with the *squaternion* Python package[3]. For each group member, the mean, standard deviation, minimum, maximum, and skewness statistics are then applied over the whole time window.

- *Synchrony among kinetic energies (G)*: This feature is computed using a modified version of an algorithm implemented in the SyncPy library (Varni et al., 2015) for extracting the S-Estimator, a measure of the total synchronization of multiple signals, relying on the eigenspectrum of the correlation matrix of such signals (Carmeli et al., 2005). Here, the signals are the kinetic energies of the three group members. The S-Estimator has a range between zero (for completely independent signals) and one (for fully synchronized signals). It is computed as follows in Equation 4.6:

$$S = 1 + \frac{\sum_{i=1}^{K} \lambda_i log(\lambda_i)}{log(K)} \tag{4.6}$$

  Where K is the number of signals, here set to the number of group members (i.e., three), and λ_i are the normalized eigenvalues of the signals' correlation matrix. We use such an S-Estimator as an approximation of the synchrony among the three group members, at each time window.

- *Touch's duration (G)*: This feature is based on the 3D distance between the hands of a group member and the upper body of another group member (it includes the chest, the head, the hip, and the arms and shoulders). We compute a binary value every second that indicates whether a touch occurred or not within the group. If one of the hands' joints is less than 15cm away from the upper body joints of another member, we approximate this distance as close enough to be considered as a touch. We choose a conservative threshold of 15cm for touch detection as sensors are not located at the fingertips but rather on the top of the hand. Also, such a threshold enables to capture the touches that occur at areas around the sensor (e.g. touching a participant's elbow instead of their forearm). We apply mean, standard deviation, minimum, maximum, and skewness statistics on the time at which touches occurred, over the whole time window.

[3] https://pypi.org/project/squaternion

Once the kinetic energy and posture expansion-related features were computed on a specific window size, we applied a Savitzky-Golay filter (Savitzky and Golay, 1964) with a polynomial order of five and a coefficient of three to reduce noise. Figure 4.4 shows an example of a Savitzky-Golay filter applied on the mean over 20s of the longitudinal expansion of a group member along the five tasks.

Figure 4.4: Example of denoising using a Savitzky-Golay filter on the average longitudinal expansion of a group member computed over windows of 20s. Here, we display the feature's values for the whole interaction (i.e., across the five tasks of GAME-ON), resulting in a sequence of 125 windows. The blue signal is the raw feature while the orange one is the filtered feature.

4.2.2 Audio Feature Extraction

As detailed in Chapter 3, audio was recorded for each group member at 48kHz using wireless headsets microphones. Each audio file was inspected and external noises were reduced to improve the voice quality and clarity using the Audacity software[4]. Concretely, during the speaker turn, we used the Noise Reduction feature[5] to reduce constant background sounds by selecting a region in the waveform that is characteristic of the noise to reduce. Otherwise, we muted the section where the person is not speaking since the audio of the three group members is available separately, hence, avoiding unrelated noises (e.g., someone else speaking while the person of interest is not). The following audio features are either automatically extracted using the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) (Eyben et al., 2015) or using hand-crafted algorithms to compute features related to turn-taking. In both cases, these features are particularly relevant to measure interactive behavior and capture affective processes such as group emotion (e.g., Ringeval et al., 2016) and cohesion (e.g., Hung and Gatica-Perez, 2010) as well as other social processes (e.g., leadership, Scherer et al., 2012).

[4] https://www.audacityteam.org/
[5] https://manual.audacityteam.org/man/noise_reduction.html

4.2.2.1 GeMAPS features

Features of the GeMAPS (Eyben et al., 2015) were extracted using the OpenSmile software (Eyben et al., 2010). We chose the GeMAPS minimalistic acoustic parameter set since it has been successfully used in many affect-related prediction tasks (e.g., emotion prediction, Ringeval et al., 2016; Chao et al., 2015). Moreover, it has been experimentally proven useful for predicting various other processes (e.g., amusement and interest, Goudbeek and Scherer, 2010). GeMAPS focuses on a set of 18 features. Some are related to the voice frequency (i.e., pitch, jitter and Formant 1,2 and 3 frequency) and are particularly relevant for describing vocal affective expressions and, in particular, anger and sadness (Goudbeek and Scherer, 2010). Others are related to the voice energy and amplitude (i.e., shimmer, loudness, and harmonics-to-noise ratio) and are pertinent to detect, for example, stress (Weninger et al., 2013). Finally, some features are related to the spectral balance of the voice (i.e., alpha ratio, Hammarberg index, spectral slope 0-500Hz and 500-1500Hz, formant 1,2 and 3 relative energy, harmonic difference H1-H2, and H1-A3) and had been successfully used for the detection of angry speech (Tahon and Devillers, 2010), and are also important for vocal valence and arousal (Goudbeek and Scherer, 2010).

As described by Eyben et al. (2015), pitch, harmonic differences, HNR, jitter, and shimmer are computed from overlapping windows that are 60 ms long and 10 ms apart. The windows are multiplied with a Gaussian window (with $\sigma = 0.4$), in the time domain prior to the transformation to the frequency domain with a Fast Fourrier Transform (FFT). No window function is, however, applied for the jitter and shimmer since they are computed in the time domain. Loudness, spectral slope, spectral energy proportions, Formants, Harmonics, Hammarberg Index, and Alpha Ratio are computed from 20 ms windows that are 10 ms apart and a Hamming function is applied to these windows. Zero-padding is applied to all windows to the next power-of-2 (samples) frame size in order to be able to efficiently perform the FFT. Then, all these features are smoothed over time with a symmetric moving average filter over three windows long. Pitch, jitter, and shimmer are, however, only smoothed within voiced regions. For each feature, the mean is applied over the whole time window. In addition, the mean of the Alpha Ratio, the Hammarberg Index, and the spectral slopes from 0-500 Hz and 500-1500 Hz over all unvoiced segments are included, yielding a total of 22 individual features.

The brief description of the features computations are taken from the work of Eyben et al. (2010):

Frequency related features:

- *Pitch (I)*: This feature is based on the fundamental frequency (F_0) which is computed via subharmonic summation (SHS) in the spectral domain, as described by Hermes (1988). This value is converted from its linear Hz-scale to a logarithmic scale. Thus, the starting semitone frequency (i.e., semitone 0) starts at 27.5 Hz. Every value below semitone 1 (i.e., 29.136 Hz) is, however, set to one as zero is reserved for unvoiced regions.

- *Jitter (I)*: It is computed as the average (over one 60 ms frame) of the absolute period to period local jitter $J_{pp}(n')$ scaled by the average fundamental period length

as described by Hermes (1988), following Equation 4.7:

$$J_{pp}(n') = |T_0(n') - T_0(n'-1)| \; for \; n' > 1 \tag{4.7}$$

With $T_0(n'-1)$ and $T_0(n')$, the length of two consecutive periods $n'-1$ and n', respectively. To make the jitter value independent of the underlying pitch period length, the average of each J_{pp} value is finally scaled by the average pitch period length.

- *F1 (I), F2 (I)* and *F3 (I) frequencies* and *F1 bandwidth (I)*: All of these four features are computed from the roots of Linear Predictor (LP) coefficient polynomial (Makhoul, 1975). Full details of implementation are available in Boersma and Weenink (2001)'s study.

Energy and amplitude related features:

- *Shimmer (I)*: It is computed as the average (over one 60 ms frame) of the relative peak amplitude differences, expressed in dB. Similarly to the jitter feature, the local period to period shimmer is computed as follows in Equation 4.8:

$$S_{pp}(n') = |A(n') - A(n'-1)| \; for \; n' > 1 \tag{4.8}$$

With the peak to peak amplitude difference $A(n') = x_{max,n'} - x_{min,n'}$ (i.e., the maximum and minimum amplitude of the pitch period n'). Finally, the relative shimmer values are averaged and normalized by the per frame average peak amplitude.

- *Loudness (I)*: This feature is an estimate of the perceived signal intensity from an auditory spectrum. First, a non-linear Mel-band spectrum is constructed by applying 26 triangular filters equidistantly distributed on the Mel-frequency scale from 20-8000 Hz to a power spectrum computed from a 25 ms frame. An auditory weighting with an equal loudness curve is then performed. Then, it is followed by a cubic root amplitude compression, applied for each band b of the equal loudness weighted Mel-band power spectrum, resulting in a spectrum that we refer to as the auditory spectrum. Loudness is then computed as the sum over all the bands of the auditory spectrum.

- *Harmonic-to-Noise Ratio (HNR) (I)*: It gives the energy ratio of the harmonic signal parts to the noise signal parts in dB. It is estimated from the short-time autocorrelation function (ACF) on a 60 ms window as the logarithmic ratio of the ACF amplitude at F_0 and the total frame energy. As in Schuller (2013)'s study, HNR ratio is computed following Equation 4.9:

$$HNR_{acf,log} = 10log_{10}\left(\frac{ACF_{T_0}}{ACF_0 - ACF_{T_0}}\right) \tag{4.9}$$

Where ACF_{T_0} is the amplitude of the autocorrelation peak at F_0 and ACF_0 is the 0^{th} ACF coefficient (equivalent to the quadratic frame energy). The logarithmic HNR value is floored to -100 dB to avoid highly negative and varying values for low-energy noise.

Spectral (balance) related features:

- *Spectral Slope (0-500 Hz and 500-1500 Hz) (I)*: These two features are computed from a logarithmic power spectrum by linear least squares approximation (Tamarit et al., 2008). Both are extracted over the low- and high-frequency regions (i.e., 50-1000 Hz and 1-5 kHz, respectively), resulting in four individual features.
- *Alpha Ratio (I)*: It corresponds to the ratio between the energies in the low-frequency region (i.e., 50-1000 Hz) and in the high-frequency region (i.e., 1-5 kHz). Such ratios are computed every 20ms, resulting in two individual features.
- *Hammarberg Index (I)*: It is the ratio of the strongest energy peak within a particular region of the spectrum. Here, such an index is computed for the 0-2 kHz and 2-5 kHz regions, respectively. Thus, two individual features are extracted from this index.
- *F1, F2, and F3 relative energies (I)*: These three individual features are computed from the linear frequency scale power spectrum by summing the energy of all bins in the bands 0-500 Hz and 0-1000 Hz, normalized by the total frame energy (i.e., the sum of all the power spectrum bins).
- *Harmonic Difference (H1-H2 and H1-A3) (I)*: These two individual features are computed from the amplitudes of F_0 harmonic peaks in the spectrum, normalized by the amplitude of the F_0 spectral peak. Here, the focus is on the ratio of the first to the second harmonic (i.e., H1-H2) and on the ratio of the first harmonic to the third formant's amplitude (i.e., H1-A3). The third formant's amplitude is estimated as the ratio of the amplitude of the highest F_0 harmonic peak in the range $[0.8 \cdot F_i; 1.2 \cdot F_i]$ to the amplitude of the F_0 spectral peak. F_i refers to the centre frequency of the first formant.

4.2.2.2 Turn-taking Features

The power of using turn-taking related features has been demonstrated by several studies addressing dominance (e.g., Mast, 2002; Jayagopi et al., 2009), and cohesion in particular (e.g., Hung and Gatica-Perez, 2010; Nanninga et al., 2017). Thus, we computed the average turn duration over the group. In an extremely involved conversation, turns duration of each participant is, indeed, theorized to be approximately equal (Hung and Gatica-Perez, 2010). Also, we would expect that, in highly cohesive groups, turns will tend to be shorter as everyone would freely contribute to the conversation. In addition, we kept track of the total speaking time of each group member as well as the amount of time in which their speeches overlap. Overlapping of speeches can, indeed, be symptomatic of conflict (West and Zimmerman, 2015) or be a sign of engagement (Hilton, 2016) and cooperation (Tannen, 1994) between group members. Finally, the laughter duration per group member is recorded as it is a highly social process (Provine, 1993) that is a good indicator of cohesion (Glenn, 2003; Kantharaju and Pelachaud, 2020). Except for the laughter duration feature that was extracted from the raw audio, we first used the voice activity detector (VAD) (Eyben et al., 2013) from the Opensmile software (Eyben et al., 2010) to construct a speech matrix from the three separate audio sources. Such a speech

matrix indicates who and when someone is speaking, over the whole window duration (i.e., $p_{i,t} = 1$ if person i is speaking at time time t, otherwise $p_{i,t} = 0$). Thus, we can compute the speaking duration by summing $p_{i,t}$ in the region of interest of the speech matrix. Based on the speaking duration, the average turn duration, the time of overlapping speech, and the total speaking time were computed. Figure 4.5 shows the features computed from the speech matrix. Below are the computational details of every turn-taking related feature:

- *Average turn duration (G)*: It is the average duration of all the turns occurring in the group, during the whole time window. A turn is considered over when a member stops speaking for at least 1s. Such a value was chosen because a great proportion of turn transitions fall between 100ms before one stops speaking and 500ms after the end of one's turn (Levinson and Torreira, 2015), hence, with 1s, we ensure that a turn is over.

- *Total speaking time (I)*. This feature is computed, for each group member, as the total time she is speaking over the whole time window. To avoid counting the small utterances, we assume that a member is speaking if she speaks for at least 1s. Small utterances, indeed, usually last between 740ms (e.g., when naming two nouns) and 900ms, the time required for three-word utterances (Schnur et al., 2006).

- *Time of overlapping speech (G)*: It is the total time for which at least two group members speak simultaneously. This feature is computed over the whole time window. As opposed to the total speaking time feature, here, we also account for the short segments in which a few milliseconds of overlap occurs.

- *Laughter duration (I)*: It is computed from the raw audio by using the automatic laugh detector developed in Ryokai et al. (2018) to automatically extract the laughs in the window. Once all the laughs are extracted, the total time of laughing is computed for each member. As a laugh might be brief (a few milliseconds), we do not put any time constraint on the detected length of the laugh.

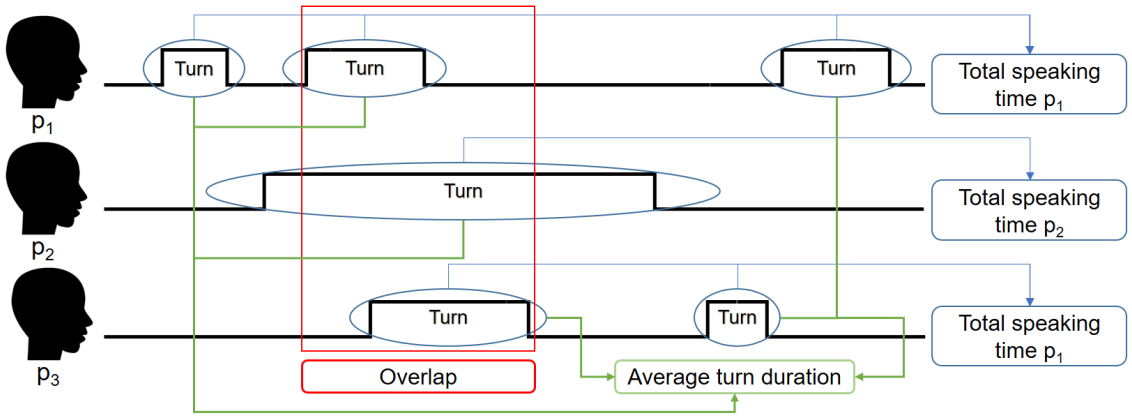


Figure 4.5: Example of the turn-taking related features that are computed over a speech matrix (i.e., total speaking time, average turn duration, and time of overlapping speech).

4.3 Conclusion

IN this Chapter, we presented the most important nonverbal features used in previous studies on automated analysis of cohesion. Then, we motivated and provided computational details of the multimodal nonverbal features that we extracted from both motion capture data and audio data. Features calculated from the motion capture data are related to kinesics and proxemics, while the ones computed from the audio data are from the GeMAPS features set and related to turn-taking. For both motion capture data and audio data, features are either computed from individuals or from the group as a whole.

Chapter 5

Computational Models of Cohesion

Contents

THIS Chapter presents a collection of computational models of cohesion. The designs of our models are thought to implement approaches following the four research axes presented in Chapter 2. First, the aim of the models is presented

and the settings that are shared by each one of them are explained. These include the data used as input, the data augmentation procedures, and the labeling strategy. In addition, we describe how we evaluated and compared the different models among them. Then, the computational models are presented. Except for the baseline that follows current literature approaches, they were all designed to investigate at least one of the research axes. In fact, a first model addressing the temporal nature of cohesion (i.e., RA1) is detailed, followed by a model that, in addition to RA1, investigates the group modeling (i.e., RA2) and another model that addresses both RA1 and RA3 (i.e., the interplay between the Social and Task dimensions of cohesion). Then, a set of four models that integrate RA1, RA2, and RA3 following different approaches are described. Finally, the results are discussed.

In particular, I co-designed the labeling strategy and a first version of the baseline and I designed and implemented all of the other computational models of cohesion. In addition, I ran all the analysis (see Maman, 2020; Walocha et al., 2020; Maman et al., 2021b, for the resulting publications).

5.1 Introduction

As of today, there is a limited literature on the automated analysis of cohesion (cf. Chapter 2). Existing computational models show the potential of using machine learning and deep learning models for such a task and set the path for more complex models, able to integrate the complexity of cohesion. Thus, following the research axes, we designed and implemented various computational models that range from a simple but consolidated state-of-the-art approach to more sophisticated approaches that increasingly address the temporal nature of cohesion (i.e., RA1), group contributions (i.e., RA2) and the interplay between the Social and Task dimensions of cohesion (i.e., RA3).

All of our computational models share the same goal which is to predict the dynamics (i.e. *decrease* and *no-decrease*) of cohesion across an interaction. In our configuration, it means that our models predict the variations of cohesion, focusing on its Social and/or Task dimensions, for each of the tasks of the escape game.

5.2 Common Settings of the Models

5.2.1 Input Data

We extracted multimodal features for 15 groups of the GAME-ON dataset. As groups interact for a long duration (i.e., 35mn 45s $\pm$4mn 2s on average to complete the five tasks of the escape game), all the models, except one, only use the two last minutes of each task. This choice was motivated by the fact that we use the self-assessments provided by the group members. As reported in several studies carried out in different contexts, self-assessments collected through questionnaires are, indeed, likely influenced by the last recalled behavior (e.g., Lord et al., 1978; Kamper et al., 2010).

5.2.2 Data Augmentation

To avoid overfitting and make the models more robust to noise, we augment the training data following two strategies. The first one consists of creating synthetic groups by adding Gaussian noise to all the features (μ= 0, $\sigma \in$ {0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5}), for each group. We tested each value of sigma to investigate the effect of such settings on the performances of the models. Thus, for each model, we only select the value of σ that maximizes performances on the validation set. This approach results in augmenting the data by a factor of four. Also, as some features are computed from individuals, the second strategy involves creating synthetic groups by computing the six permutations of the order of the three group members. This strategy augments the training data by a factor of six. Only the first two models presented in this Chapter do not apply such a data augmentation, hence, the training data for these models is augmented only by a factor of four only, whereas other models' training data is augmented by a factor of 24.

5.2.3 Labeling Strategy

We considered the task of predicting cohesion dynamics as a binary classification problem (*decrease* vs *no-decrease*). Starting from the self-assessments of cohesion rated by each group member, we built labels for *decrease* vs *no-decrease* of the Social and Task dimensions. We first focused on these assessments as they reflect the *true* internal state of each group member (Uleman et al., 2008). The labeling strategy is computed as follows. Let's consider the GEQ-Social and GEQ-Task scores, computed from the self-assessments of cohesion in Section 3.3, for two consecutive tasks (e.g., Task 1 and Task 2). This results, for each cohesion dimension, in six values: two scores for each of the three members. These scores were then ranked in ascending order to limit the potential bias introduced by the inter-member variance. Next, for each dimension, we computed, the difference between the ranks associated with the two GEQ scores of each group member. Finally, we took the average of these rank differences, resulting in the group score (see Equation 5.1):

$$GS_{Tx} = \frac{1}{n}\sum_{i=1}^{n}\left(rank_{Tx}^{(i)} - rank_{Tx-1}^{(i)}\right) \tag{5.1}$$

with GS_{Tx}, the group score computed for a transition between the tasks T_x and T_{x-1} with x $\in \{1, 2, 3, 4, 5\}$ (Transition T0-T1 is equivalent to transition Start-T1 in Table 3.2); n, the number of group members (here set to 3) and $rank^{(i)}$, the rank corresponding to the associated GEQ score given by group member i. The group score indicates whether cohesion decreased or not for a specific dimension. Finally, this GS score was binarized: a value equal to zero was assigned when the group score was negative (i.e. a decrease in cohesion occurred), whereas a value equal to one was assigned when the group score was zero or positive (i.e. no change or an increase in cohesion occurred).

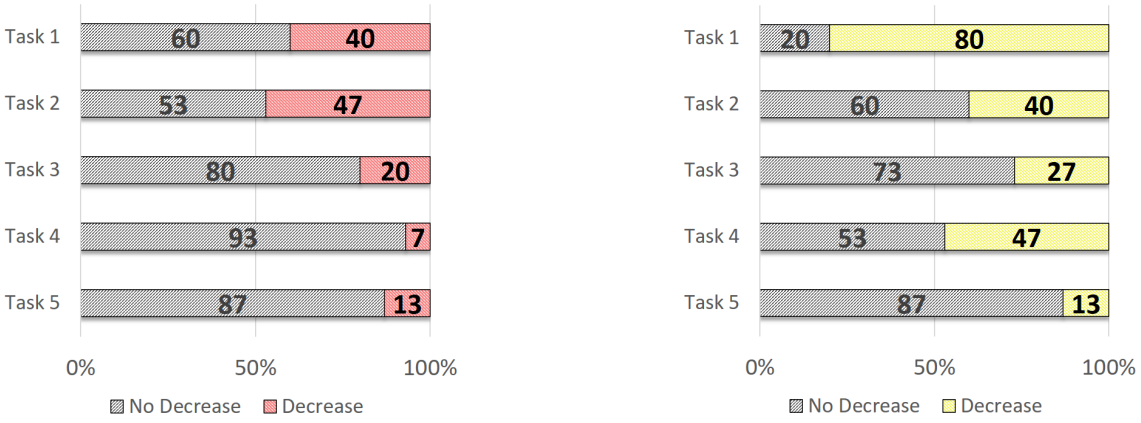
Let's take an example to illustrate how to obtain the GS score for a specific dimension (i.e., Social or Task). Table 5.1 provides fictive GEQ scores given by three persons across two consecutive tasks (here Task 1 and Task 2). These scores range from 23 to 48, hence, their associated ranks, in ascending order, range from 1 to 6, respectively. Then, the rank differences are computed, for each person between Task 2 and Task 1. We obtain: -1 for

Table 5.1: A fictive example of the GEQ scores provided by three persons across two consecutive tasks as well as their corresponding ranks and ranks difference.

	GEQ Score		Rank		Ranks
	Task 1	Task 2	Task 1	Task 2	difference
p_1	41	34	4	3	-1
p_2	45	24	5	2	-3
p_3	48	23	6	1	-5

p_1 (i.e., 3-4), -3 for p_2 (i.e., 2-5) and -5 for p_3 (i.e., 1-6). Thus, the GS score (i.e., the average rank differences for the group) is -3 which is binarized to 0. This is in line with the scores decreasing, for each member, across the two tasks.

Overall, labeling data in this way led to an imbalanced distribution for the Social dimension (i.e., 75% of *no-decrease* labels vs 25% of *decrease* labels) and a balanced distribution for the Task dimension (i.e., 59% of *no-decrease* labels vs 41% of *decrease* labels). The distributions of the labels, for each task and for each dimension, are depicted in Figure 5.1. Such an imbalance in the tasks was, however, expected due to how GAME-ON was conceived (see Table 3.2). The different strategies to address this point are described in the remaining of the Chapter.



(a) Social partition (overall: 75/25%)

(b) Task partition (overall: 59/41%)

Figure 5.1: Labels distributions resulting from our labeling strategy based on self-assessments of cohesion for the Social and Task dimensions of cohesion (see Figure 5.1a and Figure 5.1b, respectively). Overall, Social cohesion is imbalanced (i.e., 75% of *no-decrease* labels) while Task cohesion is fairly balanced (i.e., 59% of *no-decrease* labels). A high imbalance is observed when looking at each task independently.

5.3 Evaluation Methodology and Models' Comparison

A nested Leave-One-Group-Out (LOGO) cross-validation was carried out to account for the high diversity of groups. We split the 15 groups into training, validation, and test sets with ten, four, and one group(s), respectively. Then, we augmented the training set either by a factor of four (if only the strategy using Gaussian noise is applied), resulting in 40 groups, or by a factor of 24 (if both strategies are applied), leading to 240 groups.

In the case of a Deep Neural Network architecture, the models were trained up to 500 epochs with a fixed learning rate of 0.001, and the weights of the models were updated at every mini-batch composed of four groups[1]. Every 10 epochs, the models were evaluated on the validation set. In that way, we determined the optimal number of epochs based on the performances across the five tasks. Then, we retrained the models (merging both training and validation sets), based on the optimal number of epochs.

For all of the Deep Neural Network architectures, we used the same loss during the training phase. This loss accounts for the data imbalance by weighting the binary cross-entropy loss function as in Equation 5.2:

$$\begin{aligned} L_{dim,T_x} = w_{0,dim,T_x} \times [y_{dim,T_x} \log(\hat{y}_{dim,T_x})] \\ + w_{1,dim,T_x} \times [(1 - y_{dim,T_x}) \log(1 - \hat{y}_{dim,T_x})] \end{aligned} \tag{5.2}$$

where L is the loss computed for the dimension $dim \in \{Social, Task\}$ at the task T_x with $x \in \{1, 2, 3, 4, 5\}$, w_{i,dim,T_x} is the weight for class i (i.e., *decrease* vs *no-decrease*) of the corresponding dimension (i.e., Social or Task cohesion) and task (i.e., from Task 1 to Task 5), computed in an inversely proportional way to the class frequency as in Equation 5.3, while y_{dim,T_x} and $\hat{y}_{dim,T_x}$ are the target label and the scalar value in the model output for the corresponding dimension and task, respectively.

$$w_{i,dim,T_x} = \frac{n_g}{n_c * n_{i,dim,T_x}} \tag{5.3}$$

where n_g is the number of groups; n_c, the total number of classes (i.e., *decrease* and *no-decrease*) and n_{i,dim,T_x}, the number of occurrences for class i of dimension dim in task T_x. The heuristic for computing the class weights in such a way is inspired by King and Zeng (2001).
Finally, the total loss is computed as in Equation 5.4:

$$L_{total} = -\frac{1}{\lambda_1 + \lambda_2} \sum_{t \in T_x} (\lambda_1 L_{Social,t} + \lambda_2 L_{Task,t}) \tag{5.4}$$

with $\lambda_i \in \{0, 1\}$. $\lambda_1 = 0$ if Social cohesion is not predicted by the model while $\lambda_2 = 0$ if Task cohesion is not predicted by the model.

Furthermore, to limit the randomness present in the models (e.g., due to the initialization of the weights and biases, and in regularization like dropouts), we followed recommendations from Colas et al. (2018) that suggest evaluating models on several seeds (between five and 25 depending on the data and algorithms) to obtain a reliable assessment of the models' performances. Thus, we used 15 random seeds and we averaged the performances over the seeds. Performances were evaluated using F1-score as this metric accounts for the label imbalance (e.g., Goutte and Gaussier, 2005; Hammerla et al., 2016). More specifically, we compute the average F1-score for each of the dimensions predicted by the model and for each task, independently, across the 15 rounds of the LOGO and the 15 seeds. Finally, to explore the impact of different values of σ during

[1]These values were fixed building on preliminary studies.

the data augmentation, we used a similar value for the 15 rounds of the LOGO. Thus, we stored the performances on both the validation and test sets so we can select the values of σ with which we maximized performances on the validation set. Algorithm 1 shows the pseudo-code for the full LOGO cross-validation procedure for a single seed.

We assessed potential significant differences between the performances through a computationally intensive randomization test. In detail, we performed a k-sample permutation test using the *perm* package developed in R (Fay and Shaw, 2010). Such a test performs exact calculations using the Monte Carlo method during the permutation test. It is a non-parametric test avoiding the independence assumption between the results being compared and that is suitable for non-linear measures such as F1-score (Yeh, 2000). The significance level α was at 0.05. In case of multiple comparisons (i.e. comparisons between the performances of more than two models or between the performances over the five tasks in the same model), a post-hoc analysis was carried out using pairwise permutation with an FDR adjusted p-value (Benjamini and Hochberg, 1995). Such a p-value correction controls the false discovery rate (i.e., the expected proportion of false discoveries among the rejected hypotheses), hence, integrating the rate of Type-I errors in the p-value computation.

All the models presented in this study were developed and trained using Python 3.7 and Tensorflow 2.6 on NVIDIA V100 GPUs.

5.4 Feature Subsets

We grouped the features described in Chapter 4 in different subsets. In that way, we enable the computational models to differentiate between the features to investigate different research axes. Thus, we composed the following subsets:

- The "*Full features set*" (FFS): it takes all of the features, without differentiation.
- The "*Individual features set*" (IFS): it contains only the features computed from individuals (e.g., kinetic energy).
- The "*Group features set*" (GFS): it gathers only the features computed from the group as a whole (e.g., time of overlapping speech)
- The "*Task-specific*" (TFS): it regroups all the features that are particularly relevant for Task cohesion.
- The "*Social-specific*" (SFS): it contains all the features that are relevant for Social cohesion.
- The "*Common features set*" (CFS): it has all the features that are not in SFS nor in TFS, hence, that are relevant for both dimensions.

IFS and GFS are used by the models addressing RA2, while TFS, SFS, and CFS are exploited by two models investigating RA3. While composing IFS and GFS is straightforward, we describe below the methodology to obtain TFS, SFS, and CFS. Such a methodology requires a pre-trained model of cohesion predicting both Social and Task dimensions to run a post-hoc analysis.

Algorithm 1: Leave-One-Group-Out cross-validation procedure.

Input : Multimodal nonverbal features for the two last minutes of each task
Output: Average F1-score for each dimension and each task
$\mathcal{G}$ = the list of the groups used in the model
$list_\sigma = \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$
$perfs_\sigma$ = [] ; `//Stores performances for each` $\sigma \in list_\sigma$
foreach $\sigma \in list_\sigma$ **do**
  $perfs_{split}$ = [] ; `//Stores performances at each split`
  foreach $g \in \mathcal{G}$ **do**
    test_set = g
    validation_set = four random groups in $\mathcal{G}$[$\sim$ test_set]
    training_set = remaining groups in $\mathcal{G}$
    best_epoch = 0 ; `//For determining the best epoch number`
    $\mathcal{M}$ = [] ; `//For saving models`
    $F1_{val}$ = [] ; `//For saving F1-scores obtained on validation_set`
    augment_data(training_set) ; `//x4 or x24`

    while *train_model(training_set)* **do**
      Save model m in $\mathcal{M}$ every 10 epochs until 500
    end
    foreach $m \in \mathcal{M}$ **do**
      Evaluate m on validation_set
      Store overall F1-score in $F1_{val}$
    end
    Select best_epoch according to the best average F1-score in $F1_{val}$
    augment_data(validation_set)
    train_val_set = training_set + validation_set
    m = train_model(train_val_set) on best_epoch
    Evaluate m on test_set
    Store performances in $perfs_{split}$
  end
  Store average performances of $perfs_{split}$ in $perfs_\sigma$
end
Select performances in $perfs_\sigma$ that achieved the best average F1-score on the validation_set

We exploited Shapley values (Shapley, 1953), a method from coalitional game theory, to explain the predictions of the pre-trained model on the Social and Task dimension of cohesion. A prediction can, indeed, be explained by assuming that each feature value, for a given window, is a "*player*" in a game where the prediction is the "*payout*" (i.e., the prediction of the Social or Task cohesion's dynamics). Shapley values provide insights on how the "*payout*" is distributed among the features (Molnar, 2022), reflecting the marginal contribution of the features' value across all possible coalitions. We computed Shapley values using the SHAP library (Lundberg and Lee, 2017). Thus, the Shapley value explanation is represented as an additive feature attribution method as follows in Equation 5.5:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{5.5}$$

where g is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector, M is the maximum coalition size and $\phi_j \in \mathbb{R}$ is the feature attribution for a feature j. To compute Shapley values, the algorithm simulates that only some feature values are playing (i.e., $z'_j = 1$) and some are not (i.e., $z'_j = 0$).

We computed such Shapley values for each feature, at each round of the LOGO cross-validation. Then, we averaged the Shapley values of each feature across the five tasks and the 15 rounds of the LOGO and we selected the features that obtained the highest ones until we reach 70% of the *payout* explained. Such a threshold was empirically determined to ensure that TFS, SFS and CFS were not empty and approximately similar in size, and was inspired by the method used in factor analysis to determine the validity of a factor (Hair et al., 2010).

Next, we removed the features that are correlated across the SFS and TFS subsets as the SHAP algorithm is randomly selecting one of the correlated features when computing its importance score, meaning that these features could be present in both SFS and TFS. We considered that two features were strongly correlated when their Pearson correlation was over 0.70 with a $p - value < .05$ (Akoglu, 2018). Finally, for the remaining features of each subset, we also included their strongly correlated features into the subset, as suggested by Molnar (2022). In that way, we ensure that all of the important features were retained.

While this methodology enabled the discrimination of the important features, results must be analyzed with care. The obtained Shapley values are, indeed, specific to each model. In fact, different (and sometimes contradictory) subsets of features were obtained for our collection of models. Thus, we will clarify what model had been used and what are the features that constitute SFS, TFS and CFS when describing the results for the models using these subsets.

5.5 Architectures

5.5.1 A Tree-Based Approach as a Baseline

We used a Random Forest classifier (RFC) as a baseline to predict the dynamics of cohesion. As stated by Wainberg et al. (2016), such a classifier is, indeed, one of the most powerful algorithms for solving binary classification problems.

According to the structured survey presented in Chapter 2, this architecture reproduces, at the Model level, the strategies employed in most of the previous computational studies of cohesion (e.g., Hung and Gatica-Perez, 2010; Gonzales et al., 2010; Wang et al., 2012; Nanninga et al., 2017; Fang and Achard, 2018; Zhang et al., 2018; Sharma et al., 2019; Kantharaju et al., 2020). In fact, RFC takes the FFS as input and processes each thin slice independently, without modeling the time dependencies between them nor between the tasks. While such an architecture enables the prediction of the Social and Task dimensions of cohesion, it does not particularly address the relationships between its dimensions. In fact, the same architecture is used to predict both dimensions. Finally, it does not investigate how to model a group as it processes all the features similarly.

At each round of the LOGO cross-validation, a feature selection algorithm based on Kolmogorov-Smirnov statistic (Kolmogorov, 1933; Smirnov, 1948) was applied to ensure that only features which are potentially meaningful for the model were taken into account (Nilsson et al., 2007). We used the "*ks_2samp*" function from the Scipy library (Virtanen et al., 2020) to only select the features whose distribution, over both the Social and Task dimensions, are significantly different over the *decrease* and *no-decrease* classes with a significance level of $p < 0.01$.

In detail, a Gini impurity function to measure the quality of a split was used and the estimated hyper-parameters on the validation set were: the number of trees (in {100, 200, 300, 400, 500}), the maximum depth of the tree (in {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}), the minimum number of samples required to split an internal node (in {1, 2, 3, 4, 5}), and the minimum number of samples required to be at a leaf node (in {1, 2, 3, 4, 5, 6, 7}).

Finally, RFC was designed to predict cohesion dynamics for each thin slice. A majority voting was then applied over the six predictions of each of the thin slices composing a particular task to determine the overall prediction of the task, for each dimension. In case of a tie, we selected the *no-decrease* class.

5.5.2 A DNN Approach to Integrate Time

To model the time dependencies between the thin slices through a whole interaction (i.e., the five tasks of the escape game), we designed the *Full Interaction-LSTM* (FI-LSTM) model. This architecture integrates time dependencies between the thin slices with an LSTM layer. Since it processes thin slices of the whole interaction (i.e., the two last minutes of each of the five tasks), it also, to a certain extent, take into account the time dependencies that may exist between the various tasks of the interaction. This is, to the best of our knowledge, the first attempt to address the dynamics of cohesion.

This DNN architecture integrates the time by inputting the features from the FFS subset to an LSTM layer with 30 units This layer is followed by a Dropout layer with a dropout rate of 0.2 and by two fully connected (FC) layers with 16 and 8 units, respectively, and a ReLu activation function. FI-LSTM predicts the dynamics of Social and/or Task cohesion for each of the five tasks of an interaction thanks to a final FC layer with a Sigmoid activation function and one unit if the model predicts only 1 dimension or two units if the model predicts both dimensions, in a multilabel setting, and for each task. Figure 5.2 shows the FI-LSTM architecture. It has 27362 trainable weights in a multilabel

setting and 27357 when it predicts a single dimension only. Table 5.2 recapitulates the number of trainable weights for each model that we presented.

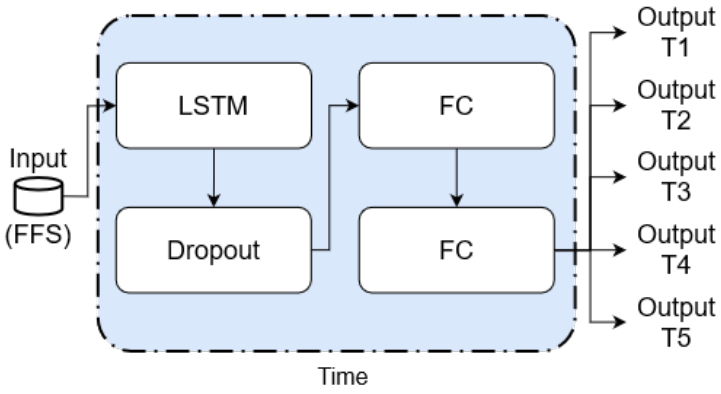


Figure 5.2: "*Full Interaction-LSTM*" (FI-LSTM) architecture. It uses the FFS features as input and integrates time dependencies between the thin slices of the last 2mn of each task using an LSTM layer.

5.5.3 A DNN Approach to Integrate Time and Group Modeling

To address RA1 and RA2, we designed the "*from Individual to Group*" (fItG) architecture (see Figure 5.3 for its architecture). It is rawly inspired by the Team LSTM model developed by Kasparova et al. (2020) that learns specific patterns of behavior to predict student engagement, using individual video-based features. FItG uses both IFS and GFS to learn a higher joint representation of the group behavior, merging individual and group representations to predict cohesion.

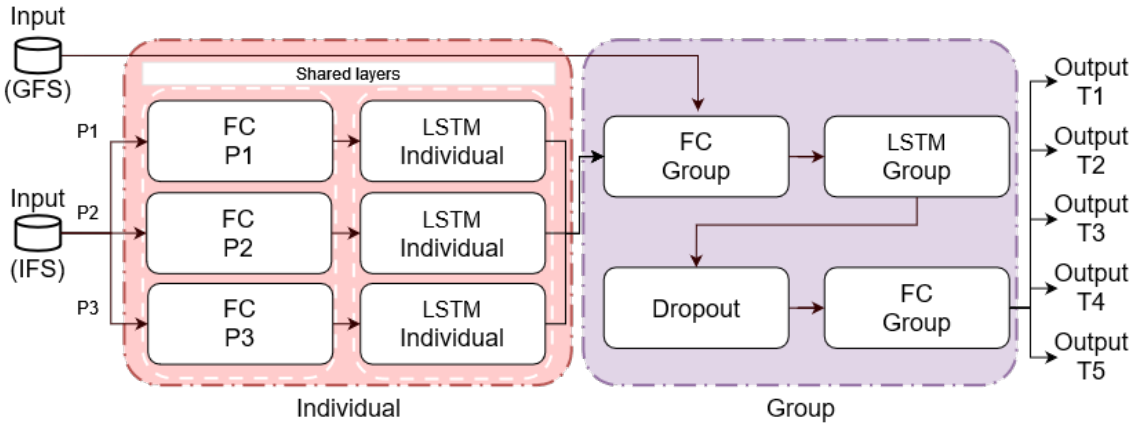


Figure 5.3: "*From Individual to Group*" (fItG) architecture. It is composed of two modules: the *Individual* module takes the IFS features as input and learns a representation of an individual; the *Group* module takes the three outputs of the Individual module, concatenated with the GFS features.

FItG is composed of two modules. The *Individual* module takes the IFS subset of features as input. It is made of three branches, each one composed of an FC layer with a

ReLu activation function and 50 units, followed by an LSTM layer. This structure enables the integration of the time dependencies between the thin slices of the interaction. This module aims at learning a higher-level representation of an individual. The model might learn undesired patterns related to the order in which each individual is processed by it across the multiple groups in the training set (e.g., learning a pattern specific to all the first group members seen by the model). To avoid this issue, we shared the weights of each layer of the three individual branches of the *Individual* module (i.e., the FC and LSTM layers). A common representation is learned, for each layer, as follows in Equation 5.6:

$$\boldsymbol{Y}_i = \phi_i \left(\sum_{j=1}^{n} (W \boldsymbol{X}_j) \right) \tag{5.6}$$

where Y_i is the output of layer i, ϕ_i, the activation function of the layer i, W, the matrix of parameters common to every group member, and X_j, the input related to the group member j. As groups are composed of three persons, n was here set equal to three.

The three outputs of the shared individual LSTM layers from the Individual module are then concatenated with the group features from the GFS subset as input of the *Group* module. This module is aimed at learning the temporal dynamics of cohesion from the group. The module is made of a first FC layer with a ReLu activation function and 64 units, followed by an LSTM layer to integrate the time dependencies. Next, a Dropout layer with a rate of 0.2 is used to prevent the model from overfitting. This is followed by another FC layer with a ReLu activation function and 16 units. Finally, as in the FI-LSTM architecture, the output consists of an FC layer with a Sigmoid activation function and one or two unit(s) depending on the number of dimensions predicted by the model (i.e. one unit for each cohesion's dimension predicted), for each task. fItG has 48152 trainable weights in a multilabel setting, and 48147 when predicting only one dimension (see Table 5.2).

As for FI-LSTM, this architecture integrates the time dependencies between the thin slices and the tasks using LSTM layers in both Individual and Group modules. The novelty introduced in fItG, resides in the strategy employed to address group modeling. It processes individual and group features differently, in related modules. This means that individual features are first processed and, combined with the group features, inform the model to learn a group behavior representation.

5.5.4 A DNN Approach to Integrate Time and the Interplay between Dimensions

We designed the "*Specific To Entwined*" transformer-based architecture (STE) to address temporal dependencies as well as the interplay between the Social and Task dimensions of cohesion (i.e., RA1 and RA3). Its architecture is displayed in Figure 5.4. Here, the main difference resides in the fact that we use all of the thin slices available in an interaction (instead of only using the two last minutes of each task) to explore how such an interplay evolves over time.

In STE, we hypothesize that, at the beginning of an interaction, both dimensions are distinct while, throughout the interaction, both dimensions converge to become interlaced.

Thus, the architecture is composed of two modules: *Specific* and *Entwined*.

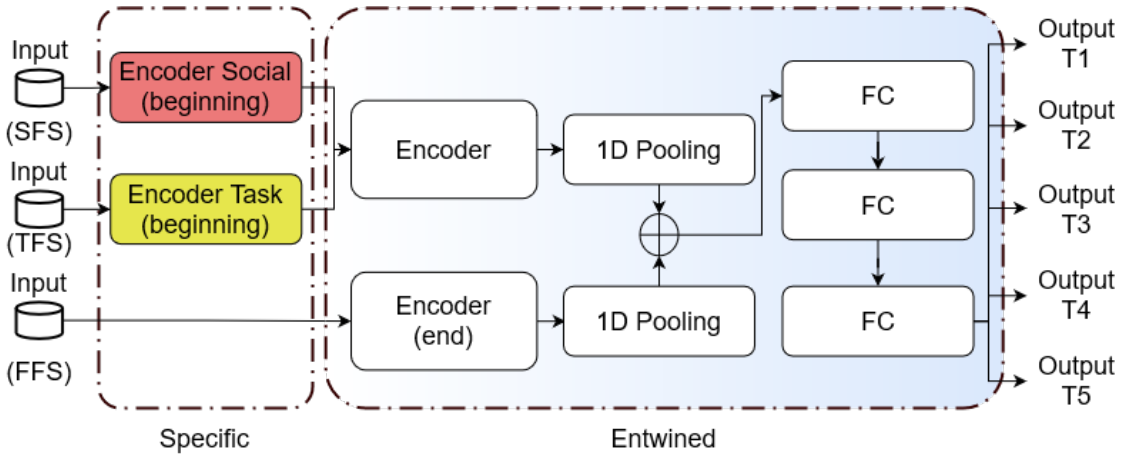


Figure 5.4: "*Specific To Entwined*" (STE) architecture. It is composed of two modules: *Specific* and *Entwined*. The first one takes SFS and TFS extracted in the first task and the first half of the second task to learn dimension-specific representations of behavior, while the second one learns the joint representation from the remaining of the interaction. Encoders refer to a transformer-encoder module (see Figure 5.5). STE outputs the dynamics of the Social and Task dimensions of cohesion in a multilabel setting.

The Specific module takes, as input, the SFS and TFS subsets computed over the thin slices of the first task and half of the second task. Facing a lack of insights with respect to the way these dimensions interplay over time, we empirically tested different values for selecting what we consider the beginning of the tasks, ranging from 5% to 90% of the thin slices of each task. Each subset is processed, separately in an encoder to learn a dimension-specific representation of group behavior. In STE, an encoder refers to a transformer-encoder block as introduced by Vaswani et al. (2017). Such a block is composed of a normalization layer, followed by a multi-head attention layer with five heads of 64 units and a dropout rate of 0.2. The output of the multi-head layer is added to the input features (referred as add). Then, it has another normalization layer followed by a 1D convolutional layer with a filter kernel size of five and a ReLu activation function as well as a dropout layer with a dropout rate of 0.2. Another 1D convolutional layer with a number of filters equal to the number of input features without activation function completes the transformer-encoder block. Finally, the block returns the sum of add and the output of the last convolutional layer. Figure 5.5 describes such a transformer-encoder block.

Then, the outputs of each encoder of the Specific module are concatenated and processed into another encoder on the Entwined module. In parallel, the FFS subset of features computed over the second half of the second task through the end of the interaction goes through a similar encoder in the Entwined module. Next, the outputs of each of the two last encoders are processed into a 1D average pooling layer. Then, the outputs of the pooling layers are concatenated and followed by three consecutive FC layers with a ReLu activation function and 128, 64, and 8 units, respectively.

Finally, as for the previously mentioned architectures in a multilabel setting (e.g., RFC, FI-LSTM, and fItG), the output consists of five distinct branches in which an FC layer with a Sigmoid activation function and two units (i.e., one for predicting each di-

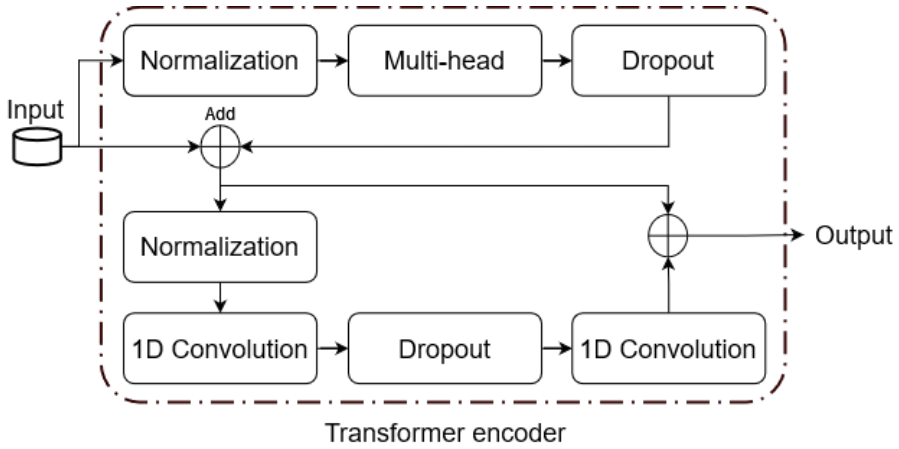


Figure 5.5: Architecture of the transformer-encoder (Vaswani et al., 2017).

mension), hence enabling the prediction of the Social and Task dynamics of cohesion. In total STE has 421478 trainable weights (see Table 5.2).

STE integrates time dependencies through transformer-encoders. Furthermore, features are processed differently depending on the time in the interaction at which they have been extracted. The interplay between Social and Task cohesion is also taken into account as, for each dimension, dimension-specific representations of behavior are learned from the beginning of the interaction and help the model learn a joint representation of cohesion, implying that both dimensions mingle as the interaction progresses.

5.5.5 DNN Approaches to Integrate Time, Group Modeling and the Interplay between Dimensions

The "*Common to Specific*" (CTS) architecture

In CTS, we explore another assumption with respect to the development of cohesion over time. We started from the hypothesis that each dimension has its own specificity, especially at the beginning of the interaction, but also shares similarities with the other one (i.e., they tend to converge over time). Thus, CTS processes each dimension independently and differently (i.e., an extra fully connected layer is added before the output for the Task dimension) while integrating shared information that is common to both dimensions. Such a design implies that both dimensions are, overall, distinct all along the interaction, as opposed to the assumption taken in the STE architecture. Similarly to STE, CTS uses the different subsets of features extracted from previous models' insights (i.e., the Social-specific, Task-specific, and Common features subsets).

CTS is aimed at learning a group behavior representation from the Common features set to combine it with the ones learned for each dimension, separately. By doing so, we expect to enrich each representation with complementary information to fully capture each dimension. Thus, CTS is composed of three modules (i.e., *Common*, *Task* and *Social*), as depicted in Figure 5.6: the Common module takes the Common features as inputs. These are processed into an LSTM layer followed by two FC layers with a ReLu activation function and 32 and 16 units, respectively. The output of this module (i.e., O_c) is then used in the Social and Task modules. The Social and Task modules use as in-

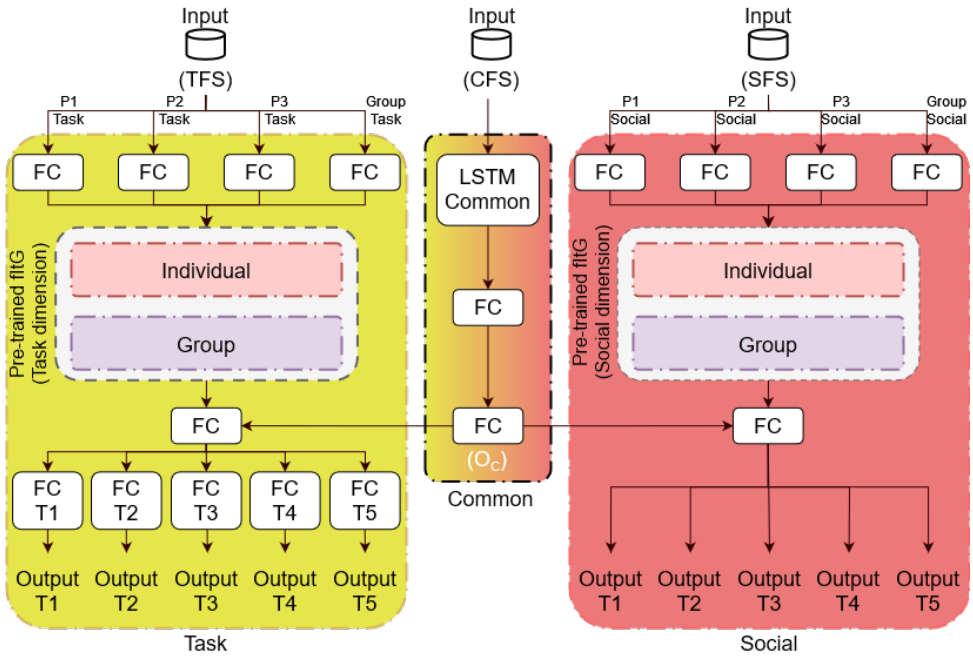


Figure 5.6: "*Common to Specific*" (CTS) architecture. It is composed of three modules (i.e., *Task*, *Social* and *Common*). Social and Task modules independently process features from Social- and Task-specific subsets, respectively to learn a dimension-specific representation of behavior, enriched by the Common module. CTS predicts, for each task, both dimensions in a multitask setting.

puts, the Social- and Task-specific features subsets (i.e., SFS and TFS), respectively, and process individual and group features into four parallel FC layers with a Relu activation function and 50 units for the individual features and 41 units for the group features. Then, a pre-trained version of the fItG model is used in each module. In the Social module, it has been specifically trained to predict Social cohesion only, while in the Task module, it was pre-trained on Task cohesion only. The output of the pre-trained fItG is then concatenated with O_c into an FC layer with a ReLu activation function and eight units in each module. While in the Social module, the prediction of Social cohesion dynamics, for each of the five tasks, is done through five parallel Dense layers with a Sigmoid activation function and one unit straight after the previous FC layer, the Task module contains an extra layer in each of the five branches used for the prediction that consists of an FC layer with a ReLu activation function and four units. This extra layer was implemented as we noticed that the Task dimension was harder to predict than Social cohesion. In total, CTS has 213705 trainable weights (see Table 5.2).

CTS exploits various subsets as inputs for the different modules and leverages a transfer learning approach by using versions of the fItG pre-trained on both dimensions, independently. Thus, it is, to the best of our knowledge, the first attempt to integrate the interplay between dimensions while addressing both time dependencies and group modeling. The methodology to define the features present in the three subsets of features is, however, highly dependent on the model used to compute the Shapley values. Thus, it

remains to be seen to what extent CTS generalizes to subsets built from other models.

Inspired by the Social Sciences theories on group development presented in Section 2.2.5, we designed three DNN architectures that each integrate one of the three identified ways to study the interplay between the Social and Task dimensions of cohesion. These architectures are based on the fItG architecture. Thus, they are, *de facto*, integrating time and group modeling.

The "*Transfer Between Dimensions*" (TBD) architecture

Due to the contradictory views on which of the two dimensions of cohesion emerges first and affects the other one, we designed two different architectures: *TBD-Social* (TBD-S) and *TBD-Task* (TBD-T). Both TBDs use a transfer learning approach to take advantage of the behavior representation learned beforehand by a pre-trained model (here the fItG) for a specific dimension to predict the other one. More specifically, TBD-S predicts the dynamics of Social cohesion using a pre-trained fItG for Task cohesion, whereas TBD-T predicts the dynamics of Task cohesion using a pre-trained fItG for Social cohesion. Figure 5.7 sketches the general architecture of the TBDs.

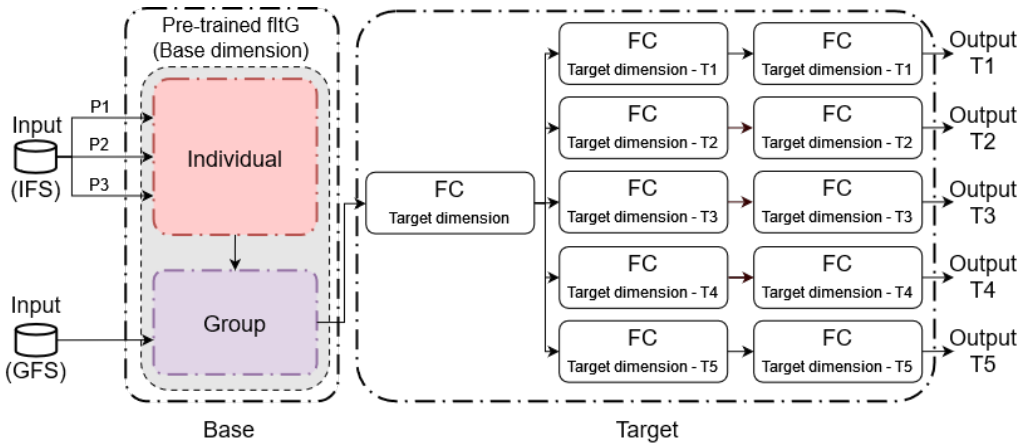


Figure 5.7: "*Transfer Between Dimensions*" (TBD) architecture from which TBD-S and TBD-T are implemented. It is composed of the *Base* and the *Target* modules. The Base module uses a pre-trained version of the fItG to learn a representation of behavior for a specific dimension (e.g., Social for TBD-T) to inform the Target module that learns a representation of the group behavior for the target dimension (e.g., Task for TBD-T).

It is composed of two modules: *Base* and *Target* detailed in the following. Both TBD-S and TBD-T have a total of 49139 trainable weights (see Table 5.2).

Leveraging a transfer learning approach, the Base module learns a representation of the group behavior for a dimension (i.e., Social for TBD-T and Task for TBD-S) from which a group behavior representation for the targeted dimension (i.e., the predicted dimension) will be learned. The Base module takes as input both the Individual and Group features subsets (i.e., IFS and GFS), and it outputs the representation of the group behavior learned for the specific dimension from the last layer of the Group module of the fItG model.

The Target module learns the group behavior representation of the targeted dimension (i.e., Social or Task cohesion). It consists of an FC layer with a ReLu activation function and 16 units that takes the output of the Base module as input. This FC layer is followed by five branches (one for each task). Each branch is composed of two consecutive FC layers with a ReLu activation function and eight and four units, respectively.

Finally, TBD is designed to predict only one dimension (i.e., the target dimension). Thus, the output consists of the prediction of the cohesion dynamics for the Social dimension (in the case of TBD-S) or the Task dimension (in the case of TBD-T), across the five tasks. It is composed of five branches (one for each task). Each branch consists of an FC layer with a Sigmoid activation function and one unit, predicting the dynamics of one dimension for a specific task.

Both TBDs integrate the Social and Task interplay unidirectionally (i.e., from Social to Task cohesion with TBD-T and from Task to Social cohesion with TBD-S). They, however, do not integrate the reciprocal impact of the two dimensions on each other.

The "*Transfer Between Dimension-Reciprocal impact*" (TBD-RI) architecture

To try to integrate this reciprocity, we designed the TBD-RI architecture. It is built on top of both the TBD-S and TBD-T architectures, hence, TBD-RI also takes advantage of a transfer learning approach to learn a group behavior representation for each dimension before concatenating them and jointly learning the Social and Task cohesion dynamics. Figure 5.8 shows the TBD-RI architecture. It is composed of two different modules: *Dimension specific* and *Reciprocal impact* and has 99002 trainable weights (see Table 5.2).

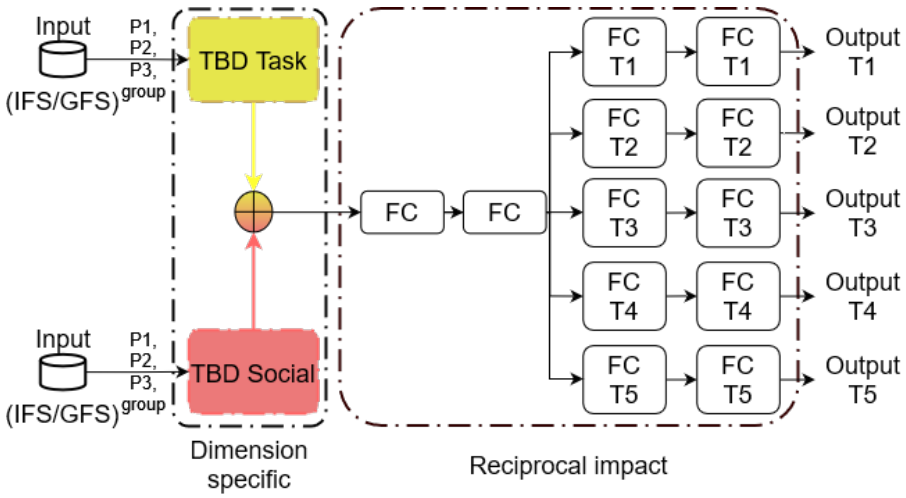


Figure 5.8: The "*Transfer Between Dimension-Reciprocal impact*" (TBD-RI) architecture. It is built on top of TBD-S and TBD-T and learns, in the *Dimension specific* module, a specific representation of the group behavior for each dimension. Both representations are concatenated and processed by the *Reciprocal impact* module. As in fItG, the *Output* module predicts the Social and Task cohesion dynamics in a multilabel setting.

The Dimension Specific module learns a representation of the group behavior for both the Social and Task dimensions of cohesion, independently. This module first splits into two branches (i.e., one for each dimension). Each branch takes both the Individual and

Group features subsets as input and uses the Base module as well as the first FC layer from the Target module to learn dimension-specific group behavior representations. Then, the outputs from each branch are concatenated, resulting in a tensor of shape $[B \times T, 2 \times F]$ with *B*, the batch size (i.e., the number of groups processed per batch), *T*, the number of timesteps and *F*, the size of the features representation of each dimension. This tensor is then processed by the Reciprocal impact module.

The Reciprocal impact module learns the reciprocal impact that the Social and Task dimensions of cohesion has on each other over time using as input the concatenation of the representations learned by the Dimension Specific module. It consists of a first FC layer with a ReLu activation function and 32 units, followed by another FC layer with a ReLu activation function and 16 units. Similar to the TBD architectures, there is a split into five branches (one for each task) with two FC layers with a ReLu activation function and with eight and four units, respectively, in each branch.

As for RFC, FI-LSTM, and fItG architectures in a multilabel setting, the output of TBD-RI consists of an FC layer, for each branch, with a Sigmoid activation function and two units. This enables the TBD-RI to predict the dynamics of the Social and Task dimensions of cohesion in a multilabel setting.

TBD-S, TBD-T, and TBD-RI all leverage a transfer learning approach based on fItG, hence, they all address RA1 and RA2 and are specifically designed to tackle RA3.

Table 5.2: Number of trainable weights per model. FI-LSTM and fItG could be designed for predicting only one dimension, hence, having five trainable weights less than the multilabel version presented here.

Number of trainable weights					
FI-LSTM	**fItG**	**STE**	**CTS**	**TBD**	**TBD-RI**
27362	48152	421478	213705	49139	99002

5.6 Analysis of the Computational Models' Performances

5.6.1 Window Size for Feature Extraction

All the features described in Chapter 4 can be computed over various lengths of time windows. It remains to be seen, however, what is the best granularity to automatically study cohesion dynamics. Thus, we explore the impact of various window sizes (i.e., 5s, 10s, 15s, and 20s) on the performances of RFC in a multilabel setting. Let's consider RFC_5, RFC_10, RFC_15, and RFC_20, the Random Forest Classifiers in a multilabel setting, using features computed on window sizes of 5s, 10s, 15s, and 20s, respectively. Table 5.3 shows the results obtained by each version of the RFC, for both dimension.

Table 5.3: Summary of the average F1-scores obtained for each dimension by the RFC in a multilabel settings with features computed on 5s, 10s, 15s, and 20s.

F1-score ±std							
RFC_5		**RFC_10**		**RFC_15**		**RFC_15**	
Social	**Task**	**Social**	**Task**	**Social**	**Task**	**Social**	**Task**
0.67 ±0.17	0.49 ±0.27	0.60 ±0.11	0.53 ±0.20	0.60 ±0.07	0.50 ±0.24	0.62 ±0.19	0.53 ±0.12

As for the Social dimension, RFC_5 achieves an average F1-Score, for the five tasks and over the 15 seeds of 0.67 ±0.17 while RFC_10, RFC_15, and RFC_20 reach 0.60 ±0.11, 0.60 ±0.07 and 0.62 ±0.19, respectively. A permutation test shows a significant difference between the models ($p = .001$). A post-hoc analysis indicates that no significant difference exists between RFC_10 and RFC_15. RFC_20 obtains, however, significantly better performances than these two ($p = .003$ for both pairs) but is also significantly outperformed by RFC_5 ($p = .003$). Thus, RFC_5 and RFC_20 are the most performing models for the Social dimension.

Concerning the Task dimension, the average F1-scores obtained are 0.49 ±0.27 for RFC_5, 0.53 ±0.20 for RFC_10, 0.50 ±0.24 for RFC_15 and 0.53 ±0.12 for RFC_20. A permutation test shows that a significant difference between these performances exists ($p = .001$). A post-hoc analysis reveals that there is no significant difference between RFC_5 and RFC_10 nor between RFC_10 and RFC_15. Performances are, however, significantly better for RFC_10 and RFC_20 than for RFC_5 and RFC_15 ($p = .006$ for each pair). These results indicate that RFC_10 and RFC_20 are the most performing models for the Task dimension.

Since there is no model outperforming all the other ones in both dimensions, we average the F1-scores obtained for the Social and Task dimensions to reflect the overall performance of the models. This results in the following performances: RFC_5 and RFC_10 achieve an averaged F1-score of 0.58 ±0.14 and 0.56 ±0.14, while RFC_15 and RFC_20 perform an averaged F1-score of 0.55 ±0.12 and 0.57 ±0.14, respectively. As previously, a permutation test shows a significant difference between the models ($p = .001$). No significant differences are found between RFC_10 and RFC_15 nor between RFC_5 and RFC_20. Performances of RFC_5 are significantly better than RFC_10 and RFC_15 ($p = .039$ and $p = .006$, respectively) and similar conclusions are drown for RFC_20 (i.e., $p = .024$ between RFC_20 and RFC_10 and $p = .006$ between RFC_20 and RFC_15). Figure 5.9 highlights the results of the RFC models for each dimension and shows the significant differences between the models.

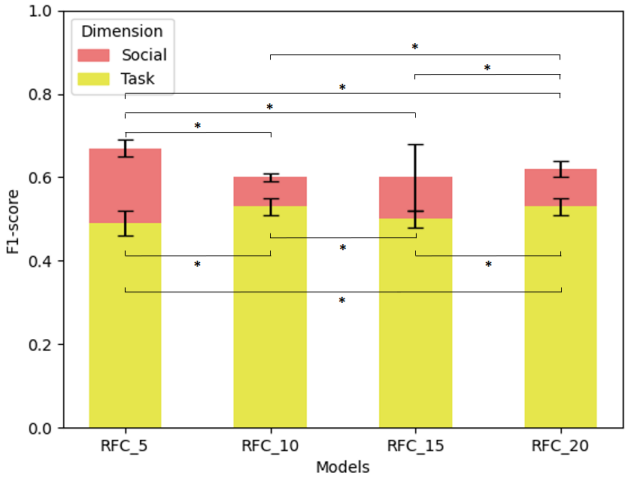


Figure 5.9: F1-score of the RFC models using various window sizes (i.e., 5s for RF_5, 10s for RFC_10, 15s for RFC_15, and 20s for RFC_20) for the Social and Task dimensions of cohesion (in pink and yellow, respectively). P-values of significant differences are displayed between the models, for the Social and Task dimension, respectively.

To summarize, results show that, for the Social dimension, the RFC reaches its best performances with a 5s window size. Performances with 20s window sizes are, however, significantly better than those of the models using 10s and 15s window sizes. For the Task dimension, RFC achieves significantly better results with window sizes of 10s and 20s. When averaging the performances of the model for both the Social and Task dimensions of cohesion, there is, however, no significant difference between 5s and 20s window sizes that both lead to higher performances than models using 10s and 15s window sizes. Given the fact that no model is significantly better than all the other ones and that the Task dimension of cohesion is usually harder to predict (e.g., Nanninga et al., 2017), we hypothesize that a 20s window would be more appropriate. Also, such a duration corroborates previous work on group interaction (Gatica-Perez et al., 2005) and cohesion perception (Ceccaldi et al., 2019).

5.6.2 Selecting the Reference Model

RFC, FI-LSTM, and fItG architectures are relatively simple compared to STE, CTS and TBD-S, TBD-T, and TBD-RI that either or both leverage transfer learning and more advanced approaches to integrate time dependencies and the interplay between dimensions. Thus, we compare the first three models, in a multilabel setting, to determine the reference model against which the other ones will be evaluated. Table 5.4 shows the details of F1-scores for the RFC, FI-LSTM and fItG models, in a multilabel setting.

Table 5.4: Average F1-scores on the 15 seeds for each task and for each dimension for the RFC, FI-LSTM, and fItG models.

	F1-score ±std					
	RFC		**FI-LSTM**		**fItG**	
	Social	**Task**	**Social**	**Task**	**Social**	**Task**
T1	0.47 ±0.06	0.42 ±0.06	0.50 ±0.11	0.56 ±0.08	0.52 ±0.10	0.65 ±0.07
T2	0.23 ±0.04	0.35 ±0.03	0.41 ±0.11	0.46 ±0.12	0.51 ±0.13	0.56 ±0.12
T3	0.70 ±0.02	0.54 ±0.03	0.69 ±0.08	0.54 ±0.11	0.65 ±0.07	0.57 ±0.13
T4	0.86 ±0.00	0.61 ±0.02	0.84 ±0.07	0.50 ±0.13	0.87 ±0.04	0.66 ±0.14
T5	0.83 ±0.05	0.73 ±0.00	0.78 ±0.05	0.76 ±0.09	0.80 ±0.04	0.74 ±0.07
Average	**0.62 ±0.02**	**0.53 ±0.02**	**0.64 ±0.04**	**0.56 ±0.06**	**0.67 ±0.03**	**0.64 ±0.02**

The RFC model achieves, for the 15 seeds, an average F1-score of 0.62 ±0.02 for the Social dimension and 0.53 ±0.02 for the Task dimension. Statistical analysis shows that there are significant differences in performances for the RFC with respect to the FI-LSTM and the fItG models for both Social ($p = .002$) and Task ($p = .001$) dimensions. A post-hoc analysis using pairwise permutation t-tests is carried out and shows that both the FI-LSTM and the fItG models outperform the RFC for the Social ($p = .006$ for both) and Task ($p = .048$ and $p = .003$, respectively) dimensions. Indeed, for the Social dimension, the FI-LSTM model reaches an average F1-score across the 15 seeds of 0.64 ±0.04 while the fItG obtained an average F1-score of 0.67 ±0.03. Such difference in the performances is, however, not significant. Regarding the Task dimension, the FI-LSTM achieves an average F1-score of 0.56 ±0.06 while the fItG significantly outperforms it ($p = .003$), reaching an average F1-score of 0.64 ±0.02.

To summarize, the fItG model is the most performing one with a F1-score of 0.67 ±0.03 and 0.64 ±0.02 for the Social and Task dimensions, respectively. Thus, for the remaining of the analysis, the other models will be compared to the performances of the fItG in a multilabel setting. These results highlight the benefits for a model to integrate the temporal nature of cohesion and to learn higher representations of both individuals and group to predict the dynamics of cohesion, especially for the Task dimension. For this dimension, fItG, indeed, significantly outperforms the other two models, confirming the importance of modeling groups from both individuals and group perspectives.

5.6.3 Evaluating the Impact of Addressing RA3

5.6.3.1 Uni vs Multilabel

A first approach to address the interplay between the Social and Task dimensions (i.e., RA3) consists of using the same architecture to predict each dimension independently (i.e., unilabel setting) and to compare the performances with the ones from the same architecture, in a multilabel setting, as suggested in the structured survey in Chapter 2. In a multilabel setting, both dimensions are, indeed, both predicted from the same node or layer and equally contribute to cohesion (e.g., by summing the losses of both dimensions for the DNN architectures), hence, addressing RA3.

RFC, in a multilabel setting, significantly improves the predictions of the Social dimension with respect to RFC in a unilabel setting (from 0.61 ±0.01 to 0.62 ±0.02, $p = .044$), while it significantly decreases the ones of the Task dimension (from 0.55 ±0.02 to 0.53 ±0.02, $p = .002$). Concerning FI-LSTM and fItG, no significant difference is found for the Social dimension. Both unilabel and multilabel settings, indeed, reach similar performances. Multilabel classification, however, significantly improves the predictions of the fItG, for the Task dimension: it, indeed, achieves 0.64 ±0.02 in a multilabel setting vs 0.61 ±0.05 in a unilabel setting ($p = .042$).

These results show that a simple approach to integrating the interplay of the Social and Task dimensions (i.e., using multilabel classification) partially improves the performances of the models predicting a single dimension. In particular, improvements mainly concern Task cohesion. Such a kind of approach, however, neglects the insights from the extensive research in Social Sciences that we expect to be beneficial for the model.

5.6.3.2 Exploiting Models' Insights

Both CTS and STE models exploit different subsets of features as inputs (i.e., TFS, SFS, and CFS). Since the fItG in a multilabel setting is the reference model, we computed the Shapley values on this model to build SFS, TFS, and CFS, as described in Section 5.4. Here below the features that were retained for the subsets:

- SFS: the lateral expansion (skewness), touch's duration (maximum, average, and standard deviation), occupied volume (skewness), kinetic energy (minimum, maximum, and standard deviation), synchrony of kinetic energies, Public and Social spaces as well as the harmonic difference (H1-H2 and H1-A3).

- TFS: the total speaking time, overlap, loudness, shimmer, F1, F2, and F3 relative energies, F1 bandwidth and frequency, F2 and F3 frequencies, Alpha Ratio and Hammarberg Index (both in the high-frequency region), Spectral Slopes (0-500 Hz and 500-1500 Hz) for both low and high-frequency regions, occupied volume (maximum and standard deviation), group amount of hands movement while not moving from translations and rotations (skewness and minimum), time in F-formation (maximum and standard deviation), total distance traveled, longitudinal (standard deviation) and latitudinal (average, maximum and minimum) posture expansion, kinetic energy (skewness and average) and group amount of motion for both the mean (average, maximum) and group ratio (average).

- CFS: all of the remaining features that are not included in SFS nor TFS, which consist of laughter duration, average turn duration, pitch, jitter, HNR, alpha ratio (low-frequency region), Hammarberg Index (high-frequency region), distance from group barycenter (average, standard deviation, maximum, minimum, and skewness), personal space, maximum of interpersonal distances (average, standard deviation, maximum, minimum, and skewness), time in F-formation (average, minimum, and skewness), longitudinal expansion (minimum, maximum, skewness, and average), lateral expansion (standard deviation), occupied volume (minimum, and average), group amount of motion (minimum, standard deviation, and skewness), group amount of hands movement (average, standard deviation, and maximum), touches' duration (minimum, and skewness).

In terms of performance, CTS achieves, over the 15 seeds, an average overall F1-score over the Social and Task dimensions of 0.63 $\pm$0.03 while STE reaches 0.64 $\pm$0.02. A permutation test shows that there are no significant differences in performances between these two models and the fItG. Similarly, no significant difference exists between the performances over the Social dimension for CTS (i.e., 0.67 $\pm$0.05), STE (i.e., 0.68 $\pm$0.02), and fItG (i.e., 0.67 $\pm$0.03). With respect to the Task dimension, a permutation test reveals that a significant difference exists between these models ($p = .007$). A post-hoc analysis shows that fItG outperforms both CTS ($p = .015$) and STE ($p = .006$) but there is no significant difference between CTS and STE. FItG, indeed, obtains an average F1-score of 0.64 $\pm$0.02 while CTS and STE achieve 0.59 $\pm$0.04 and 0.60 $\pm$0.03, respectively.

Results show that these approaches do not improve performances with respect to the fItG. This could be due to multiple reasons. The methodology used to select SFS, TFS, and CTS features is based on Shapley values of a different model (i.e., fItG). They are, however, very specific to this model and might not generalize to other models. Only a few features, indeed, overlap when defining the subsets of features between RFC, FI-LSTM, and fItG. Thus, it is possible that the subsets of features are not optimal for learning a dimension-specific representation of behavior. Moreover, as cohesion manifests differently depending on the groups' strategies, these features sets might be different for each group and might change depending on the task. Also, both CTS and STE significantly increase the number of weights compared to the fItG (i.e., 213705 and 421478, respectively vs 48152). Thus, overfitting in some tasks is more likely to happen despite the various strategies employed to avoid it.

5.6.3.3 Leveraging Social Sciences' Insights

Performances of TBDs and TBD-RI are first compared against those ones of the fItG. Then, we compare the performances obtained between each of the tasks. Table 5.5 summarizes all the performances obtained by the fItG in a multilabel setting, TBD-S, TBD-T, and TBD-RI, for each task and for each dimension. In addition, Figure 5.10 shows the box-plots of the tasks' performances over the 15 seeds, for these models.

Table 5.5: Average F1-scores on the 15 seeds for each task, and each dimension, obtained by fItG, TBD-S, TBD-T, and TBD-RI. For the Social dimension, TBD-RI is the most performing model while for the Task dimension, TBD-T outperforms other models.

	F1-score ±std					
	fItG		**TBD-S/T**		**TBD-RI**	
	Social	**Task**	**Social**	**Task**	**Social**	**Task**
T1	0.52 ±0.10	0.65 ±0.07	0.50 ±0.11	0.63 ±0.07	0.56 ±0.10	0.64 ±0.09
T2	0.51 ±0.13	0.56 ±0.12	0.49 ±0.11	0.59 ±0.09	0.61 ±0.08	0.61 ±0.09
T3	0.65 ±0.07	0.57 ±0.13	0.66 ±0.06	0.69 ±0.10	0.69 ±0.06	0.62 ±0.11
T4	0.87 ±0.04	0.66 ±0.14	0.83 ±0.09	0.65 ±0.09	0.85 ±0.05	0.57 ±0.10
T5	0.80 ±0.04	0.74 ±0.07	0.80 ±0.05	0.76 ±0.09	0.79 ±0.05	0.78 ±0.05
Average	**0.67 ±0.03**	**0.64 ±0.02**	**0.66 ±0.04**	**0.66 ±0.02**	**0.70 ±0.03**	**0.64 ±0.03**

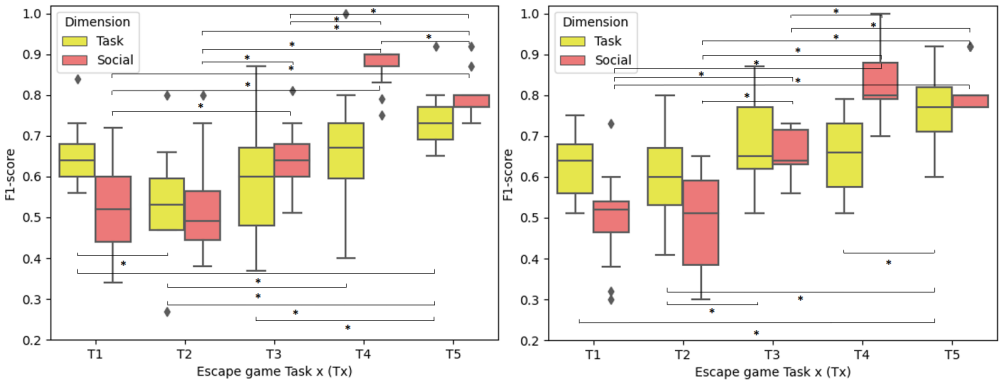
Regarding the Social dimension of cohesion, fItG reaches, an average F1-score over the 15 seeds of 0.67 ±0.03, while TBD-S obtains 0.66 ±0.04 and TBD-RI achieves 0.70 ±0.03. A permutation test shows that there are significant differences in performances between these three models ($p = .018$). A post-hoc analysis reveals that TBD-RI significantly outperformed both TBD-S ($p = .012$) and fItG ($p = .036$). Such an improvement in performance is partially explained by the significant improvement in T2 ($p = .044$). This is, indeed, the only task in which TBD-RI significantly outperforms fItG (from 0.51 ±0.13 to 0.61 ±0.08). This task remains, however, the hardest task to predict for all the models. A permutation test run across the five tasks shows a significant difference between the tasks' performances for each model ($p = .001$ for every model). No significant difference is found between T1 and T2 across the models. These two tasks are, indeed, the ones for which all the models obtained the lowest prediction performances. Models achieve significantly better performances on T3 than on T1 ($p = .003$, for every model) and on T3 than on T2 ($p = .004$, $p = .003$ and $p = .013$ for the fItG, TBD-S, and TBD-RI, respectively). T4 and T5 are the tasks in which all the models reach the best performances ($p = .003$ between T3-T4 and $p = .003$ between T3-T5, for every model).

To summarize, TBD-RI is the most performing model for the Social dimension. It significantly improves fItG and TBD-S performances, especially on T2. Also, there is a similar pattern of the performances obtained for each task across all the models: T1 and T2 are those ones for which all the models obtained the lowest performances, T3 is better predicted than the two first tasks, while the performances achieved in T4 and T5 are the highest ones.

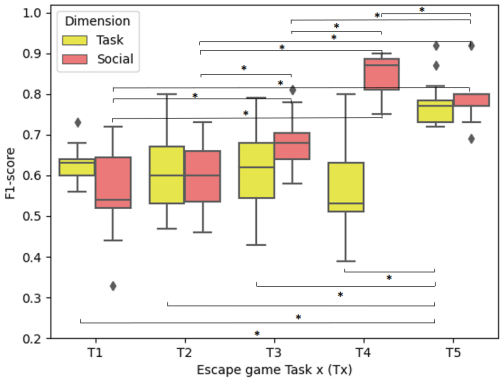
Concerning the Task dimension, a permutation test shows a significant difference of performances ($p = .014$) between fItG (0.64 ±0.02 F1-score), TBD-T (0.66 ±0.02 F1-score) and TBD-RI (0.64 ±0.03 F1-score). A post-hoc analysis reveals that the difference

in performances obtained by TBD-T is significant only with respect to the ones of fItG ($p = .018$) but not with respect to the ones of TBD-RI. Similarly to the Social dimension, only one of the worst predicted tasks is significantly improved as opposed to the fItG. In fact, TBD-T significantly outperforms fItG in T3 ($p = .034$). TBD-T, indeed, reaches an average F1-score over the 15 seeds of 0.69 $\pm$0.10 in T3 compared to 0.57 $\pm$0.13 obtained with the fItG. Such improvement indicates a change in the ability of the models to predict a subset of tasks. Statistical analysis carried out through a permutation test shows a significant difference between the performances of the five tasks of every model ($p = .001$, for the fItG, TBD-T, and TBD-RI, respectively). A post-hoc analysis shows that, for Task cohesion, T2 is always among the worst predicted tasks: T2 is significantly worst predicted than T4 ($p = .024$) and T5 ($p = .007$) for the fItG, significantly worst predicted than T3 ($p = .025$) and T5 ($p = .010$) for TBD-T, and significantly worst predicted than T5 for TBD-RI ($p = .005$). T5 remains significantly better predicted across all the tasks and models (except for T3 in TBD-T which reaches similar performances). TBD-RI obtains fewer variations across the tasks. There is only a significant difference between T5 and the other tasks ($p = .005$ for each pair of tasks T1-T5, T2-T5, T3-T5, and T4-T5), while no significant differences are found between the other pairs of tasks, meaning that performances in T1, T2, T3, and T4 are equivalent.

To summarize, only TBD-T outperforms fItG for the Task dimension, especially due to the significant improvement in T3. Also, T2 remains among the worst predicted tasks across all the models, while T5 is always the task in which the models achieve significantly better F1-scores.



(a) fItG (b) TBDs



(c) TBD-RI

Figure 5.10: Box-plots of the tasks' performances over the 15 seeds for fItG, TBD-T and TBD-S, and TBD-RI. Significant differences between the tasks are marked with a "*".

5.7 Conclusion

IN summary, we first presented the settings that are shared by every computational model of cohesion and we explained in depth the evaluation procedure and the statistical analysis we performed to compare models' performances. All the models were developed following at least one of the three first research axes presented in Chapter 2. This includes the integration of time between thin slices (and, by extension, between tasks) with the FI-LSTM, fItG, STE, CTS, TBD-S, TBD-T, and TBD-RI models; the integration of group modeling through the fItG, CTS, TBD-S, TBD-T, and TBD-RI models; and the integration of the interplay between the Social and Task dimensions of cohesion based on model' insights with the STE and CTS models, and based on Social Sciences' insights with the TBD-S, TBD-T, and TBD-RI models.

Among the simpler models, fItG is the most performing, showing the relevance of integrating individual and group contributions. Based on this model, we showed that TBD-RI was the most performing model for the Social dimension. As for the Task dimension, TBD-T outperforms other models.

To conclude, these results highlight the benefits and the potential of designing computational models of cohesion driven by Social Sciences' theories and insights.

Chapter 6

Integrating Other Group Processes

Contents

THIS Chapter presents the approaches we implemented to integrate other group processes (i.e., group emotion and emergent leadership) into computational models of cohesion. These approaches address the fourth research axis presented in Chapter 2 (i.e., "Relationships with other group processes"), hence, helping answering RQ2. First, we present two DNN architectures to explore the links between cohesion and group emotion, inspired by the Bottom-up and Top-down approaches (Barsade and Gibson, 1998). Then, we introduce two families of approaches for studying the relationships between cohesion and emergent leadership. One family acts at the Input level of the computational model of cohesion, by amplifying differences between the leader(s)' features and the ones from their follower(s). The other family modifies the architecture of the models (i.e., at Model level) by injecting leadership's knowledge. All the performances are evaluated against the fItG (see Chapter 5), and the results are discussed.

I designed and implemented the work on the integration of group emotions into computational models of cohesion (see Maman et al., 2021a, for the resulting publication) and I co-designed both families of approaches for integrating emergent leadership into DNNs (see Sabry et al., 2021, for the resulting publication).

6.1 Introduction

As stated by Severt and Estrada (2015), various relationships between cohesion and other group processes may be observed depending on the function (e.g., instrumental), the dimension (e.g., Social), or the level of analysis of cohesion (e.g., horizontal) that is being investigated. Therefore, it is expected that each function, dimension, or level of cohesion will be associated more or less strongly with different group processes. While group performance has been one of the most commonly studied group outcome of cohesion in the Social Sciences literature (e.g., Carron et al., 2002; Beal et al., 2003; Evans and Dion, 2012), Severt and Estrada (2015) also suggest examining relationships between cohesion and more specific group processes (e.g., group trust) in different contexts. Thus, we specifically focus on group emotion and emergent leadership as they are both associated with the emergence and development of cohesion (e.g., Fox et al., 2000; Xie et al., 2019, respectively). In this Chapter, multiple approaches integrating how to integrate the links between cohesion and these group processes are investigated to answer RQ2.

6.2 Cohesion and Emotion

6.2.1 Emotion in Groups

Emotion can either bind or splinter a group (Magee and Tiedens, 2006) and are emergent processes (Scherer, 2009; Coan and Gonzalez, 2015). Thus, they appear crucial for studying group dynamics. Barsade and Gibson (1998) highlighted two approaches to characterize group emotions.

Top-down focuses on the group as a whole. This means that group dynamics influence the feelings and behaviors of members of the group. Following this approach, scholars in Social Sciences characterized group emotions as (1) forces which shape individual emotional response (e.g., Le Bon, 1897), (2) social norms (e.g., Gibson, 1997), (3) the interpersonal glue that keeps groups together (e.g., Festinger et al., 1950) and (4) a display of group's maturity and development (e.g., Bales and Strodtbeck, 1951). In this Section, we follow the first characterization of group emotion.

Bottom-up investigates how the emotions of group members combine to create a group emotion, approximating the group as the sum of its parts. This approach led researchers to examine the group through a variety of compositional perspectives such as the mean of the group's members, the degree of emotional variance within the group, and the influence of the most emotionally extreme members of the group.

There is, however, an open debate on defining the best approach. As both the Top-down and the Bottom-up approaches bring different characterizations of group emotion, Barsade and Gibson (1998) recommend exploring methods following both these

approaches to have a complete picture of this group process. Furthermore, literature on cohesion and group emotion highlighted the importance to consider these processes from both individuals and the group as a whole (Braun et al., 2021).

6.2.2 Links between Cohesion and Group Emotion

The links between cohesion, an affective emergent state, and individual and group emotion had been particularly studied in Social Sciences, showing that cohesion and emotions influence each other (e.g., Barsade and Gibson, 1998; Lawler et al., 2000; Vanhove and Herian, 2015). For example, highly cohesive teams likely promote positive emotions such as happiness among group members. Reciprocally, positive individuals likely create a climate conducive to cohesion, hence, solidifying and strengthening the group bonds. This cohesion-emotion relationship has been studied through different angles (e.g., subjective well-being) among various types of groups such as sport groups (García-Calvo et al., 2014), student project groups (Picazo et al., 2015) or Antarctic station crews (Sarris and Kirby, 2005). Taken together, these results support the interconnectedness of these processes.

From a computational perspective, as already mentioned in Chapter 2, a few existing attempts to take advantage of the cohesion-emotion relationships exist, with the aim of improving the performances of their computational models on group emotion (i.e., Guo et al., 2019; Xuan Dang et al., 2019; Sharma et al., 2019; Gavrikov and Savchenko, 2020; Ghosh et al., 2022; Zou et al., 2020; Tien et al., 2021). These studies, however, did not specifically investigate how emotion could be related to a specific dimension of cohesion. Also, despite suggesting new models and methods to jointly predict both cohesion and emotion, they did not explore how different approaches of group emotion (e.g., Top-down or Bottom-up approaches) could impact their models' performances.

In contrast to these studies, we investigate how group emotion affects the Social and Task dimensions of cohesion. We also explore multiple approaches to characterize group emotion (i.e., Top-down, Bottom-up) to improve the joint prediction of cohesion and group emotion.

6.2.3 Experimental Settings

We extended the fItG model for jointly studying both cohesion and group emotion in small groups interactions. As opposed to RFC and FI-LSTM, fItG has the particularity to model both the individuals and the group in interrelated modules (see Figure 5.3), hence, making it suitable for studying the influence of group emotion following the Bottom-up approach (i.e., individual emotions influence the group one) and the Top-down approach (i.e., group emotion influence the individuals' one). FItG performances, however, differ from the ones presented in Chapter 5 as they were obtained using a previous training methodology that did not apply the data augmentation strategy consisting of adding Gaussian noise. Retraining all the models with the latest training methodology was not feasible in the remaining time of the Thesis.

6.2.3.1 Labeling Strategy for Group Emotion

To build our labels of group emotion, we first analyzed the labels of emotion provided by each group member over the five tasks (i.e., they could choose among the following labels: Admiring, Angry, Proud, Ashamed, Happy, and Frustrated. See Section 3.2.1.4 for more details). Figure 6.1a shows the percentages of the six labels of emotion per task. The two most dominant labels of emotion chosen were "*Happy*" and "*Frustrated*". The "*Other*" category includes 19 different labels of emotion provided by the participants. In the tasks eliciting an increase of cohesion in both dimensions (i.e., Tasks 4 and Task 5), happiness was the most dominant feeling, corresponding to 34% and 54% of the answers, respectively. In Task 1, the feeling of happiness was probably influenced by participants' excitation at the start of the game. The following three other emotions related to Task 1 were, however, chosen: *Proud*, *Frustrated* and *Admiring*. A participant was more likely to feel proud or frustrated depending on whether she found an object or not. Arguably, as participants were friends, one would more easily feel admiration toward one's group members.

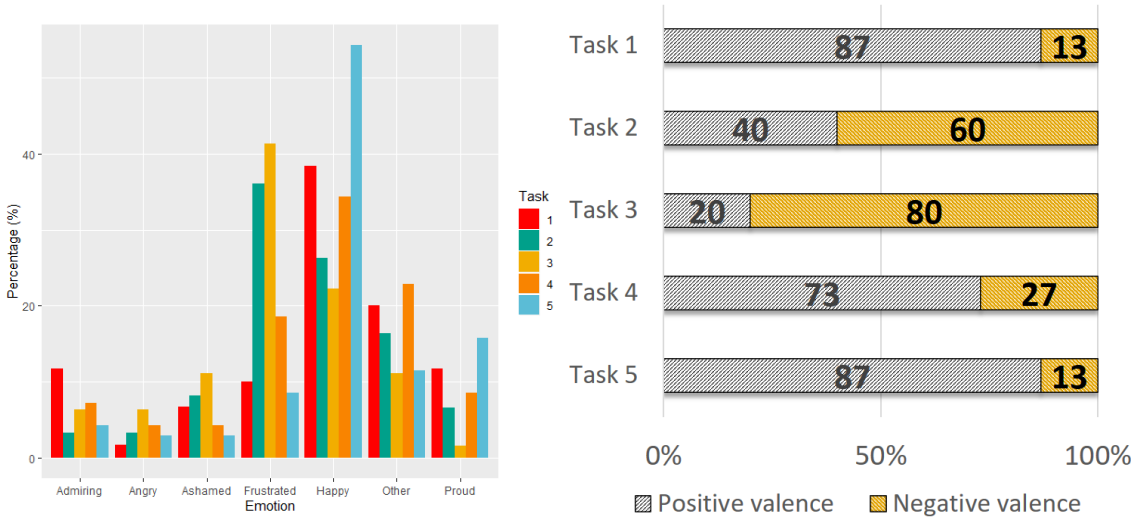
In Task 2 and Task 3, participants felt frustrated (36% and 41% respectively). These two tasks were intentionally made difficult (or impossible) to complete. In Task 2, however, we observed a higher diversity in the answers. This is probably related to participants' appreciation of the quality of their own performance. We also noticed that happiness was either the first or the second most dominant emotion in every task of the game.

Building upon evidence showing that positively or negatively valenced emotions could affect cohesion in tasks requiring group decision making or creativity (Barsade and Knight, 2015), independently of the approach investigated (i.e., Top-down or Bottom-up, Vanhove and Herian, 2015), emotion is here addressed in terms of its valence.

Valence labels are obtained in the following way. We first assigned a valence (positive or negative) to every emotion picked up by each group member, after each task (more than one emotion could be provided per group member). Then, for each task and each group, we summed up $+1$ if a group member chose an emotion with a positive valence (e.g., happy) or -1 if a group member chose an emotion with a negative valence (e.g., ashamed). Depending on the sign of this sum, we defined the label as "*Positive valence*" or "*Negative valence*". This labeling strategy resulted in a slightly imbalanced distribution (61% of *Positive valence*). Similarly to the cohesion labels, high imbalances for each task occurred (see Figure 6.1b).

6.2.3.2 DNN Approaches to Integrate the Cohesion-emotion Relationships

We designed two architectures, starting from the fItG: fItG_Bu and fItG_Td, implementing the Bottom-up and Top-down approaches, respectively. In these architectures group emotion was integrated using multitask learning, taking inspiration from the work of Parthasarathy and Busso (2017) that designed a framework to jointly predict arousal, valence and dominance using multitask learning. They proved that a primary task (i.e., predicting arousal) could benefit from multitask learning by taking advantage of the shared representation of the features jointly learned with the secondary tasks (i.e., predicting valence and dominance). Similarly, according to Vanhove and Herian (2015) stating that relationships exist between emotion and the Social and the Task dimensions, we



(a) Distribution of individual emotion labels per task (b) Group emotion partition (overall: 61/39%)

Figure 6.1: Percentages of the six emotion labels provided by each group member (Figure 6.1a), for the five tasks, and the resulting group emotion labels distributions based on the valence of group emotion (Figure 6.1b). Overall, 61% of the labels are *Positive valence* while 39% are *Negative valence*. A high imbalance is observed for each task.

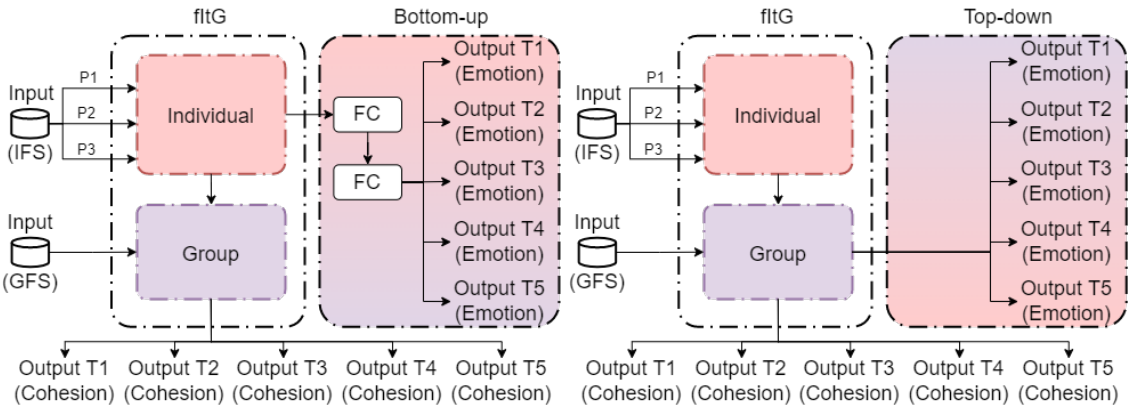
expect that the prediction of the dynamics of these cohesion dimensions (taken as the primary task) will be improved by the knowledge extracted from the prediction of group emotion (taken as the secondary task).

Figure 6.2 depicts both fItG_Bu (Figure 6.2a) and fItG_Td (Figure 6.2b). In the fItG_Bu, the three combined outputs from the Individual module are taken as input for the Bottom-up module. This input feeds two FC layers with a ReLu activation function and 64 and 16 units, respectively. These layers are followed by an FC layer with a Sigmoid activation function and one unit, for each task. These final layers predict the valence of group emotion for each task. As valence is predicted from the output of the Individual module of the fItG, it has, during training, a direct impact on the common shared representation of an individual. The Individual module being part of the input of the Group module, integrating emotion following the Bottom-up approach also affects the group representation.

In the fItG_Td, the output of the Group module is taken as input. An FC layer with a Sigmoid activation function and one unit for each of the five tasks is used. In this way, the group and individual representations will both be impacted by the valence prediction during back-propagation.

6.2.4 Results and Discussion

As in Chapter 5, we apply the same procedure to compare the performances obtained by the three architectures and we chose a significance level of $\alpha = 0.05$ for the statistical tests. For the sake of brevity, only the significant results are detailed.



(a) fItG_Bu (b) fItG_Td

Figure 6.2: FItG_Bu (Figure 6.2a) predicts group valence emotion after the *Individual* module of the fItG and reflects a Bottom-up approach (i.e., individuals influence the group). FItG_TD (Figure 6.2b) predicts group valence emotion after the *Group* module of the fItG and implements a Top-down approach (i.e., the group influences each individual).

As reported in our previous study (Maman et al., 2021a), the fItG model obtains an average F1-score of 0.69 $\pm$0.03 for the Social dimension and 0.61 $\pm$0.03 for the Task dimension, over the five tasks and over the 15 seeds. FItG_Bu reaches an average F1-score of 0.67 $\pm$0.03 for the Social dimension and 0.65 $\pm$0.04 for the Task dimension, while it achieved 0.65 $\pm$0.03 for valence. FItG_Td, meanwhile, obtained an average F1-score of 0.68 $\pm$0.03, 0.63 $\pm$0.03 and 0.64 $\pm$0.04 for the Social and Task dimensions and for the valence, respectively.

A permutation test shows a significant difference in performances between the three architectures, for the Task dimension only ($p = .016$). A possible explanation is that positive emotions maintain a particularly strong relationship with social cohesion (Vanhove and Herian, 2015), making it more difficult for the model to differentiate these processes. A post-hoc analysis using pairwise permutation t-tests is also carried out. These tests reveal that only fItG_Bu reaches significance ($p = .012$). This improvement in performances (from 0.61 $\pm$0.03 to 0.65 $\pm$0.04) indicates that integrating valence in a Bottom-up fashion helps the model to learn a better representation of an individual, leading to a more accurate representation of the group as well. This result is in line with the Social Sciences literature stating that emotions convey attributes such as intentions, and capabilities (Magee and Tiedens, 2006), which are also relevant for the instrumental property of cohesion and more specifically for the Task dimension (Severt and Estrada, 2015). Table 6.1 summarizes the performances of the three architectures for the Social and Task dimensions of cohesion as well as for the valence of group emotion.

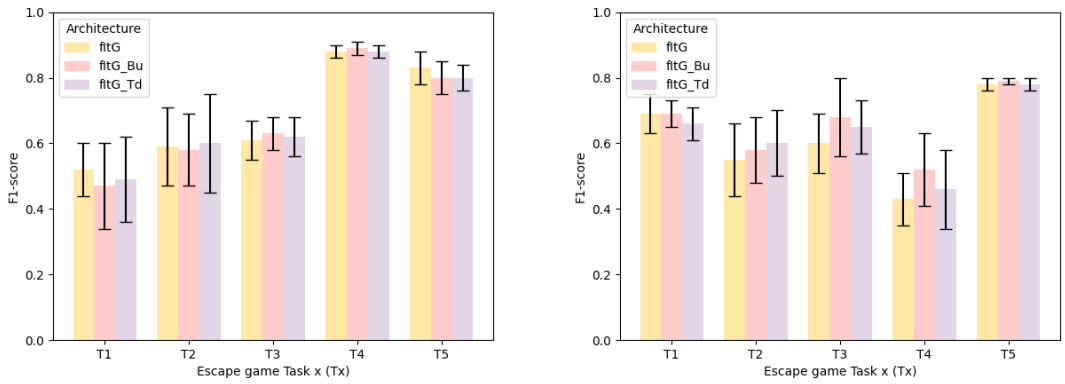
Then, we analyze the details of the tasks' performances, for each dimension. Task 1 for the Social dimension and Task 4 for the Task dimension are particularly miss-predicted for the three models (see Figure 6.3). FItG_Bu, however, significantly improves the performances of Task 4 concerning the Task dimension. It obtains a significantly higher average F1-score (i.e., 0.52 $\pm$0.10 instead of 0.43 $\pm$0.08, $p = 0.022$). Regarding the prediction of the secondary task, that is the prediction of valence, the model reaches, on

average, an F1-score of 0.65 ±0.03. We can explain these performances by the fact that the models' selection, for each seed, is based on the highest F1-score of the primary task (i.e., the prediction of the Social and Task dimensions). This requires a trade-off in terms of performance for the secondary task. These results indicate that the features contain enough information to describe both group processes. It also highlights the difficulty to predict both group processes within the same model.

To summarize, only integrating valence following the Bottom-up approach significantly improves the performances of the fItG model for the Task dimension, and especially for Task 4. This result confirms that jointly predicting the dynamics of cohesion and the valence helps to learn a shared representation of the features that brings additional information to the prediction of the Task dimension of cohesion.

Table 6.1: Summary of the average F1-scores over the 15 seeds for the primary and secondary tasks (predicting cohesion's dynamics and valence of group emotion, respectively), per task and per dimension for the fItG, the fItG_Bu, and the fItG_Td models.

	F1-scores ±std							
	fItG		**fItG_Bu**			**fItG_Td**		
	Social	**Task**	**Social**	**Task**	**Group emotion**	**Social**	**Task**	**Group emotion**
T1	0.52 ±0.08	0.69 ±0.06	0.47 ±0.13	0.69 ±0.04	0.76 ±0.05	0.49 ±0.13	0.66 ±0.06	0.78 ±0.07
T2	0.59 ±0.12	0.55 ±0.11	0.58 ±0.11	0.58 ±0.10	0.55 ±0.10	0.60 ±0.15	0.60 ±0.10	0.40 ±0.13
T3	0.61 ±0.06	0.60 ±0.09	0.63 ±0.05	0.67 ±0.12	0.66 ±0.03	0.62 ±0.06	0.65 ±0.08	0.67 ±0.05
T4	0.88 ±0.03	0.43 ±0.08	0.88 ±0.02	0.52 ±0.10	0.47 ±0.08	0.88 ±0.02	0.46 ±0.12	0.57 ±0.08
T5	0.84 ±0.05	0.78 ±0.02	0.81 ±0.05	0.79 ±0.02	0.79 ±0.02	0.80 ±0.04	0.78 ±0.02	0.78 ±0.02
Average	**0.69 ±0.03**	**0.61 ±0.03**	**0.67 ±0.03**	**0.65 ±0.04**	**0.65 ±0.03**	**0.68 ±0.03**	**0.63 ±0.03**	**0.64 ±0.04**



(a) Social cohesion

(b) Task cohesion

Figure 6.3: Average F1-score per task over the 15 seeds for Social (Figure 6.3a) and Task (Figure 6.3b) cohesion. FItG is in yellow, fItG_Bu in pink and fItG_Td in purple.

6.3 Cohesion and Emergent Leadership

6.3.1 Emergent Leadership

There has been increasing attention to examining informal (i.e., horizontal), rather than formal (i.e., vertical), approaches to leadership over the last several decade (Hanna et al., 2021). This follows the recent trends of flattening organizational hierarchies and self-managed teams (e.g., Zaccaro et al., 1991; McClean et al., 2018), leading to new types of informal leadership such as emergent leadership and shared (or collective) leadership. These kinds of leadership arise naturally from group interaction, rather than from a higher authority such as managers (Hanna et al., 2021). While emergent leadership is defined as "*the degree to which an individual with no formal status or authority is perceived by one or more team members as exhibiting leaderlike influence*" (Hanna et al., 2021), shared leadership is conceptualized as a group dynamic process in which group members interchangeably "*utilize skills and expertise within a network, effectively distributing elements of the leadership role as the situation or problem at hand requires*" (Friedrich et al., 2009).

In our work, we focus on emergent leadership, an individual emergent state that evolves over time (Gerpott et al., 2019) and that has been positively linked to cohesion and team performance (e.g., De Souza and Klein, 1995). Previous studies show that a team with an emergent leader can outperform teams with a formally designed leader (e.g., De Souza and Klein, 1995; Taggar et al., 1999; Spisak et al., 2015). The role of an emergent leader is, however, never settled. It depends on the person's abilities, the need of the group, and the team task (Seers, 1989). Thus, the nature of the task impacts its emergence (i.e., an emergent leader may appear in a team for a particular task but not for another, Taggar et al., 1999).

Despite the complexity of defining emergent leadership, the SSP community started investigating its automated analysis by collecting several datasets specifically designed for it such as the ELEA (Sanchez-Cortes et al., 2011a) dataset. This facilitated the development of computational models for detecting emergent leadership (e.g., Sanchez-Cortes et al., 2011b; Beyan et al., 2016a, 2019; Muller and Bulling, 2019) that range from simple (e.g., using an SVM to predict the most and the least emergent leader in a group from video only, Beyan et al., 2016b) to more complex approaches (e.g., using unsupervised learning with both video and audio data, Beyan et al., 2019). Features used in these studies are related to the speaking activity (e.g., total speaking time, total time of silence) and the visual focus of attention (e.g., looking someone with no mutual engagement, total time being looked at). Such features are used in our computational models and are described in Appendix B. These automated approaches are, however, centered on emergent leadership only. Such an emergent state has relationships with other group processes (e.g., dominance, Kalma et al., 1993), hence, its automated analysis could benefit from other group processes as they may simultaneously occur.

6.3.2 Links between Emergent Leadership and Cohesion

Previous studies from Sociology and Psychology reveal that a link between emergent leadership and cohesion exists (e.g., Light Shields et al., 1997; Stashevsky and Koslowsky, 2006; López-Zafra et al., 2008; Callow et al., 2009; Vincer and Loughead, 2010; Tung and

Chang, 2011). For example, Callow et al. (2009) empirically show a positive correlation between some of the emergent leadership behaviors (e.g., fostering acceptance of group goals, promoting teamwork) and the Social and the Task dimensions of cohesion by analyzing leadership and cohesion questionnaires from 309 clubs standard ultimate Frisbee players in the United Kingdom. Also, cohesion has been proven to mediate the relationship between emergent leadership and team performance (Dionne et al., 2004; Tung and Chang, 2011). In particular, Xie et al. (2019) investigated college student group work in an online class and showed a strong correlation between emergent leadership and cohesion. Yamaguchi and Maehr (2004) also found that emergent leadership leads to stronger cohesion in elementary classrooms where students collaborate in math activities.

From a computational perspective, there is, however, to the best of our knowledge, only Wang et al. (2012)'s study (see Chapter 2) that integrates the links between cohesion and leadership (without differentiating between formal and informal leadership). In their study, they predicted cohesion based on leadership and (dis)agreement between group members. This study, however, only uses audio-verbal features with a logistic regression model and does not take into account the dynamic aspects of both processes. It also does not explore if their model manages to predict cohesion without leadership, making it hard to evaluate the impact of the integration of the links between leadership and cohesion.

6.3.3 Labeling Strategy for Emergent Leadership Detection

Since more than one person can exhibit leadership in small groups (Taggar et al., 1999), we made the assumption that a group of three persons can either be composed of zero, one or two emergent leaders.

To build our labels for the emergent leadership detection, we considered such a task as a binary classification problem (i.e., zero means a person is not an emergent leader while one indicates it is an emergent leader). The procedure to obtain the labels is explained as follows. First, we define a leadership score for each group member to compare them and determine whether an emergent leader exists in the group.

As described in Chapter 3, group members provided self- and external assessments of leadership in a round-robin rating, by answering a set of five questions (see Appendix A, for the details of the questions). Both these assessments have pros and cons (Vinciarelli and Mohammadi, 2014). With self-assessment, persons are inclined to judge their performance favorably, while external assessment tends to limit such a bias (Koopmans et al., 2013) but might not reflect the *true* internal state of the person (Uleman et al., 2008). Therefore, for each of the five questions, we choose to emphasize the differences between external and self-assessments of leadership by multiplying the external ratings by 0.4 and multiplying self-assessments by 0.2. Such values (i.e., 0.2 and 0.4) were empirically chosen so they add up to 1. Then, to compute the score for each of the six items and for each group member, the self- and external assessments were summed and normalized by six (i.e., the maximum score possible on the Likert scale).

Based on the five scores obtained (i.e., one for each item) for each task and each group member, we computed their mean, as it is usually done in automated studies on leadership, based on self-assessments (e.g., Sanchez-Cortes et al., 2010, 2011a,b). We also computed their median. In this way, we aim to capture possible disagreement between group members' leadership scores, as suggested in the work of Hanna et al. (2021).

Afterward, for both the mean and median scores (i.e., $mean_L$ and med_L, respectively), we applied a similar strategy to detect the number of emergent leaders in each group, as follows.

For each task and each score (i.e., $mean_L$ or med_L), if the difference between the minimum and the maximum scores obtained in the group is higher than a threshold of 0.1[1], it is indicative of the presence of at least one emergent leader for this particular task. In this case, we also compute the difference between the scores of the two potential leaders. If this difference is smaller than a second threshold of 0.05[1], it means that two emergent leaders exist for this particular task as they were both perceived almost as influential as the other leader.

Applying such a strategy to both $mean_L$ and med_L results in two (potentially different) labels distributions. The final label is obtained as follows: for each task and each group member, if the labels from both distributions are similar, we select it as the final label. Otherwise, we first test the reliability of the threshold obtained with med_L. We, indeed, noticed that the med_L labels distribution was more sensitive to the variations of thresholds. In fact, a diminution of 0.01 and 0.005 for the first and second thresholds, respectively, resulted in 8% of different labels. Thus, for such edge cases, the label obtained with $mean_L$ is conserved. Otherwise, the label obtained with med_L is held. This labeling strategy results in a slightly imbalanced labels distribution (i.e., 60% of "*leaders*" and 40% of "*non-leaders*") and was validated by an expert in groups that watched random samples of videos from the GAME-ON dataset to assess leadership.

6.3.4 Families of Approaches

6.3.4.1 Features Based Leadership

Scholars in Social Psychology state that an emergent leader is the most influential and active person in the group, who talks and moves the most (Baird Jr, 1977; Stein and Heller, 1979; Darioly and Mast, 2014). The two following approaches are based on these insights and suggest amplifying the features that characterize emergent leaders. The first one called *Normalization*, consists of normalizing the individual features of each group member regarding the ones of the leader(s). The second one named *Weighting*, gives a particular weight to the features that are relevant for describing leaders' behavior.

Normalization

We amplified the differences between the leader(s) and the follower(s) by normalizing each individual's features with respect to the ones of the leader(s). Leaders were identified based on the labeling strategy previously described in Section 6.3.3. Concretely, we applied the Min-Max scaling method to each individual feature, taking the minimum and maximum values from the feature vector(s) of the leader(s), i.e., $min(feat_{leader})$ and $max(feat_{leader})$, respectively, as follows in Equation 6.1:

$$X\_norm_i = \frac{X_i - min(feat_{leader})}{max(feat_{leader}) - min(feat_{leader})} \tag{6.1}$$

[1]This threshold was empirically determined.

Where X_i is the feature vector of the group member i, and X_norm_i is the same feature vector X_i normalized according to $min(feat_{leader})$ and $max(feat_{leader})$. In the case $min(feat_{leader})$ and $max(feat_{leader})$ are not the extremes values for each group member, it implies that some values of the feature vector of the follower(s) are not in the standard range of normalization (i.e., between zero and one). For that reason, all the values greater than one and less than zero were set to one and zero, respectively.

Weighting
For this specific approach, we defined a new subset of features, i.e., the "*Leadership features set*" (LFS), obtained from FFS, that is composed of features that are relevant to studying emergent leaders' behavior. Table 6.2 shows the features present in LFS. To constitute such a features set, we took inspiration from the fact that an emergent leader is perceived by his peers as a dominant person with the most active body language (Gerpott et al., 2018). In more detail, the emergent leader is perceived as the person who walks and talks the most, has an active posture, and is also the person who has the longest variation in the tone of voice and energy (Sanchez-Cortes et al., 2011b; Gerpott et al., 2018). Thus, the Weighting approach only weights the LFS features before inputting them into the fItG. In that way, the differences between the features of the leader(s) and its follower(s) are amplified. We empirically tested multiple weighting values (i.e., from 1.5 to 5) to observe whether and how amplifying the differences between the emergent leader and its followers impacted the fItG performances.

Table 6.2: The "*Leadership features set*" (LFS). Each feature was selected as it is associated to emergent leadership behavior. In the *Weighting* approach, these features are weighted to amplify the differences between emergent leader(s) and follower(s) in the fItG model. Features with a "$\star$" indicates that their functionals were selected.

Features	
Motion capture-based	**Audio-based**
Maximum of the interpersonal distances$\star$	Pitch
Distances from	Jitter
group barycenter$\star$	Shimmer
Total distance traveled	Loudness
Occupied volume$\star$	HNR
Kinetic energy$\star$	Total speaking time

6.3.4.2 Representation Based Leadership

This family of approaches aims to integrate a behavior representation that incorporates leadership information into a DNN. Here, we specifically focus on modifying fItG's individual module since emergent leadership is an individual-level emergent state (Taggar et al., 1999; Hanna et al., 2021). To this aim, we present two approaches: the first one, named *Extracted from Assessments*, directly uses the leadership scores obtained through the labeling strategy (see Section 6.3.3). The second one, called *Automatically Learned*, uses a behavior representation learned by a pre-trained model that predicts emergent leadership. In both approaches, the behavior representation is concatenated with the outputs of the individual module of the fItG and is the input of another FC layer shared among the three group members. This extra layer allows the model to learn a higher-level represen-

tation of the individual behavior that integrates leadership knowledge. Figure 6.4 shows how both approaches from this family are integrated into the fItG.

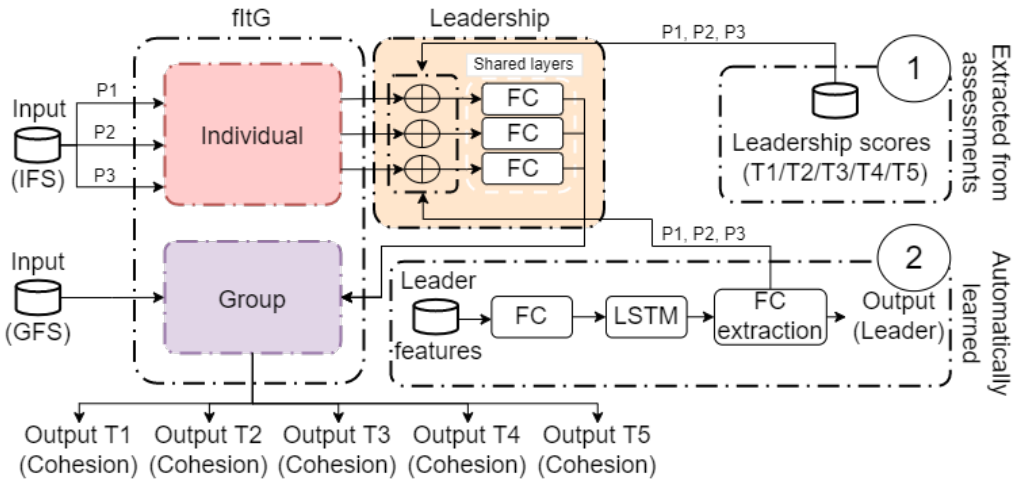


Figure 6.4: Integration of the approaches of the Representation Based Leadership family into the fItG model. The *Extracted from Assessments* approach is using leadership scores while the *Automatically Learned* approach uses a DNN to generate a representation of behavior based on leadership. Both are integrated into the fItG model after the individual module through shared FC layers. Each approach is used independently.

Extracted from Assessments

This approach is straightforward and consists of concatenating the leadership scores obtained for each group member, at each task, with the outputs of the fItG's Individual module, into the extra shared FC layer. These leadership scores represent the degree of a person to be perceived as a leader by her peers through the escape game. By providing such scores, we aim to guide the model in learning a more efficient representation of individual behaviors that integrates leadership information.

Automatically Learned

Regarding the second approach, the goal is also to inject new leadership information by creating an automatically generated representation of leadership behavior. To this aim, we developed and pre-trained a DNN model aimed at predicting emergent leadership for each individual to be used as a feature extractor Using DNNs as features extractor is, indeed, a common and robust practice (e.g., Schroff et al., 2015; Liu et al., 2017). We studied the detection of emergent leadership for each particular individual as a binary classification problem (i.e., zero means that the individual is not an emergent leader while one indicates she is an emergent leader). Also, we developed a set of features, grounding on studies addressing the automated detection of emergent leadership (i.e., Sanchez-Cortes et al., 2011b; Beyan et al., 2016a,b). These features are either related to the speaking activity (SA) or related to the Visual Focus Of Attention (VFOA), as presented in Table 6.3. Details about the computation of these features are provided in Appendix B. SA features correspond to the speaking length, the interruption between individuals, and the turn-

Table 6.3: List of nonverbal features used in the *Automatically Learned* leadership representation approach. These features are extracted for each individual, independently, and are related to their speaking activity (SA) and their visual focus of attention (VFOA).

Features	
Speaking activity	
Total speaking time when at least one group member is speaking (Tss)	Average of speaking turns duration
Total speaking time when no one is speaking (Tsn)	Total number of time being un/successfully interrupted (TIunsuc/TIsuc)
Total number of times a person speaks first right after another one	Total number of time un/successfully interrupting other turns (TIOunsuc/TIOsuc)
Ratio between Tss and Tsn	Total number of speaking turns (Tst)
Total time of silence (Tsil)	Ratio between TIOsuc and TIsuc
Ratio between total speaking time (Tss+Tsn) and Tsil	Ratio between TIOsuc and Tst
	Ratio between TIunsuc and Tst
Visual focus of attention	
Looking someone with no ME (LnoME)	Total time being looked at
Being looked with no ME (BLnoME)	Number of times one initiates a ME
ME with any member	Ratio between BLnoME and LnoME

taking of each person, while VFOA features essentially relate to mutual engagement (ME) that is happening when two persons are looking at each other at the same time.

This set of features serves as the input of the pre-trained leadership model. This model is composed of a Fully Connected layer with a ReLu activation function and 24 units followed by an LSTM layer and an FC layer with a ReLu activation function and 16 units. The output of this layer is used to (1) make the final prediction (i.e., leader or not leader) during training, thanks to an FC layer with a Sigmoid activation function and one unit, and (2) integrate the learned representation of leadership into the fItG model during the fItG training phase. This pre-trained model is designed to predict if a group member is an emergent leader for a specific task, independently of her group. We trained it using a 5-fold cross-validation and a fixed learning rate of 0.0001 coupled with an early stopping regularization technique on the epochs to avoid over-fitting. Performances were evaluated using the average F1-score, allowing us to compare them with the fItG. We run this model on 1000 randomly extracted seeds and we averaged its performances, reaching an average F1-score of 0.64 $\pm$0.02. Considering the variety of tasks on which the model is evaluated, such a performance is acceptable. For the purpose of using it as a pre-trained model in the fItG, we selected the most performing model that reached a 0.71 F1-score.

6.3.5 Results and Discussion

Results presented in this Section aim to show whether and how applying our approaches improved the fItG performances on both the Social and Task dimensions of cohesion. We first test if there are significant differences between each family of approaches with respect to the fItG. Then, the best approach from each family is compared to each other. The same procedure to compare the performances of the models described in Chapter 5 is applied in this Section. Performances of the fItG are the same as the ones presented in Section 6.2

(i.e., an average F1-score of 0.69 ±0.03 for the Social dimension and 0.61 ±0.03 for the Task dimension, over the tasks and over the 15 seeds).

Regarding the Social dimension of cohesion, there is no significant difference in performances between the approaches from the same family and the fItG. Significant improvements of the F1-Score for the Task dimension of cohesion are, however, achieved for each family of approaches. For the sake of clarity, only the significant results are reported, hence, the remaining of the analysis focuses on the Task dimension of cohesion. Table 6.4 summarizes the F1-scores obtained by each approach of both families for the Social and Task dimensions of cohesion.

This result is in line with Social Psychology's insights stating that, when a team is working under a time constraint, emergent leaders focus on the task by assigning roles to the group member and developing strategies to improve team performance (De Souza and Klein, 1995; Taggar et al., 1999).

Table 6.4: Summary of the average F1-scores for the fItG model and each approach from the Features Based Leadership and Representation Based Leadership families. Performances with a "*" indicate a significant difference with respect to the fItG.

		F1-score ±std	
Family	**Approach**	**Social**	**Task**
Baseline	fItG	0.69 ±0.03	0.61 ±0.03
Features Based Leadership	Normalization	0.66 ±0.04	0.62 ±0.04
	Weighting (by 1.5)	0.68 ±0.03	0.64 ±0.04*
Representation Based Leadership	Extracted from Assessments	0.67 ±0.04	0.65 ±0.03*
	Automatically Learned	0.67 ±0.03	0.67 ±0.04*

6.3.5.1 Approaches from the Features Based Leadership Family

Concerning the Task dimension of cohesion, statistical tests show a significant difference between this family of approaches and the fItG ($p = .010$). Post-hoc analysis reveals that only the *Weighting* approach significantly improved fItG performances (i.e., from 0.61 ±0.03 to 0.64 ±0.04, $p = .006$). These findings suggest that amplifying a leader's behavior might be beneficial to a certain extent. In fact, the *Normalization* approach amplifies the differences between the leader's features and its followers' features by 21%. This amplification may give the emergent leader(s) too much importance making followers insignificant to the model. In comparison, the *Weighting* approach (with a weight set to 1.5) amplifies the differences by 4%. We empirically confirmed this effect by using different weights (from 1.5 to 5), which corresponds to an amplification of differences ranging from 4% to 13%. As displayed by Figure 6.5, amplifying leaders' behavior significantly improves performances (see p-values in the yellow boxes) until a weight of three (i.e., an amplification of differences of 11%). Augmenting the weight to a higher value does not improve fItG performances.

These results show that highly amplifying the differences between an emergent leader and its followers might go against the emergence of an informal leader. Particularly, a highly amplified leader can be perceived as autocratic (i.e., a leader who has too much control over the task), which has been shown to be less effective during task completion (Lewin, 1939).

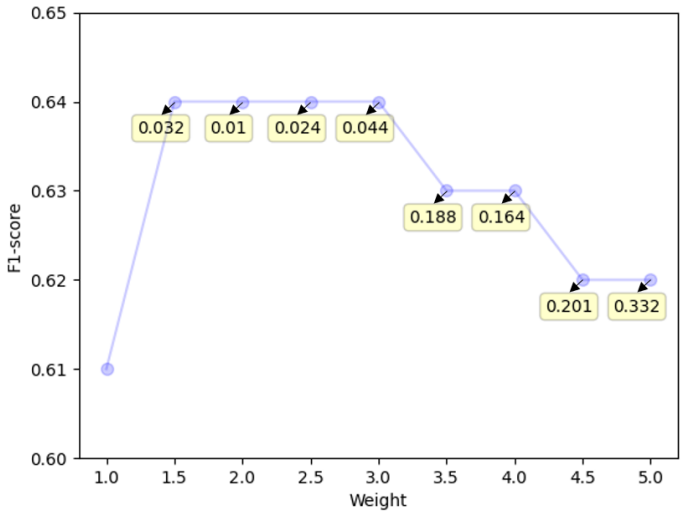


Figure 6.5: Average F1-score over 15 randomly extracted seeds of the fItG model, for the prediction of the Task dimension of cohesion with the *Weighting* approach. Weights ranging from one to five are applied to the features. For each weight, p-values (in yellow) indicate whether or not there is a significant difference with the baseline (weight 1).

6.3.5.2 Approaches from the Representation Based Leadership Family

About this family of approaches, statistical tests show a significant difference in performances between both approaches and the fItG for the Task dimension only ($p = .001$). Post-hoc analysis reveals that both approaches significantly improve fItG performances. In particular, the *Extracted from Assessments* approach based on the computed leadership scores significantly improves fItG from 0.61 $\pm$0.03 to 0.65 $\pm$0.03 ($p = .003$). The *Automatically Learned* approach based on the pre-trained model also reaches a significantly better F1-score of 0.67 $\pm$0.04 ($p = .003$). Finally, this approach also significantly outperforms the *Extracted from Assessments* approach ($p = .014$). These improvements highlight the benefits of integrating leadership information directly into the computational model architecture.

Such approaches help the model to learn a high-level representation of individuals that integrates leadership's characteristics. Furthermore, the fact that the Automatically Learned approach outperforms the Extracted from Assessments indicates that the cohesion model is sensitive to the variety of information added. In fact, the Automatically Learned approach has a leadership representation for each group member and each task while the Extracted from Assessments approach only provides a representation of leadership that is the same across the tasks, for each person.

6.3.5.3 Comparing Both Families of Approaches

Lastly, we compare the best-performing approaches of each family, for the Task dimension of cohesion. The *Automatically Learned* approach achieves an F1-score of 0.67 $\pm$0.04. Thus, it significantly outperforms the performances of the *Weighting* approach that reaches an F1-score of 0.64 $\pm$0.04 ($p = .010$). This result shows that the

Representation Based Leadership family of approaches is more effective than the *Features Based Leadership*.

It highlights the benefits of adding extra information for learning a representation of individuals instead of solely relying on amplifying existing features. The best performing approach is, indeed, using additional information from other features to automatically detect emergent leaders. Such an approach helps the fItG learn a more complex representation of individuals since it merges two sources of information (instead of one for the other family of approaches). In that way, the fItG learns new patterns that improve the prediction of the dynamics of the Social and Task dimensions of cohesion.

6.4 Conclusion

THIS Chapter presented our work as part of the "Relationships with other group processes" research axis (i.e., RA4), helping us answering RQ2. In summary, we first investigated the relationships between cohesion and group emotion (addressed as its valence), by introducing fItG_Bu and fItG_Td, two DNN architectures based on the fItG, that respectively implement the Bottom-up and Top-down approaches for characterizing group emotion. We showed that only integrating emotion following the Bottom-up approach significantly improved the performances of the fItG model for the Task dimension. This result corroborates previous findings with respect to the link between emotions and Task cohesion in particular (Vanhove and Herian, 2015).

As for the relationships between cohesion and emergent leadership, we introduced two families of approaches to integrate leadership information into DNN architectures designed for automatically studying cohesion. The first one focuses on amplifying leaders' features to increase the gap with their follower(s). The second one injects additional leadership information into the architecture to help the model learn a representation of behavior that takes such relationships with cohesion into account. Similarly, results showed that these approaches were only improving performances for the Task dimension of cohesion. Due to the nature of the interaction (i.e., collaborating under a time constraint), it is also in line with Social Psychology's insights regarding the emergence of task-focused leaders (De Souza and Klein, 1995; Taggar et al., 1999). Also, the most efficient approach consists of injecting a leadership representation based on leadership-specific features in the cohesion model, hence, underlining the importance of extracting diverse behavioral information.

To conclude, both works highlight the benefits of integrating other processes into computational models of cohesion and show the potential of implementing strategies following the "Relationships with other processes" research axis.

Chapter 7

Conclusions, Limitations and Future Work

Contents

In the work presented in this Thesis, we contributed to the development of computational models of cohesion. Their architectures were designed following four research axes that were identified through a survey of the Social Sciences' literature on emergent states, and cohesion. In particular, our computational models range from a simple but consolidated state-of-the-art approach to more sophisticated approaches that increasingly address the temporal nature of cohesion, the group modeling, the interplay between its Social and Task dimensions, and the links with other group processes.

In this Chapter we conclude this Thesis by summarizing the main contributions, identifying the limitations of this work, and suggesting future work from both short- and long-term perspectives.

7.1 Summary of Contributions

First contribution: *A structured survey on cohesion for supporting the automated analysis of cohesion in small groups interactions.*

During the first phase of our work, we analyzed the Social Sciences literature and we identified multiple challenges specific to the study of cohesion. Four research axes (RA) stemmed from this analysis. We built a structured survey for supporting the automated analysis of cohesion around them. The four research axes are the following ones:

RA1: The temporal nature of cohesion

RA2: The group modeling

RA3: The interplay between its dimensions

RA4: The relationships with other group processes

Inspired by the "Input–Process–Output" (IPO) theoretical framework (Hackman and Morris, 1975) for conceptualizing teams in Social Sciences, we clustered the approaches employed in the literature on the automated analysis of cohesion according to the following two criteria: the research axis that they address and the level at which they are applied (i.e., Input, Model or Output) in the computational model of cohesion (see Figure 2.6). In addition, we also introduced approaches that are, in our opinion, worth investigating. This structured survey helped us organizing our work and served as a basis for designing our computational models of cohesion.

Second contribution: *Multimodal dataset for the automated cohesion analysis.*

To the best of our knowledge, at the time of the data collection, there was no existing dataset explicitly addressing cohesion. Thus, we collected GAME-ON (Group Analysis of Multimodal Expression of cohesiON), a multimodal dataset specifically designed for the study of cohesion dynamics. It is composed of more than 11 hours of video, audio and motion capture data from the interaction of 17 groups of three friends playing an escape game. The escape game was thought to elicit variations of the Social and Task dimensions of cohesion across five tasks that require different skills to solve a murder (i.e., finding the murderer, its weapon, and the location of the murder). In addition, we also collected repetitive self-assessments (before and after each task) of cohesion and other group processes such as leadership and emotion as well as external assessments of cohesion.

Such a dataset enables the automated study of cohesion through the four research axes previously mentioned. It, indeed, elicited variations, for multiple groups, of the Social and Task dimensions of cohesion, an affective group emergent state, over time (RA1, RA2, and RA3). In addition to the self- and external assessments of cohesion, it also contains self-assessments of diverse group processes such as leadership and emotion. Thus, facilitating the study of the relationships between cohesion and other group processes (RA4).

Third contribution: *Design and implementation of computational models of cohesion.*

We designed and implemented a set of computational models of cohesion that gradually address the four research axes. First, we showed a Random Forest classifier, following the approaches employed in the current literature on the automated analysis of cohesion. Then, we explicitly investigated the first research axis with the FI-LSTM model and the first two research axes with the fItG model. Analytical results show that fItG is the most performing model. Moreover, we designed more complex approaches based on it to address RA3. In fact, we implemented STE and CTS based on an approach inspired by coalitional game theory. These models, however, did not significantly improved fItG's performances. Also, inspired by Social Sciences' insights on the way the Social and Task dimensions of cohesion interplay over time, we built three additional DNNs architecture leveraging a transfer learning approach that each reflects a different theory regarding the way the Social and Task dimensions interplay. In particular, TBD-S assumes that Task cohesion sets the stage for Social cohesion while TBD-T suggests the opposite (i.e., Social cohesion informs Task cohesion). TBD-RI integrates the reciprocal interplay of these two dimensions over time. TBD-T and TBD-RI are the most performing models with respect to the Task and Social dimensions, respectively. Finally, we integrated the link between cohesion and emotion (i.e., RA4) into the fItG architecture. We designed two approaches to characterize group emotion, that are grounded on the Top-down and Bottom-up views of such a group process (Barsade and Gibson, 1998). Furthermore, we explored whether or not predicting cohesion and emotion in a multi-task setting could improve fItG's performances. Only the Bottom-up approach (from the individuals to the group) significantly improves them, for the Task dimension. We also designed two families of approaches to integrate the link between cohesion and emergent leadership. One directly impacted the features of the model (e.g., by weighting the ones of the emergent leader) and the other one focused on integrating emergent leadership representation into the model (e.g., from the labels of emergent leader). Approaches from both families improve the fItG performances.

With this collection of computational models of cohesion, we addressed each research axis, hence, answering both RQ1 and RQ2.

7.2 Limitations

The contributions presented in this Thesis are not exempted from limitations. In the following, we discuss such limitations from the input of the computational models (i.e., data and features), the model (i.e., according to each research axis), and the output (i.e., labeling strategy) perspectives.

7.2.1 Input

One of the main limitations of our work is the fact that we only used the GAME-ON dataset to evaluate our models. We, however, made this choice as it is, to the best of our knowledge, the only dataset that is specifically designed for cohesion and that provides

self-assessments of such an emergent state. It remains to be explored if similar conclusions apply in different contexts of interaction and types of groups.

GAME-ON also has some limitations. While it provides more than 11h of multimodal data from 17 groups composed of three persons (i.e., 51 individuals in total), it remains a relatively small amount of data, preventing us from developing deeper neural networks and limiting the generalization of the results. Furthermore, it was designed to explore only two dimensions of cohesion (i.e., Social and Task) over the four presented in (Severt and Estrada, 2015)'s framework. In addition, the relatively short duration of each data collection session (i.e., around 1 hour) is likely to have constrained the range of variations of cohesion we could observe.

As for the features, we focused on extracting them from the audio and motion capture data. Features computed from further signals (e.g., EEG) and modalities (e.g., face) could be used to enrich the features set. Also, we used a similar window size for both audio- and motion capture-based features. Thus, further analysis would help identify the optimal window sizes according to the signal and/or the modality.

7.2.2 Model

There are several limitations at the Model level that we categorized according to the four research axes.

Firstly, to investigate RA1, all of the models are predicting the dynamics of cohesion for the whole interaction, once all the thin slices used for each task (i.e., the last 2mn) are processed. Leaning toward the development of "real-time" applications, models relying solely on the thin slices of the previous and/or current task(s), instead of the whole interaction, should be investigated.

Secondly, to address RA2, all the DNNs were designed to integrate a pre-fixed number of person. Here, we tested the architectures on groups of three persons. Adding a new person to a group would imply retraining the models. Thus, designing architectures able to dynamically self-adapt to various sizes of groups would provide more flexibility to analyze more diverse groups.

While we identified multiple ways to explore RA3, the computational models focused on predicting the Social and Task dimensions of cohesion only. A new and open challenge would be to build computational models that can also take into account other dimensions (e.g., group pride) and their interplay.

Finally, with respect to RA4, we separately investigated the links between cohesion and group emotion and cohesion and emergent leadership on the fItG. Thus, approaches to integrate multiple group processes within the same computational model should be designed to fully integrate external influences.

7.2.3 Output

The labeling strategy we employed across all of our computational models of cohesion was designed to consider the dynamics of the Social and Task dimensions of cohesion as a binary problem (i.e., decrease vs no-decrease). More complex strategies could be conceived, for example, to integrate more granularities in the categorization of the cohesion's dynamics (e.g., *decrease* vs *static* vs *increase*). Another improvement would be to

account for the potential disagreements within the group (e.g., two "increase" vs one "decrease" in cohesion). Moreover, labels could also result from the combination of self- and external assessments of cohesion to minimize the cons introduced by both ratings (Vinciarelli and Mohammadi, 2014). Finally, the same limitations apply to the labeling strategy employed to characterize group emotion based on its valence. In fact, arousal could be used to complement valence, hence, providing more fine-grained information for moving from a binary to a multiclass approach.

7.3 Future Work

7.3.1 External Assessments of Cohesion

As previously mentioned in Section 7.2.3, more complex labeling strategies could be investigated. We are currently working on designing a *true* label of cohesion that mixes both self- and external assessments. This would be a first step towards the development of more robust computational models. Thus, we consider two approaches. The first one is straightforward and consists of mixing both types of assessments and applying the same labeling strategy described in Chapter 5. With this approach, each assessment (from a group member or an external annotator) is equally considered. The second strategy requires computing labels from both self- and external assessments, independently. Then, we select the most reliable label according to the Social Sciences literature. In fact, in a binary classification setting, three cases are possible: (1) both labels are similar (i.e., *decrease/decrease* or *not-decrease/not-decrease*), (2) the label extracted from self-assessments shows *not-decrease* and the label from external assessments is *decrease*, and (3) the label extracted from self-assessments indicates a *decrease* while the label form external assessments results in *not-decrease*. In case (1), both types of assessments lead to the same label, hence, we retain it. In case (2), we retain the label produced by external assessment. We ground this choice based on Vinciarelli and Mohammadi (2014)'s study stating that, when persons assess themselves, they tend to provide ratings towards socially desirable characteristics (here the presence of cohesion). Thus, in this particular case, we select the "*decrease*" label from external assessments. In case (3), we retain the label produced by self-assessments as external raters did not have the full context and outcomes of the interaction. In fact, they did not have information about the success or failure of the task. They only had a brief description of the task and did not know if the group succeeded, hence, potentially biasing external ratings. According to Mullen and Copper (1994), performance, indeed, has a stronger effect on cohesion than cohesion on performance. Furthermore, Boone et al. (1997) showed that failure affects cohesion more than success, especially for the Social and Task dimensions. Successes only maintain the initial level of cohesion. Thus, because of the negative impact of performance on cohesion, we select the label *decrease* from self-assessments, relying on group members' feelings and knowledge about the task's success.

7.3.2 Cohesion in a Virtual Environment

As briefly introduced in the structured survey (see Chapter 2), with the advent of new technologies and the actual world context (e.g., health crisis, climate change), more and more tools are developed to encourage people to meet and gather virtually (e.g., virtual and hybrid conferences). Among these new technologies, virtual reality (VR) applications are a promising medium for supporting distributed groups through a broad range of activities such as gaming (e.g., Star Trek: Bridge Crew[1]), building and socializing in virtual social communities (e.g., VRChat[2]), for educational purposes (e.g., Dunleavy et al., 2009) and many more. Thus, in a long-term perspective, understanding how cohesion manifests in a virtual environment would provide another angle of research and enrich our comprehension of cohesion, leading to the development of more robust multimodal, and potentially hybrid, systems, able to adapt to teams of humans and, eventually, mixed team of humans and robots or virtual agents.

Inspired by the GAME-ON scenario, we designed a VR application to study groups with members connected remotely in a virtual environment. As of today, we tested the application and ran a pilot study. Such a kind of application will enable the collection of data to study the interaction of diverse groups (e.g., various sizes, cultural differences, relationships among group members).

[1] https://www.ubisoft.com/fr-fr/game/star-trek/bridge-crew
[2] https://hello.vrchat.com/

Appendices

Appendix A

Questionnaires

A.1 Adapted GEQ

We adapted the GEQ with the following items on a 9-point Likert scale answering format (from 1: "*Strongly disagree*" to 9: "*Strongly agree*"). Six items are related to the Social dimension of cohesion while eight items are related to the Task dimension of cohesion.

A.1.1 Items Related to the Social Dimension of Cohesion

1. I did not enjoy socially interacting with the team.
2. I do not want to continue playing with this team.
3. I would rather solve the enigmas on my own than together.
4. We did not have fun during the task.
5. I would like to spend more moments like the previous one with this team.
6. I wish I was on a different team.

A.1.2 Items Related to the Task Dimension of Cohesion

7. I was unhappy with my team's level of desire to win.
8. This team did not give me enough opportunities to use my abilities when we shared the enigmas.
9. Our team was united in trying to solve as many enigmas as possible.
10. We all took responsibility for any loss or poor performance.
11. Our team members had conflicting aspirations for solving the enigmas.
12. If members of our group had problems while trying to resolve a problem, everyone wanted to help them.

13. Our team members did not communicate freely about each member's responsibilities during our task.

14. Our team did not work well together.

A.2 Leadership

We assessed leadership with the following six items on a 6-point Likert ranging from 1 ("*Completely disagree*") to 6 ("*Completely agree*"). We used a round-robin rating, hence, each participant answered each item three times (i.e., the number of group members, including herself).

The following items were retained, with GM a specific group member:

1. GM decided what shall be done and how it will be done.

2. GM assigned group members to particular tasks.

3. GM tried out his ideas in the group.

4. GM took a leadership role in our team.

5. GM provided direction for the team.

6. GM set goals for the team.

A.3 Emotion

Participants answered the following question: "*How do you feel?*". They could pick one label of emotion among the three positive and three negative labels below as well as provide their own label:

- Admiring
- Angry
- Proud
- Ashamed
- Happy
- Frustrated

The labels were chosen according to Roseman (2001)'s Emotion Theory. Among these items, two of them result from an "other-caused" causal attribution (admiration and anger), two from a self-caused causal attribution (pride and shame), and the other two from a circumstances-caused causal attribution (happiness and frustration). We selected these specific labels as they were the most relevant given the context of the game.

Appendix B

Computational Details of Emergent Leaders' Features

Below are the computational details of the emergent leaders' features described in Chapter 6 (see Table 6.3). Every feature is computed from individuals. Thus, we explain the procedure to extract each feature for a person i (i.e., p_i).

B.1 Features Related to Speaking Activity

All of these "Speaking Activity" (SA) features are extracted using the same speech matrix computed in Chapter 4. Thus, we know who is speaking at each point in time.
The following SA features were extracted:

- Total speaking time when at least one other group member is speaking: it consists of summing all the frames in which p_i is speaking while p_j or p_k is speaking (i.e., for a time t, $p_{i,t} = 1$ and $p_{j,t} = 1$ or $p_{k,t} = 1$).
- Total speaking time when no one is speaking: it is computed as in the previous feature, with the condition that for a time t, $p_{i,t} = 1$ and $p_{j,t} = p_{k,t} = 0$.
- Ratio between the total speaking time when at least one other group member is speaking and the total speaking time when no one is speaking.
- Total number of times a person speaks first right after another one: every time the turn of another member is about to finish (i.e., 10 frames before the end of the turn), we check if p_i is the first one to speak for at least 1s. If that is the case, the counter is increased by 1.
- Total time of silence: it is computed by summing all the frames where $p_i = p_j = p_k = 0$.
- Ratio between the total time of speaking for p_i and the total time of silence.
- Total number of speaking turns: we only account for the turns that last at least 2s (i.e., the number of times p_i speaks for at least 2s in the time window).

- Average of speaking turns duration: it is computed by summing the duration of all the p_i's turns and dividing it by the total number of speaking turns of p_i.
- Total number of times being un/successfully interrupted: it is computed every time p_i is speaking. If another group member starts speaking during p_i's turn, for at least 1s, making p_i's stop, then we increment the number of successful interruptions. Otherwise, it is considered as an unsuccessful interruption.
- Total number of times un/successfully interrupting other turns: as for the previous feature, we also quantify the number of times p_i interrupted (successfully or not) the other members.
- Ratio between the total number of time successfully interrupting other turns and the total number of times being successfully interrupted
- Ratio between the total number of time successfully interrupting other turns and the total number of speaking turns
- Ratio between the total number of time unsuccessfully interrupting other turns and the total number of speaking turns

B.2 Features Related to Visual Focus of Attention

All the "Visual Focus of Attention" (VFOA) features are computed from each group member (p_i). These rely on the concept of *Mutual Engagement* (ME) which is approximated as two people looking at each other. Thus, as for the F-formation feature (see Chapter 4), we also computed, for each person, a cone of attention starting from their head. Based on these cones, we extracted a "*facing*" matrix that indicates who looks who, at each point in time. A visual focus occurs when p_i is looking towards another person for at least 0.25s. A ME occurs when the cones of attention of two persons intersect during a visual focus (i.e., for at least 0.25s). Below are the VFOA features we extracted:

- Looking someone with no ME (LnoME): it is the total time at which p_i is looking at another member (for a minimum of 0.25s), without being looked at.
- Being looked at with no ME (BLnoME): it is the total time at which p_i is being looked by at least another member, for at least 0.25s), without looking at any of them.
- ME with any member (ME): it is the total time at which a ME occurs between p_i and any other member.
- Total time being looked at: it is computed by summing all the frames in which p_i is looked at for at least 0.25s. Thus, it results in the sum of BLnoME and ME.
- Number of times one initiates a ME: it is computed by looking at the ME that occur between p_i and another member. Based on the facing matrix, we know who started looking at the other one. If p_i initiated the ME, the counter is increased by 1.
- Ratio between BLnoME and LnoME.

Bibliography

Abrams, A. M. and der Pütten, A. M. (2020). I-c-e framework: Concepts for group dynamics research in human-robot interaction. *International Journal of Social Robotics* 12, 1213–1229

Aigrain, J., Spodenkiewicz, M., Dubuisson, S., Detyniecki, M., Cohen, D., and Chetouani, M. (2018). Multimodal stress detection from multiple assessments. *IEEE Transactions on Affective Computing* 9, 491–506

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine* 18, 91–93

Alameda-Pineda, X., Subramanian, R., Ricci, E., Lanz, O., and Sebe, N. (2017). Salsa: A multimodal dataset for the automated analysis of free-standing social interactions. In *Group and Crowd Behavior for Computer Vision* (Elsevier). 321–340

Albrecht, K. (2006). *Social intelligence: The new science of success* (John Wiley & Sons)

Allen, J., Hung, H., Keyton, J., Murray, G., Oertel, C., and Varni, G. (2021). Insights on group and team dynamics. In *Proceedings of the 23rd International Conference on Multimodal Interaction*. 855–856

Alsulami, A. M. (2021). Towards building group cohesion and learning outcomes based on nonverbal immediacy behavior. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies* 12, 1–12

Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin* 111, 256–274

Andreaggi, G., Robazza, C., and Bortoli, L. (2000). Coesione sociale e sul compito negli sport di squadra: il "group environment questionnaire". *Giornale Italiano di Psicologia dello Sport* 2, 19–23

Aragonés, J. I., Poggio, L., Sevillano, V., Pérez-López, R., and Sánchez-Bernardos, M.-L. (2015). Measuring warmth and competence at inter-group, interpersonal and individual levels / medición de la cordialidad y la competencia en los niveles intergrupal, interindividual e individual. *International Journal of Social Psychology* 30, 407–438

Ashton, N. L., Shaw, M. E., and Worsham, A. P. (1980). Affective reactions to interpersonal distances by friends and strangers. *Bulletin of the Psychonomic Society* 15, 306–308

Axelsson, A.-S. (2002). The digital divide: Status differences in virtual environments. In *The Social Life of Avatars* (Springer). 188–204

Back, K. W. (1951). Influence through social communication. *The Journal of Abnormal and Social Psychology* 46, 9–23

Baird Jr, J. E. (1977). Some nonverbal elements of leadership emergence. *Southern Speech Communication Journal* 42, 352–361

Balaguer, I., Castillo, I., and Duda, J. L. (2003). Interrelationships between motivational climate and cohesion in cadet football. *EduPsykhe* 2, 243–58

Bales, R. F. and Strodtbeck, F. L. (1951). Phases in group problem-solving. *The Journal of Abnormal and Social Psychology* 46, 485–495

Barsade, S. G. and Gibson, D. E. (1998). Group emotion: A view from top and bottom. *Research on Managing Groups and Teams* , 81–102

Barsade, S. G. and Knight, A. P. (2015). Group affect. *Annual Review of Organizational Psychology and Organizational Behavior* 2, 21–46

Bartone, P. T. and Adler, A. B. (1999). Cohesion over time in a peacekeeping medical task force. *Military Psychology* 11, 85–107

Baumeister, R. F. and Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin* 117, 497–529

Beal, D. J., Cohen, R. R., Burke, M. J., and McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology* 88, 989–1004

Bendermacher, N. (2010). Beyond alpha: Lower bounds for the reliability of tests. *Journal of Modern Applied Statistical Methods* 9, 95–102

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 289–300

Bennis, W. G. and Shepard, H. A. (1956). A theory of group development. *Human relations* 9, 415–437

Berg, S., Neubauer, C., Robison, C., Kroninger, C., Schaefer, K. E., and Krausman, A. (2021). Exploring resilience and cohesion in human-autonomy teams: Models and measurement. In *Proceedings of the 12th International Conference on Applied Human Factors and Ergonomics* (Springer), 121–127

Beyan, C., Capozzi, F., Becchio, C., and Murino, V. (2016a). Identification of emergent leaders in a meeting scenario using multiple kernel learning. In *Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction*. 3–10

Beyan, C., Carissimi, N., Capozzi, F., Vascon, S., Bustreo, M., Pierro, A., Becchio, C., and Murino, V. (2016b). Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 317–324

Beyan, C., Katsageorgiou, V.-M., and Murino, V. (2019). A sequential data analysis approach to detect emergent leaders in small groups. *IEEE Transactions on Multimedia* 21, 2107–2116

Biancardi, B., Maisonnave-Couterou, L., Renault, P., Ravenet, B., Mancini, M., and Varni, G. (2020). The wonowa dataset: Investigating the transactive memory system in small group interactions. In *Proceedings of the 22nd International Conference on Multimodal Interaction*. 528–537

Bird, A. M., Foster, C. D., and Maruyama, G. (1980). Convergent and incremental effects of cohesion on attributions for self and team. *Journal of Sport and Exercise Psychology* 2, 181–194

Birdwhistell, R. L. (2010). *Kinesics and context* (University of Pennsylvania press)

Birmingham, C. and McCord, M. (2002). Group process research: Implications for using learning groups. *Team based learning: A transformative use of small groups* , 77–97

Bliese, P. D. and Halverson, R. R. (1996). Individual and nomothetic models of job stress: An examination of work hours, cohesion, and well-being 1. *Journal of Applied Social Psychology* 26, 1171–1189

Boersma, P. and Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot international* 5, 341–345

Bollen, K. A. and Hoyle, R. H. (1990). Perceived cohesion: A conceptual and empirical examination. *Social Forces* 69, 479–504

Bonillo, C., Romão, T., and Cerezo, E. (2019). Persuasive games in interactive spaces: The hidden treasure game. In *Proceedings of the 20th International Conference on Human Computer Interaction*. 1–8

Boone, K. S., Beitel, P., and Kuhlman, J. S. (1997). The effects of the win/loss record on cohesion. *Journal of Sport Behavior* 20, 125–134

Borja, L. F., Azorin-Lopez, J., and Saval-Calvo, M. (2017). A compilation of methods and datasets for group and crowd action recognition. *International Journal of Computer Vision and Image Processing (IJCVIP)* 7, 40–53

Boyd, M., Kim, M.-S., Ensari, N., and Yin, Z. (2014). Perceived motivational team climate in relation to task and social cohesion among male college athletes. *Journal of Applied Social Psychology* 44, 115–123

Braun, M. T., Kozlowski, S. W., and Kuljanin, G. (2021). Multilevel theory, methods, and analyses in management. In *Oxford Research Encyclopedia of Business and Management* (Oxford University Press)

Brunet, P. M., Cowie, R., Heylen, D., Nijholt, A., and Schröder, M. (2012). Conceptual frameworks for multimodal social signal processing. *Journal on multimodal user interfaces* 6, 95–99

Burgoon, J. K., Magnenat-Thalmann, N., Pantic, M., and Vinciarelli, A. (2017). *Social signal processing* (Cambridge University Press)

Buton, F., Fontayne, P., and Heuzé, J.-P. (2006). La cohésion des groupes sportifs: évolutions conceptuelles, mesures et relations avec la performance. *Movement Sport Sciences* , 9–45

Cabrera-Quiros, L., Demetriou, A., Gedik, E., Meij, L. v. d., and Hung, H. (2021). The matchnmingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing* 12, 113–130

Calbris, G. (2011). *Elements of meaning in gesture*, vol. 5 (John Benjamins Publishing)

Callow, N., Smith, M. J., Hardy, L., Arthur, C. A., and Hardy, J. (2009). Measurement of transformational leadership and its relationship with team cohesion and performance level. *Journal of applied sport psychology* 21, 395–412

Calvo, R. A., D'Mello, S., Gratch, J. M., and Kappas, A. (2015). *The Oxford handbook of affective computing* (Oxford Library of Psychology)

Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. (2017). Affective computing and sentiment analysis. In *A practical guide to sentiment analysis* (Springer). 1–10

Campbell, N. (2008). Multimodal processing of discourse information; the effect of synchrony. In *Proceedings of the 2nd International Symposium on Universal Communication* (IEEE), 12–15

Carless, S. A. and De Paola, C. (2000). The measurement of cohesion in work teams. *Small Group Research* 31, 71–88

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, eds. S. Renals and S. Bengio (Springer), 28–39

Carmeli, C., Knyazeva, M. G., Innocenti, G. M., and De Feo, O. (2005). Assessment of eeg synchronization based on state-space analysis. *Neuroimage* 25, 339–354

Carron, A. V. and Brawley, L. R. (2000). Cohesion: Conceptual and measurement issues. *Small Group Research* 31, 89–106

Carron, A. V., Colman, M. M., Wheeler, J., and Stevens, D. (2002). Cohesion and performance in sport: A meta analysis. *Journal of sport and exercise psychology* 24, 168–188

Carron, A. V., Widmeyer, W. N., and Brawley, L. R. (1985). The development of an instrument to assess cohesion in sport teams: The group environment questionnaire. *Journal of Sport Psychology* 7, 244–266

Casey-Campbell, M. and Martens, M. L. (2009). Sticking it all together: A critical assessment of the group cohesion-performance literature. *International Journal of Management Reviews* 11, 223–246

Ceccaldi, E., Lehmann-Willenbrock, N., Volta, E., Chetouani, M., Volpe, G., and Varni, G. (2019). How unitizing affects annotation of cohesion. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–7

Čereković, A. (2014). An insight into multimodal databases for social signal processing: acquisition, efforts, and directions. *Artificial Intelligence Review* 42, 663–692

Chao, L., Tao, J., Yang, M., Li, Y., and Wen, Z. (2015). Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. 65–72

Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T. X., Tu, J., Huang, Z., Harper, M., Quek, F., Xiong, Y., McNeill, D., Tuttle, R., and Huang, T. (2006). Vace multimodal meeting corpus. In *Proceedings of the 2nd Machine Learning for Multimodal Interaction. Lecture Notes in Computer Science*, eds. S. Renals and S. Bengio (Springer), vol. 3869, 40–51

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment* 6, 284–290

Coan, J. A. and Gonzalez, M. Z. (2015). *Emotions as emergent variables.* (The Guilford Press), chap. 9. 209–225

Codish, D. and Ravid, G. (2014). Personality based gamification-educational gamification for extroverts and introverts. In *Proceedings of the 9th CHAIS Conference for the Study of Innovation and Learning Technologies: Learning in the Technological Era*. 36–44

Colas, C., Sigaud, O., and Oudeyer, P. (2018). How many random seeds? statistical power analysis in deep reinforcement learning experiments. *CoRR* abs/1806.08295

Collins, S. H., Adamczyk, P. G., and Kuo, A. D. (2009). Dynamic arm swinging in human walking. *Proceedings of the Royal Society B: Biological Sciences* 276, 3679–3688

Converse, S., Cannon-Bowers, J., and Salas, E. (1993). Shared mental models in expert team decision making. *Individual and group decision making: Current issues* 221, 221–246

Cota, A. A., Evans, C. R., Dion, K. L., Kilik, L., and Longman, R. S. (1995). The structure of group cohesion. *Personality and social psychology bulletin* 21, 572–580

Coultas, C. W., Driskell, T., Shawn Burke, C., and Salas, E. (2014). A conceptual review of emergent state measurement. *Small Group Research* 45, 671–703

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine* 18, 32–80

Cronin, M. A., Weingart, L. R., and Todorova, G. (2011). Dynamics in groups: Are we there yet? *Academy of Management Annals* 5, 571–612

Curşeu, P. L. (2006). Emergent states in virtual teams: a complex adaptive systems perspective. *Journal of Information Technology* 21, 249–261

Darioly, A. and Mast, M. S. (2014). The role of nonverbal behavior in leadership: An integrative review. *R. E. Riggio & S. J. Tan (Eds.)* Leader interpersonal and influence skills: The soft skills of leadership, 73–100

De Souza, G. and Klein, H. J. (1995). Emergent leadership in the group goal-setting process. *Small group research* 26, 475–496

DeChurch, L. A. and Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: a meta-analysis. *Journal of applied psychology* 95, 32–53

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing* 3, 349–365

Dhall, A. (2019). Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*. 546–550

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., and Gedeon, T. (2017). From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 524–528

Dion, K. L. (2000). Group cohesion: From field of forces to multidimensional construct. *Group Dynamics: Theory, Research, and Practice* 4, 7–26

Dionne, S. D., Yammarino, F. J., Atwater, L. E., and Spangler, W. D. (2004). Transformational leadership and team performance. *Journal of organizational change management* 17, 177–193

Doyran, M., Schimmel, A., Baki, P., Ergin, K., Türkmen, B., Salah, A. A., Bakkes, S. C., Kaya, H., Poppe, R., and Salah, A. A. (2021). Mumbai: multi-person, multimodal board game affect and interaction analysis dataset. *Journal on Multimodal User Interfaces* 15, 373–391

Dulebohn, J. H. and Hoch, J. E. (2017). Virtual teams in organizations. *Human Resource Management Review* 27, 569–574

Dunleavy, M., Dede, C., and Mitchell, R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of science Education and Technology* 18, 7–22

Edmondson, A. C. and Lei, Z. (2014). Psychological safety: The history, renaissance, and future of an interpersonal construct. *Annual review of organizational psychology and organizational behavior* 1, 23–43

Ekman, P., Davidson, R. J., and Friesen, W. V. (1990). The duchenne smile: Emotional expression and brain physiology: Ii. *Journal of personality and social psychology* 58, 342–353

Ens, B., Lanir, J., Tang, A., Bateman, S., Lee, G., Piumsomboon, T., and Billinghurst, M. (2019). Revisiting collaboration through mixed reality: The evolution of groupware. *International Journal of Human-Computer Studies* 131, 81–98

Estabrooks, P. A. and Carron, A. V. (2000a). The physical activity group environment questionnaire: An instrument for the assessment of cohesion in exercise classes. *Group Dynamics* 4, 230–243

Estabrooks, P. A. and Carron, A. V. (2000b). Predicting scheduling self-efficacy in older adult exercisers: The role of task cohesion. *Journal of Aging and Physical Activity* 8, 41–50

Evans, C. R. and Dion, K. L. (2012). Group cohesion and performance: A meta-analysis. *Small Group Research* 43, 690–701

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing* 7, 190–202

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. 835–838

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods* 4, 272–299

Fang, S. and Achard, C. (2018). Estimation of Cohesion with Feature Categorization on Small Scale Groups. In *Proceedings of WACAI* (Ile de Porquerolles, France)

Fay, M. P. and Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software* 36, 1–34

Feese, S., Arnrich, B., Tröster, G., Meyer, B., and Jonas, K. (2012). Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion. In *Proceedings of International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*. 520–525

Festinger, L., Schachter, S., and Back, K. (1950). *Social pressures in informal groups* (Harper)

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics* (Springer). 66–70

Fiske, S., Cuddy, A., and Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences* 11, 77–83

Forsyth, D. (2012). *Group Dynamics* (Wadsworth Publishing), 6 edn.

Fox, L. D., Rejeski, W. J., and Gauvin, L. (2000). Effects of leadership style and group dynamics on enjoyment of physical activity. *American Journal of Health Promotion* 14, 277–283

Freedman, G. and Flanagan, M. (2017). From dictators to avatars: Furthering social and personality psychology through game methods. *Social and personality psychology compass* 11, e12368

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 119–139

Friedrich, T. L., Vessey, W. B., Schuelke, M. J., Ruark, G. A., and Mumford, M. D. (2009). A framework for understanding collective leadership: The selective utilization of leader and team expertise within networks. *The Leadership Quarterly* 20, 933–958

García-Calvo, T., Leo, F. M., Gonzalez-Ponce, I., Sánchez-Miguel, P. A., Mouratidis, A., and Ntoumanis, N. (2014). Perceived coach-created and peer-created motivational climates and their associations with team cohesion and athlete satisfaction: Evidence from a longitudinal study. *Journal of sports sciences* 32, 1738–1750

Gatica-Perez, D., McCowan, I., Zhang, D., and Bengio, S. (2005). Detecting group interest-level in meetings. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. vol. 1, 489–492

Gavrikov, I. and Savchenko, A. V. (2020). Efficient group-based cohesion prediction in images using facial descriptors. In *Proceedings of the 9th International Conference on Analysis of Images, Social Networks and Texts* (Springer), 140–148

George, D. and Mallery, P. (2016). *SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 update , 2003* (Boston: Allyn & Bacon)

Gerpott, F. H., Lehmann-Willenbrock, N., Silvis, J. D., and Van Vugt, M. (2018). In the eye of the beholder? an eye-tracking experiment on emergent leadership in team interactions. *The Leadership Quarterly* 29, 523–532

Gerpott, F. H., Lehmann-Willenbrock, N., Voelpel, S. C., and Van Vugt, M. (2019). It's not just what is said, but when it's said: A temporal account of verbal behaviors and emergent leadership in self-managed teams. *Academy of Management Journal* 62, 717–738

Ghosh, S., Dhall, A., Sebe, N., and Gedeon, T. (2022). Automatic prediction of group cohesiveness in images. *IEEE Transactions on Affective Computing* 13, 1677–1690

Gibson, D. E. (1997). The struggle for reason: The sociology of emotions in organizations. *Social perspectives on emotion* 4, 211–256

Gilbert, N. (2004). Agent-based social simulation: dealing with complexity. *The Complex Systems Network of Excellence* 9, 1–14

[Python package] Gillies, S. et al. (2007). *Shapely: manipulation and analysis of geometric objects*

Glenn, P. (2003). *Laughter in interaction*, vol. 18 (Cambridge University Press)

Goldin-Meadow, S. and Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual review of psychology* 64, 257–283

Goleman, D. (2006). *Social intelligence* (Hutchinson)

Gonzales, A. L., Hancock, J. T., and Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research* 37, 3–19

Gordon, I., Gilboa, A., Cohen, S., Milstein, N., Haimovich, N., Pinhasi, S., and Siegman, S. (2020). Physiological and behavioral synchrony predict group cohesion and performance. *Scientific Reports* 10, 1–12

Goudbeek, M. and Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America* 128, 1322–1336

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval* (Springer), 345–359

Griffith, J. (1988). Measurement of group cohesion in u.s. army units. *Basic and Applied Social Psychology* 9, 149–171

Grossman, R., Friedman, S. B., and Kalra, S. (2017). Teamwork processes and emergent states. *The Wiley Blackwell handbook of the psychology of team working and collaborative processes* 42, 243–269

Grossman, R., Rosch, Z., Mazer, D., and Salas, E. (2015). What matters for team cohesion measurement? A Synthesis. *Research on Managing Groups and Teams* 17, 147–180

Gully, S. M., Devine, D. J., and Whitney, D. J. (2012). A meta-analysis of cohesion and performance: Effects of level of analysis and task interdependence. *Small Group Research* 43, 702–725

Gunes, H. and Hung, H. (2015). Emotional and social signals: A neglected frontier in multimedia computing? *IEEE MultiMedia* 22, 76–85

Guo, D., Wang, K., Yang, J., Zhang, K., Peng, X., and Qiao, Y. (2019). Exploring regularizations with face, body and image cues for group cohesion prediction. In *Proceedings of the 21st International Conference on Multimodal Interaction*. 557–561

Hackman, J. R. and Morris, C. G. (1975). Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. *Advances in experimental social psychology* 8, 45–99

Hagstrom, W. O. and Selvin, H. C. (1965). Two dimensions of cohesiveness in small groups. *Sociometry* , 30–43

Hair, J., Black, W., Babin, B., Anderson, R., and Tatham, R. (2010). Multivariate data analysis. 6th (ed.) prentice-hall. *Upper Saddle River NJ*

Hall, E. T. (1966). *The hidden dimension* (Garden City, NY: Doubleday)

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* 8, 23–34

Hamari, J., Koivisto, J., and Sarsa, H. (2014). Does gamification work?-a literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii international conference on system sciences* (IEEE), 3025–3034

Hammerla, N. Y., Halloran, S., and Ploetz, T. (2016). Hdeep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 1533–1540

Hanna, A. A., Smith, T. A., Kirkman, B. L., and Griffin, R. W. (2021). The emergence of emergent leadership: a comprehensive framework and directions for future research. *Journal of Management* 47, 76–104

Hans, A. and Hans, E. (2015). Kinesics, haptics and proxemics: Aspects of non-verbal communication. *IOSR Journal of Humanities and Social Science* 20, 47–52

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Springer)

Hermes, D. J. (1988). Measurement of pitch by subharmonic summation. *The journal of the acoustical society of America* 83, 257–264

Hersey, P. and Blanchard, K. H. (1969). *Management of organizational behavior: Utilizing human resources* (Academy of Management Briarcliff Manor, NY 10510)

Heuzé, J.-P. and Fontayne, P. (2002). Questionnaire sur l'ambiance du groupe: A french-language instrument for measuring group cohesion. *Journal of Sport and Exercise Psychology* 24, 42–67

Hilton, K. (2016). The perception of overlapping speech: Effects of speaker prosody and listener attitudes. In *Proceedings of the Interspeech 2016*. 1260–1264

Hoch, J. E. and Kozlowski, S. W. (2014). Leading virtual teams: Hierarchical leadership, structural supports, and shared team leadership. *Journal of applied psychology* 99, 390–403

Hogg, M. A. and Hardie, E. A. (1991). Social attraction, personal attraction, and self-categorization-, a field study. *Personality and Social Psychology Bulletin* 17, 175–180

Hughes, J. F., Van Dam, A., McGuire, M., Foley, J. D., Sklar, D., Feiner, S. K., and Akeley, K. (2014). *Computer graphics: principles and practice* (Pearson Education)

Hung, H. and Chittaranjan, G. (2010). The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *Proceedings of the 18th ACM international conference on Multimedia*. 879–882

Hung, H. and Gatica-Perez, D. (2010). Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia* 12, 563–575

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., and Jundt, D. (2005). Teams in organizations: From input-process-output models to imoi models. *Annual Revision Psychologist* 56, 517–543

Jackson, P. H. and Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika* 42, 567–578

Jayagopi, D. B., Hung, H., Yeo, C., and Gatica-Perez, D. (2009). Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 501–513

John, O. P. (1990). The "big five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. *Handbook of personality: Theory and research* , 66–100

Jokinen, K., Furukawa, H., Nishida, M., and Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 1–30

Jones, S. E. and LeBaron, C. D. (2002). Research on the relationship between verbal and nonverbal communication: Emerging integrations. *Journal of communication* 52, 499–521

Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2019). Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 190–204

Kalma, A. P., Visser, L., and Peeters, A. (1993). Sociable and aggressive dominance: Personality differences in leadership style? *The Leadership Quarterly* 4, 45–64

Kamper, S. J., Ostelo, R. W., Knol, D. L., Maher, C. G., de Vet, H. C., and Hancock, M. J. (2010). Global perceived effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *Journal of clinical epidemiology* 63, 760–766

Kantharaju, R. B., Langlet, C., Barange, M., Clavel, C., and Pelachaud, C. (2020). Multimodal analysis of cohesion in multi-party interactions. In *Proceedings of The 12th Language Resources and Evaluation Conference* (European Language Resources Association), 498–507

Kantharaju, R. B. and Pelachaud, C. (2020). Analysis of Laughter in Cohesive Groups. In *Proceedings of the Laughter and Other Non-Verbal Vocalisations Workshop* (Bielefeld, Germany), 74–76

Kantharaju, R. B. and Pelachaud, C. (2021). Social signals of cohesion in multi-party interactions. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 9–16

Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., and Driessen, P. F. (2005). Gesture-based affective computing on motion capture data. In *International conference on affective computing and intelligent interaction* (Springer), 1–7

Kasparova, A., Celiktutan, O., and Cukurova, M. (2020). Inferring student engagement in collaborative problem solving from visual cues. In *Companion Publication of the of the 22nd International Conference on Multimodal Interaction*. 177–181

Katz, D. (1960). The functional approach to the study of attitudes. *Public Opinion Quarterly* 24, 163–204

Kauffeld, S., Lehmann-Willenbrock, N., and Meinecke, A. L. (2018). *The Advanced Interaction Analysis for Teams (act4teams) Coding Scheme* (Cambridge University Press), chap. 21. Cambridge Handbooks in Psychology. 422–431

Keller, R. T. (1986). Predictors of the performance of project groups in R&D organizations. *Academy of Management Journal* 29, 715–726

Kendon, A. (1990). Spatial organization in social encounters: The f-formation system. *Conducting interaction: Patterns of behavior in focused encounters*

Kendon, A. (2010). Spacing and orientation in co-present interaction. In *Development of multimodal interfaces: Active listening and synchrony* (Springer). 1–15

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis* 9, 137–163

Klein, K. J. and Kozlowski, S. W. (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions.* (Jossey-Bass)

Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal communication in human interaction* (Cengage Learning)

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* 4, 83–91

Koopmans, L., Bernaards, C., Hildebrandt, V., van Buuren, S., Van der Beek, A. J., and de Vet, H. C. (2013). Development of an individual work performance questionnaire. *International journal of productivity and performance management* 62, 6–28

Koutsombogera, M. and Vogel, C. (2018). Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2945–2951

Kozlowski, S. W. (2015). Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review* 5, 270–299

Kozlowski, S. W. and Chao, G. T. (2018). Unpacking team process dynamics and emergent phenomena: Challenges, conceptual advances, and innovative methods. *American Psychologist* 73, 576–592

Kozlowski, S. W., Gully, S. M., Nason, E. R., and Smith, E. M. (1999). Developing adaptive teams: A theory of compilation and performance across levels and time. *The changing nature of work performance: Implications for staffing, personnel actions, and development* , 240–292

Kozlowski, S. W. J. and Chao, G. T. (2012). The dynamics of emergence: Cognition and cohesion in work teams. *Managerial and Decision Economics* 33, 335–354

Kozlowski, S. W. J. and Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest* 7, 77–124

Kozub, S. A. and McDonnell, J. F. (2000). Exploring the relationship between cohesion and collective efficacy in rugby teams. *Journal of sport behavior* 23, 120–129

Kubasova, U., Murray, G., and Braley, M. (2019). Analyzing verbal and nonverbal features for predicting group performance. In *Proceedings of the 46th Interspeech* (ISCA), 1896–1900

Lai, C. and Murray, G. (2018). Predicting group satisfaction in meeting discussions. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*. 1–8

Lakhmani, S. G., Neubauer, C., Krausman, A., Fitzhugh, S. M., Berg, S. K., Wright, J. L., Rovira, E., Blackman, J. J., and Schaefer, K. E. (2022). Cohesion in human-autonomy teams: an approach for future research. *Theoretical Issues in Ergonomics Science* , 1–38

Lakin, J. L. and Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological science* 14, 334–339

Lanaj, K. and Hollenbeck, J. R. (2015). Leadership over-emergence in self-managing teams: The role of gender and countervailing biases. *Academy of Management Journal* 58, 1476–1494

Lawler, E. J., Thye, S. R., and Yoon, J. (2000). Emotion and group cohesion in productive exchange. *American Journal of Sociology* 106, 616–657

Lawler, E. J. and Yoon, J. (1996). Commitment in exchange relations: Test of a theory of relational cohesion. *American sociological review* , 89–108

Le Bon, G. (1897). *The crowd: A study of the popular mind* (TF Unwin)

Lehmann-Willenbrock, N. and Allen, J. A. (2018). Modeling temporal interaction dynamics in organizational settings. *Journal of business and psychology* 33, 325–344

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., and Saul, J. R. (2008). A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel psychology* 61, 273–307

Levi, D. (2001). *Group dynamics for teams* (Thousand Oaks, CA: Sage)

Levinson, S. C. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology* 6, 731

Lewin, K. (1939). Field theory and experiment in social psychology: Concepts and methods. *American Journal of Sociology* 44, 868–896

Lewin, K. (1951). *Field theory in social science: selected theoretical papers* (Harpers)

Light Shields, D. L., Gardner, D. E., Light Bredemeier, B. J., and Bostro, A. (1997). The relationship between leadership behaviors and group cohesion in team sports. *The Journal of Psychology* 131, 196–210

Liu, B., Yu, X., Zhang, P., Yu, A., Fu, Q., and Wei, X. (2017). Supervised deep feature extraction for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* 56, 1909–1921

López-Zafra, E., Garcia-Retamero, R., and Landa, J. M. A. (2008). The role of transformational leadership, emotional intelligence, and group cohesiveness on leadership emergence. *Journal of Leadership Studies* 2, 37–49

Lord, R. G., Binning, J. F., Rush, M. C., and Thomas, J. C. (1978). The effect of performance cues and leader behavior on questionnaire ratings of leadership behavior. *Organizational Behavior and Human Performance* 21, 27–39

Lott, A. J. and Lott, B. E. (1965). Group cohesiveness as interpersonal attraction: A review of relationships with antecedent and consequent variables. *Psychological Bulletin* 64, 259–309

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the Advances in neural information processing systems*. 4765–4774

Magee, J. C. and Tiedens, L. Z. (2006). Emotional ties that bind: The roles of valence and consistency of group emotion in inferences of cohesiveness and common fate. *Personality and Social Psychology Bulletin* 32, 1703–1715

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE* 63, 561–580

Maman, L. (2020). Multimodal groups' analysis for automated cohesion estimation. In *Proceedings of the 22nd International Conference on Multimodal Interaction*. 713–717

Maman, L., Ceccaldi, E., Lehmann-Willenbrock, N., Likforman-Sulem, L., Chetouani, M., Volpe, G., and Varni, G. (2020). Game-on: A multimodal dataset for cohesion and group analysis. *IEEE Access* 8, 124185–124203

Maman, L., Chetouani, M., Likforman-Sulem, L., and Varni, G. (2021a). Using valence emotion to predict group cohesion's dynamics: Top-down and bottom-up approaches. In *Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8

Maman, L., Likforman-Sulem, L., Chetouani, M., and Varni, G. (2021b). Exploiting the interplay between social and task dimensions of cohesion to predict its dynamics leveraging social sciences. In *Proceedings of the 23rd International Conference on Multimodal Interaction*. 16–24

Maman, L. and Varni, G. (2020). GRACE : Un projet portant sur l'étude automatique de la cohésion dans les petits groupes d'humains. In *Proceedings of WACAI* (Ile d'Oléron, France)

Marks, M. A., Mathieu, J. E., and Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *The Academy of Management Review* 26, 356–376

Maslow, A. H. (1943). A theory of human motivation. *Psychological review* 50, 370–396

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

Maynard, M. T., Kennedy, D. M., Sommer, S. A., and Passos, A. M. (2015). Team cohesion: A theoretical consideration of its reciprocal relationships within the team adaptation nomological network. In *Team cohesion: Advances in psychological theory, methods and practice* (Emerald Group Publishing Limited), vol. 17. 83–111

McAuley, E., Duncan, T., and Tammen, V. V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport* 60, 48–58

McClean, E. J., Martin, S. R., Emich, K. J., and Woodruff, C. T. (2018). The social consequences of voice: An examination of voice type and gender on status and subsequent leader emergence. *Academy of Management Journal* 61, 1869–1891

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

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods* 23, 412–433

Menesini, E., Tassi, F., and Nocentini, A. (2018). The competitive attitude scale (CAS): a multidimensional measure of competitiveness in adolescence. *Journal of Psychology & Clinical Psychiatry* 9, 240–244

Michalisin, M. D., Karau, S. J., and Tangpong, C. (2004). Top management team cohesion and superior industry returns: An empirical study of the resource-based view. *Group & Organization Management* 29, 125–140

Miranda Correa, J. A., Abadi, M. K., Sebe, N., and Patras, I. (2018). Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* , 479–493

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edn. Https://christophm.github.io/interpretable-ml-book

Moreno, J. L. (1934). *Who shall survive?: A new approach to the problem of human interrelations.* (Nervous and Mental Disease Publishing Co)

Morgeson, F. P., DeRue, D. S., and Karam, E. P. (2010). Leadership in teams: A functional approach to understanding leadership structures and processes. *Journal of Management* 36, 5–39

Moustafa, F. and Steed, A. (2018). A longitudinal study of small group interaction in social virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*. 1–10

Mullen, B. and Copper, C. (1994). The relation between group cohesiveness and performance: An integration. *Psychological Bulletin* 115, 210–227

Müller, P., Huang, M. X., and Bulling, A. (2018). Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*. 153–164

Muller, P. M. and Bulling, A. (2019). Emergent leadership detection across datasets. In *Proceedings of the 21st International Conference on Multimodal Interaction*. 274–278

Murray, G. and Oertel, C. (2018). Predicting group performance in task-based interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 14–20

Nanninga, M. C., Zhang, Y., Lehmann-Willenbrock, N., Szlávik, Z., and Hung, H. (2017). Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 206–215

Niewiadomski, R., Mancini, M., Baur, T., Varni, G., Griffin, H., and Aung, M. S. (2013). Mmli: Multimodal multiperson corpus of laughter in interaction. In *Proceedings of the International Workshop on Human Behavior Understanding* (Springer), 184–195

Nilsson, R., Pena, J. M., Björkegren, J., and Tegnér, J. (2007). Consistent feature selection for pattern recognition in polynomial time. *The Journal of Machine Learning Research* 8, 589–612

Nov, O. and Arazy, O. (2013). Personality-targeted design: theory, experimental procedure, and preliminary results. In *Proceedings of the 16th conference on Computer supported cooperative work*. 977–984

Okada, S., Aran, O., and Gatica-Perez, D. (2015). Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 17th ACM on International Conference on Multimodal Interaction*. 15–22

Pantic, M., Cowie, R., D'Errico, F., Heylen, D., Mehu, M., Pelachaud, C., Poggi, I., Schroeder, M., and Vinciarelli, A. (2011). Social signal processing: the research agenda. In *Visual analysis of humans* (Springer). 511–538

Pantic, M., Nijholt, A., Pentland, A., and Huanag, T. S. (2008). Human-centred intelligent human? computer interaction (hci^2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems* 1, 168–187

Pantic, M., Pentland, A., Nijholt, A., and Huang, T. S. (2007). Human computing and machine understanding of human behavior: A survey. In *Artifical intelligence for human computing* (Springer). 47–71

Pantic, M. and Vinciarelli, A. (2014). Social signal processing. *The Oxford handbook of affective computing* , 84–93

Parthasarathy, S. and Busso, C. (2017). Jointly predicting arousal, valence and dominance with multi-task learning. In *Interspeech*. vol. 2017, 1103–1107

Paskevich, D. M., Brawley, L. R., Dorsch, K. D., and Widmeyer, W. N. (1999). Relationship between collective efficacy and team cohesion: Conceptual and measurement issues. *Group Dynamics: Theory, Research, and Practice* 3, 210–222

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

Pentland, A. (2007). Social signal processing [exploratory dsp]. *IEEE Signal Processing Magazine* 24, 108–111

Peters, G.-J. (2014). The alpha and the omega of scale reliability and validity: why and how to abandon cronbach's alpha. *European Health Psychologist* 16, 56–69

Piana, S., Mancini, M., Camurri, A., Varni, G., and Volpe, G. (2013). Automated analysis of non-verbal expressive gesture. In *Human Aspects in Ambient Intelligence* (Springer). 41–54

Picard, R. W. (1999). Affective computing for hci. In *HCI (1)* (Citeseer), 829–833

Picard, R. W. (2000). *Affective computing* (MIT press)

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies* 59, 55–64

Picazo, C., Gamero, N., Zornoza, A., and Peiró, J. M. (2015). Testing relations between group cohesion and satisfaction in project teams: A cross-level and cross-lagged approach. *European Journal of Work and Organizational Psychology* 24, 297–307

Provine, R. R. (1993). Laughter punctuates speech: Linguistic, social and gender contexts of laughter. *Ethology* 95, 291–298

Rapp, T., Maynard, T., Domingo, M., and Klock, E. (2021). Team emergent states: What has emerged in the literature over 20 years. *Small Group Research* 52, 68–102

Redcay, E. and Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nature Reviews Neuroscience* 20, 495–505

Revelle, W. and Zinbarg, R. (2009). Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika* 74, 145

Ringeval, F., Marchi, E., Grossard, C., Xavier, J., Chetouani, M., Cohen, D., and Schuller, B. (2016). Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association* (ISCA), 1210–1214

Roseman, I. and Smith, C. (2001). Appraisal theory. *Appraisal processes in emotion: Theory, methods, research* , 3–19

Roseman, I. J. (2001). A model of appraisal in the emotion system. *Appraisal processes in emotion: Theory, methods, research* , 68–91

Roseman, I. J. (2013). Appraisal in the emotion system: Coherence in strategies for coping. *Emotion Review* 5, 141–149

Rosh, L., Offermann, L. R., and Van Diest, R. (2012). Too close for comfort? Distinguishing between team intimacy and team cohesion. *Human Resource Management Review* 22, 116–127

Runkel, P. J., Lawrence, M., Oldfield, S., Rider, M., and Clark, C. (1971). Stages of group development: An empirical test of tuckman's hypothesis. *The Journal of Applied Behavioral Science* 7, 180–193

Ryokai, K., Durán López, E., Howell, N., Gillick, J., and Bamman, D. (2018). Capturing, representing, and interacting with laughter. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–12

Saarinen, A., Harjunen, V., Jasinskaja-Lahti, I., Jääskeläinen, I. P., and Ravaja, N. (2021). Social touch experience in different contexts: A review. *Neuroscience & Biobehavioral Reviews* 131, 360–372

Sabry, S., Maman, L., and Varni, G. (2021). An exploratory computational study on the effect of emergent leadership on social and task cohesion. In *Companion Publication of the 23rd International Conference on Multimodal Interaction*. 263–272

Sahi, R. S., Dieffenbach, M. C., Gan, S., Lee, M., Hazlett, L. I., Burns, S. M., Lieberman, M. D., Shamay-Tsoory, S. G., and Eisenberger, N. I. (2021). The comfort in touch: Immediate and lasting effects of handholding on emotional pain. *PloS one* 16, 1–15

Salah, A. A., Pantic, M., and Vinciarelli, A. (2011). Recent developments in social signal processing. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* (IEEE), 380–385

Salas, E., Grossman, R., Hughes, A. M., and Coultas, C. W. (2015). Measuring team cohesion: Observations from the science. *Human Factors* 57, 365–374

Salinäs, E.-L. (2002). Collaboration in multi-modal virtual worlds: Comparing touch, text, voice and video. In *The social life of avatars* (Springer). 172–187

Sanchez-Cortes, D., Aran, O., and Gatica-Perez, D. (2011a). An audio visual corpus for emergent leader analysis. In *Proceedings of theWorkshop on multimodal corpora for machine learning: taking stock and road mapping the future, ICMI-MLMI* (Citeseer), 1–4

Sanchez-Cortes, D., Aran, O., Mast, M. S., and Gatica-Perez, D. (2010). Identifying emergent leadership in small groups using nonverbal communicative cues. In *Proceedings of the International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*. 1–4

Sanchez-Cortes, D., Aran, O., Mast, M. S., and Gatica-Perez, D. (2011b). A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia* 14, 816–832

Santamaria-Granados, L., Munoz-Organero, M., Ramirez-Gonzalez, G., Abdulhay, E., and Arunkumar, N. (2018). Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos). *IEEE Access* 7, 57–67

Sarris, A. and Kirby, N. (2005). Antarctica: A study of person-culture fit. *Australian Journal of Psychology* 57, 161–169

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* 36, 1627–1639

Schachter, S., Ellertson, N., McBride, D., and Gregory, D. (1951). An experimental study of cohesiveness and productivity. *Human Relations* 4, 229–238

Scherer, K. R. (2009). Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 3459–3474

Scherer, S., Weibel, N., Morency, L.-P., and Oviatt, S. (2012). Multimodal prediction of expertise and leadership in learning groups. In *Proceedings of the 1st International Workshop on Multimodal Learning Analytics*. 1–8

Schneider, B. (1990). The climate for service: An application of the climate construct. *Organizational climate and culture* 1, 383–412

Schneider, H. W. and Mcdougall, W. (1921). The group mind. *Journal of Philosophy* 18, 690–697

Schnur, T. T., Costa, A., and Caramazza, A. (2006). Planning at the phonological level during sentence production. *Journal of psycholinguistic research* 35, 189–213

Schroeder, R. (2002). Social interaction in virtual environments: Key issues, common themes, and a framework for research. In *The social life of avatars* (Springer). 1–18

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*. 148–152

Schuller, B. W. (2013). *Intelligent audio analysis* (Springer)

Seers, A. (1989). Team-member exchange quality: A new construct for role-making research. *Organizational behavior and human decision processes* 43, 118–135

Severt, J. B. and Estrada, A. X. (2015). On the function and structure of group cohesion. In *Team Cohesion: Advances in Psychological Theory, Methods and Practice* (Emerald Group Publishing Limited), vol. 17. 3–24

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games* 2, 307–317

Sharma, G., Ghosh, S., and Dhall, A. (2019). Automatic group level affect and cohesion prediction in videos. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (IEEE), 161–167

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 420–428

Siebold, G. L. (2006). Military group cohesion. *Military life: The psychology of serving in peace and combat* 1, 185–201

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika* 74, 107–120

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics* 19, 279–281

Smith, M. J., Arthur, C. A., Hardy, J., Callow, N., and Williams, D. (2013). Transformational leadership and task cohesion in sport: The mediating role of intrateam communication. *Psychology of sport and exercise* 14, 249–257

Soroush, M. Z., Maghooli, K., Setarehdan, S. K., and Nasrabadi, A. M. (2017). A review on eeg signals based emotion recognition. *International Clinical Neuroscience Journal* 4, 118–129

Spink, K. S. (1990). Group cohesion and collective efficacy of volleyball teams. *Journal of Sport and Exercise Psychology* 12, 301–311

Spisak, B. R., O'Brien, M. J., Nicholson, N., and van Vugt, M. (2015). Niche construction and the evolution of leadership. *Academy of Management Review* 40, 291–306

Stashevsky, S. and Koslowsky, M. (2006). Leadership team cohesiveness and team performance. *International Journal of Manpower* 27, 63–74

Stein, R. T. and Heller, T. (1979). An empirical analysis of the correlations between leadership status and participation rates reported in the literature. *Journal of Personality and Social Psychology* 37, 1993–2002

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826

Taggar, S., Hackew, R., and Saha, S. (1999). Leadership emergence in autonomous work teams: Antecedents and outcomes. *Personnel Psychology* 52, 899–926

Tahon, M. and Devillers, L. (2010). Acoustic measures characterizing anger across corpora collected in artificial or natural context. In *Proceedings of Speech Prosody*

Tamarit, L., Goudbeek, M., and Scherer, K. (2008). Spectral slope measurements in emotionally expressive speech. *Proceedings of Speech Analysis and Processing for Knowledge Discovery* , 169–183

Tannen, D. (1994). *Gender and discourse* (Oxford University Press)

Tao, J. and Tan, T. (2005). Affective computing: A review. In *Proceedings of International Conference on Affective computing and intelligent interaction* (Springer), 981–995

Tekleab, A. G., Quigley, N. R., and Tesluk, P. E. (2009). A longitudinal study of team conflict, conflict management, cohesion, and team effectiveness. *Group & Organization Management* 34, 170–205

Ten Berge, J. M. and Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika* 69, 613–625

Teyssier, M., Bailly, G., Pelachaud, C., and Lecolinet, E. (2020). Conveying emotions through device-initiated touch. *IEEE Transactions on Affective Computing*

Thorndike, E. L. (1920). Intelligence and its uses. *Harper's magazine*

Thye, S. R., Yoon, J., and Lawler, E. J. (2002). The theory of relational cohesion: Review of a research program. *Advances in group processes* , 217–244

Tien, D. X., Yang, H.-J., Lee, G.-S., and Kim, S.-H. (2021). D2c-based hybrid network for predicting group cohesion scores. *IEEE Access* 9, 84356–84363

Tomasello, M. (2014). The ultra-social animal. *European journal of social psychology* 44, 187–194

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., and Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current anthropology* 53, 673–692

Torro, O., Holopainen, J., Jalo, H., Pirkkalainen, H., and Lähtevänoja, A. (2022). How to get things done in social virtual reality-a study of team cohesion in social virtual reality-enabled teams. In *Proceedings of the 55th Hawaii International Conference on System Sciences*. 470–479

Tracy, J. L. and Robins, R. W. (2004). Show your pride: Evidence for a discrete emotion expression. *Psychological Science* 15, 194–197

Trizano-Hermosilla, I. and Alvarado, J. (2016). Best alternatives to cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology* 7

Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological bulletin* 63, 384–399

Tuckman, B. W. and Jensen, M. A. C. (1977). Stages of small-group development revisited. *Group & Organization Studies* 2, 419–427

Tung, H.-L. and Chang, Y.-H. (2011). Effects of empowering leadership on performance in management team: Mediating effects of knowledge sharing and team cohesion. *Journal of Chinese Human Resources Management* 2, 43–60

Uleman, J. S., Adil Saribay, S., and Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology* 59, 329–360

Van Baaren, R., Janssen, L., Chartrand, T. L., and Dijksterhuis, A. (2009). Where is the love? the social aspects of mimicry. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 2381–2389

Van Bergen, A. and Koekebakker, J. (1959). "Group cohesiveness" in laboratory experiments. *Acta Psychologica* 16, 81–98

Van Dijk, E., De Dreu, C. K. W., and Gross, J. (2020). Power in economic games. *Current opinion in psychology* 33, 100–104

Van Vugt, M. and Schaller, M. (2008). Evolutionary approaches to group dynamics: An introduction. *Group Dynamics: Theory, Research, and Practice* 12, 1–6

Vanhove, A. J. and Herian, M. N. (2015). Team cohesion and individual well-being: A conceptual analysis and relational framework. In *Team Cohesion: Advances in Psychological Theory, Methods and Practice* (Emerald Group Publishing Limited). 53–82

Varni, G., Avril, M., Usta, A., and Chetouani, M. (2015). Syncpy: A unified open-source analytic library for synchrony. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And InfLuence*. 41–47

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* 30, 5998–6008

Vincer, D. J. and Loughead, T. M. (2010). The relationship among athlete leadership behaviors and cohesion in team sports. *The Sport Psychologist* 24, 448–467

Vinciarelli, A. and Mohammadi, G. (2014). A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 273–291

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009a). Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 1743–1759

Vinciarelli, A., Pantic, M., Bourlard, H., and Pentland, A. (2008). Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM international conference on Multimedia*. 1061–1070

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schroeder, M. (2011). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3, 69–87

Vinciarelli, A., Salamin, H., and Pantic, M. (2009b). Social signal processing: Understanding social interactions through nonverbal behavior analysis. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops* (IEEE), 42–49

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17, 261–272

Wainberg, M., Alipanahi, B., and Frey, B. J. (2016). Are random forests truly the best classifiers? *The Journal of Machine Learning Research* 17, 3837–3841

Wallbott, H. G. (1998). Bodily expression of emotion. *European journal of social psychology* 28, 879–896

Waller, M. J., Okhuysen, G. A., and Saghafian, M. (2016). Conceptualizing emergent states: A strategy to advance the study of group dynamics. *The Academy of Management Annals* 10, 561–598

Walocha, F., Maman, L., Chetouani, M., and Varni, G. (2020). Modeling dynamics of task and social cohesion from the group perspective using nonverbal motion capture-based features. In *Companion Publication of the 22nd International Conference on Multimodal Interaction*. 182–190

Wang, W., Precoda, K., Hadsell, R., Kira, Z., Richey, C., and Jiva, G. (2012). Detecting leadership and cohesion in spoken interactions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 5105–5108

Wang, Y., Wu, J., Huang, J., Hattori, G., Takishima, Y., Wada, S., Kimura, R., Chen, J., and Kurihara, S. (2020). Ldnn: Linguistic knowledge injectable deep neural network for group cohesiveness understanding. In *Proceedings of the 22nd International Conference on Multimodal Interaction*. 343–350

Webster, J. and Staples, D. S. (2006). Comparing virtual teams to traditional teams: An identification of new research opportunities. In *Research in personnel and human resources management* (Emerald Group Publishing Limited). 181–215

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior* (Springer). 185–208

Weisfeld, G. E. and Beresford, J. M. (1982). Erectness of posture as an indicator of dominance or success in humans. *Motivation and Emotion* 6, 113–131

Welzl, E. (1991). Smallest enclosing disks (balls and ellipsoids). In *New results and new trends in computer science* (Springer). 359–370

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., and Scherer, K. R. (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology* 4, 228–239

West, C. and Zimmerman, D. H. (2015). *Small insults: A study of interruptions in cross-sex conversations between unacquainted persons.*, vol. IV (Routledge/Taylor & Francis Group)

West, M. (1996). *Reflexivity and work group effectiveness: A conceptual integration* (John Wiley & Sons, Ltd)

Wheelan, S. A. (1994). *Group processes: A developmental perspective.* (Allyn & Bacon)

Wiltermuth, S. S. and Heath, C. (2009). Synchrony and cooperation. *Psychological science* 20, 1–5

Xie, K., Hensley, L. C., Law, V., and Sun, Z. (2019). Self-regulation as a function of perceived leadership and cohesion in small group online collaborative learning. *British Journal of Educational Technology* 50, 456–468

Xuan Dang, T., Kim, S.-H., Yang, H.-J., Lee, G.-S., and Vo, T.-H. (2019). Group-level cohesion prediction using deep learning models with a multi-stream hybrid network. In *Proceedings of the 21st International Conference on Multimodal Interaction*. 572–576

Yamaguchi, R. and Maehr, M. L. (2004). Children's emergent leadership: the relationships with group characteristics and outcomes. *Small Group Research* 35, 388–406

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. *The 18th International Conference on Computational Linguistics (COLING)* 2, 947–954

Yukelson, D., Weinberg, R., and Jackson, A. (1984). A multidimensional group cohesion instrument for intercollegiate basketball teams. *Journal of Sport and Exercise Psychology* 6, 103–117

Zaccaro, S. J., Blair, V., Peterson, C., and Zazanis, M. (1995). Collective efficacy. In *Self-efficacy, adaptation, and adjustment* (Springer). 305–328

Zaccaro, S. J., Foti, R. J., and Kenny, D. A. (1991). Self-monitoring and trait-based variance in leadership: An investigation of leader flexibility across multiple group situations. *Journal of applied psychology* 76, 308–315

Zacks, J. M. and Swallow, K. M. (2007). Event segmentation. *Current directions in psychological science* 16, 80–84

Zacks, J. M. and Tversky, B. (2001). Event structure in perception and conception. *Psychological bulletin* 127, 3–21

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2008). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* 31, 39–58

Zhang, Y. (2016). Functional diversity and group creativity: The role of group longevity. *The Journal of Applied Behavioral Science* 52, 97–123

Zhang, Y., Olenick, J., Chang, C.-H., Kozlowski, S. W., and Hung, H. (2018). Teamsense: assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1–22

Zheng, D., Witt, L., Waite, E., David, E. M., van Driel, M., McDonald, D. P., Callison, K. R., and Crepeau, L. J. (2015). Effects of ethical leadership on emotional exhaustion in high moral intensity situations. *The Leadership Quarterly* 26, 732–748

Zhu, B., Guo, X., Barner, K., and Boncelet, C. (2019). Automatic group cohesiveness detection with multi-modal features. In *Proceedings of the 21st International Conference on Multimodal Interaction*. 577–581

Zou, B., Lin, Z., Wang, H., Wang, Y., Lyu, X., and Xie, H. (2020). Joint prediction of group-level emotion and cohesiveness with multi-task loss. In *Proceedings of the 5th International Conference on Mathematics and Artificial Intelligence*. 24–28

Zurcher Jr, L. A. (1969). Stages of development in poverty program neighborhood action committees. *The Journal of Applied Behavioral Science* 5, 223–258

DOCTORAL SCHOOL

Titre : Analyse Automatique de la Cohésion dans l'Interaction de Petits Groupes

Mots clés : Analyse Multimodale, Apprentissage Automatique, Apprentissage Profond, Cohésion, Traitement des Signaux Sociaux

Résumé : Au cours de la dernière décennie, un nouveau domaine de recherche multidisciplinaire appelé traitement des signaux sociaux (SSP) a émergé. Il vise à permettre aux machines de détecter, reconnaître et afficher les signaux sociaux humains. L'analyse automatisée des interactions de groupe est l'une des tâches les plus complexes abordée par ce domaine de recherche. Récemment, une attention particulière s'est portée sur l'étude automatisée des états émergents. En effet, ceux-ci jouent un rôle important dans les dynamiques d'un groupe car ils résultent des interactions entre ses membres.
Dans cette Thèse, nous abordons l'analyse automatique de la cohésion dans les interactions de petits groupes. La cohésion est un état émergent affectif multidimensionnel qui peut être défini comme un processus dynamique, reflété par la tendance d'un groupe à rester ensemble pour poursuivre des objectifs et/ou des besoins affectifs. Malgré la riche littérature disponible sur la cohésion du point de vue des Sciences Sociales, l'analyse automatique de la cohésion en est encore à ses débuts.
En s'inspirant de connaissances tirées des Sciences Sociales, cette Thèse vise à développer des modèles informatiques de cohésion suivant quatre axes de recherche, en s'appuyant sur des techniques d'apprentissage automatique et d'apprentissage profond. Ces modèles doivent en effet tenir compte de la nature temporelle de la cohésion, de sa multidimensionnalité, de la façon de modéliser la cohésion du point de vue des individus et du groupe, d'intégrer les relations entre ses dimensions et leur évolution dans le temps, ainsi que de tenir compte des relations entre la cohésion et d'autres processus de groupe. De plus, face à un manque de données disponibles publiquement, cette Thèse a contribué à la collecte d'une base de données multimodales spécifiquement conçue pour étudier la cohésion, et pour contrôler explicitement ses variations dans le temps. Une telle base de données permet, entre autres, de développer des modèles informatiques intégrant la cohésion perçue par les membres du groupe et/ou par des points de vue externes.
Nos résultats montrent la pertinence de s'inspirer des théories tirées des Sciences Sociales pour développer de nouveaux modèles computationnels de cohésion et confirment les avantages d'explorer chacun des quatre axes de recherche.

Title : Automated Analysis of Cohesion in Small Groups Interactions

Keywords : Cohesion, Deep Learning, Machine Learning, Multimodal analysis, Social Signal Processing

Abstract : Over the last decade, a new multidisciplinary research domain named Social Signal Processing (SSP) emerged. It is aimed at enabling machines to sense, recognize, and display human social signals. One of the challenging tasks addressed by SSP is the automated group interaction analysis. Recently, a particular emphasis is given to the automated study of emergent states as they play an important role in group dynamics. These are social processes that develop throughout group members' interactions. In this Thesis, we address the automated analysis of cohesion in small groups interactions. Cohesion is a multidimensional affective emergent state that can be defined as a dynamic process reflected by the tendency of a group to stick together to pursue goals and/or affective needs. Despite the rich literature available on cohesion from a Social Sciences perspective, its automated analysis is still in its infancy.
Grounding on Social Sciences' insights, this Thesis aims to develop computational models of cohesion following four axes research axes, leveraging Machine Learning and Deep Learning techniques. Computational models of cohesion, indeed, should account for the temporal nature of cohesion, the multidimensionality of this group process, take into account how to model cohesion from both individuals and group perspectives, integrate the relationships between its dimensions and their development over time, and take heed of the relationships between cohesion and other group processes. In addition, facing a lack of publicly available data, this Thesis contributed to the collection of a multimodal dataset specifically designed for studying group cohesion and for explicitly controlling its variations over time. Such a dataset enables, among other perspectives, further development of computational models integrating the perceived cohesion from group members and/or external points of view.
Our results show the relevance of leveraging Social Sciences' insights to develop new computational models of cohesion and confirm the benefits of exploring each of the four research axes.

Institut Polytechnique de Paris
91120 Palaiseau, France