Conversation Quality

Modeling in Free-Standing Conversational Groups

N. Laxminarayanan Raj Prabhu

Technische Universiteit Delft

TU Delft

Conversation Quality

Modeling in Free-Standing Conversational Groups

by

N. Laxminarayanan Raj Prabhu

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday July 30, 2020 at 10:30 AM.

Student number:	4764722	
Project duration:	November 1, 2019 – July 30, 2020	
Thesis committee:	Assoc. Prof. dr. ir. H. Hung,	TU Delft, supervisor
	Dr. C.R.M.M. Oertel,	TU Delft
	Dr. C.C.S. Liem,	TU Delft
	Chirag Raman,	TU Delft

This thesis is confidential and cannot be made public until August 30, 2020.

TU Delft

Abstract

Social interactions in general are multifaceted and there exists an wide set of factors and events that influence them. Hence, interactions as a social phenomena have been studied by researchers in the fields of psychology and social signal processing from different stand points. The common trend in literature is to perform an in-depth study on a particular aspect of social conversations. Contrast to studying a particular aspect of social interactions, in this research, we attempt to comprehensively quantify social interactions with respect to individual experiences in the interaction, particularly focusing on spontaneous interactions which are typically non-task-directed interactions. To achieve this, we design a novel perceived measure, the perceived *Conversation Quality*, which intends to quantify spontaneous interactions by accounting for several aspects of spontaneous interactions, namely *Rapport*, *Interaction Quality*, *Inter-personal liking* and *Free-for-all*. Such an attempt to quantify spontaneous interactions is a substantial contribution towards building socially intelligent system to support human-human and human-robot interaction.

To quantitatively study the perceived Conversation Quality, we devised a questionnaire which measures, at the individual- and at the group- level, the perceived Conversation Quality in spontaneous interactions. Existing literature in the field of social signal processing showed that behavioural features such as Turn-Taking and Bodily Coordination features are informative of several subtle social constructs like cohesion, rapport and interest-levels. Drawing inspiration from them, we extract Turn-Taking and Bodily Coordination features to model perceived Conversation Quality. Using a Logistic Regression, optimised using the Stochastic Gradient Descent algorithm, we were able to predict the individual- and the group- level perceived Conversation Quality with a mean AUC of 0.76 (±0.13) and 0.96 (±0.03) respectively. From the experiments performed, we see that the Synchrony and Convergence based bodily coordination features are the best performing feature sets while predicting the individual-level perceived Conversation Quality and, the Turn-Taking based features are the best performing feature sets while predicting the group-level perceived Conversation Quality.

To further study the properties of the perceived Conversation Quality measure, we performed several statistical tests which study the effect of different social factors on the measure. From theses results, we see that the perceived Conversation Quality, both the the individual- and the group- level, decreases with the increase in number of participants in the spontaneous interaction. Moreover, the equal distribution of talk time amongst participants and the duration of their silence periods have a significant effect on the perceived Conversation Quality. Another interesting finding is that successful and unsuccessful interruptions have a positive effect and a negative effect on perceived Conversation Quality respectively. Moreover, the results also show that the time factor revealing bodily coordination features (lagged correlations and convergence) have a significant effect on both the two levels of Conversation Quality, suggesting that the time factor involved in the bodily coordination is more informative than the degree of coordination in itself. An important takeaway from the statistical test and the predictive modeling experiments is that, the two forms of perceived Conversation Quality, the individual- and the group- level, are completely different from one another with little commonality in their respective statistically significant features.

Preface

Ever since I moved to The Netherlands to broaden my knowledge in Data Science and Technology, my interest and fascination towards working with human-centric behavioural data, thereby creating maximum impact, has increased in folds. I strongly believe that studying human behaviour using data science can help us progress towards solving humanity's most pressing issues.

Looking back at the last two years of my Master's Degree, I realise that I have obtained half of my required ECTS from the Socially Perceptive Computing Lab at Delft University of Technology. The experience in the lab has helped me find my passion and has motivated me to further peruse my career in the domain of Social Perception. I feel extremely grateful and fortunate for such an experience. I still remember the first lecture of the Social Signal Processing course which I had followed and where I was introduced to the domain of Social Perception. The course piqued my interests towards studying human behaviour using machine learning techniques. Little did I know that I would be soon doing an extensive study on social interactions with my master's thesis. The ten months of journey with my Master's Thesis was enjoyable through both ups and downs. I strongly believe that while doing my thesis, I have also learnt a lot in terms of research processes and on handling human-centric behavioural data.

Before you lies my thesis. This thesis has been written to complete my Master's Degree in Computer Science at Delft University of Technology, and is written in such a way that it should be understandable for anyone with a university degree in engineering. The research and writing of this thesis took place between November 2019 and July 2020 and was supervised by Dr. Hayley Hung and Chirag Raman. I hope you will enjoy reading my thesis.

N. Laxminarayanan Raj Prabhu
Delft, July 2020

Acknowledgements

First of all, I would like to thank Hayley for her excellent supervision and support during the course of my masters thesis. The involvement and interest she showed over my project helped me stay motivated throughout the thesis period. Her constant support and motivation made me push my own boundaries and realise my strengths and weaknesses. Without Hayley, I wouldn't have had the motivation and courage to research and venture into studying complex social concepts.

Equally important is Chirag Raman's support during the whole thesis period. His sharp feedbacks and criticisms has helped me often to direct my focus towards the details. The experience under his supervision has helped me mature as a researcher. Chirag has not just been a project supervisor, but also a mentor who always has looked out to protect my career growth.

I would also like to take this moment to thank Swathi Yogesh, Divya Suresh Babu and Nakul Ramachandran, for their time and patience in helping me annotate the dataset. Without them, it wouldn't have been so easy to collect the ground-truths for this research.

Also, I would like to thank the members of the Socially Perceptive Computing Lab at TU Delft, Ekin Gedik, Jose Vargas and Stephanie Tan, for their valuable feedbacks. Whenever they had an opportunity to listen to my thesis, they were very kind enough to show interest on my work and share their valuable opinions on the thesis project.

Also, I would like to thank Stavros Makrodimitris, Yeshwanth Napolean and Tim Rietveld, for their time and thoughtfulness. Their timely feedbacks and thoughts helped me sharpen my research work. I would also like to thank Gary Gilson for helping me design the cover page of this report, thanks a lot for your time and your creative ideas.

Last, but certainly not the least, I would like to thank my parents and my brother, Pranave. Without their support it wouldn't have been possible for me to come to the Netherlands and join TU Delft at the first place. Their support and motivation has always been key in helping me pursue my dreams and aspirations.

Contents

1 Introduction

Humans are social animals by nature (Aristotle, Politika ca. 328 BC), and social interactions are essential part of their day-to-day life, eventually having a key influence on their physical and mental well-being [114]. The analysis of social interactions has hence, attracted interests from social psychologists for centuries [31][63][42]. More recently, the advances in sensing technologies and the innovations in the field of machine learning have also drawn interest from computer scientists, towards the automatic analysis of human behavior in social interactions [117][36][22][60]. While current computing systems have mastered several cognitive tasks using artificial intelligence, social intelligence is still far from an achievement. Perceiving and adapting to different social behaviours are commonly labeled as *socially intelligent behaviour*. The ability to understand and manage social signals of a person we are communicating with is the core of such social intelligence [117]. Vinciarelli et al., (2008) through their famous work - *Social Signal Processing* (SSP) [117], gave rise to a new field in computer science aimed at building the next generation of socially aware computing. Vinciarelli et al., in the paper, argue for the importance of including the essence of social intelligence in the next generation computing systems.

The ability of a computing system to automatically analyse and perceive social behaviours (social intelligence) has a wide range of applications, in domains such as human-computer interaction, computer supported cooperative work, healthcare, organisation research, surveillance, education, political science and epidemiology. Another interesting application of such socially intelligent computers, often unattended to by the research community is the enabling of feedback tools that can help people receive feedbacks on their social interactions and in-turn assess their own social roles in those interactions, which would eventually help them increase their social capabilities and the quality of their relationships with their peers. Currently, people indulging in social interactions in their day-to-day life, do not have a retrospective view on their previous interactions. Such a retrospective view on interactions, if adopted in large-scale, can lead to optimal interactions, playing a positive role in the physical and mental well-being of all the members involved in the interaction. Hence, a socially intelligent system capable of automatically perceiving an individual's interaction and providing them with constructive feedbacks regarding their interactions can be a potential real-world application.

In this research, we design and develop a conversation quality measure and its modeling methodology which contributes towards enabling feedback systems in providing constructive feedbacks to individuals indulging in *spontaneous* interactions. Spontaneous interactions are typically non-task-directed interactions which are unplanned, unconstrained and natural situations [102][93][125], in contract to task-oriented conversations which are pre-planned and constrained (for e.g. team meetings). The research work from Reitter et al., (2010) [102] reveals the presence of contrasting behaviour patterns between a task-directed and a non-task-directed interaction. This requires us to study such interactions separately with their respective considerations. In this research, we specifically focus on interactions which are non-task-directed and spontaneous, focusing on enabling feedback systems in such interactions. The feedbacks are intended to improve the overall quality of the interaction, thereby positively impacting the experiences of every individual in the interaction.

For such a system, to be capable of enabling feedback systems, it has to automatically and unobtrusively perceive events and individual experiences in a social interaction. At the same time, it also has to perceive interactions in a comprehensive manner. That is, the system should, in an all-inclusive manner, be able to perceive different aspects of an interaction thereby measuring the quality of the interaction. Such a measure is not just relevant to feedback systems, but also to any socially intelligent system, in the domains of computer supported cooperative work (CSCW) and human-computer interaction (HCI), which aims to perceive the overall quality of spontaneous interactions by considering different key aspects of the interaction.

Rest of this chapter presents the research goals along with the main contributions of this research. Finally, the organization of the rest of the thesis is given.

1.1. Research Goals

The main research goal of this thesis is to define and model the measure of *Conversation Quality* in spontaneous interactions and also study its properties. With a direct focus on studying *Conversation Quality* in spontaneous interactions, we frame three sub-goals for this research. The first sub-goal of this research is to formally define the social construct of *Conversation Quality* and also identify its forms of perception in spontaneous interactions. The second sub-goal is to predictively model the *Conversation Quality* using several categories of automatically extracted group-level behavioural features. Finally, the third sub-goal is to directly study the properties of *Conversation Quality* by performing several statistical tests. We strongly believe that the accomplishment of the goals mentioned above will be a significant contribution towards building socially intelligent systems capable of perceiving the overall quality of spontaneous interactions.

Social interactions in general are multifaceted and there exists an wide set of factors and events that influence them. Hence, interactions as a social phenomena have been studied by researchers in the field of SSP from different stand points. The common trend in literature is to perform an in-depth study on a particular aspect of conversations (e.g. Interest-levels[37] and Rapport[85]). Such a trend commonly followed by researchers can be owed to the fact that perceptions of social interaction in itself are extremely complex and subjective in nature, and hence studying a particular aspect of the interaction might reduce the complexity involved in the research. But at the same time, it is also important to study the overall quality of such interactions, in a comprehensive manner, with respect to the individual experiences in the social interactions. This research, focused on defining and modeling *Conversation Quality*, is an attempt on such a study with a comprehensive viewpoint on spontaneous interactions. To the best of our knowledge, there is no existing work which has attempted to quantify the overall quality of conversation in a spontaneous free-standing group conversation setting.

1.2. Contributions

This section, presents further in detail the key contributions as the novelty of this research, with respect to the knowledge gap in existing literature.

1.2.1. Conversation Quality as a social construct

The main contribution of this research is that we design a novel measure to quantify spontaneous conversations - *Conversation Quality*. While social interaction measures, such as *cohesion* [56], *rapport* [85], *bonding* [60], *enjoyment* [80], *involvement* [92] and *interest-levels* [61][37], intend to capture a particular aspect the interaction, the measure of *Conversation Quality* intends to quantify the overall quality of conversation in a spontaneous interaction, with respect to the experiences of individuals in the interaction. With that said, the measure of *Conversation Quality* intends to embrace the complexity involved in the perception of spontaneous interactions and comprehensively quantifies them, by capturing different aspects of the interaction. To design the measure of *Conversation Quality*, we draw inspiration from several key social constructs and aspects studied in the literature, which focuses towards individual experiences, namely *Rapport*[85], *Bonding*[60], *Free-for-all*[79], *Involvement*[92], *Interest-levels*[37], and *Quality of interaction*[28]. While Rapport, Bonding and Free-for-all capture relational (relational with interacting partners) appraisal based individual experiences, the aspects such as Involvement, *Interest-levels*, and *Quality of interaction* capture individual appraisal based individual experiences. While designing the measure of *Conversation Quality*, we also define its constituents and identify it's forms of perception in spontaneous conversations. A more detailed definition of the measure of perceived *Conversation Quality* will be presented in section 3.1.

1.2.2. Modeling Conversation Quality

This particular contribution of this research corresponds to the second and the third sub-goal mentioned earlier. Post designing the measure of *Conversation Quality*, as part of this thesis, we design and implement a methodology to perform the predictive modeling of *Conversation Quality* and also study its properties. As part of the predictive modeling process, we investigate several modeling techniques and several sets of behavioural cues to model the measure of conversation quality. By performing a predictive modeling of the novel measure of *Conversation Quality*, we believe it will contribute towards the enabling of socially intelligent systems to provide constructive conversational feedbacks. To study the properties of the measure of *Conversation Quality*, we perform several statistical tests and correlation analysis with respect to the *Conversation Quality* measure and several sets of behavioural cues.

For the above mentioned tasks, behavioural cues broadly categorised into *Turn-Taking*[56] and *Bodily Coordination*[61][30] are used. Turn-taking features are cues related to the turn-taking system existing in a social interaction (explained in detail in Section-5.3.5) and, Bodily Coordination are *synchrony* and *convergence* based features related to the *coordination in body movements* of interacting partners (explained in detail in Sections-5.3.1, 5.3.3). As part of the methodology, the above mentioned feature sets are automatically extracted from manual annotations of speaking status and bodily worn tri-axial accelerometers respectively. The feature sets used in the research are inspired from literature works studying different aspects of a social interaction, namely cohesion[88][56], rapport[85], interest-levels[37][61], experience[79][70] and engagement[55].

1.3. Research Questions

With respect to the contributions of this research, the main research question this thesis intends to answer is,

> ***How can we quantify the overall quality of spontaneous interaction with respect to the experiences of individuals in the interaction, and, what are the behavioural cues which influence such spontaneous interactions?***

While the above research question explains in a high-level the research goals of this thesis, to further direct this research we framed three sub-research questions and are presented below,

1. Quantifying spontaneous conversations

> *How can we quantify the overall quality of spontaneous interactions, with respect to the experiences of individuals in the interaction? What are the constituents of the measure which quantifies spontaneous conversations, and, what are its forms of perception?*

2. Properties of spontaneous conversations

> *What are the properties of the measure which quantifies spontaneous interactions? What are the differences, with respect to the behavioural cues studied, between interactions at the two extremes of the measure scale?*

3. Predictive modeling in spontaneous conversations

> *How can we model and predict the measure which quantifies spontaneous interactions? What are its best defining features?*

1.4. Research Context

This study was performed in an independent fashion at the TU Delft *Socially Perceptive Computing Lab* and can be seen as a subset of the larger MatchNMingle project. No external parties were involved or had any influence over any of the decisions taken during the course of this project. Access to the data that was used is restricted to protect the privacy of the participants in the experiments. For more information on accessing the dataset, please visit the MatchNMingle project website.

1.5. Outline

The rest of the thesis is organized as follows,

1. In Chapter-2, a detailed literature review on works which have studied different facets of a social interactions are presented. We intend to draw inspirations from such literature to design and model the measure of spontaneous interaction, the *Conversation Quality*.

2. In Chapter-3, we present to the reader the novel measure of *Conversation Quality*. In the chapter, we define *Conversation Quality*, its constituents and also its forms of perception. Finally, we also devise a questionnaire to measure *Conversation Quality* in spontaneous interactions.

3. In Chapter-4, the dataset used to study spontaneous interactions and their *Conversation Quality* is discussed. In the chapter, we also present some preliminary analysis on the *Conversation Quality* annotations (ground-truths) collected for the research.

4. In Chapter-5, we discuss in detail the methodology used to model *Conversation Quality* in spontaneous interactions. The chapter explains the pre-processing, feature extraction, statistical tests and predictive modeling techniques used in the study of *Conversation Quality*.

5. In Chapter-6, we present the experiments performed to predicatively model and study the properties of *Conversation Quality*. In the chapter, along with the explanation on the experimental setup, we also present the results of the experiments.

6. In Chapter-7, we discuss, in an all-embracing manner, the results from the experiments performed. While discussing the experiments and their results, we also highlight several key findings of the research. In the process of doing so, we intend to answer the thesis' research questions (Section-1.3).

7. Finally, in Chapter-8, concluding remarks on our study of *Conversation Quality* is presented. In the chapter, the limitations of the study along with some suggestions for future research are also discussed.

2

Quality of social interactions and Individual experiences: *A Literature Review*

In this chapter, a literature review on works which are key in studying the quality of social interactions with respect to individual experiences is presented. The automatic analysis of social behavior has been of interest to both social scientists and computer scientists. Therefore, studies from both research fields are reviewed here. Firstly, in this chapter, we present an overview on several key concepts which are vital in studying and analysis social conversations. Secondly, we review a series of literature studying several different aspects of a social conversations and individual experiences. Then, we discuss a number of social cues and modeling techniques, which have been successfully used in literature to model group behaviour. Finally, we review some of the social constructs which are similar to *Conversation Quality*, to understand how researchers have handled different constructs under different scenarios.

The aim of the literature review is to review literature which have studied different facets of a social conversation and draw inspirations from them to design and model the comprehensive measure of *Conversation Quality*.

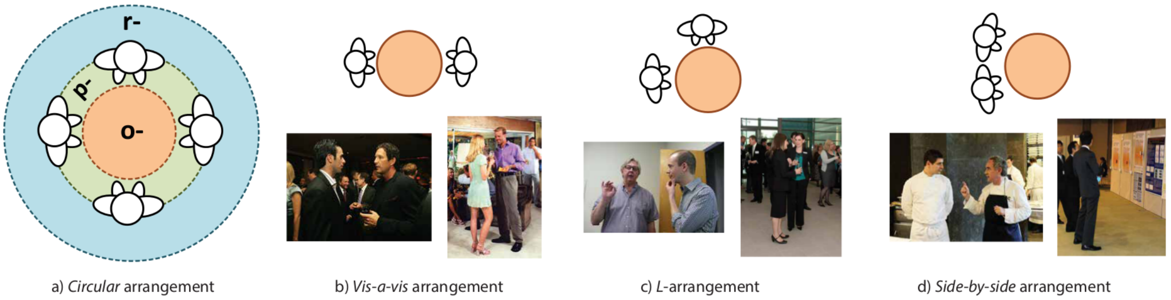
2.1. Conversation Analysis

Groups are intriguing social phenomena. They are at the core of organizational functioning across all sectors and society at large. Therefore, studying group conversations are essential research items for social psychologists across the research community. Similarly, computer scientists in the research field of SSP have also followed the same path into the automatic analysis and perception of social conversations. Small-group face-to-face conversations are one such dynamic social conversation setting, where a wide range of *inter-personal relationships* and *social constructs* emerge from within. Fundamental research on social interactions was pioneered by Goffman [42], whose symbolic interaction perspective explains society via the everyday behavior of people and their interactions. Face-to-face conversations are the most common form of social interactions, and free-standing conversational groups (FCGs) denote small groups of two or more co-existing persons engaged in ad-hoc interactions [43].

This section introduces several concepts involved in studying conversations. Firstly, it introduces key concepts involved in studying conversations with respect to the *spatial* component of a social conversation. Subsequently, we also introduce several key concepts involved in studying conversations with respect to the *temporal* component of a social conversation

2.1.1. The Spatial Perspective

To perform analysis and study conversations, it is important to understand how group conversations are spatially formed and maintained in structure. For this purpose, in this section, we introduce the concept of *F-Formations* and *Proxemics.*



a) *Circular* arrangement b) *Vis-a-vis* arrangement c) *L*-arrangement d) *Side-by-side* arrangement

Figure 2.1: A graphical illustration of F-Formation (a), along with some of the arrangements (b)-(d). The illustration was taken from [27].

F-Formations An F-formation is the most fundamental concept to conceptualize the spatial nature of an FCG. Kendon (1990) [63, p.210] defined f-formations to arise whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access. An F-Formation was further defined to consist of three spatial spaces within it, as seen in Fig-2.1, the o-space, p-space and r-space. The o-space in the f-formation is the convex space that is jointly shared amongst the participants of the f-formation. An f-formation begins to exist whenever two or members form a stable o-space. The o-space, in an FCG, is controlled by all participants and tends to shut out non-participants of the interaction from the participants of the interaction. The p-space is the space surrounding the o-space, where all the participants of the interaction are standing. The r-space is simply defined as all other space beyond the p-space and o-space.

While an f-formation is a spatial perspective on FCGs, it is not static in nature. While defining f-formations, Kendon also points out the relationship between the dynamic behaviour of the participants and the maintenance of the o-space. Kendon defines the f-formation system as, "the system of behavioural organisation by which certain spatial-orientational patterns are established and sustained in free-standing conversations" [63]. This definition implies that the members of a group must actively cooperate to keep the F-Formation together, hence there exists an intention intended by participants in the f-formation to sustain the o-space. Works such as [128], have modeled social involvement by solely relying on the spatial nature of the f-formations. Zhang et al. (2016) [128] in this work, learn a frustum of attention from the spatial context of an f-formation to analyse differing levels of social involvement in FCGs. With that said, to study social constructs emerging from FCGs, it is important to also consider the spatial context explained by f-formations. This research considers *F-Formations* to be the most fundamental spatial concepts in defining and studying free-standing conversation groups.

Proxemics Proxemics is the study of the dynamic process by which people position themselves in face to face interaction, that affects the behaviour, communication, and social interaction of participants in the interaction [45] [48]. The concept of *Proxemics* was pioneered by Edward Hall, who defined it as "the interrelated observations and theories of man's use of space as a specialized elaboration of culture". The study of proxemics is not only relevant to FCGs, but also used in several other fields, such as studying the effect of population density and culture on behaviour, communication, and social interaction. With respect to the spatial context of FCGs, the study of proxemics involves studying non-verbal cues such as, interpersonal distance, orientation of feet, hips and shoulders and head pose. Such non-verbal cues are widely used in studying different aspects of social interactions. For example, these cues have been key in studying several social concepts, right from detecting f-formations [119] and initiation/termination of interaction [82] to social involvement [128] and interest-levels [37]. Given that proxemic related cues such as head and body poses can be used in tasks such as detecting f-formations, interest-levels and in estimating social involvement, it is also possible that this same information could be valuable source in studying several social constructs emerging from an f-formation, which in turn could be important to study the measure of conversation quality.

Existing work in psychology and social sciences literature has shown a strong relationship between the non-verbal cue of lower-body orientation and social involvement [64][62]. Vasquez et al. (2015) [119] building on the above motivation, propose an alternating optimization procedure that estimates lower body orientations and detects groups of interacting people. The authors show that show that the proposed f-formation detection methodology based on lower body orientations, can improve over the state-of-the-art detection, in-particular the detection of non-interacting people without sacrificing group detection accuracy. This is an interesting insight as in crowded mingling scenarios, the lower body orientation adds more value towards detecting interaction group that that of interpersonal distance.

Similarly, Mead et al. (2013) [82] proposed a psychophysical framework to autonomously estimate proxemic features such as human–robot distance and orientation based nonverbal cues. The estimated features were then modeled using a Hidden Markov Model (HMM) to enable autonomous and mobile sociable robots to perceive, join and maintain face-to-face social interactions with human users while reliably recognizing natural spatio-temporal human behaviors such as pose and body gestures. The results showed interesting potentials of proxemic cues. The results show that modeling by relying on psychophysical features, which encode the sensory experience of each interacting agent, outperform those trained on physical features, which only encode spatial relationships.

There exists strong literature backing to the fact that head and body poses are a window into ones involvement. They are often used by researchers as a proxy for an individual's directed attention and for their degree of involvement. When researching group-level interest in meetings, Gatica et al. (2013) [37] use basic pose estimates as visual cues to model interest-levels. Group interest-level is a similar social construct to that of conversation quality and works such as this can be direct inspirations for designing and modeling conversation quality. Body orientation information has also been used to estimate social involvement. Hung and Zhang (2016) [128] propose a method to detect f-formation associates, using body orientation alone as a discriminative proxemic feature. F-formation associates are individuals who are part of an f-formation but do not have full member status and therefore have a lower degree of involvement. A relationship between involvement and body orientation in turn implies a possible relationship between body orientation and conversation quality.

2.1.2. The Temporal Perspective

Conversations are fundamentally dynamic in nature and evolve over time. For example, individual experiences in a conversation are time evolving and dependent on several temporal factors. This dynamic nature of social interactions are commonly referred to as social dynamics. This section presents concepts which are key in studying the temporal characteristics of group conversations - namely *Floors*, *Turns* and *Emerging Social Constructs*.

Floors Edelsky's (1981) work [31], used conversation analysis to examine social interaction during a series of meetings, and reviewed several key units of face-to-face conversations - "*turn*", "*topic*", "*floor*", and "*gap*". According to Edelsky,

> The floor is defined as the acknowledged what's-going-on within a psychological titnelspace. What's going on can be the development of a topic or a function (teasing, soliciting a response, etc.) or an interaction of the two. It can be developed or controlled by one person at a time or by several simultaneously or in quick succession. It is official or acknowledged in that, if questioned, participants could describe what's going on as "he's talking about grades" or "she's making a suggestion" or "we're all answering her." [31, p.405]

Crucially, she observed two contrasting styles of conversation, which have since been termed *cooperative floors* and *exclusive floors*. The cooperative floor is typified by a feeling of participants being "*on the same wavelength*" in a conversation that is a "*free-for-all*" ([31], p. 384); here there is a sense that no one owns the floor, it being perfectly acceptable for everyone to talk at once. In contrast, the exclusive floor is characterised by a sense of orderliness, with only one person owning the floor at a time and turns rarely overlapping. The cooperative floor seems to capture the sense of cohesiveness and engagement that is associated with positive experiences in conversational scenarios. This research, specifically concentrates on *cooperative floors* of conversations, with a goal to model the quality of such conversations.

Turn-Taking System A conversation floor is an interesting hostspot for several events, with the individuals in the floor interacting and sharing views with one another. A *Turn* is one such key event in a floor, which is one of the key units operationalised by Edelsky for conversation floor analysis. Edelsky defines a turn in conversation floor as,

> an on-record "speaking" (which may include nonverbal activities) behind which lies an intention
> to convey a message that is both referential and functional.[31, p.403]

A turn is a continuous time interval when someone is speaking or their binary speaking activity is 1. Several researchers have handled it differently. For example, Lai et al. (2013) [71] used *spurts* to measure turn-taking systems, where spurts are defined as speaking segments separated by at least 500ms silence. Similarly, Levitan et al. (2011) [76] used *IPU*s (inter-pausal unit), defined as a pause-free segment of speech from a single speaker, where pause is defined as a silence of 50ms or more from the same speaker. The turn durations are hypothesized to be approximately equal for all participants in highly involved conversations [56].

With respect to the above definition, the current speaker is the person currently speaking, currently holds the turn. Turns form the basic unit which individuals in the conversation use to carry out the conversation. The collection of turns exchanged among participants in the interaction, leads to the existence of a turn-taking system. The turn-taking system is a combination of turn-taking based events, which include sub-events such as silence, overlaps, back-channels, interruptions and simultaneous turns. An example illustrating different possible events in a turn-taking system can be seen in Figure-2.2.

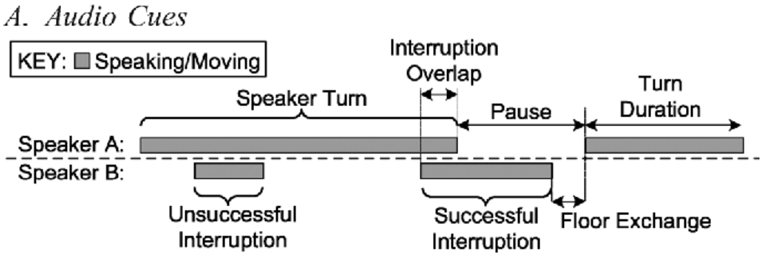


Figure 2.2: An overview of a sample turn-taking system. This illustration is an example illustrating an example turn-taking system in a dyadic setting, between participants - *Speaker A* and *Speaker B*. The illustration 6 particular sub-events in a turn-taking system. This illustration was adopted from [56].

Here below, we explain in detail several key components of a turn-taking system (The concepts explained below are illustrated in Figure-2.2),

1. *Floor Exchange* They are events which lead to a transfer of turns from one individual to another in a social interaction. A shorter floor exchange duration is hypothesised to occur more often in conversations of high quality, with quality conversation flow.

2. *Pause* - They are events which capture the degree of individual participation and conversation flow in social interactions. Pauses can occur at different levels, *Pauses Between Individual Turns*, which measures the individual-level silence event between the individual's consecutive turns, and, *Pauses Between Floor Exchanges* measure the time between exchanges of the floor to see how quickly turns are passed between participants [56].

3. *Back-channels* - Sacks (1982) [106] define them as, "short segments of speech uttered to signal continued interest and understanding without taking the floor". A back-channel response can be verbal, non-verbal, or both. The term was first coined by Victor Yngve (1970) [127]. Back-channels are generally considered to be shorter turns intended to give feedback to the current speaker, and tends to induce turn-overlap. An analysis on the Switchboard Corpus of English telephone conversations by Levinson and Torreira [74] reveals that among all overlaps measured, 73% of them were induced by back-channels.

4. *Overlap* - Overlapping Speech occurs when more than one individual in a conversation group speak at the same time [56]. Overlapping speech can be an indication of both - conflict [123] and of engagement between individuals [113]. Heldner and Edlund (2011) [50], while studying overlaps, found that on average 40% of floor exchange in their corpora involved overlaps, considering only overlaps of over 10ms. Overlap of shorter durations are generally considered to be back-channels.

5. *Successful Interruption* - Levinson and Torreira (2015) [74] differentiate between types of overlaps, where *between-overlaps* (Successful Interruption) refer to overlaps where the floor was transferred without a silent gap between speakers.

6. *Unsuccessful Interruption* - Levinson and Torreira (2015) [74] differentiate between types of overlaps, where *within-overlaps* (Unsuccessful Interruption) refer to the overlapping speech occured in between a speaking turn and did not result in a transfer of floor.

All the components of a turn-taking system, as explained above, have a direct effect on the quality of a conversation and individual experiences in a social interaction. Hence, it is important to consider these concepts while studying the measure of conversation quality.

Speaking turn is the most basic measure to be extracted before extracting other key components and measures of a turn-taking system. Turns or speaking status can be automatically extracted from audio signals. For example, works such as, Hung et al.'s (2008) [57] and Wyatt et al.'s (2011) [126] use a sliding window approach on the audio energy levels do detect the speaking status of a particular individual. Similarly, Alex Pentland (2005) [98] employs an alternative way of detecting speaker activity. A two-level Hidden Markov Model is used to segment the stream of audio into voiced and non-voiced segments. The voiced segments are subsequently grouped into speaking and non-speaking. This was largely based on previous work in which turn-taking behaviour was modelled as a Markov Process [59]. Some works also rely on external annotators to annotate for speaking status using the audio or audio recordings [17].

More recently, Gedik et al. (2016) [38] proposed a methodology to automatically extract speaking status of an individual from triaxial accelerometer readings. With the assumption that body movements might be an indicator of an individual's speaking status, the authors used the readings from a bodily worn (around the neck) triaxial accelerometer to extract speaking status. However, this methodology might be prone to lot of noise from inter-personal difference and other bodily movements unrealated to speaking status. To tackle this, the authors propose a transfer learning based approach, based on Transductive Parameter Transfer to detect speaking status.

Social Constructs A social construct is a dynamic phenomena that exists not in objective reality, but emerges from human interactions and behaviour. It exists because humans agree that it exists. Helmhout et al. (2005) [51] states that, in organizations and small groups,

> *social constructs take the form of, for instance shared stories, shared institutions (behavior rule systems), shared designs, shared plans, and shared artefacts.*

These social constructs support habits of action aimed at cooperation and coordinated behavior. Social constructs are dynamic phenomena which emerge from within a social group depending on the collective behaviours of its individuals in the group, which in-turn affects the behaviour of the individuals. Hence, they tend become multilevel systems [67], where the high-level social construct emerges from the interactions of low-level entities, where the high-level construct in turn affects the interactions of low-level entities [107].

Social constructs are of different types depending on the levels at which the construct manifests in. For example, individual-level and group-level social construct. When studying small groups and FCGs, researchers tend to include individual-level and/or group-level phenomena in their research design. Many constructs of interest to small group scholars are influenced by group process, which means data related to the construct of interest are collected from members following the group interaction [12]. In small groups and FCCGs, the individual-level social construct is a dynamic phenomena results from the individual's interaction with their interaction partners and manifests (exists) only within the perception of the individual. The individual-level social construct can also manifests within the perception of an external observer observing the individual and their interactions. But, at the same time the individual-level construct manifesting within the individual themselves and within the external observer can be qualitatively and quantitatively different from one another. Some example of individual-level constructs are - *Individual Affect*, *Individual Experience*, *Involvement*, *Received Rapport Individual's group cohesion perception* and *Individual's group climate perception*. Similarly, in small-groups and FCGs, a group-level social construct is a dynamic phenomena resulting from the collective behaviour of the group and its members and manifests (exists) over the group as a whole. Similar to the individual-level construct, group-level construct can also manifest within the external observer observing the group and the within group interactions. Group-level constructs are often derived from individual-level data, using a composition model [115]. Some example of group-level constructs are - *Group Affect*, *Group Cohesion*, *Group Climate*, *Team Knowledge*[66] and *Group Interest-levels*[37].

Mierlo et al. (2009) [115], in their research work, study the different ways in which the group-level constructs can be derived from self-reported individual-level constructs. In the organizational sciences, the most common approach is to collect individual survey responses and aggregate those to the group level [115] [65] [84]. The authors present a methodological framework for addressing the distinction between and the baseline psychometric quality of composed group constructs, by empirically analysing the two common composition methods. Firstly, the composition method of *direct consensus composition*, where group-level constructs are considered as *compositions of individual-level* data, e.g. aggregation of individual level ratings on perception on their own self, and secondly, the composition method of *referent-shift consensus composition*, where group-level constructs are considered as *compositions of group-level* data, e.g. aggregation of individual level ratings on perception on the whole group.

2.2. Aspects of Conversation Experience

As discussed in the previous sections, the analysis of conversations and social interactions has been handled by researchers in terms of several events within the conversation. Similarly, we also discussed the individual- and group- level constructs which emerge from within. Several researchers in the SSP community have successfully attempted to model such social constructs like bonding[60], experience[79], rapport[85], cohesion[88] and involvement[92], using a myriad of nonverbal cues. The quality of a conversation, a similar social construct, is a highly subjective measure and intends to measure the overall quality of the conversation. At the same time, in this research, we specifically focus on enabling feedback systems. With that said, it is not just enough to provide feedbacks on a specific aspect of the conversation, rather, a feedback on more broader measure which covers different aspect of a conversation is required. Hence, to define and design such a measure, requires us to look at a conversation in a broader manner, by including several facets of a conversation to define it's quality. This has motivated us to study the measure of *Conversation Quality* in a more inclusive manner, covering different aspects of a social interaction. Hence, the measure of conversation quality draws inspirations from several important aspects of conversations studied in literature.

In this section, we present several aspects of conversation experience studied in literature, which are key in defining the quality of the conversation. This research work treats conversations the same way Cuperman and Ickes (2009) [28] have handled perceptions in social interactions. Cuperman and Ickes, used the unstructured dyadic interaction paradigm to examine the effects of gender and the Big Five personality traits on the members' behaviors and perceptions of the interaction. For this purpose, the authors introduced the *Perception of Interaction* (POI) questionnaire to collected self-reported measures of a participant's perception of the conversation quality. This questionnaire contained 27 items that required the participants to rate their interaction experience, with respect to several aspects of the conversation. These aspects covered by POI include aspects such as, *quality of the interaction* (e.g., "To what degree did the interaction seem smooth, natural, and relaxed to you?"), the *degree of rapport* they felt they had with the other person (e.g., "To what degree did you feel accepted and respected by the other person?"), and the *degree to which they liked* the other person (e.g., "To what extent would you like to interact more with the other person in the future?"). This measure of interaction has been successfully adopted by several other research to study the bonding and interaction experience in a human-agent interaction, such as Jaques et al. (2016) [60] and Cerekovic et al. (2014) [22].

Similar to Cuperman and Ickes [28], Lindley and Monk [79] studied several behavioral process measures of conversational equality, freedom, fluency and enjoyment to develop the *Thin Slice Enjoyment Scale* as a measure of experience and empathised enjoyment in social conversations. The thin-slice scale specifically concentrates of four aspects of a social conversation - *conversation equality, conversation freedom, conversation fluency* and *conversation enjoyment.* Inspired by the work on cooperative floor by Edelsky [31], the authors identified a number of quantitative measures which resonate particularly strongly with Edelsky's (1981) views on cooperative floors of social interactions. The identified measures were used to measure different aspects of conversation. Specifically, the authors adopted the measures of conversational equality and freedom [19][79], measures of conversational fluency through frequent turns, turn overlap and turn duration [29], and measures of turn synchronisation [122], seem to develop the thin-slice scale to measure experience and empathised enjoyment in social conversations.

Looking at the aspects of conversation studied by [28] and [79], we can realise that the aspects considered tap into similar spectrums of conversational and individual experience. For example, the quality of interaction aspect studied by [28] constitutes of the conversation equality and freedom studied by [79]. At the same time, the POI study by Cuperman and Ickes [28] is very similar to our study of Conversation Quality, which also attempts to study the overall quality of conversations with respect to individual experiences. On the other hand, the Thin Slice Enjoyment study by Lindley and Monk [79] studies individual experiences with a direct focus on the turn-taking system of the interaction. Nevertheless, both these works are important to draw inspirations from while studying the overall quality of a conversation.

The following sections are categorised with respect to the three aspects of perception of interaction, suggested by Cuperman and Ickes [28]. A literature review on how different researchers have handled and studied these three aspects of a conversation is presented below.

2.2.1. Quality of Interaction

This particular aspect of a social interaction or conversation captures an individual's experience in the interaction with their partner. For example, the degree to which the individual's interaction was smooth and relaxed or the degree to which the individual's interaction was forced and awkward. In social sciences literature, measures of conversational equality and freedom [19][79], seem to capture well an individual's experience in the interaction.

Conversational equality is one concept based on turn-taking systems which well explains ones experience in an interaction (quality of interaction). The measure is based on the amount of time that individuals within a group talk for, and how evenly this is distributed within the group [79]. Lindley and Monk [79], in the process of measuring social experience, adopt the measure of conversation equality proposed by Carletta et al. (1998) [19], which is purely based on the even distribution of talk-duration among individuals in the group. The mathematical representation of the equality of conversation used by Lai et al.,(2013) [71] to model participant affect in meetings is as follows,

$$P_{eq} = 1 - \frac{\sum_i^N (T_i - T)^2 / T}{E} \tag{2.1}$$

where N is the number of participants, T_i is the total turn-duration for individual i, $T = (\sum_i^N (T_i)/N)$ (i.e. equal participation). E, represents the maximum possible value of the term under the sum - the average distance from equal participation (so E represents the case when only one participant speaks for the entire meeting). Values closer to 1 indicate greater equality.

Such a measure of conversation equality to model quality of interaction makes a strong assumption that an individual has a good experience only if he shares an equal talk-duration as other individuals in the group. At the same time, works such as [85], [46], [93], modeling rapport and involvement, have shown that there exists more social cues other than talk-duration to fully explain an individual's experience in an interaction. Oertel et al. [93] introduced the measure of equal distribution of *mutual-gaze* to model involvement in FCGs. The authors calculate mutual gaze to be the proportion of the duration in which a two individuals simultaneously look at each other. The results presented shows a very high correlation between mutual gaze and involvement. To include mutual-gaze as the measure of involvement, the hypothesis behind it is that when individuals in the conversation have an equal distribution of mutual-gaze with all other individuals, it implies a high quality of interaction.

Mutual-gaze as a social cue, has been successfully used by several other researchers to model different facets of social behaviour, such as, human-robot engagement [110], emergent leaders [10] and interaction quality [9]. Notable is the work done by Berry et al.[9], who studied the quality of interaction in female dyads. The authors specifically study the correlation between the big-5 [54] personality traits, several nonverbal behavioural cues and the quality of interaction. The authors manually coded 18 nonverbal behaviors that were likely to reflect the primary relational dimensions of involvement, differential status, dominance, and emotional valence. In particular, postural cues such as gestural frequency, body openness, and body orientation, and gaze-related cues such as gaze initiations, gaze terminations, amount of direct gaze were coded. The correlation study showed a strong link between nonverbal behaviors, and observer ratings of interaction quality. For example, gaze-related cues and body openness were positively correlated with the perceived interaction quality with a partial correlation of r=+0.27 and p<0.05. From successful works in literature, we see that it is important to design the aspect of conversation equality to include all the measures of equal distribution of talk, gaze-related cues and body pose and orientation. While conversation equality and mutual gaze reveals the involved, smooth and relaxed interaction of individuals in the interaction, cues like body pose and gestures reveal the degree of awkwardness, self-consciousness and uncomfortableness of individuals in the interaction.

Another measure of quality of interaction, similar to that of equality is *Conversation Freedom*. Conversational freedom is the measure derived from patterns of turn taking within a group, and indicates how frequently individuals within it take turns in the conversation immediately after specific other group members [79]. For example, if person Y only speaks after person X and never after person Z, freedom will be low. Lindley and Monk [79], in the process of measuring social experience, adopt the measure of conversation equality proposed by Carletta et al. [19], which is a direct measure of the interactivity of the group conversation. A higher degree of conversation freedom experienced by individual leads to a higher higher quality of conversation and the experience of the individual in itself.

Measures of freedom in turn-taking patterns in the conversation have been adopted by different researchers, and successfully used to model social behaviour in different scenarios from spontaneous conversations to task-related meetings and group-discussions. For instance, Muller et al., [85], used the probability of an individual taking the turn at a turn transition - ProbTurn/TurnTrans, to predict the received rapport of the individual, in a spontaneous conversation scenario. The results presented shows that such a measure of speech activity achieves an average precision (AP) of 0.44. This is a relatively good predictor of received rapport, when in comparison with hand motion which achieves 0.37 AP, prosodic feature which achieves 0.30 AP and the baseline (random prediction) which achieves 0.25 AP. Lai et al., [71] modelled participant's affect in meetings, by employing a mathematical equation of conversation freedom, similar to equation-2.1. The equation used by Lai et al., is as follow,

$$F_{cond} = 1 - \frac{H_{max}(Y|X) - H(Y|X)}{H_{max}(Y|X)} \tag{2.2}$$

where, $H(Y|X)$ is the conditional entropy of speaker Y being the next participant to speak after participant X's turn, with $H_{max}(Y|X)$ representing the maximal possible value for groups of a given size. So, F_{cond} is 0 when turn-taking follows a strict order, i.e. only speaker y follows x, and is 1 when every speaker follows everyone else in equal proportion. The measure of conversation freedom is similar to the free-for-all concept prevalent in cooperative floors of interaction studied by Edelsky [31]. In such a free-for-all floor there is a sense that no one owns the floor, it being perfectly acceptable for everyone to talk at once. The ability of interaction partners to uphold this sense of free-for-all floor has a direct link on to the quality of the interaction and the individual's experience in the conversation. Several literature has shown successful usage of the measure of conversation freedom and its implications on group constructs and individual experience [70][71][87][79].

With that said, from literature, we realise that the *quality of interaction* has been extensively studied in different forms, using different social cues and modeling techniques. Hence, it is important to draw inspirations from works studying aspects of interaction such as *individual experience*, *free-for-all*, *equality* and *freedom*, to design the measure of *Quality of Interaction*.

2.2.2. Degree of Rapport

Rapport is the close and harmonious relationship in which interaction partners are "in sync" with each other [85]. Rapport, is widely acknowledged to result in smoother social interactions, improved collaboration, and improved interpersonal outcomes. Cuperman and Ickes [28] defined this aspect of a conversation to be the degree of rapport an individual in the interaction feel towards the other interaction partners in the social interaction. For example, the degree to which an individual was accepted and respected by other individuals in the group or the degree to which the other individuals were paying attention to the individual. In social sciences literature, measures of conversational fluency through frequent turns, turn overlap and turn duration [29], and measures of turn synchronisation [122] seem to capture well the rapport shared among individuals in the social interaction.

Conversation Fluency is the measure which indicates a group's ability to coordinate speaking turn. The coordinated turn-taking system enables the fluency and continuity of natural conversation and entails regulation of the timings of turns at talk [52]. Entrainment, the phenomenon of dialogue partners becoming more similar to each other, is widely believed to be crucial to dialogue coordination [76]. At the same time, measures of conversation fluency seem to resonate particularly strongly with Edelsky's description of cooperative floors (same wavelength and free-for-all) [29]. Lindley and Monk (2008) [78] use a measure of the number of turns in the conversation, which measures how interactive turn taking is, as an indicator of conversational fluency. More recently [79], in the process of measuring social experience, the authors combine number of turns, turn-duration, turn-overlap and turn-synchronisation, to measure conversation fluency. The fluency of a conversation is vital in studying the quality of a conversation as the ability of the conversation partners to hold a fluent conversation directly implies a higher degree of rapport and thereby a higher quality of conversation.

Levitan et al., (2011) [76], studied the entrainment in speech preceding backchannels. The results show that entrainment increases over the course of a dialogue and is also associated with measures of dialogue coordination and task success. Levitan et al., [76], used two measures of dialogue coordination, namely periods of silence and interruptions. Long latencies (periods of silence) before backchannels were found to be a sign of poor coordination. Similarly, interruptions signal poor coordination, as when a speaker has not finished what he has to say, but his partner thinks it is her turn to speak.

Turn overlap, measured as the amount of time that group members engage in overlapping turns, is often taken as an indicator of low conversational fluency, and indeed, groups that are unable to coordinate turn taking may experience overlapping speech as disruptive or frustrating. However, in Edelsky's description of the cooperative floor [31], it is emphasised that the floor belongs to no one, and that everyone may talk at once. Hence, overlapping speech can be treated as periods of dominance assertion [123] and also collaboration [113]. Similarly, Hung et al., [56], present an interesting result where the *TotalOverlap* feature was positively correlated with cohesion. The results conclude that more turn-overlap is a reliable sign of high cohesion, in the AMI corpus [20]. This also aligns with findings in social psychology that interruptions are indicative of good rapport such as when people are able to finish each other's sentences [113].

Similar results have been obtained by Lindley and Monk [79], where the number of turns correlates significantly with turn overlap, lending support to the idea that overlapping turns might also be taken as an indicator of conversational fluency. To over come this uncertainty, the consider turn-overlap alongside an additional measure of turn synchronisation [122]. Turn synchronisation is calculated as the expected proportion of overlap minus the observed proportion, and so a higher score indicates a more coordinated conversation. Overlap and synchronisation thus reflect potentially different aspects of the conversation and synchronisation is to be preferred as a measure of coordination that is not confounded with the amount of time that group members spend talking [79].

Literature has an abundance of research with respect to coordination of conversations, its resulting fluency and its relationship with rapport shared among conversation partners. At the same time, an individual's received rapport has a direct effect on the individual's experience in the conversation, making it a vital aspect of conversations to be studied. Hence, degree of rapport becomes one of the most important aspect of a conversation involved in designing the measure of *Conversation Quality*.

2.2.3. Degree of Likeness

Cuperman and Ickes [28] defined this aspect of a conversation to be the degree to which an individual liked the other individuals in the social interaction. For example, the extent to which an individual would like to interact more with their interaction partners or the extent to which an individual liked the other individuals in the interaction. Several research works such as Jacques et al.'s [60] work on human-agent bonding and Lindley and Monk's [79] show that degree of likeness and the degree of rapport in an interaction are inter-linked with one another. Below, we will review some research works related to modeling constructs like likeness, enjoyment and involvement of individuals.

Lindley and Monk [79] in their work, to measure the *Empathised Enjoyment* design a composite score based on *Thin Slice Enjoyment Scale.* Various process measures of group behaviour and experience was also explored along with the new rating scale of empathised enjoyment. To test the validity of the *Thin Slice Enjoyment* rating scale (annotated by external annotators using video clips), the authors used the process measures (e.g. number of turns, mean turn duration, turn-overlap duration, number of smiles, turn freedom, turn-synchronisation etc..,) to test their correlations with that of the rating scale. The correlation results obtained, shows some interesting insights into the link between the process measures used and empathised enjoyment. For instance, increasing synchronisation significantly correlates with empathised enjoyment i.e. coordinated conversations are rated as more enjoyable. One of the key conclusions presented by the authors is that enjoyment and likeness of individuals is indicator of good coordinated conversations.

On the other hand, instead of studying enjoyment based on coordination of conversations, Lingenfelser et al., (2014) [80] define enjoyment events as the ones displayed explicitly by an individual in the form of behavioural cues. Lingenfelser et al., (2014) [80], present a real-time modality fusion approach to detect enjoyment-episodes within the audiovisual *Belfast Story-Telling Corpus* based on the 16 *Enjoyable Emotions Induction Task* [53]. Lingenfelser et al., define enjoyment segments as, "*an episode of positive emotion, indicated by visual and auditory cues of enjoyment, such as smiles and voiced laughters*". Though conclusions in [79][78] show that likeness and enjoyment are directly related to coordinated conversations, results in [80] show that positive emotion and enjoyment are also indicated by multi-modal (audio-visual) cues such as laughter and voice.

Proxemic based features such as body orientation information has also been used to model *degree of likeness* based constructs such as social involvement and interest-levels. Existing work in psychology and social sciences literature has shown a strong relationship between the non-verbal cue of lower-body orientation and social involvement [64] [15]. Building on this motivation, Vasquez et al. (2015) [119] propose an alternating optimization procedure that estimates lower body orientations and detects groups of interacting people. Similarly, Hung and Zhang (2016) [128] propose a method to detect f-formation associates, using body orientation alone as a discriminative proxemic feature. F-formation associates are individuals who are part of an f-formation but do not have full member status and therefore have a lower degree of involvement. A relationship between involvement and body orientation in turn implies a possible relationship between body orientation and conversation quality.

Another social construct very much similar to *degree of likeness* is *interest-level*. Gatica et al. (2013) [37], define group interest as the degree of engagement that meeting participants display as a group during their interaction, as perceived through the audio and visual modalities by an external observer. The authors use proxemic based *body pose* as one of the discriminative features to classify high and neutral group interest-levels. Building on [37]'s works on interest-levels, Kapcak et al. (2019) [61] propose a methodology to estimate romantic interest-levels and attraction in a dyadic speed date setting, using features based on bodily coordinations, extracted from body-worn tri-axial accelerometers. The results show that romantic, social and sexual interest (or attraction) can be predicted using coordination features such as synchrony and mimicry.

With that said, from literature, we realise that the *degree of likeness* has been extensively studied in different forms, using different social cues and modeling techniques. Hence, it is important to draw inspirations from such works, which study the likeness of an individual in forms of *enjoyment*, *involvement*, *interest-level* and *attraction*, to design the measure of *Degree of Likeness*.

2.3. Modeling Conversation Quality

In the previous sections, we discussed an array of research works which focused on studying different aspects of a social conversation. In this section, we specifically concentrate on how conversational groups are modeled in literature. In this review, the following topics will be discussed,

- What are some of the social behavioural cues used to capture and quantify an individual's behaviours in groups?
- How high-level group constructs are modeled using the low-level behavioural cues?

In this section we present three unique, widely used, non-verbal behavioural cues that has been good indicators of several social constructs. Specifically, we discuss *turn-taking*, *prosody* and *coordination* based cues. Finally, in this section, we also discuss how group-level constructs are handled and modeled, using the behavioural cues extracted from individuals.

2.3.1. Social Signals

Burgoon et al. (2017) [16] defines a *social signal* to have properties like, they are observable behaviours that individuals display during a social interaction, they reveal the intention of an individual and produces change in the behaviour of the individual's interaction partners, and the changes produced on interaction partners are not random, but, follow certain principals and laws. Few examples of social signals include attention, empathy and dominance.

With social signals defined above, *behavioural cues* are nothing but the behaviours displayed by individuals as part of the social signal. Behavioural cues are the most fundamental concept in the modeling of human behaviour in the field of SSP. Almost all modeling techniques proposed in by research in the field of SSP, rely on extracted features based on social signals, backed by research works in psychology. Behavioural cues can be broadly classified in to verbal and non-verbal cues. Predominantly, works in literature concentrate on modelling social constructs using non-verbal cues. Some of the examples of such non-verbal behavioural cues include turn-taking, prosodic, body pose and gestures. The link between social signals and non-verbal behavioural cues is illustrated in Figure-2.3.

Turn-Taking The turn-taking system prevalent in a social interaction and its implications on individual experience was discussed earlier in Section-2.1.2. In this section, we will see works which have successfully used turn-taking to model group-level behaviour.

Turn-taking and its relevance towards social constructs is one of the most widely studied concepts in research fields of social sciences and SSP. Turn-taking based features have been used to model social constructs from user-satisfaction in Spoken Dialog Systems (SDS) [24] and engagement in conversations [94][98] to cohesion [56] and entrainment [77].

Chowdhury et al. (2016) [24], propose an approach to model *user satisfaction* as it unfolds, in an SDS. The authors use turn-taking based features as one of their set of features to classify user-satisfaction as postive, negative or neutral. The authors use Inter-Pausal Units (IPUs), consecutive positive speaking status between 50ms of pausal gaps, as speaking turns to extract other turn-taking features. In specific, the authors extract turn-taking features such as *participation equality*, *turn-taking freedom*, *percentage overlap* , *competitive overlaps*, *non-competitive overlaps*, *turn-duration* etc., to model user-satisfaction. The comparative performance analysis results shows that the contribution of the turn-taking features outperforms both prosodic

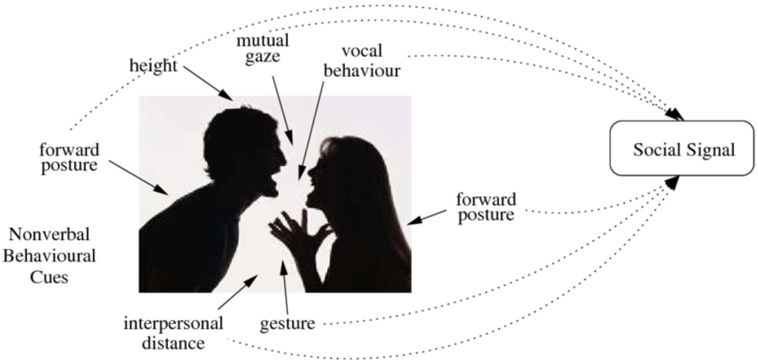


Figure 2.3: Examples of non-verbal behavioural cues. Social Signals are the collection of several behavioural cues. The illustration is adopted from [117]

and lexical features, highlighting the capability of turn-taking feautures. Further analysis reveals that, the use of non-competitive turns increase the chance of a positive user experience, whereas competitive turns tend to decrease the chance of positive experience.

Similar to [24], Oertel et al. (2015) [71], propose a technique to estimate a participant's affect in meetings, by using turn-taking and lexical features. The authors extracted turn-taking features such as *participation equality*, *turn-taking freedom*, *participant speaking time* and *number of back-channels* (less than 500ms of continuous speaking activity) to model participants' affect. The results indicate that participants gave higher satisfaction and cohesiveness ratings to meetings with greater group turn-taking freedom and individual very short utterance rates, while lower ratings were associated with more silence and speaker overlap.

One of the first works to model team cohesion using turn-taking features was from Hung et al. (2010) [56], who proposed an array of non-verbal turn-taking based features to quantify cohesion in small groups (e.g. meetings). An extensive feature extraction task was performed to extract *group-level* turn-taking features which can be broadly classified into *Pauses Between Individual Turns*, *Pauses Between Floor Exchanges*, *Turn Lengths* and *Overlapping Speech*, to model team cohesion. While analysing feature importance, it was revealed that *TotalPauseTime* (90% accuracy), *MaxOverlapSpeakingRate* (89% accuracy) and *MinMeanTurnDuration* (87% accuracy) were the top performing features. An interesting result was also obtained with the *TotalOverlap* feature, where *TotalOverlap* was positively correlated with cohesion levels. This aligns with findings in social psychology that interruptions are indicative of good rapport such as when people are able to finish each other's sentences [113].

From the literature review, we see that turn-taking has been considered to be the state-of-the-art technique to model several social phenomena. And turn-taking's potential to explain several complex social phenomena looks promising. Turn-taking features have been used as baselines for comparisons by researchers, to validate their methodology [129][61]. With that said, it is important to also include turn-taking features for the study of conversation quality. A statistical study to analyse relationship between turn-taking features and the measure of conversation quality will be a solid contribution to the research community.

Prosody In linguistics, prosody is concerned with those elements of speech that are not individual phonetic segments (vowels and consonants) but are properties of syllables and larger units of speech, including linguistic functions such as intonation, tone, stress, and rhythm. Prosody based features have been used to model social constructs from rapport [85] and group performance [87] to cohesion [88] and entrainment [75].

Rapport is a complex social construct which only manifests in subtle non-verbal signals. Muller et al. (2018) [85] model the phenomena of rapport using a rich set of non-verbal signals which includes speech prosody. Here, the authors used a set of 768 prosodic features based on IS09 challenge feature set from openSMILE [33]. The features are extracted from individual segments when the participant speaks, and then aggregated over all segments of a speaker by taking the mean and the standard deviation, resulting in 768 features. With a baseline model performance of average precision - 0.25 AP, the prosodic features out-performed the baseline with an average precision of 0.30 AP. The results also show that prosody when fused with facial expression features drops in performance, only matching the baseline at 0.25 AP.

Murray et al. (2018) propose a methodology to automatically predicting group performance using acoustic features extracted from the speech signal, and linguistic features derived from the conversation transcripts. Similar to [85], to extract prosodic features, the authors use the openSMILE software. But, the authors use only a subset of features (76 speech features) which includes mel-frequency cepstral coefficients (MFCCs), associated delta features, jitter, shimmer, PCM loudness, F0 envelope, F0 contour, voicing probability, and log power of Mel-frequency bands. The results shows us that the best-performing models utilize both linguistic and acoustic features, and that linguistic features alone can also yield good performance on this task.

Nanninga et al., (2017) [88] proposed a novel method for estimating task and social cohesion by quantifying the dynamic alignment of Paralinguistic speech cues. Paralinguistics is also similar to speech-based prosodic cues. By solely extracting features only related to paralinguistic mimicry (dynamic alignment) on a 2-min time window, the authors were able to achieve a prediction performance (social cohesion) of 0.71 area under the ROC curve performing on par with the state-of-the-art where turn-taking features were used. At the same time, the proposed method outperforms the state-of-the-art method with a prediction performance of 0.64 area under the ROC curve while predicting task cohesion. The results presented by the authors here, are initial findings in the capabilities of prosodic mimicry features.

Though prosody and its importance in modeling social behaviour is widely recognised by several researchers in SSP domain, it requires availability of audio data (atleast low-frequency audio). And unlike turn-taking features, it cannot be annotated for, and requires to be collected during the data collection process. Nevertheless, prosodic cues as potential features for modeling social behaviour cannot be ignored.

Coordination and Congruence More recently, researchers from the field of SSP have also successfully used several other non-verbal features such as gaze, prosody and activity-levels [91] [98] [92] [86] [37] [69], to model group constructs. Coordination and Congruence between interacting partners take up several unique forms and can be measured in different ways. For example, synchrony, a form of coordination, can measured using correlation, mutual information and mimicry, and convergence, a form of time-lagged coordination, can be measured using the correlations of individual behaviour with that of the temporal factor, thereby capturing the time evolving attempt to synchronize or coordinate.

Synchrony is the dynamic and reciprocal adaptation of the temporal structure of behaviors between interactive partners [30]. Rapport building, the smoothness of a social conversations, and cooperation efficiency are closely linked to the ability to synchronize (similar phenomenons include - mimicry, social resonance, coordination, entrainment, attunement, chameleon effect, etc.) with a partner [30]. Nanninga et al.'s (2017)[88] work to model group cohesion using paralinguistic mimicry, Cambell's (2009)[18] work in modeling of active listening and engagement using synchrony in discourse, Levitan et al.'s (2011)[76] study on back-channeling entraiment's correlation with dialogue coordination and task success and Kapcak's (2019)[61] work to model attraction using bodily coordination are few examples of successful implementation of mimicry and synchrony based social behaviour modeling. These works have successfully linked synchrony with task-related outcomes. Delaherche et al.'s (2012) [30] review on interpersonal synchrony presents more such works, along with an in-depth analysis on the functions and the measures of synchrony.

More importantly, *mimicry* and *synchrony* based features have been has proved be better predictors of several social constructs in different social scenarios, in comparison with the state-of-the-art turn-taking features [88] [61]. With that said, studying the implications of coordinations on the measure of conversation quality is something which we cannot ignore.

2.3.2. Modeling group-level Behaviour

In the previous sections, we had discussed the social signals and behaviour cues involved in the modeling of social behaviour. Much of these cues are fundamentally individual-level cues and to describe a group-level construct, we will have to extend these individual-level cues to form group-level descriptive features. For example, synchrony is a window into pairwise and dyadic alignment and a group-level synchrony measure should be derived from the pairwise measures to describe the overall group's synchrony/coordination. Researchers, while modeling group-level constructs have extracted the overall group's feature using several assumptions on the composition of groups. In this section, we review two such assumptions and techniques, *Groups as a collection of individuals* and *Groups as a collection of pairs*.

The first and the most basic approach is to consider a *group as collection of individuals*. In this approach, features and cues are extracted for each individual in the group and the group-level descriptive features are computed as the aggregation (statistical aggregations) of individual-level features. This approach makes an assumption that the group-level descriptor is an equally weighted linear combination of individual-level descriptors. This approach is aligned with the socio-evolutionary perspective, which treats groups as aggregate of individuals and views group behavior as the product of individual behaviors that scale up to the group level [99]. The statistical aggregations (for group-level descriptors) are generally calculated as the sum, mean, variance, median, min and max over all the individual-level features. A more data-driven approach to this assumption is the social network perspective which uses individual-level attributes and social network properties as inputs and treats group-level descriptors as the result/output of a linear function over the individual-level attributes and social network properties.

Such an analysis of both the perspective explain above were studied by Olguin et al. (2009) [95]. Olguin et al. extracted individual-level descriptors such as speech activity features and showed that these low level descriptions of individual behaviour can not only be used to capture their own behaviour, but also that of the group-, organization- level descriptors. The study was performed using a correlation analysis on the results of the multilinear regression model. While the authors could not predict the individual perception of job performance from sensor data for each individual, they were able to estimate the overall group performance by aggregating (socio-evolutionary perspective) the daily sensor features across subjects.

The first and a more recent approach is to consider a *group as collection of pairs or dyads*. Due to the fact that synchrony is a window into pair-wise and dyadic alignment, researchers working on synchrony-based features tend to follow this approach [88]. Similar to the statistical aggregation followed in the previous approach, in this approach as well, the statistical aggregations (for group-level descriptors) are generally calculated as the sum, mean, variance, median, min and max over all the pairwise features. In the research work by Nanninga et al. (2017) [88], the authors use aggregation functions such as min, max, mean, median, variance and range. The assumption here is that the aggregations calculated as the group-level descriptors explain the overall distribution of pairwise features for the group, and might be descriptive of group-level constructs. One possible advantage of this approach is that, same modelling technique can be used for groups of all sizes. At the same time, this approach compromises and disregards the fact that group behaviours vary with respect to group sizes. Hence, this approach may not help our model to generalise across group sizes. One way to handle this was demonstrated by Gedik et al. (2018) [39]. The authors to detect f-formations, learn different models for different group size, thereby also capturing the effect of group-sizes on group-level behaviour.

2.4. Similar Social Constructs

Social conversation and interaction have been studied extensively by the research community, with respect to the different constructs which emerge from with the social groups. In this research, we concentrate on one such construct - the *Conversation Quality*. At the same time, it also important to review other constructs studied in literature, which might be closely linked to the quality of a conversation. Hence, in this section, we review works in literature performed on such similar social constructs. In the subsequent sections, we review three such unique social constructs - *Involvement*, *Rapport* and *Cohesion*. The following sections will be categorised with respect to the three constructs.

2.4.1. Involvement and Engagement

Gatica-Perez et al., (2013) [37] define group involvement or interest-levels as, "the perceived degree of interest or involvement of the majority of the group". The authors used a public corpus consisting of 50 five-minute, four-participant meetings (scripted for turn-taking patterns) recorded using three cameras and 12 microphones, to study interest-levels in meetings. To classify high and neutral group interest-levels, the authors extracted three set of features - audio-only (activity, pitch etc..,), video-only (skin-color head, orientation etc..,) and audio-visual fused. To automatically detect segments of high interest-levels, the authors used single- and multi-stream Hidden Markov Models (HMM). The results presented, prove the successful detection of group interest using fused audio-visual cues. Furthermore, the work provided an initial result on the capability of modality fusion for performance improvement.

John H. Antil (1984) [3] define involvement as "the level of perceived personal importance and/or interest evoked by a stimulus (or stimuli) within a specific situation". Following this definition, Oertel et al. (2011) [93] study participants' degree of involvement in social conversations. The authors model the degree of involvement as a scalar phenomenon categorised under 4 levels of involvement. The authors study different levels of involvement with respect to an array of prosodic and body movement based behavioural cues. The initial results from statistical tests, presented by the authors, show that, voice span and intensity and body and face movement features are reliable cues for the detection of distinct levels of participants involvement in conversation.

Similar to [93], Oertel et al. (2013) [91], propose a methodology to model individual-level engagement and group-level involvement, and also study the relationship between the two social constructs in a 8-party corpus. It is important, at this point, to note that Oertel et al. in [91], have similar goals as our research here, where we aim to model individual-level experience and group-level conversation quality and study its relationship. Hence, it is important to draw inspirations from this particular work. The authors solely rely on eye-gaze based features, such as *presence, entropy, symmetry and max gaze*, to capture individual-level behaviour in social interaction. The individual-level gaze based cues were used as a proxy to quantify individual-level engagement. The results presented show that the individual-level engagement features were able to predict group-level features with a prediction accuracy of 70%.

Similar to [91], Bednarik and Hradis (2012) [7], also investigated conversational engagement in multi-party conversation, emphasising their analysis based only on gaze based features. The authors investigate 6 different levels of engagement, namely, no interest, following, responding, conversing, influencing and managing, and their relationship with different gaze based cues. The levels of engagement were annotated by at least two annotators for a 15 second intervals. Unlike [37], the authors here, use an annotation scheme that intends to measure the individual-level engagement within a group, rather than a group-level construct as a whole. The results presented, shows that the levels of engagement were classified between high or low levels with an accuracy of 74%, by relying on the previously extracted gaze-cues. Furthermore, the authors also describe the specific nuances of the gaze features with respect to distinct levels of engagement. For example, the authors distinguished between number of switches between faces, number of unique faces attended and number of faces attended across different levels of engagement.

2.4.2. Rapport

Rapport, is widely acknowledged to result in smoother social interactions, improved collaboration, and improved interpersonal outcomes. With degree of rapport as one of the key aspects of *Conversation Quality* and *Perception of Interaction*, as studied in [28][60][22], rapport is closely linked to the construct of *Conversation Quality* and requires an extensive literature review in it.

Müller et al. (2018) [85], define *Rapport* as, "the close and harmonious relationship in which interaction partners are “in sync” with each other". The authors in their research of rapport, model and classify participants' *received rapport ratings* into low versus medium-to-high rapport. Here, received rapport ratings are collected from each participant who rated *other individuals* in the group in-terms of rapport, using the questionnaire in [8]. The authors use a rich set of non-verbal cues for classifying rapport, namely facial expressions, hand motion, gaze, speaker turns, and speech prosody. The results and feature importance analysis reveals that facial expression features are the best performing features with an average precision of 0.7. The results also show that using features of participants' personalities can even achieve early prediction without a drop in performance.

Features related to synchrony and congruence has also been used to estimate *rapport*. In a dyadic interaction setting, Hagad et al. (2011) [47] predicted levels of rapport through automatic detection of posture congruence (similar to mimicry). For the detection of posture, a Support Vector Machines (SVM) classification model was deployed on *Histogram of Oriented Gradient* (HOG) features extracted from head and arm regions. The congruence features were extracted based on the similarity of the detected postures of interacting participants. Four such congruence defining features are extracted, namely, *Time spent in congruent postures*, *Time spent not in congruent postures*, *Number of posture changes* and *Posture changes that led to congruence*, where congruence information are calculated. Various machine learning techniques were used to create rapport models. Among the models used, SVM and Multilayer Perceptron performed best, at around 71% average accuracy.

2.4.3. Social Cohesion

Cohesion is generally considered as the group members' inclinations to forge social bonds, resulting in the group sticking together and remaining united [21]. Team cohesion has been directly linked to team effectiveness and productivity. Thus, the automatic estimation of team cohesion can be a useful tool to assess meeting quality and team functioning. Literature recognises team cohesion as a combinations of *social* and *task* team cohesion. While, social cohesion refers to the attractiveness of group members towards each other, task cohesion reflects the extent to which a team is united in trying to reach its goals.

One of the first works in team cohesion was from Hung et al.(2010) [56], who proposed an array of audio-visual non-verbal turn-taking based features to quantify cohesion in small groups (e.g. meetings). The proposed model was deployed on the AMI meeting corpus, annotated (by a group of annotators) for cohesion using a psychology literature questionnaire. Based on the results of the kappa measure [26] (inter-annotator and dataset agreement), 61 points with a score greater than 0.3 was used for the subsequent experiments for estimating team cohesion. An extensive feature extraction task was performed to extract *turn-taking* (state-of-the-art for estimating cohesion) features broadly classified as *Audio, Video* and *Audio-Video* features. A supervised binary classification problem was solved using a linear kernel *Support Vector Machine* (SVM) coupled with sampling techniques (to fix class imbalance) and cross-validation. The authors report results of - 90%, 83%, 82% achieved by audio cues, video cues and audio-visual cues respectively. Thus, showing that automatically extracted behavioral cues (audio-visual) can be valid low-level features to estimate perceived levels of cohesion in meetings.

Nanninga et al., (2017) [88] proposed a novel method for estimating task and social cohesion by quantifying the dynamic alignment of nonverbal behaviors in speech - *Paralinguistic Mimicry*. The authors address the usage of two different types of paralinguistic mimicry features - *synchrony* (defines relative similarities) and *convergence* (defines converging behaviour getting more similar in time), which are group-level features extracted from pairwise mimicry metrics (forming a pair-wise mimicry matrix). A binary classification problem for task and social cohesion was handled using a *Gaussian Naive Bayes classifier*. By solely extracting features only related to paralinguistic mimicry (dynamic alignment) on a 2-min time window, the authors were able to achieve a prediction performance (*social cohesion*) of 0.71 area under the ROC curve performing on par with the state-of-the-art where turn-taking features were used. At the same time, the proposed method outperforms the state-of-the-art method with a prediction performance of 0.64 area under the ROC curve while predicting *task cohesion*. The results validated the hypothesis that mimicry based behavioral features can be good indicatives of social and task cohesion.

Zhang et al. (2018) [129] proposed a methodology (Team Sense) to predict the group cohesion (as a binary classification problem) using group-level features which are aggregations (e.g. mean, median, min and max) of individual-level affect features (extracted using dyadic interaction and individual activities). The authors deployed three supervised learning models - linear kernel SVM, random forest and a logistic regression to detect individual affects, with the SVM results used for detecting group-level cohesion. To detect social and task cohesion, the same three supervised models were used with a combiner, resulting in a Majority Classifier (MC). The Majority Classifier was able to predict social cohesion with an AUC of 0.65 and predict task cohesion with a AUC of 0.80. These results validated the hypothesis that behavior features extracted from individuals' wearable data can be aggregated to group level and are effective in assessing group cohesion.

2.5. Concluding Remarks

In this chapter, existing research works related to the study and modeling of several aspects of social interactions, with respect to conversational and individual experiences, was reviewed. Additionally, we also reviewed how different nonverbal social cues are highly informative of several subtle social constructs. Finally, we also saw how several social constructs, similar to that of Conversation Quality, has been studied by researchers in the field of SSP.

Spontaneous interactions are multi-faceted and there exists a wide set of factors influencing them. Nevertheless, several social psychologists have operationalised the study of such interactions with different concepts. Particularly notable is Edelsky's work on cooperative conversation floors [31]. Spontaneous interactions are cooperative floors of interactions by nature, and Edelsky's work [31] can be a potential starting point to design the measure which quantifies such spontaneous interactions. At the same time, to quantify such a multi-faceted social phenomena, it is also important to consider different aspects of the interaction. While current works in existing literature tend to study a particular construct in interactions, an attempt to study the overall quality of conversation in spontaneous interactions is still a knowledge gap. Works such as [79] and [28] have attempted such a similar study in different social settings and for different purposes. With quantifying spontaneous interactions being a knowledge gap in literature, the above mentioned literature works can be key in contributing towards designing the measure which quantifies the overall quality of spontaneous interactions.

Social constructs which quantify different aspects of social interactions, such as Rapport, Cohesion, Interest-levels and Engagement, have been successfully modeled in literature using several sets of nonverbal social cues. Such social cues have been used to predict the social constructs and also study its properties. The sets of nonverbal social cues, as discussed in the review, includes turn-taking, proxemics, prosody and coordination based feature sets. The choice of nonverbal cues to study a social construct is completely dependent on the availability of the respective modalities in the dataset. Nevertheless, turn-taking based features have been the state-of-the-art technique to model several subtle social constructs. At the same time, more recently, coordination based features, extracted at different modalities, are proven by several researcher to have out-performed the state-of-the-art turn-taking features in predicting several social constructs. Hence, the literature review motivates us to study the measure of spontaneous interactions using the two nonverbal feature sets of turn-taking and coordination.

As a conclusion, to design a measure which quantifies the overall quality of spontaneous interactions with respect to the individual experiences, it is important draw inspiration from several research works which study unique aspects of social interactions. Especially, it important to draw inspiration from works which study similar social settings (e.g. [31]) and operate with a similar purpose of quantifying the overall quality of an interaction (e.g. [28][79]). Additional, to predicatively model and study the properties of spontaneous interactions, turn-taking coordination features based nonverbal feature sets should be considered.

3 Conversation Quality

As discussed in Chapter-1, quantifying spontaneous interactions is one of the main contributions and research goals of this thesis. To achieve this, we design a novel measure which quantifies the overall quality of conversation in a spontaneous interactions and we name this measure as *Conversation Quality*. This chapter is dedicated to explain the design choices and considerations involved in designing the measure of *Conversation Quality*. In this chapter, firstly, in Section-3.1, we formally define the measure of *Conversation Quality*. Subsequently, in Section-3.2 we present the constituents of *Conversation Quality* and the unique aspects they intend to capture in spontaneous interactions, and in Section-3.3, we also describe the forms in which the measure of *Conversation Quality* occurs as a social construct in spontaneous interactions. Finally, in Section-3.4, we present the questionnaire used to measure the *Conversation Quality* and its forms.

3.1. What is *Conversation Quality?*

In this section, we formally define the measure of *Conversation Quality*. The measure of *Conversation Quality*, introduced in this research, has been inspired from Edelsky's definition of cooperative floors [31], discussed previously in the literature review chapter (Chapter-2). Considering spontaneous interactions as a form of cooperative conversational floors, Edelsky's high-level definition of cooperative floors [31, p.384] will be a suitable starting point to quantify the overall quality of spontaneous interactions.

Cooperative floors have been studied extensively in existing literature. For example, in the social sciences literature, measures of conversational equality and freedom [19][79], measures of conversational fluency through frequent turns, turn overlap and turn duration [29], and measures of turn synchronisation [122], seem to resonate particularly strongly with Edelsky's (1981) views on cooperative floors. Similarly, in the SSP literature, similar measures of turn-taking patterns have been used to model social constructs very much related to *Conversation Quality* - rapport [85], involvement [92], engagement [55], interaction quality [108], [120], cohesion [56], satisfaction [24] [70], experience [79] and performance [6]. From these works, we realise that Edelsky's cooperative floors, turn-taking patterns and several key social constructs are closely linked to one another. This motivates us further to rely solely on Edelsky's views on cooperative floor to define the measure of *Conversation Quality* and comprehensively account for individual experiences in the spontaneous interaction.

With respect to Edelsky's definition of cooperative floors [31, p.384], in this research, we define the measure of *Conversation Quality* in a spontaneous interaction as,

> *the degree to which participants in the spontaneous interaction are of the **same wavelength** and maintain a **free-for-all** floor.*

The two keywords here - *same wavelength* and *free-for-all*, are the two high-level aspects of *Conversation Quality* and are vital in defining the measure and intends to capture the individual experiences in spontaneous interactions. In a cooperative spontaneous interaction setting, the aspect of *same wavelength* tends to capture the aspects of rapport and inter-personal liking shared by individuals with their interacting partners and, the aspect of *free-for-all* intends to capture the overall quality of interaction held by the individuals in the interaction. With these two high-level aspects of *Conversation Quality*, we believe a comprehensive quantification of spontaneous interactions can be achieved. In the following sections, we will further discuss how these two high-level dimensions of *Conversation Quality* can be captured aptly using low-level constituents of *Conversation Quality*.

3.2. Constituents of Conversation Quality

In this section, we present the constituents of the measure of conversation quality. Each of these constituents intend to uniquely capture a specific aspect of individual experiences in a spontaneous interaction, thereby measuring the two high-level aspects of conversation quality (*same wavelength* and *free-for-all*). In the literature review presented previously (chapter-2), we saw two different ways in which individual experiences in a conversation were handled,

1. Cuperman and Ickes' three constituents of perception of an interaction (POI) [28].

2. Lidley and Monk's four constituents of the thin-slice empathised enjoyment scale (TES) [79].

For this research, we have adopted constituents from both these research works. Specifically, we have adopted all the constituents discussed in Cuperman and Ickes' POI [28] and have adopted one of the constituents in Lidley and Monk's TES [79]. Therefore, with respect to POI [28] and TES [79], we present the *four* constituents of conversation quality measure as *Quality of Interaction* (QoI), *Degree of Rapport* (DoR), *Degree of Likeness* (DoL) and *Free-for-All* (FfA). While QoI, DoR and DoL are adopted from POI [28], the FfA is adopted from TES [79].

With that said, while previously studied social constructs, such as *cohesion* [56], *rapport*[85], *bonding*[60], *enjoyment*[80] and *interest-levels*[37] intend to capture a particular aspect of social interactions, the measure of *Conversation Quality* intends to quantify the overall quality of conversation in a spontaneous interaction, with the four constituents mentioned above. For example, the measure of *Rapport*, the measure which is the most similar to *Conversation Quality*, captures the *interpersonal relationship* in a social interaction by measuring the degree to which interacting partners share a *close and harmonious relationship and are "in-sync" with each other* [85], but does not capture other key aspects such as *quality interaction*[28], *free-for-all*[79] and *interpersonal liking*[28]. Similarly, the measures such as *interest-levels* and *engagement* capture the *degree of involvement* displayed by individuals in the interaction, but does not capture the aspects such as *interpersonal relationship*[85], *quality of interaction*[28] and *free-for-all*[79].

With respect to the high-level aspects, the constituents of *Quality of Interaction* (QoI), *Degree of Rapport* (DoR) and *Degree of Likeness* (DoL) capture the multi-faceted "*same wavelength*" aspect of Conversation Quality, and the constituent of *Free-for-All* (FfA) exclusively captures the "*free-for-all*" aspect of Conversation Quality. These constituents of the measure of conversation quality are explained further in detail in the coming sections.

3.2.1. Quality of Interaction

The *Quality of Interaction* constituent of *Conversation Quality* captures an individual's experience in the interaction with their partner. For example, the degree to which the individual's interaction was smooth and relaxed or the degree to which the individual's interaction was forced and awkward. This particular constituent of *Conversation Quality* was adopted from Cuperman and Ickes' POI [28], and intends to capture the "*same wavelength*" aspect of Conversation Quality. The high-level aspect of "*same wavelength*" resonates well with interaction based measures such as *Conversation Fluency* and *Conversation Synchronisation* [79]. The *Quality of Interaction* constituent, by capturing behaviours related to such interaction based measures, successfully captures the "*same wavelength*" high-level aspect.

The quality of the interaction strongly depends on the nature of interaction held amongst interacting partners in the spontaneous interaction. It can be influenced by several interaction based factors right from smoothness and awkwardness in the interaction, to comfort-levels and interest-levels in the interaction. In contrast to the social construct of *Rapport*, which captures the inter-personal relationship, this particular constituent captures the nature of interaction between interaction partners.

3.2.2. Degree of Rapport

Rapport is the close and harmonious relationship in which interaction partners are "in sync" with each other [85]. The aspect of *degree of rapport* in a conversation captures the inter-personal relationship amongst interacting partners in a spontaneous interaction. For example, the degree to which an individual was accepted and respected by other individuals in the group or the degree to which the other individuals were paying attention to the individual. This particular constituent of *Conversation Quality* was adopted from Cuperman and Ickes' POI [28], and intends to capture the "*same wavelength*" aspect of Conversation Quality.

Rapport, is widely acknowledged to result in smoother social interactions, improved collaboration, and improved interpersonal outcomes. The phenomenon of *Entrainment* is one similar phenomenon, which explains the aspects of *in sync* and *same wavelength* among interlocutors in a social interaction. Levitan et al. (2011) [76] while studying entrainment in back-channels, state the importance of entrainment towards conversation quality - *Entrainment, the phenomenon of dialogue partners becoming more similar to each other, is widely believed to be crucial to conversation quality and success.* With that said, the *Degree of Rapport* is an essential constituents of *Conversation Quality*. While the *Quality of Interaction* constituent effectively captures the nature of interaction between interaction partners, the *Degree of Rapport* constituent captures the inter-personal relationship between interacting partners.

3.2.3. Degree of Likeness

The aspect of *degree of likeness* [1] of individual in a conversation captures the degree to which an individual likes their interaction partners and the ongoing conversation with them. For example, the extent to which an individual would like to interact more with their interaction partners or the extent to which an individual liked the other individuals in the interaction. Similar to *Degree of Rapport*, this particular constituent of *Conversation Quality* was also adopted from Cuperman and Ickes' POI [28], and intends to capture the "*same wavelength*" aspect of Conversation Quality.

Social constructs such as involvement and enjoyment are good indicators of degree of likeness [22]. While it is an important constituent of conversation quality, it was used originally by researchers in the form of self-reported measures. As the degree of likeness is an intimate measure of an individual's interaction, it is not possible to extend this measure to perceived measures of conversations.

3.2.4. Free-for-All

The *Free-for-All* constituent of *Conversation Quality* captures the equal opportunity given to all the individuals in the conversation group. For example, free-for-all factors like conversation freedom [70], conversation equality [79] and an individual's opportunity to take the lead in the conversation [28][60] were used by researchers as one of the measures to study bonding and experience in social interactions. This particular constituent of *Conversation Quality* was adopted from Lindley and Monk's TES [79], and intends to capture directly the "*free-for-all*" high-level aspect of Conversation Quality. The concept of *Free-for-All* was first illustrated by Edelsky [31] while differentiating cooperative floors from exclusive floors. Free-for-all is an essential aspect of cooperative floors and spontaneous conversations, and hence is an important constituent in measuring the *Conversation Quality*.

3.3. Perception of *Conversation Quality*

In the previous sections, we presented the definition and the constituents of *Conversation Quality*. In this section, we present how *Conversation Quality* as a social construct can be perceived in spontaneous interactions. While presenting the perception of *Conversation Quality*, we also discuss the design choices considered while defining the perceptional forms of *Conversation Quality*.

Unlike a task-directed social interaction (e.g. a discussion meeting [20] or a decision-making meeting [14]), spontaneous interactions are typically non-task-directed. Generally, in task-directed social interactions, social constructs associated to task outcomes (e.g. group performance and cohesion) emerge objectively from within the group [87][88]. On the other hand, in non-task-directed social interactions, social constructs do not objectively emerge from within the group, rather, subjective individual experiences and perceptions of social events exist within the consciousness of the observers related to the social interaction [92][85]. With that said, *Conversation Quality* as a social construct in spontaneous interaction should be handled with respect to the perceptions of individuals related to the interaction.

Perceptions of social constructs are particularly subjective in nature and differs hugely between each and every individual related to the interaction. While we talk about individuals related to the interaction, it is important to note that it also includes external observers along with the participants of the interaction, that

[1] Though literally *Likeness* means resemblance or similarity amongst interacting partners, in our case it means the feeling of regard or fondness (*Likeness*) amongst interacting partners. This ambiguity in the literal meaning was not solved in this research as we directly adopted this particular constituent from Cuperman and Ickes' POI [28], where the same ambiguity exists.

is, the external observers also hold a subjective perception towards a particular social construct of the group. In the social sciences literature, the perception of external observers and interaction participants are usually measured using *perceived* [37] and *self-reported* [9] measures respectively. The social construct of *Conversation Quality* can be measured by both the measures of perception, with each of them having certain traits of their own. The choice between the two measures to measure the *Conversation Quality* of spontaneous interactions depend on several factors which include, but are not limited to, data-collection strategy, application, research question and relevance to context.

Self-report studies have many advantages, but they also suffer from specific disadvantages due to the way that subjects generally behave [35]. Self-reported answers may be exaggerated that respondents may be too embarrassed to reveal private details, various biases may affect the results, like social desirability bias [90]. In cases of series of short bursts of spontaneous interaction, subjects may tend to also forget longitudinal details and require a ESM[2] based data collection. At the same time, perceived measures are only an approximation of actual perceptions of the individuals and their perceptions. But, perceived measures are free from several keys issues faced by self-reported measures, mainly the issues of egoistic biases [90][35], recall biases [89] and cognitive errors [89]. Such issues are highly relevant towards development of feedback systems in spontaneous interactions, for the following reasons,

1. In-the-wild spontaneous interactions are dynamic and generally occur as a series of short bursts of interactions and hence, self-reported measures in such scenarios might be prone to recall biases and cognitive errors.

2. Feedback systems, which measure *Conversation Quality*, should be averse to any subjective, egoistic and behavioural biases, so that it can optimally perceive social interaction without individual-level biases.

With the above consideration, for this research, we rely on *perceived measures* of *Conversation Quality* to contribute towards development of feedback systems in spontaneous interactions.

Spontaneous social interactions are multi-level systems that involve social constructs emerging from different levels of interactions [41]. For example, social constructs, in our case the *perception of Conversation Quality*, can be measured at individual-level or the group-level. Perception of *individual-level* constructs occur with a prime focus on a particular individual in the interaction and, the perception of *group-level* constructs occur with an overall focus over the whole social group which includes all the individuals participating in the interaction. In the social sciences literature, group-level constructs are generally derived from the aggregation of individual-level constructs [12]. When studying groups and teams, researchers can include individual-level and/or group-level phenomena in their research design. This research, which focuses towards enabling feedback systems requires modeling of both the levels of phenomena. The ability of a socially intelligent system to perceive and understand both the individual-level and group-level *Conversation Quality* helps the system in understanding the influence of the individual-level phenomena on the group-level phenomena, thereby providing constructive feedbacks to improve the group's conversation quality.

With respect to the above mentioned consideration, for this research, we rely on external observers to collect perceived measures of conversation quality. In that case, we consider that the social construct of *Conversation Quality* exists in the perception of external observers of the spontaneous interaction and, in two forms - *Perceived Individual's experience of Conversation Quality* (the individual-level phenomena) and *Perceived Group's Conversation Quality* (the group-level phenomena). An illustration with respect to the two perceived measures of *Conversation Quality* can be seen in Fig-3.1. The two forms of conversation quality are defined in the coming sections.

3.3.1. Perceived Group's Conversation Quality

For this research, we define the perceived *group-level* conversation quality as an external observer's perception of the conversation quality of the group as a collection of all individuals in the group. This perceived measure directly taps into the what an external observer perceives or feels about the conversation going on in the whole group. On a high-level, this measure is the answer to the question -

How do you rate *the overall experience of all the group members* in the conversation, in-terms of

[2] The experience sampling method, also referred to as a daily diary method, or ecological momentary assessment (EMA), is an intensive longitudinal research methodology that involves asking participants to report on their thoughts, feelings, behaviors, and/or environment on multiple occasions over time. [11]

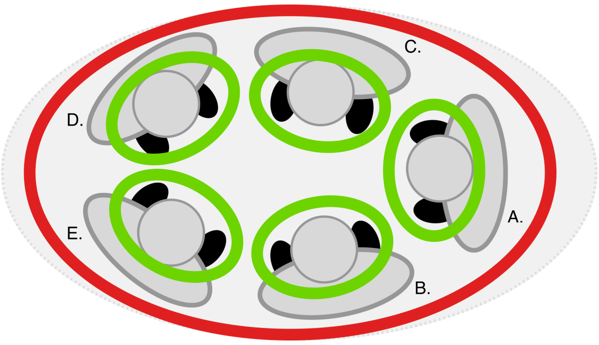
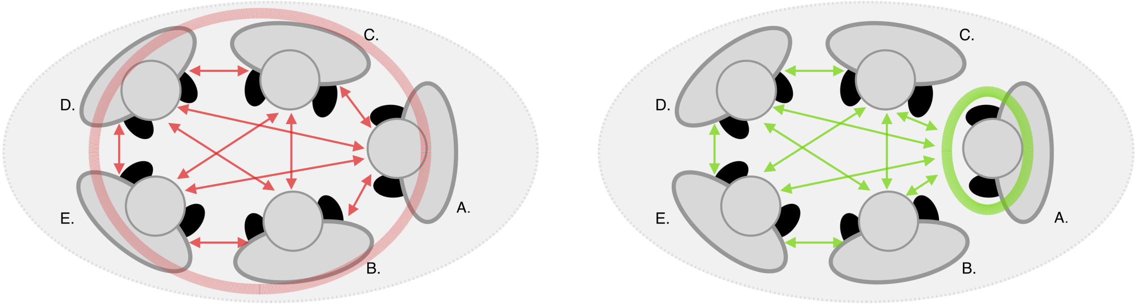


Figure 3.1: An illustration on the two forms of perceived *Conversation Quality*, as perceived by an external observer. The *Red* boundary illustrates the scope of observation to measure the group-level perceived Conversation Quality - *Perceived Group's Conversation Quality* and, the *Green* boundaries illustrate the scope of observation to measure the individual-level perceived Conversation Quality - *Perceived Individual's Experience of Conversation Quality*.



(a) The group-level *Perceived Group's Conversation Quality* and its within-group interactions which influence the social construct. Here, the scope of observation, for the external observer, is the whole group which includes all the group members and their interactions.

(b) The individual-level *Perceived Individual's Experience of Conversation Quality* (of individual A) and its within-group interactions which influence the social construct. Here, the scope of observation, for the external observer, is a particular individual (in this case Individual 'A') and their interactions.

Figure 3.2: An illustration of the two forms of perceived *Conversation Quality*, along with the within-group interaction (denoted by arrows) which might influence the social construct. The illustration above presents the forms of the perceived conversation quality as perceived by an external observer. The boundaries in the illustration (red and green oval) represent the scope of observation relied on to rate the respective forms of *Conversation Quality*.

> whether all the group members are of the *same wavelength*, with respect to the *quality of interaction*, *degree of rapport* and *degree of likeness*, and maintain a *free-for-all* floor?

A visual representation of the measure can be seen in Fig-3.2a. Such a measure, is defined in a manner in which it intends to quantify the whole group's quality of conversation as perceived by an external observer. With respect to the definition of the *Perceived Group's Conversation Quality*, we hypothesis that an external observer tends to rate this measure the least, if the observer perceives that all the group members are not having a good experience in the conversation, with the group members having a low *quality of interaction*, *degree of rapport* and *degree of likeness* with one another, and also not providing a free-for-all floor to the each other. On the other hand, the external observer tends to rate this measure the highest, if the observer perceives that all the group members had the best experience in the conversation, with the group members having a high *quality of interaction*, *degree of rapport* and *degree of likeness* with one another, and also providing a free-for-all floor to the each other. This results in *one* ratings per group, where the rating is the whole group's conversation quality as perceived by the external observer.

3.3.2. Perceived Individual's Experience of Conversation Quality

For this research, we define the perceived *individual's* experience of conversation quality as an external observer's perception of the quality of the conversation as experienced by the *individual*. This perceived measure directly taps into an external observer's perception of an individual's experience in their conversation with the group. On a high-level, this measure is the answer to the question -

> How do you rate *a particular individual's experience* in the group, in-terms of whether the individual is of the *same wavelength* as the other group members, with respect to the *quality of*

interaction, *degree of rapport* and *degree of likeness*, and experiences a *free-for-all* floor?

A visual representation of the measure can be seen in Fig-3.2b. Such a measure, is defined in a manner in which it intends to quantify an external observer's perception regarding an individual's experience in the group conversation. With respect to the definition of the *Perceived Individual's Experience of Conversation Quality*, we hypothesis that an external observer tends to rate this measure the least for an individual, if the individual does not seem to have a good experience in the conversation, by having a low *quality of interaction*, *degree of rapport* and *degree of likeness* with other group member, and also does not seem to experience a free-for-all floor. On the other hand, the external observer tends to rate this measure the highest for an individual, if the individual seems to have the best experience in the conversation, by having a high *quality of interaction*, *degree of rapport* and *degree of likeness* with other group member, and also seem to experience a free-for-all floor. This results in '*n*' ratings per group, '*n*' being the group-size, where each individual in the group receives a *Conversation Quality* rating given to them by the external observer.

3.4. Perceived Conversation Quality Questionnaires

In this subsection, we describe the two questionnaires used to measure the respective forms of *Conversation Quality* as perceived by external naive annotators. Along with the description of the questionnaire, we also provide literature backing for the components of the questionnaire and also present the questionnaires in itself.

The *Perceived Conversation Quality* (PCQ) questionnaires were drafted by drawing inspirations from several research works in literature studying different aspects of social interactions. In specific, the *Perception of Interaction* (POI) by Cuperman and Ickes (2009) [28] and the *Thin-Slice Enjoyment Scale* (TES) were the prime motivation behind several components of the questionnaire. The POI questionnaire has been successfully used in different scenarios for studying social interactions. For example, Jaques et al. (2016) [60] have successfully used the POI questionnaire to study *bonding* in conversations and Cerekovic et al. (2017) [22] have used POI to study rapport with virtual agents in a dyadic setting. Similarly, the TES has been successfully used by Lindley and Monk [78][79] to study conversation experience and enjoyment in task based small-group interactions. While the above research works have successfully adopted their respective questionnaires, for this research we deal with a different social setting - an in-the-wild free-standing spontaneous conversation setting. Hence, to make questionnaire components from POI and TES relevant to our social setting, the questionnaire elements we modified slightly. The following steps were performed to modify the items available in the questionnaires (POI and TES) to suit our social setting and construct,

1. A subset of the questionnaire items from both the [28](POI) and [79](TES) were selected and included to form the final questionnaire.

2. Some of the items used in [28](POI) and [79](TES) were excluded, as they were irrelevant to our setting. For example, the question which involved content of the conversation - "The character often said things completely out of place", was excluded.

3. The questions under the *Degree of Likeness* (2.2.3) measure were excluded. The original questionnaire (POI) was a self-reported questionnaire and included intimate questions under the category of degree of likeness. Since our research focuses on a perceived measure of conversation quality using external annotators, including intimate questions would not be suitable. For example, the question - "Did you desire to interact more with partner in the future?" was excluded as an external annotator cannot perceive an individual's degree of likeness.

4. All the questions included for the final questionnaire for *Conversation Quality* was made relevant for external annotators. The original questionnaire (POI) was a self-reported questionnaire and was directed to the participant themselves. Since our research uses external annotators, the questions were modified to be directed towards the annotator themselves. For example, the question - "I did not want to get along with the character" was modified to - "The individual seemed to have gotten along with the group pretty well".

5. The original questionnaire (POI) was devised to measure perceptions of dyadic interactions [60] and interactions with virtual agents [22]. This was modified to a small group setting which includes all group members in the group conversation. For example, the question - "I felt accepted and respected

by the character" was modified to - "The group members accepted and respected each other in the interaction".

6. Finally, two questionnaires were prepared by following the above mentioned measures. The two questionnaires are the *Questionnaire for Perceived Group's Conversation Quality* which measures, at group-level, the overall group's conversation quality (3.3.1) and the *Questionnaire for Perceived Individual's Experience of Conversation Quality* (3.3.2) which measures, at individual-level, an individual's experience of conversation quality. Both the questionnaires combine to form the *Questionnaire for Perceived Conversation Quality* (QPCQ) devised in this research.

Below are the questions that were used for the manual annotations for *Conversation Quality*. We used two different questionnaires for each of the two forms of *Conversation Quality*. The questions below have been organized in terms of the different constituents and aspects of Conversation Quality - *Quality of Interaction* (QoI), *Degree of Rapport* (DoR) and *Free-for-all Interaction* (FfA). The numbers before each question indicate the ordering of the questions in the original questionnaire. The source for each term is provided at the end of each question.

3.4.1. Questionnaire for *Perceived Group's Conversation Quality*

Instruction for the annotators: Use the set of questions below to annotate your perception of the group's conversation quality, as seen in the video. Each interaction aspect in the below questionnaire should be rated using a five-point likert scale (Disagree strongly (1) to Agree strongly (5)). Read the questions carefully and observe the whole group carefully before annotating the video. You are allowed to re-watch the video again if required.

Quality of Interaction

1 The interaction within the group was smooth, natural and relaxed. [28]

2 The group members looked to have enjoyed the interaction. [28]

3 The interaction within the group was forced, awkward, and strained. [28]

Degree of Rapport

4 The group members accepted and respected each other in the interaction. [28]

7 The group members seemed to have gotten along with each other pretty well. [28][60]

8 The group members were paying attention to their partners throughout the interaction. [28]

9 The group members attempted to get "in sync" with their partners. [28][60]

10 The group members used their partner's behavior as a guide for their own behavior. [28][60]

Free-for-all Interaction

5 The group members received equal opportunity to participate freely in the interaction. [79]

6 The interaction involves equal participation from all group members. [79]

3.4.2. Questionnaire for *Perceived Individual's Experience of Conversation Quality*

Instruction for the annotators: Use the set of questions below to annotate your perception of the individual's experience in the conversation, as seen in the video. Each individual present in the conversation has to be annotated separately with the below questions. Each interaction aspect in the questionnaire below should be rated using a five-point likert scale (Disagree strongly (1) to Agree strongly (5)). Read the questions carefully and observe the individual carefully before annotating the video. You are allowed to re-watch the video again if required.

Quality of Interaction

1 The individual looked like they had a smooth, natural, and relaxed interaction. [28]

2 The individual looked like they enjoyed the interaction. [28]

3 The individual's interaction seemed to be forced, awkward, and strained. [28]

4 The individual looked like they had a pleasant and an interesting interaction. [28]

5 The individual looked uncomfortable during the interaction. [28]

10 The individual looked to be self-conscious during the interaction. [28]

Degree of Rapport

8 The individual was paying attention to the interaction throughout. [28]

9 The individual seemed to have gotten along with the group pretty well. [28][60]

Free-for-all Interaction

6 The individual attempted to take the lead in the conversation. [60][22]

7 The individual looked like they experienced a free-for-all interaction. [79]

The above presented two questionnaires are used in this research to obtain ground-truths for the further study of perceived *Conversation Quality* at both group-level and individual-level. The following chapters explain this study in detail.

4 Data

In this thesis, to model the the perceived conversation of free-standing conversational groups, we used the data from the MatchNMingle dataset [17]. The whole dataset is collected in an ecologically valid in-the-wild setting, which contributes to the ecological validation of our study of conversation quality. Such a validation, shows the scalability of our study to real-world applications. In this chapter, we give an introduction to the MatchNMingle dataset [17]. First, we present a brief description of the MatchNMingle dataset along with the experiment context and data collection procedure. Then, we describe the annotation process of our ground truth conversation quality along with its analysis. Finally, we present the results of the data analysis task performed on the dataset.

4.1. MatchNMingle Dataset

The *MatchNMingle* is a multimodal/multisensor dataset for the analysis of free-standing conversational groups and speed-dates in-the-wild [17]. The dataset has been used extensively by several researchers in the field of social sciences and data science. The dataset was collected in an indoor in-the-wild setting instead of a lab setting. Therefore, the social interactions between participants were as natural as possible. The MatchNMingle dataset collected by Cabrera-Quiros et al. (2018) [17], leverages the use of wearable devices and overhead cameras to record a large number of in-the-wild social interactions during a real-life speed-date event and a cocktail party. The dataset consists of dynamic social interaction of 92 participants which makes it one of the largest dataset with a large number of participants and their ever-evolving social interactions. The dataset consists of 2 hours of data recording collected from wearable acceleration, binary proximity, video, audio, personality surveys, frontal pictures and speed-date responses.

4.1.1. Experiment Context

In this section, we describe the context of the experimental setup in the MatchNMingle dataset. The MatchNMingle dataset was collected during a mingling event which took place for three days in total in a public social bar. The event on each respective day, started with a speed dating session where participants of opposite sex had a three minute of speed-date interaction with each other, followed by a mingling session which lasted for approximately one hour where the participants were allowed to freely interact with one another. In this thesis, to model the perceived conversation quality, we used only the later part of the dataset which was collected during the mingling session.

Participants for the data collection were recruited from a university campus and none of the participants were acquainted before the event. Participants were aged between 18 and 30. Participants were recruited such that they fit in the criteria of being single and heterosexual. Measures were taken to keep the number of male and female participants to be equal in number. To record individual behavioural data during the event, all the 92 participants were requested to wear a sensor pack consisting of a tri-axial accelerometer and a proximity sensor around their necks. The sensor pack can be seen in Figure-4.1. Because of the malfunction of hardware, some of the devices failed recording at various times. After removing these devices, in total 72 participants had sufficient data recorded by wearable devices.

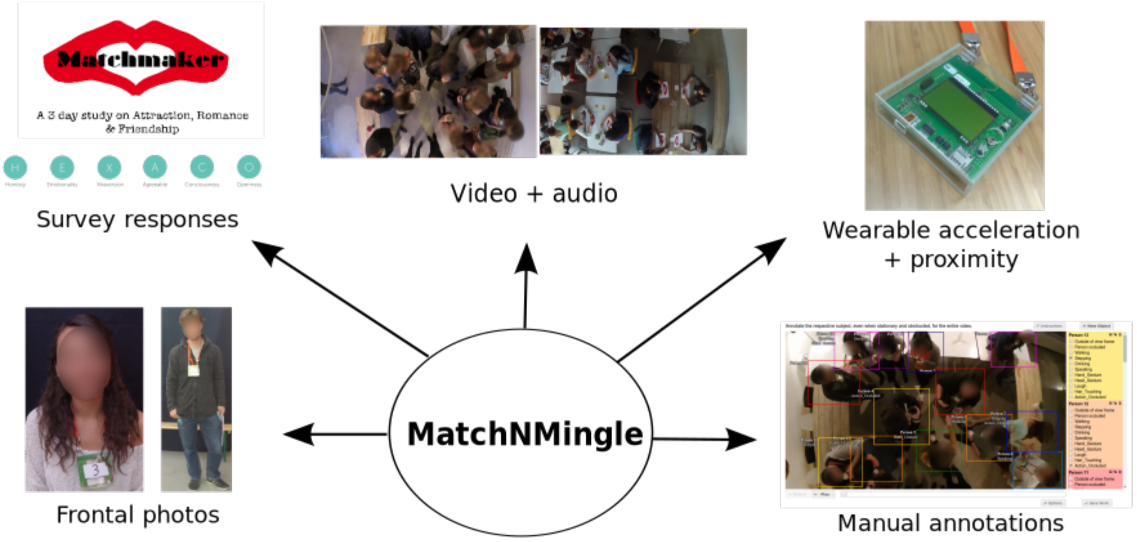


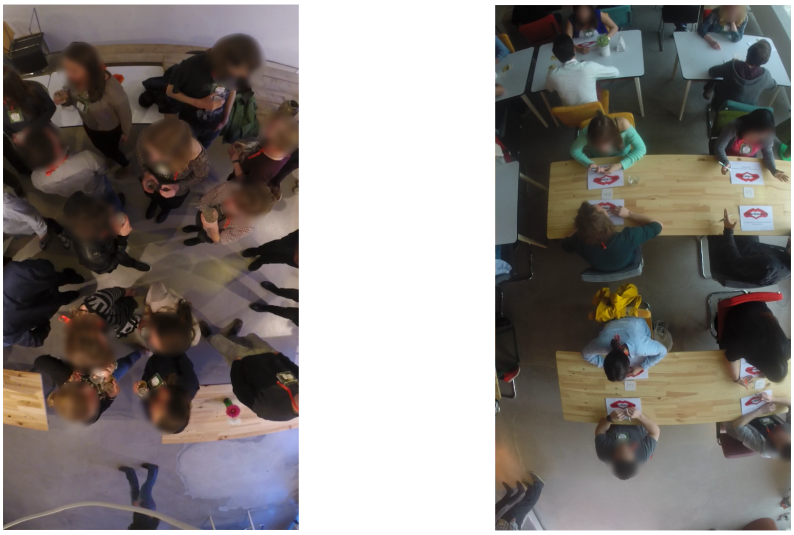
Figure 4.1: An overview of the different modalities and information available with the MatchNMingle dataset [17].

4.1.2. Data Collection

In this section, we explain in detail the data collection strategy used in the collection of the MatchNMingle dataset. To collect participant behavioural data, throughout the event, accelerometer reading were collected for each individual using tri-axial accelerometers, at a frequency of 20Hz. Along with the accelerometer data, proximity information was collected in binary values, where 1 represented two people who are in close proximity (approximately 2-3 meters) and 0 represented two people being far away from each other. In addition to the accelerometer and proximity recordings, the whole event area was captured using 9 different cameras from top, during both the speed-date and mingling sessions. The mingling sessions were filmed using GoPro cameras that were fitted into the ceiling. In total 5 cameras wered used to film the mingling part of the event. These cameras filmed at a resolution of (1920 ×1080 or 16:9), sample rate of 30 fps and with a ultra-wide field of view. Few snapshots from the recorded videos can be seen in Figure-4.2. The dataset collected also consists of several self-reported responses filled by participants post the event and also before the event. At the online registration for the event, before the event, participants were also asked to fill different questionnaires to test individual personality differences. These personality questionnaires included the HEXACO personality inventory [5], the Self Control Scale (SCS)[112], the Sociosexual Orientation Inventory (SOI) [97]. In addition to this, in the beginning of the events hair samples from each participant are collected to gather hormonal baselines. And also, post the speed-date session participants were given 1 minute to fill a match booklet which has a questionnaire about their date partner indicating their interest on each other. An overview of the available information in the MatchNMingle dataset can be seen in Figure-4.1.

As mentioned earlier, in this thesis, to model the perceived conversation quality, we use only the data collected during the mingling session. Thereby, at our disposal are the accelerometer readings of participants and video recording of interaction during all three days of the one hour mingling session. In addition to the above mentioned two information, we also had manual annotations of several key events in the social interaction. For each participant in the mingling session, nine different actions were manually annotated. The actions annotated included speaking turns, walking, stepping, drinking, speaking, hand gesture, head gesture, laughing, hair touching and action occlusion. All these annotations were performed manually at a frequency of 20Hz using the video recording of the mingling session. In the MatchNMingle dataset this set of annotations is available as a 36000 × 828 matrix. Annotations for F-Formations were also made for a limited subset of the data. In total 10 minutes worth of F-Formation annotations were made, while there were 90 minutes of video available. For this research, we use this F-Formation annotation as the basic element to define an independent group and its interaction.

Although we have these varied data items, the ground truth for our research was still missing from the existing MatchNMingle dataset. The ground truth for this research work is the *Perceived Conversation Quality* score for both its manifestation forms, the *Perceived Group's Conversation Quality* (3.3.1) and the *Perceived Individual's Experience of Conversation Quality* (3.3.2). For this purpose of collecting ground truth for the



(a) Snapshot from the dynamic mingling session.

(b) Snapshot from the dyadic speed-date session.

Figure 4.2: Snapshots of the MathNMingle dataset's video recording, filmed using over-head GoPro cameras that were fitted into the ceiling.

modeling of perceived conversation quality, we manually annotated the social interactions (the f-formations) in the mingling session for the manifestations of the measure of conversation quality.

4.2. Annotation of Conversation Quality

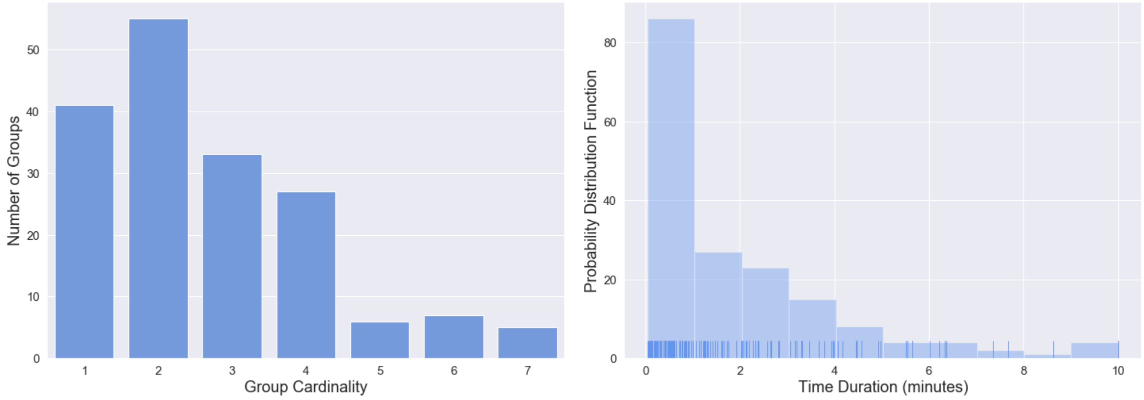
In the chapter-3, the measure of *Conversation Quality* was defined along with its manifestation forms. And as we discussed before, defining of such a measure is one of the novel contribution of this research work. Similarly, annotating the MatchNMingle dataset was another task executed during the research work. The annotations of the measure of *Conversation Quality* and its manifestations constitutes for the ground truth for the experiments in this thesis. In this section, firstly, we begin with a detailed explanation of the annotation strategy used in the process and finally, we perform some data analysis on the annotations collected.

4.2.1. Annotation Procedure

In this subsection, we define the annotation procedure which includes a detailed reporting regarding the resources utilised for the annotation process and the strategy deployed to collect valid annotations for *Conversation Quality* and its manifestations.

In this research, we use the data recorded during the mingling session, available in the MatchNMingle dataset. The mingling session was recorded using five GoPro cameras that were fitted into the ceiling. This video recording was the only resource used for the manual annotation of the *Conversation Quality*. No audio data was used for the annotation process. Several other research works in literature have successfully collected rich annotation data by relying completely on video clips without relying on audio recordings. Audio recordings in most of the conversation scenarios are unavailable due to privacy reasons. Moreover, manual annotations by also time consuming as the audio recordings are generally prone to noise, lack of clarity, requires speaker diarisation and language constraints. On the other hand, manual annotations using only video recording are easier and less time consuming. At the same time, video recordings have the capability to capture rich non-verbal behaviours of participants in the social interaction.

The MatchNMingle dataset consists of f-formation annotations, annotated using the spatial position for all participants during the mingling session. In total 30-minutes across three days of the mingling session were annotated for f-formations. The sub-sampled 30-minutes of video segments were chosen randomly with an aim to eliminate the possible effects of acclimatization, and to maximize the density of participants and the number of social actions that could occur in the whole scene. In the 30-minutes segments of annotated f-formations in the mingling session, there we in total *174* f-formations which included singletons, dyads and larger groups. The distribution of the f-formations with respect to the group size can be seen in Figure-4.3a. For this research, we assume a group to be a f-formation and the group members to be all the participants present in the particular f-formation. Raman et al.'s (2019) [100] work in the MatchNMingle dataset found evidences that prove existence of multiple floors and conversations within an f-formation. But as the MatchNMingle dataset did not have annotations of the conversation floors within f-formations and for the sake of reducing complexity, for this research on *Conversation Quality*, we consider a group to be an f-formation.

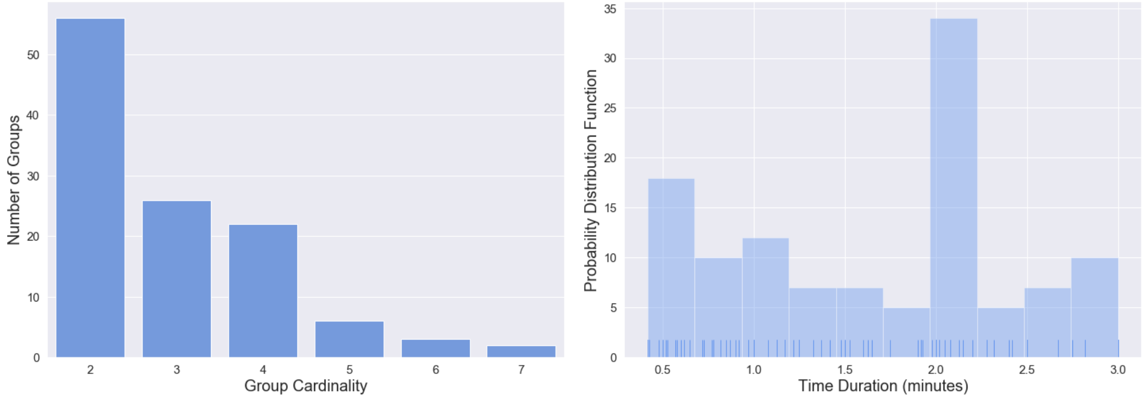


(a) Distribution of f-formations with respect to the group size.

(b) Distribution of f-formations with respect to the duration of interaction.

Figure 4.3: Distributions of annotated f-formations with respect to the group-size and duration, in the MatchNMingle dataset. (Before splitting of f-formations)

Before the f-formation groups were given to the annotators for annotation, we cropped the f-formation out of the video recording in the dataset. This was done in order to prevent annotators from getting distracted away from the current f-formation in focus. Apart from cropping out f-formations from the video recordings, the f-formations were also split in to multiple smaller segments and presented to annotators as independent videos of social conversations. This was done in order to collected more granular annotations for group conversations.



(a) Distribution of final f-formations which are to be annotated, with respect to the group size.

(b) Distribution of final f-formations which are to be annotated, with respect to the duration of interaction.

Figure 4.4: Distribution of final f-formations which are to be annotated, with respect to the group-size and duration, in the MatchNMingle dataset. (After omission and splitting of f-formations)

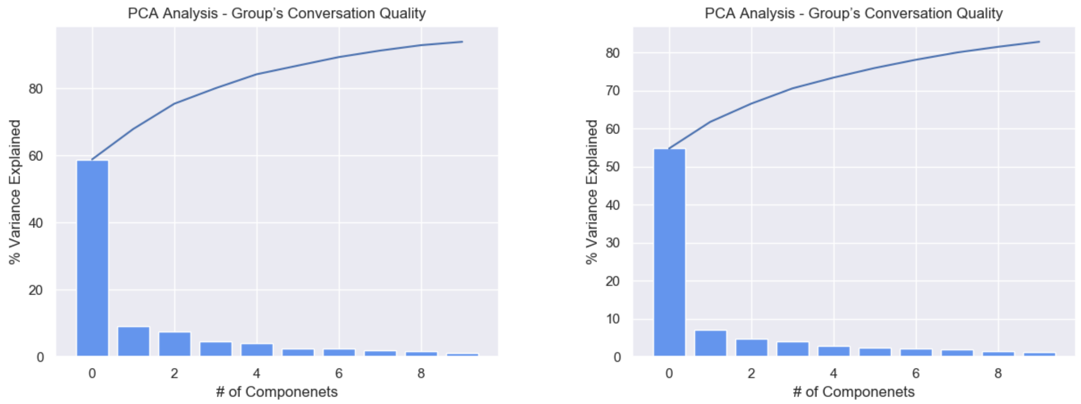
The duration of f-formations in the mingling session varied largely from few seconds to several minutes. Some of the f-formations even lasted throughout the segment annotated for. The distribution of f-formations with respect to its duration can be seen in Figure-4.3b. From the figure, we see that a good number f-formations even lasts for more than several minutes, while at the same time, a majority of the f-formation had lasted for only couple of minutes or even for a minute. In this case, it is not fair or reliable enough to have one label annotation to define the conversation quality for both the f-formations that lasted for less than a minute and that which lasted for more than several minutes. For this purpose, the splitting process on longer lasting f-formations explained above was performed. The duration of the annotated f-formations were distributed with a mean of *1.91*, standard deviation of *2.13*, median of *1.10* and a mode of *0.42*. With the above distribution in consideration, we decided to split f-formations of interaction duration of length greater that *3* minutes to independent groups of 2-3 minutes each. From Figure-4.3b, we also see that there are a good number of f-formation with interaction duration less that *30 seconds*. For the same reason for which we split the longer lasting f-formations, we omitted the f-formations with durations less than *30 seconds*. Post the omission and splitting processes, the total number of f-formation groups was 115. The distribution of those groups with respect to the group size and interaction duration can be seen in Figures - 4.4a and 4.4b respectively. These f-formation groups were provided to the annotators in randomised order for each annotator, to prevent any annotator bias which might occur in case a strict f-formation clips order is followed.

With the f-formation annotations and video clips of f-formations, for this research we decided to request naive annotators to help us in the annotation of *perceived Conversation Quality* for every group (f-formation) in the mingling session. For this purpose, we chose three naive annotators and requested them to help us with the annotation process. The three annotators were aged between 22-30 years. Out of the three annotators, two were females and one was male. The annotators were provided with video clippings of f-formations present in the mingling session and were asked to fill out the QPCQ two questionnaires (explained in Section-3.4) to measure the *Perceived Group's Conversation Quality* and the *Perceived Individual's Experience of Conversation Quality* respectively.

4.2.2. Annotation Analysis

In this subsection, we present the results of the data analysis performed on the annotation responses collected through the annotation procedure explained in the above subsection. The data analysis on the annotations include the inter-annotator agreeability tests and the analysis on the distribution of annotations. The annotation analysis for both the group-level(3.4.1) and individual-level(3.4.2) questionnaires are presented below. In the rest of this section, the two questionnaires - *Perceived Group's Conversation Quality* and *Perceived Individual's Experience of Conversation Quality* will be referred to as group-level and individual-level questionnaires respectively.

Distribution analysis To analyze the distribution of the annotations, we first carried out principal component analysis on the data. This analysis showed that 71% and 65.2% of the variance in the group-level and individual-level annotation data respectively, could be explained by the first principal component. And the first four principal components are capable of explaining over 80% of the data variance. The eigenvalue bar-chart can be seen in Figure-4.5.

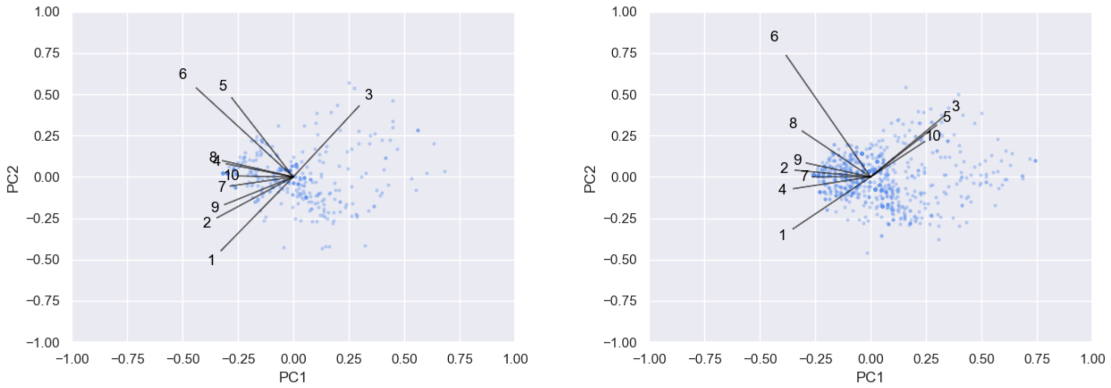


(a) Eigenvalue distribution for the annotations of Perceived Group's Conversation Quality.

(b) Eigenvalue distribution for the annotations of Perceived Individual's Experience of Conversation Quality.

Figure 4.5: Eigenvalue for each component (bar chart) and the cumulative percentage of the explained variability (line plot), for the annotations of Conversation Quality.

For further analysis of the annotations data, we plotted the data in a two dimensional plot where the two dimensions represented the first two principal components of the annotations data. The two dimensional plot along with the factor loadings can be seen in Figure-4.6. Each line shown in the plots are the magnitude of loading of each question in the principal component space. A longer line indicates a larger variability of the vector in the two components and vice-versa. Every line is also labeled with the number which corresponds to the question number in the respective questionnaire.



(a) Factor loadings for the annotations of Perceived Group's Conversation Quality.

(b) Factor loadings for the annotations of Perceived Individual's Experience of Conversation Quality.

Figure 4.6: Plot showing the factor loadings (black lines) and scores (gray dots) for the data when observed in relation to the first and second principal components of the annotation data, for the annotations of Conversation Quality. The numberings labeled on each loading line corresponds to the respective question item in the respective questionnaire. Longer the line larger is the variability of annotation of the questionnaire, with respect two the top two principal components.

In the group-level annotations plots (4.6a), we see that questions are particularly clustered into two clusters, one cluster where questions show high variance towards the negative scale of the first principal component and the second cluster where questions show high variance towards the positive scale of the first principal component. A similar two cluster scenario is also observed in the individual-level annotations plot (4.6b). On further analysis, we found that the questions in the two cluster corresponds respectively to the orientation of the scale for each question. For example, in the figure-4.6b, the questions 5,3 and 10 are reversed in scale orientation from the rest of the questions. Similarly, in the figure-4.6a, the question 3 is reversed in scale orientation from the rest of the questions. Thus, question items that were expected to have high scores resulting in high conversation quality appeared in one cluster while those which were expected to have low scores resulting in low conversation quality appeared in the other cluster. This observation suggests that very similar scoring patterns were occurring for a significant number of different meetings. We also see from the plots (4.6a, 4.6b) that few question items are strongly loaded with respect to the other questions. For example, question 6 in both the annotations (4.6a, 4.6b) and also question 5 in (4.6a). In was interesting to note that, all these above mentioned highly loaded questions belong to the *Free-for-All* part of the questionnaire. This suggests us that the annotations for the free-for-all question items had the highest variance (between groups) in comparison with the other segments of the questionnaire. The distribution of annotation scores with respect to each questionnaire item can be found in the Appendix-A.

Post the basic analysis of the annotations data, we performed the inter-rater agreeability tests to statistically test the agreement amongst annotators across group conversations. In statistics, inter-rater reliability (similar names include inter-rater agreement, inter-rater concordance and inter-observer reliability) is the degree of agreement among annotators. It is a score of how much homogeneity or consensus exists in the ratings given by various judges. Such a consensus is required in order to reliably validate the annotation ground truths and average the scores of the annotators. There are a number of statistics that can be used to determine inter-rater reliability. Different statistics are appropriate for different types of measurement and the choice of statistic should be purely made with respect to the type of annotation data at hand. Some of these statistics which are capable of measuring inter-rater agreeability are, joint-probability of agreement, Cohen's kappa, Scott's pi and the related Fleiss' kappa, inter-rater pearson correlation, inter-rater spearman correlation, concordance correlation coefficient, intra-class correlation, and Krippendorff's alpha. Using our QPCQ questionnaires, we collected annotations in a 5 point likert scale for each question items as explained in the previous section. Thus, we had a set of ordinal data annotations. Ordinal data is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories is not known [73]. The ordinal scale is distinguished from the nominal scale by having a ranking. It also differs from interval and ratio scales by not having category widths that represent equal increments of the underlying attribute [111].

Cohen's Kappa measure With a set of ordinal annotations data, we chose Cohen's *weighted kappa measure* [25] as the statistic to measure the agreement amongst annotators and across group conversations. While Cohen's kappa coefficient (κ) is a statistic that is used to measure inter-rater reliability for qualitative categorical items, the *weighted* Cohen's kappa coefficient (κ) is a statistic that is used to measure inter-rater reliability for qualitative ordinal items. Cohen's kappa measures the pairwise agreement between two raters who each classify N items into C mutually exclusive categories. The formula used to calculate the kappa coefficient (κ) is,

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \tag{4.1}$$

In the above equation, p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. The values of p_o and p_e are calculated using the confusion matrix of the pairwise categorical annotations. For k categories, N observations, and n_{ki} number of times category k is annotated by annotator i, the hypothetical probability of chance agreement p_e and the relative observed agreement among raters p_o are calculated using the formulas,

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

$$p_o = \frac{1}{N} \sum_{k1=k2} n_{k1} + n_{k2}$$

If the raters are in complete agreement then κ=1 and if there is no agreement among the raters other than what would be expected by chance (as given by p_e), κ=0. It is possible for the statistic to be negative, which implies that there is no effective agreement between the two raters or the agreement is worse than random.

Cohen's *Weighted* Kappa measure The weighted kappa is a slight modification on the kappa measure explained above. The weighted kappa allows disagreements to be weighted differently and is especially useful when the annotation data are ordinal in nature. Three matrices are involved, the matrix of observed scores, the matrix of expected scores based on chance agreement, and the weight matrix. Weight matrix cells located on the diagonal (upper-left to bottom-right) represent agreement and thus contain zeros. Off-diagonal cells contain weights indicating the seriousness of that disagreement. Often, cells one off the diagonal are weighted 1, those two off 2, etc. For this research, we used the quadratic weighted matrix where the relationship between weights of cells one off and of cells two off are quadratic. That is, in the weight matrix there exists a quadratic decay from the diagonal. The weighted kappa is calculated as below,

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}} \tag{4.2}$$

where k=number of categories and w_{ij}, x_{ij}, and m_{ij} are elements in the weight, observed, and expected matrices, respectively. When diagonal cells contain weights of 0 and all off-diagonal cells weights of 1, this formula produces the same value of kappa as the calculation given above. For each meeting, the weighted kappa was computed for each pair-wise combination of the three annotators. Then, the average of all three kappa values was taken to be the mean weighted kappa agreement for that meeting.

Inter-annotator agreeability The inter-annotator agreeability tests for both the group-level (3.4.1) and individual-level (3.4.2) questionnaires were performed. For each and every individual and group, the weighted kappa was calculated for each pair-wise combination of the three annotators and the average of all the possible pairwise scores was calculated to be the final kappa agreement for that individual or the group. To analyse the group-wise and individual-wise mean kappa scores further, we plotted the mean kappa score against the mean conversation quality score. The mean conversation quality score for an individual and a group was calculated as below,

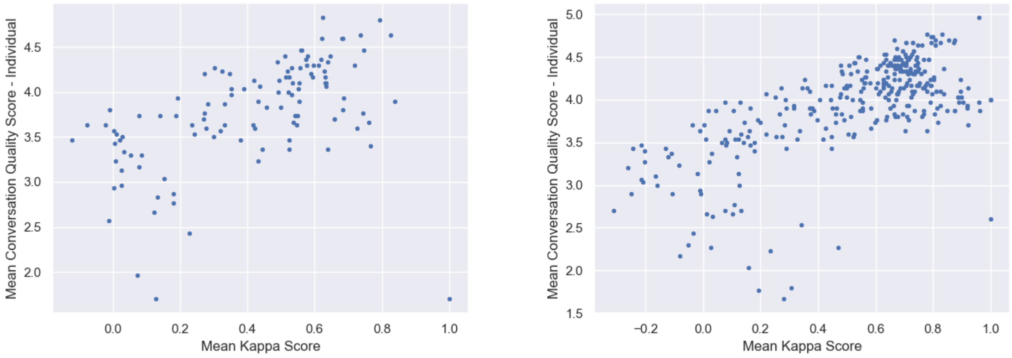
1. Calculate the sum of all annotation ratings across questionnaire items in the respective questionnaire.

2. Perform above step for each annotator response.

3. Calculate the mean of the summed annotator-wise ratings. The mean value is the mean conversation quality score for that particular individual or group.

4. This results in conversation quality scores that range between 0 and 5.

The scatter plots between the mean kappa score and the mean conversation quality score can be seen in Figure-4.7. A similar plot was used by Hung et al. (2010) [56] to analyse the inter-rater agreeability scores for small-group meetings of different levels of cohesion.



(a) Group-level mean weighted kappa scores (κ) with respect to the group's mean conversation quality score

(b) Individual-level mean weighted kappa scores (κ) with respect to the individual's mean conversation quality score

Figure 4.7: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses from all 3 annotators

From the figures above (4.7a, 4.7b), we see that there exists a linear relationship between mean kappa scores and mean conversation quality scores. That is inter-annotator agreeability decreases as conversation quality scores decrease. This suggests that annotators agree better on conversations of higher quality when compared to conversations of lower quality. We also observe that samples exhibiting scores of higher conversation quality tend to be more numerous than those exhibiting lower conversation quality scores. Thus, the final dataset might be an unbalanced in terms of class distribution. This is the same case for both the group-level and individual-level annotations. But, this was not expected by us. We expected similar results as seen in [56], where inter-annotator agreements on cohesion levels for meetings were higher at the two extremes of the scale. That is, in [56] annotators tend to *agree more* on extremely low and extremely high scores of cohesion and *tend to agree less* on mid-level of cohesion scores.

	Minimum	Maximum	Mean	Variance
Annotator 1	1.4	5.0	3.88	0.53
Annotator 2	1.7	4.8	3.90	0.29
Annotator 3	1.2	4.8	3.44	0.62

Table 4.1: Annotator-wise distribution of conversation quality annotations, with respect to the annotations of Perceived Group's Conversation Quality.

	Minimum	Maximum	Mean	Variance
Annotator 1	1.0	5.0	3.90	0.71
Annotator 2	2.0	5.0	3.98	0.24
Annotator 3	1.1	4.9	3.65	0.48

Table 4.2: Annotator-wise distribution of conversation quality annotations, with respect to the annotations of Perceived Individual's Experience of Conversation Quality.

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	-	0.3346	**0.4915**
Annotator 2	0.3346	-	0.4133
Annotator 3	**0.4915**	0.4133	-

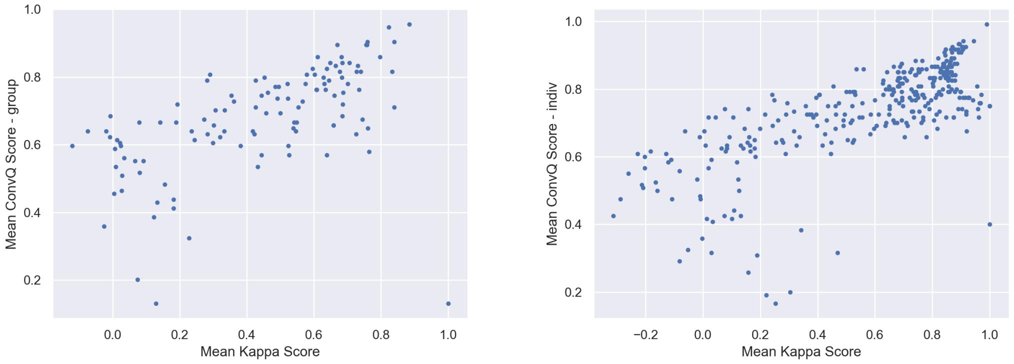
Table 4.3: Pairwise-wise inter-annotator agreeability, with respect to the annotations of Perceived Group's Conversation Quality. The overall agreement of all 3 annotators is 0.4132.

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	-	0.4682	**0.6190**
Annotator 2	0.4682	-	0.5065
Annotator 3	**0.6190**	0.5065	-

Table 4.4: Pairwise-wise inter-annotator agreeability, with respect to the annotations of Perceived Individual's Experience of Conversation Quality. The overall agreement of all 3 annotators is 0.5313.

To better analyse the above results, we decided to further perform similar tests on each of the pairs of annotators. The pairwise agreements of the annotators (in terms of the weighted kappa scores) can be seen in Tables-4.3 and 4.4. The annotator-wise annotation distribution was also calculated and reported in Tables-4.1 and 4.2. The best agreeing pair, as seen in Tables-4.3 and 4.4, is the pair of *Annotator 1* and *Annotator 3*. From the Tables-4.1 and 4.2, we see that Annotator-2 (the annotator missing from the best pair) has the highest mean, lowest variance and the highest minimum. This distribution of *Annotator-2* responses suggests that the annotator has mostly tended to score higher for samples and was reluctant to score less, in relatively comparison with other two annotators.

Similar, situations of low inter-annotator agreeability has been handled by researchers in several fashion. The most widely used method is where the annotators were made to discuss together to reach a consensus on samples of lower agreeability. But, unfortunately we weren't able to do this due to unavailability of few annotators. As our next strategy, we checked for mean-shift between annotators in their annotations responses. A similar approach was used by Ringeval et al. (2013) [104] to account for variabilities in human judgements. The authors used a zero-mean (ZM) local normalization technique i.e., for each annotated item and for each annotator to remove an eventual bias in the annotation values, e.g., shifted toward positive or negative values. We performed a similar test and a similar plot to that of Figure-4.7 was plotted and can be seen in Figure-4.8. The resulting plots show that no major changes are seen post the ZM technique. This suggests that there exists no mean shift between annotators but there exists a basic difference in annotator judgements.

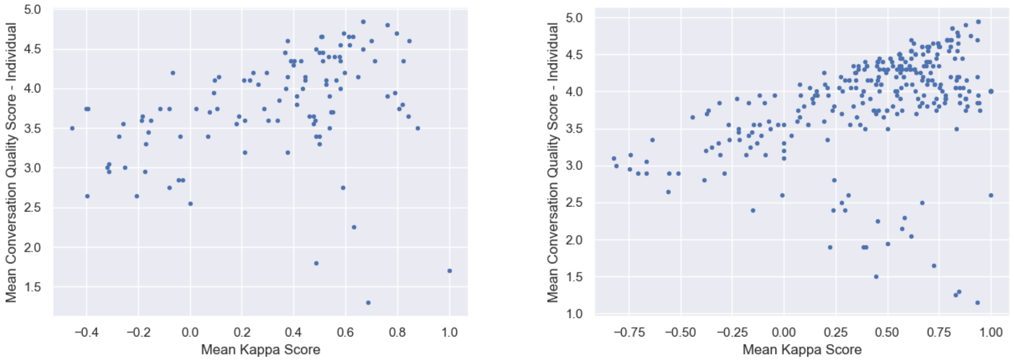


(a) Group-level mean weighted kappa scores (κ) with respect to the group's mean conversation quality score (After ZM adjustment)

(b) Individual-level mean weighted kappa scores (κ) with respect to the individual's mean conversation quality score (After ZM adjustment)

Figure 4.8: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses from all 3 annotators (After ZM adjustment)

Similar plots to that of Figure-4.7 were plotted for each of the pairs. The best agreeing pair's conversation quality score vs kappa score plot can be seen in Figure-4.9. Similarly, the other pairwise plots can be found in the Appendix-C, only the best agreeing pair is shown here below. From the Figure-4.9, we see that, unlike in



(a) Group-level mean weighted kappa scores (κ) with respect to the group's mean conversation quality score

(b) Individual-level mean weighted kappa scores (κ) with respect to the individual's mean conversation quality score

Figure 4.9: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses from the 2 annotators with maximum agreeability

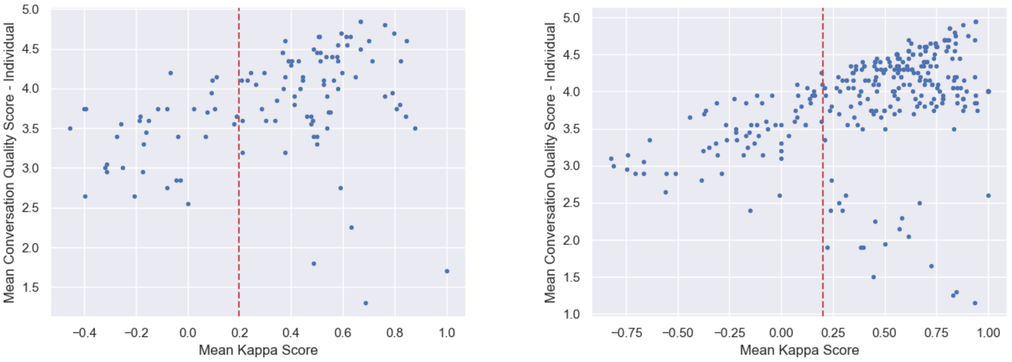
Figure-4.7, similar results to that of Hung et al. (2010) [56] are obtained. We see that *Annotator 1* and *Annotator 3* tend to agree more on conversation scores on both extremes of the scale and agree less on conversation scores that exist between both the extremes of the scale (min-range 3.0 to 3.5). In this case also, we observe that samples exhibiting scores of higher conversation quality tended to be more numerous than those exhibiting lower conversation quality scores. This is the same case for both the group-level and individual-level annotations. *With the above results in consideration, for further analysis we only considered the annotation responses received from the best agreeing pairs of annotators.* Due to the fact that the *Annotator-2*'s responses had the least variance across samples and the least agreeability amongst other annotators, we omit the responses from Annotator-2 for defining our final conversation quality ground-truth.

It is also important to note that results of high agreement in extremes of scale and low agreements on mid-ranges, are more evident in individual-level annotations (4.9b, 4.7b) than that in group-level annotations (4.9a, 4.7a). Only a very few groups of low conversation quality score have high agreement. This difference between group-level and individual-level annotations can also be seen in the pairwise and overall agreements seen in Tables-4.3 and 4.4. We see that individual-level annotations have a higher agreement than that of group-level annotations. This suggests that external annotators are more suitable for annotating individual-level social constructs than annotating group-level social constructs. In a group-level annotation the annotator is exposed to a larger hot-spot of events (hence larger data to be handled by annotators) than that of a individual-level annotation. Hence, annotators tend to differ amongst one another in the way in which they handle the larger data exposure, e.g. the way in which annotators handle the *aggregation of individuals* to annotate the group might differ largely amongst annotators. For further research in *conversation quality*, this is an important factor to consider with respect to external annotators.

4.3. Final Dataset and Analysis

In this section, we discuss the characteristics of the final dataset drafted for further experiments in this research. Firstly, we explain the design choices involved in drafting the final dataset with respect to the annotation analysis and missing data, which were explained earlier. Additionally, this section also presents the results of the data analysis performed on the final dataset.

Post the analysis of *Conversation Quality* annotations, we decided to decide on the final dataset based on the inter-annotator score for each sample annotations. The final conversation quality score for each sample (an individual or a group) is the average conversation quality scores given by each annotator for that particular sample. But, to have reliable ground-truths of conversation quality, it is important to consider the inter-annotator reliability scores calculated using the weighted kappa κ measure. The kappa measure ranges between +1.0 and -1.0. Because it is used as a measure of agreement, only positive values would be expected in most situations, negative values would indicate systematic disagreement. Several researchers have offered *rules of thumb* for interpreting the kappa score and the level of agreement, many of which agree in the gist even though the thresholds defined to explain the reliability vary amongst each other. Perhaps the first was Landis and Koch (1977) [72], who characterized kappa values *less than 0* as indicating *no agreement*, *from 0* *to 0.20 as slight, from 0.21 to 0.40 as fair, from 0.41 to 0.60 as moderate, from 0.61 to 0.80 as substantial,* and *from 0.81 to 1 as almost perfect agreement.*



(a) Kappa Threshold for Perceived Group's Conversation Quality.

(b) Kappa Threshold for Perceived Individual's Experience of Conversation Quality.

Figure 4.10: Mean Conversation Quality score Vs Mean Kappa score - Kappa Threshold at *0.2* (indicated by red dotted line).

For this research, though we used three external annotators to annotate the ground-truth, we defined the ground-truth for conversation quality only using the annotations collected from *Annotator 1* and *Annotator 3*. To form the final dataset, We used the threshold kappa agreement to be *0.2*. With respect to [72], a inter-annotator agreement score of *0.2* and above corresponds to an agreement of fair and above. Samples below this threshold were ignored and only samples with agreement greater than the threshold were considered for the final dataset. The dataset split with respect to the kappa threshold can be seen in Figure-4.10 (shown in red dotted lines). After applying the kappa threshold, we were left with *273* samples (reduced from *340*) for individual-level conversation quality and *81* samples (reduced from *115*) for group-level conversation quality.

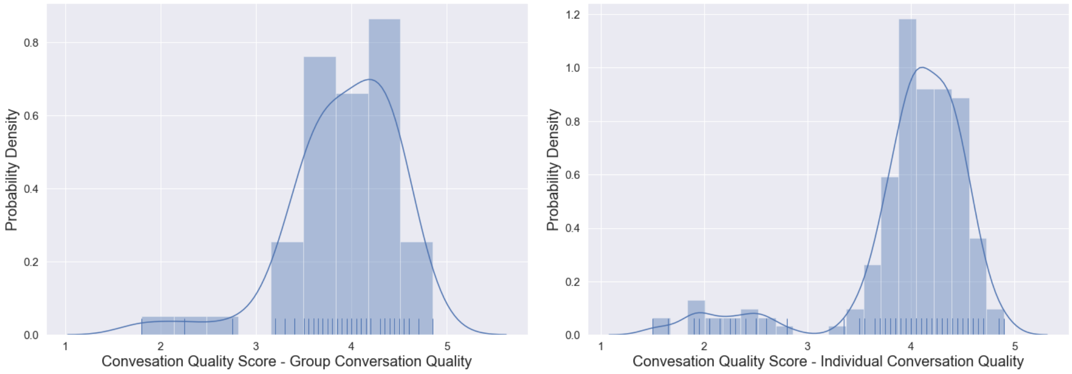
As mentioned earlier, there were few missing acceleration data for individuals in the MatchNMingle dataset (during mingling session). These missing acceleration data will affect the modeling of both the individual-level and group-level conversation quality. In individual-level modeling, as we are modeling group behaviour, incomplete group data affects the modeling of the individual in itself. In group-level modeling, incomplete group members might affect the modeling of the group-level phenomenon of conversation quality (annotated by annotators analysing the whole group). There are two ways in which we could handle this missing data,

1. Only use the F-formations (groups) if every participant has data. But, this might result in eliminating a large number of the data samples.

2. Use the F-formation (groups) if there is *enough* data. This can be done by using a threshold on the percentage of missing data, where threshold defines the concept of *enough* data. Samples with a high percentage (high number of individuals whose data is missing) should be eliminated and rest of the samples can be retained. For example, in cases of dyadic interaction and 50% of the data is missing (one person is missing), we should eliminate the interaction sample. At the same time, in a case of 5 individuals in the F-formation and only one person is missing, we have enough data to consider the sample.

For this research, we decided used the second strategy listed above. We used a missing data threshold of 50% for group-level modeling. That is, we only consider f-formation samples with *more than half of the data* (more than half of the population in the f-formation has available accelerometer data). After applying the missing data threshold, we were left with *179* samples (reduced from *273*) for individual-level conversation quality and *58* samples (reduced from *81*) for group-level conversation quality.

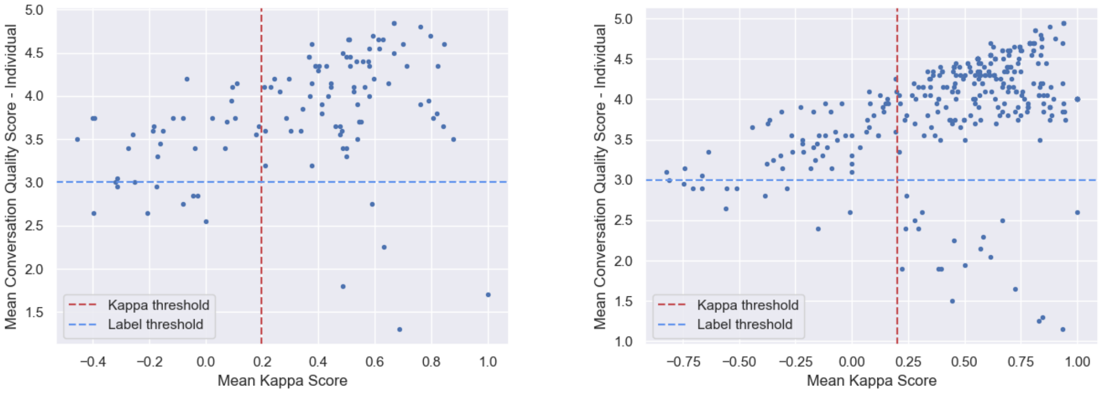
After filtering the dataset using the *kappa agreement* threshold and the *missing data* threshold, we drafted our final dataset. The final dataset consisted of - *179* samples (reduced from *340*) for individual-level conversation quality and *58* samples (reduced from 115) for group-level conversation quality. The distribution of the ground-truth labels of conversation quality, with respect to the final dataset prepared, can be seen in Figure-4.11. From the distribution of conversation quality labels in both the cases of individual-level and group-level modeling, we see that the data could be class imbalanced towards higher scales of conversation quality.



(a) Distribution of the ground-truth labels of Perceived Group's Conversation Quality.

(b) Distribution of the ground-truth labels of Perceived Individual's Experience of Conversation Quality.

Figure 4.11: Distribution of the ground-truth labels of Conversation Quality, in the final dataset.



(a) Kappa and Conversation Quality Threshold for Perceived Group's Conversation Quality.

(b) Kappa and Conversation Quality Threshold for Perceived Individual's Experience of Conversation Quality.

Figure 4.12: Mean Conversation Quality score Vs Mean Kappa score - Kappa Threshold at 0.2 (indicated by red dotted line) and Conversation Quality Threshold at 3.0 (indicated by blue dotted line).

For the predictive modeling of Conversation Quality, further explained in sections-5.4.2, we intend to classify the levels of conversation quality, that is, we intend to classify low conversation quality score samples from the high conversation quality scores. For this purpose, we used a Conversation Quality threshold of 3.0 (indicated by blue dotted line in Figure-4.12) and considered samples with scores less than 3.0 to Low Conversation Quality samples and the samples with scores greater than 3.0 to be High Conversation Quality samples. This results in a binary classification task and the results of such tasks can be seen in section-6.2. The final dataset and its class distribution, with respect to the binary classification task can be in seen in Table-4.5.

Low GroupCQ (≤ 3.0)	High GroupCQ (> 3.0)
3	55

(a) Class Distribution between High and Low *Perceived Group's Conversation Quality* (GroupCQ) scores. (Total = 58 samples)

Low IndivCQ (≤ 3.0)	High IndivCQ (> 3.0)
16	163

(b) Class Distribution between High and Low *Perceived Individual's Experience of Conversation Quality* (IndivCQ) scores. (Total = 179 samples)

Table 4.5: Class Distribution between High and Low Conversation Quality samples, after Kappa Threshold at 0.2 and Conversation Quality Threshold at 0.3. This is the final dataset used for the predictive modeling.

5
Methodology

In this chapter, we present the techniques involved in the predictive modeling and the analysis of the individual and group -level conversation quality measures. Firstly, we present an outline of the complete methodology by explaining the different processes and modules involved. Subsequently, we explain the data preprocessing and feature extraction techniques deployed in detail. Finally, we explain in detail the classification strategy used for the predictive modeling of the conversation quality measures. In the following sections, each of the modules and stages in the methodology will be discussed in more detail.

5.1. Introduction to Methodology

The prime goal of the methodology devised is to predict the measures of conversation quality (both group-level and individual-level) collected via annotations, by extracting pairwise movement synchrony features from bodily worn tri-axial accelerometers. The sub-goal of the methodology is to perform a quantitative study on the conversation quality ground-truths to study its properties and its relationship with the several features extracted from individuals in the conversation. An overview of the methodology used can be seen in Figure-5.1.

From the Figure-5.1, we see that our methodology, similar to many other machine learning projects, has four major modules - the *Preprocessing* module, the *Feature extraction* module, the *Modeling* module and the *Evaluation* module.

The first module - the *Preprocessing* module is an initial step to convert the noisy raw tri-axial accelerometer readings to low-level features. The module uses techniques like channel extraction and sliding window features to achieve this. This is further explained in Section-5.4. The second module is the *Feature extraction* module where higher-level behavioural features are extracted from the previously generated low-level features. In particular, in this module, we extract features that can be categorised broadly into three different categories of behavioural features - *Synchrony*, *Convergence* and *Turn-taking*. These features are further transformed into group-level features before modeling *Conversation Quality* (a measure of *group* behaviour involving several interlocutors). This is further explained in Section-5.3. The third module - the *Modeling* module is where the high-level behavioural features, which are rich representations of individual behaviour, are modeled using several techniques to predict the measure of *Conversation Quality* and also study its properties. This is further explained in Section-5.4. The final module is the *Evaluation* module where the results of the modeling techniques are validated and inferred, using several qualitative and quantitative analyses. This is further explained in Section-5.4.2.

5.2. Preprocessing

In this section, the techniques used in the preprocessing module that were applied to the low-level accelerometer data before feature extraction are presented and motivated. The preprocessing technique is intended to reduce the noise inherent in the accelerometer signals before behavioural features are extracted from them. The preprocessing module performs an initial step in data modeling where low-level features are extracted from the raw tri-axial acceleration signals recorded by the accelerometers that participants were wearing. Two key steps are involved in this module, firstly several new data channels (raw values, absolute values and

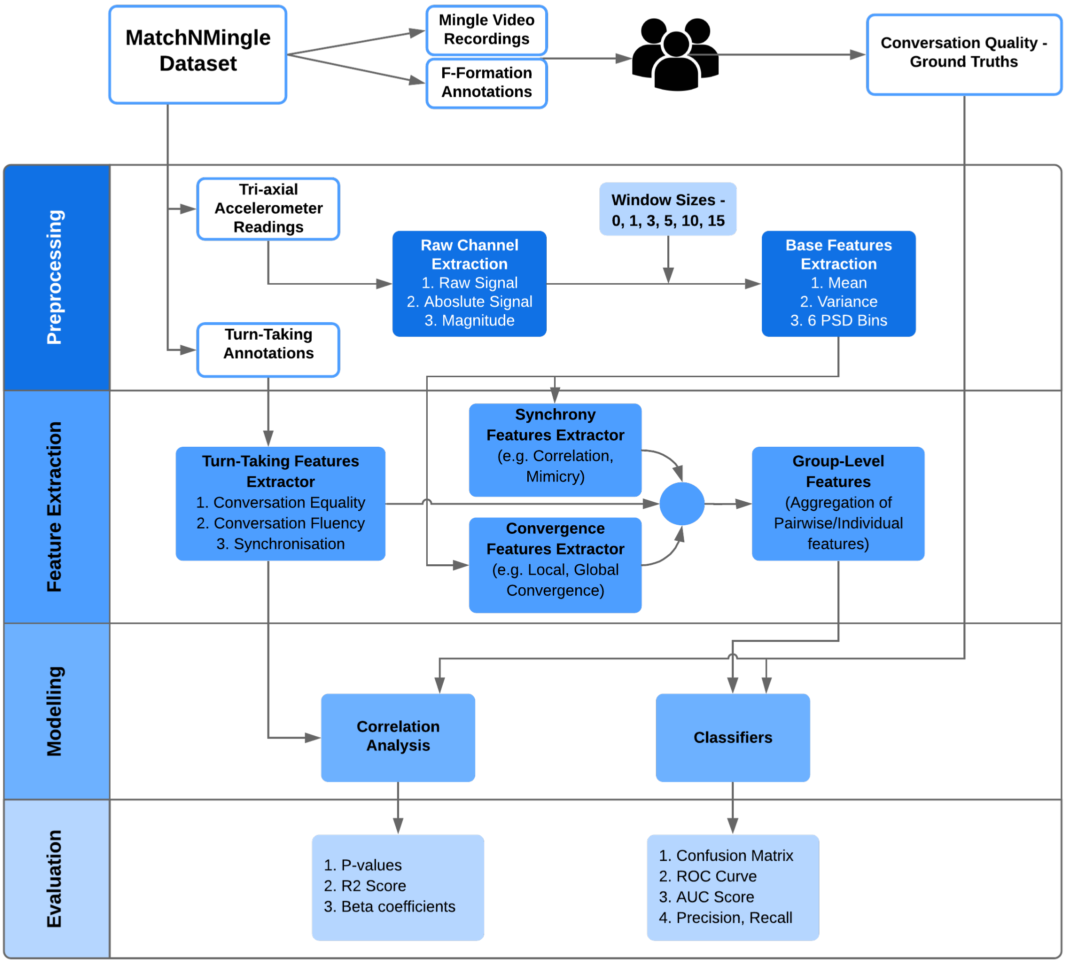


Figure 5.1: An overview of the methodology. The overview illustrates all the modules and techniques involved in the modeling of individual-level and group-level *Conversation Quality*. The methodology takes in the *MatchNMingle* dataset and *Conversation Quality annotations* as inputs, processes them across four different modules (explained further in Sections-5.4 to 5.4.2) and outputs predictions of *Conversation Quality* along with inferences with respect to the measure's properties.

its magnitude) are extracted from the raw accelerometer channels (x, y, z). Secondly, from the different channels extracted earlier, basic features (statistical and spectral features) are extracted from the signal using a sliding window approach. For this research, we use a similar preprocessing technique as [61] who modelled romantic interests in dyadic group setting, using the speed date data available in the MatchNMingle dataset.

The MatchNMingle dataset used tri-axial accelerometers, worn by all participants, to record their bodily movements. The accelerometer recorder bodily movements along 3-direction (3 dimensions), specifically, along the X axis - capturing the horizontal (left-right) movements, along the Y axis - capturing the vertical (up-down) movements and along the Z axis - capturing the 3rd dimensional forward-backward movements. The tri-axial accelerometer recorded the bodily movements at a sampling frequency of 20Hz, recording 20 fine-grained samples every second. The accelerometer readings are prone to noise and requires preprocessing.

First and foremost, each axis recording from the tri-axial accelerometer is normalized by calculating the z-score for each individual (along each axis). This technique help in removing the interpersonal differences in movement intensity, across participants. The z-scores for an individual is calculated using the mean and standard deviation values calculated from the whole participants data and not only within the f-formation duration. The normalized raw z-scores are then transformed to different data channels, each capturing an unique aspect of the individuals bodily movements. These different channels have been successfully used in several research works in existing literature to model human behavior using bodily worn accelerometer [61][39]. The channels extracted are,

1. The *raw* z-scores - captures the bodily movements of individuals along *with* the information regarding the direction of body movement along an axis. In free-standing conversation groups and f-formations, where individuals face each other, the direction of movement could provide valuable insights on the quality of the conversation.

2. The *absolute* z-scores - captures the bodily movements of individuals *without* the information regarding the direction of body movement along an axis. Sometimes, the direction of movement along an axis may not contribute much to the quality of the conversation, but a sense of movement along an axis might add values. And that is the reason why absolute values were one of the data channels extracted.

3. The *magnitude* of z-scores - calculated as $\sqrt{(x^2 + y^2 + z^2)}$, captures the overall magnitude of the bodily movements across directions and axes.

Each channel adds a new dimension of information to the low level accelerometer reading and it is important to use all these different interpretations (channels) of signal recordings for the modeling of *Conversation Quality*. The extraction of the above three channels from tri-axial accelerometer recordings results in seven different accelerometer channels (3 raw channels, 3 absolute channels and 1 magnitude channel).

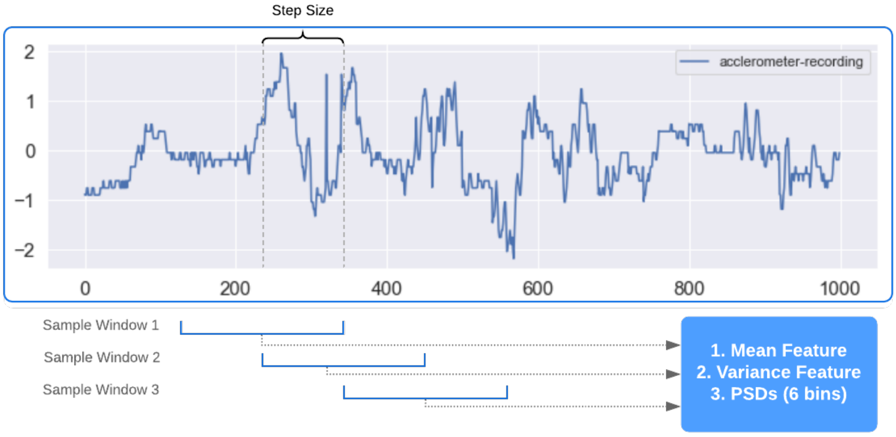


Figure 5.2: An illustration of a sliding window technique, performed as the preprocessing step. A sliding window of chosen window size is used to split the accelerometer channel into multiple segments. From each of these segments two statistical (mean and variance) and six spectral features (PSDs in 6 bins) are extracted and they are further used for extracting complex behavioral features.

The next step of preprocessing, before the feature extraction stage is the extraction of basic statistical and spectral features from the different data channels extracted previously. The data channels can be treated as they are and behavioral features can be extracted directly from them. However, the accelerometer reading might be heavily prone to noise. To handle this, in the literature, the studies that focused on extracting body movements from accelerometer based data have successfully used statistical and spectral features extracted from raw data using a sliding window approach [39][61][81]. Such an approach has been extensively used in literature for tasks like detecting social actions using similar accelerometer data [58]. The idea to extract features from these windows of interactions is based on a concept from the field of psychology known as *thin-slicing*, which was first described by Ambady and Rosenthal [2]. This concept describes the ability to find patterns in event based only on short windows or narrow slices. With such strong evidences, we decided to use this approach in this research. The above mentioned works in literature, used short window lengths to perform their experiments, as short window lengths are more suited for tasks like action detection. But in our case, we eventually are going to extract behavioural features based on coordination of body movements, and we cannot use a short window length as the coordinations might occur at different time points. A shorter window length might not capture these body movements and their coordinations. And a very large window length might capture more information on movements than required. Hence, the task of selecting a window size becomes harder in our case. To tackle this, similar to [61], we decided to use try varying time lengths for sliding windows and take a data-driven approach to select the best sliding window length. We also perform a post comparison with respect to the performances of the features that are extracted using different window lengths. Additionally, we also decided to hold the seven raw channels, without applying the sliding window approach also as one of the set of data channels for feature extraction. This is done in order to prevent any possible data loss which might happen during the sliding-window approach.

In Figure-5.2 we illustrate the sliding window process in detail. The sliding window technique to extract spectral and statistical features are performed on all the seven data channels extracted earlier. Each of the seven data channel, is divided into n-second windows using a sliding-window approach, with n/2 second shifts between each window. Since the optimal window size that captures necessary information is not known, the possible values of n are chosen as $[1, 3, 5, 10, 15]$ seconds. Further, the effect of different window sizes can be evaluated during the modeling phase.

Similar to [61], statistical and spectral features were extracted from each window extracted using the above mentioned process. The following features are extracted from the aforementioned seven data channels,

1. The *Statistical Features - mean and variance* over each of the window segments are computed. The mean is the sum of a collection of numbers divided by the count of numbers in the collection. And the variance is the expectation of the squared deviation of a random variable from its mean. Both these features are intended to reduce the noise present over the data channels.

2. The *Statistical Features - power spectral density (PSD)* over each of the window segments are computed using six logarithmically spaced bins between 0-10 Hz. Computing PSD is a method to convert a continues signal into a discrete form and shows how power of a signal is distributed over different frequencies. By dividing them into logarithmically spaced bins, the resolution at low frequencies is increased because lower frequencies contribute more to the power of the signal, but the dimensionality is kept low. Each bin gives information about the characteristic of behavior of the person at that time window, therefore each bin is treated as a single feature.

The combination of the aforementioned features results in *eight* feature dimensions per window segment. And the calculation of eight sliding window features for each of the seven data channels extracted previously results in 56 dimensions per window segment. Since the complex features are computed with the sliding-window approach with 4 different window sizes, in total 224 different feature categories are extracted, that will further be used to extract the behavioral coordination features. Additionally, we also decided hold on to the raw version of the seven data channels, for reasons explained earlier. This results in a total of $224+7 = 231$ base feature dimensions after preprocessing. While computing synchrony based behavioral features, each of the *231* dimension will be treated separately for feature extraction. The feature extraction technique performed on the *231* dimension preprocessed accelerometer signal is explained in detail in the subsequent section.

5.3. Feature Extraction

This section presents in detail the individual and pairwise behavioural features extracted from the *231* dimension preprocessed accelerometer data to model the measure of conversation quality. The features are inspired from works in literature which have successfully used the features to model different aspect of social interactions. Specifically, in this research, we use three unique set of bodily coordination based features - *synchrony, convergence and causality*, extracted from bodily worn accelerometers. The synchrony, convergence and causality features are extracted for all the *231* dimensional previously extracted preprocessed accelerometer signal. The three coordination based features mentioned above, capture unique phenomena of any signal pairs. From the literature review presented earlier, we saw that turn-taking features are also capable of explaining subtle social constructs. Hence, for further analysis, in addition to the bodily coordination based features, we also extract several turn-taking features from the speaking status annotations available along with the MatchNMingle dataset.

The subsequent sections will explain all the features extracted - *Synchrony, Convergence, Causality* and *Turn-Taking* features in detail.

5.3.1. Synchrony

Delahercehe et al. (2012) [30] define synchrony as the, "dynamic and reciprocal adaptation of the temporal structure of behaviors between interactive partners". In this work, the authors presented a review on the different techniques used to quantify the interpersonal measure of synchrony. Delaherche et al. state that the measures of synchrony is capable of giving a global snapshot of the interaction, basing their statement on the results of several correlation analysis. In this research, we use synchrony as one of the measures of bodily movement coordination between individuals in a conversational group. Synchrony based features have been successful in modeling several social constructs such as cohesion [88], romantic interest [61] and rapport [47]. At the same, constructs such as cohesion and rapport are closely related to *Conversation Quality*. In fact, the *degree of rapport* is one of the constituents of *Conversation Quality*. The aforementioned facts are the motivation to adopt synchrony as one the set of features to model conversation quality.

In this research, to study conversation quality, we extract four unique measures of interpersonal synchrony. These measures have been inspired from [88] (modeling cohesion), [61] (modeling romantic interest), [75] (studying entrainment), [39] (detecting social groups) and [40] (modeling social experience). Specifically, we extract four unique measures of interpersonal synchrony, such as,

1. *Correlation* - Statistically, correlation captures the causal relationship between two random variables or bivariate data. In our case, we use correlation to capture the causal relationship between the bodily movements of two individuals indulging in a social interaction.

2. *Time Lagged Correlation* - Similar to correlation, the time-lagged version of it, captures the same causal relationship between the bodily movements but in a time step lagged manner.

3. *Mutual Information* - Statistically, the mutual information between two random variables is a measure of the mutual dependence between the two variables. In our case, we use this measure to capture the dependence in bodily movements from one individual to another.

4. *Mimicry* - This particular measure of synchrony captures the extent to which the data samples for one individual follows the distribution of their interaction partners.

All the above introduced measures of synchrony are explained further in detail along with their method of implementation in the subsequent sections.

Correlation The measure of *correlation* has been extensively used in literature, to capture the pairwise similarity of body motion between two people [101][61]. As a measure of correlation, in this research, we use the *Pearson correaltion coefficient* (using the *pearsonr* method available in the scipy package [118]) to measure the correlation. The pearson correlation coefficient is calculated as follows,

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sigma(X)\sigma(Y)} \tag{5.1}$$

where, x and y are the preprocessed accelerometer data of person X and Y, x_i and $y+i$ are the data values of x and y respectively at time-step i, μ_x and μ_y are the means of x and y respectively, and σ_y and σ_y are the standard-deviations of x and y respectively.

The Pearson correlation coefficient, by capturing the linear relationship between two signals, intends to capture the degree of similarity in the bodily movements of two interacting individuals. The pearson correlation varies between -1 and +1, with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y (directly proposition). Negative correlations imply that as x increases, y decreases (inversely proposition). With that said, it is important to note that even a high negative correlation will be capable of explaining the phenomenon of coordination. With respect to the measure of correlation between interlocutors, we hypothesis that, larger the absolute value of the correlation coefficient, larger the degree of coordination is, thereby, will result in high *Conversation Quality*.

Time-Lagged Correlation The previously presented measure of correlation coefficient, measures similarity between signals which are exactly in synchrony (with respect to time) with one another. But, there might be cases where similarity between signals exist in a time-lagged fashion. To capture this phenomenon, we use the measure of *Time-Lagged Correlation*. The time-lagged correlation is computed using pearson correlation coefficients at different time lags, as follows,

$$\rho_{XY} = \frac{\sum_{i=1}^{N-\tau}(x_i - \mu_x)(y_{i+\tau} - \mu_y)}{\sigma(X)\sigma(Y)} \tag{5.2}$$

where, X and Y are the preprocessed accelerometer data of person X and Y, x_i and $y+i$ are the data values of x and y respectively at time-step i, μ_x and μ_y are the means of X and Y respectively, and σ_y and σ_y are the standard-deviations of X and Y respectively. More importantly, the variable τ denotes the time-lag, that is, the positive time-lag in terms of time steps between X and Y.

As features, to capture the time-lagged correlation, we use four unique measures,

1. Maximum Correlation - This particular measure captures the maximum *positive* pearson correlation coefficient between x an y, irrespective of the time-lag it happens in. With respect to this feature, we hypothesise that the maximum *positive* correlation captures the maximum similarity between two signals irrespective of the time-lag, and larger its value implies a larger coordination, thereby, will result in high *Conversation Quality*.

2. Minimum Correlation - This particular measure captures the maximum *negative* pearson correlation coefficient between x an y, irrespective of the time-lag it happens in. With respect to this feature, we hypothesise that the maximum *negative* correlation captures the maximum inversely propositional similarity between two signals irrespective of the time-lag, and larger its value implies a larger coordination, thereby, will result in high *Conversation Quality*.

3. Time-Lag of Max. Correlation - This particular measure captures the time-lag required to achieve the maximum *positive* correlation. With respect to this feature, we hypothesise that shorter the time-lag used by interlocutors to achieve high positive correlation, better the coordination, thereby, will result in high *Conversation Quality*.

4. Time-Lag of Min. Correlation - This particular measure captures the time-lag required to achieve the maximum *negative* correlation. With respect to this feature, we hypothesise that shorter the time-lag used by interlocutors to achieve high negative correlation, better the coordination, thereby, will result in high *Conversation Quality*.

To calculate the above four features, we use the *correlate* method available in the *numpy* package [121]. The function when used with mode parameter as "full" performs the cross correlation with all possible time-lags. Then, by using aggregation functions such as max, min, argmax and argmin on the cross-correlations, we compute Maximum Correlation, Minimum Correlation, Time-Lag of Max.Correlation and Time-Lag of Min.Correlation features respectively.

Mutual Information The Mutual Information (MI) is a measure of synchrony which captures the mutual dependence between two signals or variables. In other words, it measures the "amount of information" gained on one signal by observing another signal. The measure of MI has been used widely in literature to measure the co-occurrences between two people's behavior and model constructs such as enjoyment [81] and romantic interest [61]. In our case, we use this measure of synchrony to capture the degree of dependence of signal values between two interlocutors. It is calculated as follows,

$$MI(X,Y) = \frac{H(X) + H(Y) - H(X,Y)}{\sqrt{H(X)H(Y)}} \tag{5.3}$$

where H(X) and H(Y) denotes the entropy of preprocessed accelerometer data of person X and person Y, and H(X, Y) represents the joint entropy of these preprocessed accelerometer data X and Y. The mutual information measure ranges between values of 0 and 1. A score closer to 1 is obtained when interlocutors have an influence on each other's behavior and a score of 0 is obtained when interlocutors have an absolutely no influence on each other's behavior. With respect to this particular feature, we hypothesise that closer the value of MI is to zero, larger the mutual dependence and coordination, thereby, will result in high *Conversation Quality*.

To calculate this feature, we use the Mutual Information calculator available in the *SyncPy* package [116]. This particular calculates mutual information through a clustering based approach. That is, it computes mutual information from entropy values (H(X), Y(X)) from estimates of k-nearest-neighbours distances. This particular approach was first proposed by Kraskov et al. (2004) [68].

Mimicry Feese et al. (2012) [34] define mimicry as the event when a person follows the behavior of their interaction partner by displaying the same nonverbal cue just after their interaction partner had displayed it. In this measure, we use the similar mimicry features as of Nanninga et al. [88]. The mimicry metric used in [88] was originally extracted from paralinguistic signals, in our case, we extract these features from the preprocessed accelerometer data.

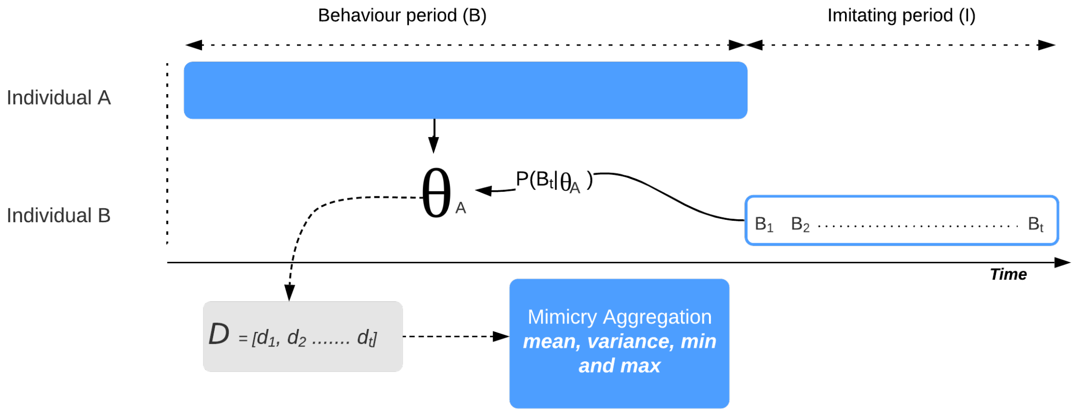


Figure 5.3: An illustration of the mimicry based feature extraction. The illustration shows the extraction of Lagged Mimicry for Individual B and Lead Mimicry for Individual A. The θ is the learnt model from A's behaviour period and $B_1, B_2.....B_t$ are the data samples at each time stamp of Individual B. A distance vector D is computed with respected to the probability values $P(B_t|\theta_A)$, which is later used to extract aggregate based mimicry features.

With respect to [34] and [88], the goal mimicry feature extraction is to capture the mirroring phenomenon (two interlocutors imitate each other) between interaction partner pairs. The illustration of the mimicry feature extraction technique can be seen in Figure-5.3. As seen in the image, the accelerometer data of interacting partner pairs are split into a behaviour period (B) and imitating period (I). With B and I, we intend to capture the degree of imitation at I with respect to the action at B. To achieve this, we fit Gaussian Mixture Models (GMM) on *each individual in a interacting pair*, by only using the behaviour period (B) of a particular person (In Figure-5.3 it is Person A). Similar to [88], model complexity was empirically optimised using the Bayes Information Criterion (BIC). The best fit model is a GMM with 4 Gaussian components and full covariance matrix. With the fit GMMs (using data from B), as shown in Figure-5.3, likelihood scores for *every sample-point* at I (I period of the interacting partner in the pair. In Figure-5.3 it is Person B) was calculated, resulting in an array of likelihood scores for each sample-point at I. Then, aggregation functions namely *min, max, variance and mean*, were used to compute the final mimicry scores. It is important to note that mimicry features are asymmetric, hence, the above mentioned process was performed twice for a pair using the periods of B and I from both the individuals in the interacting pair (Lead and Lagged Mimicry). That is for an individual in the interacting pair, both the Lead Mimicry (when the individual's data is considered as the behaviour period, *B*) and Lagged Mimicry (when the individual's data is considered as the imitation period, *I*) are calculated. As can be seen in Figure-5.3, the above process of extracting mimicry features, results in *four* features (min, max, variance and mean), and while also computing the asymmetric mimicry features it results in a total of *eight* features for an interacting pair.

With respect to the min, max, variance and mean features of mimicry, we hypothesis that, for an interacting pair, larger the min, max and mean values of mimicry scores, and smaller the variance of mimicry score, higher the degree of coordination is, thereby, will result in high *Conversation Quality*.

An overview of all the synchrony-based features extracted can be seen in Table-5.1.

	Feature Set Type	**Feature Variant**	**Number of Features**
1	Correlation	correlation coefficient (ρ_{xy})	1
2	Time-lagged Correlation	min, max, argmin, argmax	4
3	Mutual Information	min, max, mean, variance	4
4	Mimicry	lag_min, lag_max, lag_mean, lag_variance, lead_min, lead_max, lead_mean, lead_variance	8

Table 5.1: An overview of the 4 sets of synchrony based features extracted for an interacting pair.

5.3.2. Causality

In statistics, the phrase "correlation does not imply causation" refers to the inability of correlation statistics to legitimately capture the causal effect (cause-and-effect relationship) between two signals solely on the basis of an observed association or correlation between them [1]. Hence, it is important to also capture the causal effect using a different measure of coordination and synchrony. In this section, we discuss the two causality features extracted in this study to study the causal effect based coordination,

1. *Coherence* - In signal processing, under certain conditions, coherence can be used to capture the causality between the two signals. This particular measure is hence used to capture the causality in body coordination of two interlocutors.

2. *Granger's Causality* - The Granger's causality test is a statistical test which, similar to coherence, captures the causality of one signal over another but in a different manner. This particular measures capture the causality by estimating whether one signal is useful in forecasting the other signal.

Coherence In signal processing, under certain conditions, the coherence metric can be used to capture the causality between the two signals. Unlike mimicry and correlation metrics which characterise the temporal elements of signals, the coherence metric characterises the signals using the frequency domain information of a signal. Coherence in literature, has been widely implemented to study the dynamic functional connectivity in the brain networks. In studying social signal processing, Richardson et al. (2005) [103] have used coherence based methods to study discourse comprehension of speakers and listeners. The coherence between two signals X and Y can be measured as follows,

$$C_{XY}(f) = \frac{|G_{XY}(f)|^2}{G_{XX}(f)G_{YY}(f)} \tag{5.4}$$

where, X and Y are the preprocessed accelerometer data of person X and Y, $G_{XY}(f)$ corresponds to the cross-spectral density of a signal and $G_{XX}(f)$ and $G_{YY}(f)$ correspond to the auto-spectral density of signals X and Y respectively. Values of coherence will always satisfy the property: $0 \leq C_{XY}(f) \leq 1$. To calculate this feature, we use the Coherence calculator available in the *SyncPy* package [116]. With respect to this particular measure of synchrony, we hypothesis that, larger the value of the coherence metric, larger the degree of cause-and-effect relationship and coordination, thereby, will result in high *Conversation Quality*.

Granger's Causality The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another, first proposed by Clive Granger (1969) [44]. Unlike coherence which captures causality in the frequency domain, granger's causality captures the "predictive causality", that is, it captures the causality by estimating whether one signal is useful in forecasting the other signal. The Granger causality analysis is usually performed by fitting a series of vector autoregressive model (VAR) to the signals, with varying lags (between 1 and maximum lag possible). In particular, let $X(t) \in \mathbb{R}^{d\times 1}$ for $t = 1, \ldots, T$ be a d-dimensional multivariate signal. Granger causality is performed by fitting a VAR model with L time-lags as follows,

$$X(t) = \sum_{\tau=1}^{L} A_\tau X(t-\tau) + \varepsilon(t) \tag{5.5}$$

where $\varepsilon(t)$ is a white Gaussian random vector, and A_τ is a matrix for every τ. A signal X_i is called a granger cause of another time series X_j, if at least one of the elements $A_\tau(j,i)$ for $\tau = 1, \ldots, L$ is significantly larger than zero (in absolute value). In other words, an *f-test* is performed on the Ordinary Least Squares (OLD) model with the optimal lag (estimated using the BIC criterion), resulting in a f-value and a p-value which is open for interpretation.

For this research, we use the Granger Causality calculator available in the *SyncPy* package [116]. The resulting *f-value* obtained from the granger's causality test is directly used as the feature for further modeling. With respect to this particular measure of synchrony, we hypothesis that larger the absolute f-value metric, larger the degree of predictive causality and coordination, thereby, will result in high *Conversation Quality*.

An overview of all the causality-based features extracted can be seen in Table-5.2.

	Feature Set Type	Feature Variants	Number of Features
1	Coherence	min, max	2
2	Granger's Causality	f_value	1

Table 5.2: An overview of the 2 sets of causality based features extracted for an interacting pair.

5.3.3. Convergence

Edlund et al. (2009) [32] considers *convergence* as the increasing similarity between interacting partners with time, in a spoken dialogue system (SDS). Unlike synchrony, which captures the presence of coordination, convergence intends to capture the dynamics (the decrease/increase in coordination across time) in coordination. In our research, we use three specific metrics of convergence,

1. *Symmetric Convergence* - The symmetric convergence captures the decrease or increase in similarity between any two signals along time, *without any lag* between the two signals. In our case, it captures the change in similarity (in body movements) across time, between two interacting partners.

2. *Asymmetric Convergence* The asymmetric convergence captures the decrease or increase in similarity between any two signals along time, *with a time-lag* between the two signals. In our case, it captures the change in similarity (in body movements) across time, between two interacting partners, with one individual's movement data time-lagged from their interacting partner.

3. *Global Convergence* - The global convergence captures the change in similarity between two signals, specifically between its initial time-segments and its later time-segments. In our case, it captures the overall change in similarity (in body movements) between interacting partners, specifically between the initial segments of their conversation and the later segments of their conversation.

These three convergence features are inspired from [61] [88][83]. All the above introduced measures of convergence are explained further in detail along with their method of implementation in the subsequent sections.

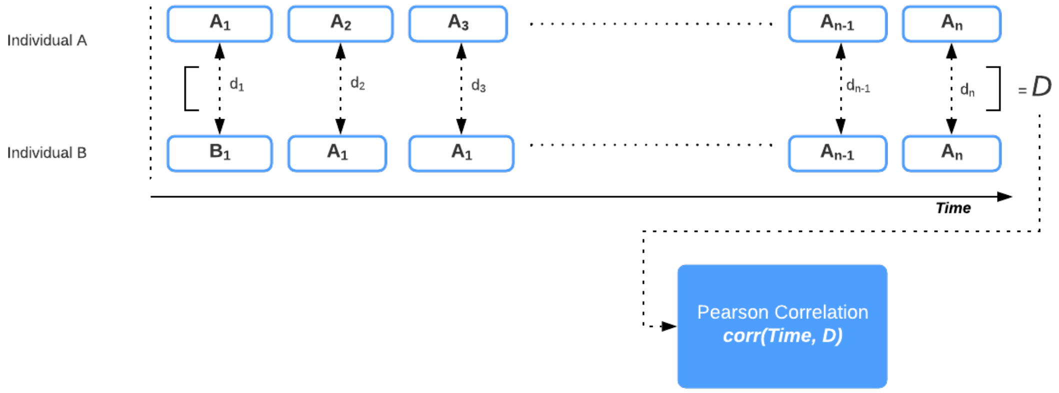


Figure 5.4: An illustration of the symmetric convergence based feature extraction. The illustration shows the extraction of symmetric convergence between interacting partners Individual A and B. $A_1, A_2....A_n$ is the vector representing Individual A's accelerometer channel and similarly, $B_1, B_2....B_n$ is the vector representing Individual B's accelerometer channel. A distance vector D is calculated using squared distance between the A and B's data samples. Finally, D is used to computer the correlation with *time*, to capture the evolving similarity.

Symmetric Convergence For an interacting pair, the measure of symmetric convergence captures the increase or decrease of similarity in their behavior at each time step over time. An overview of the methodology implemented to extract symmetric convergence between an interacting pair can be seen in Figure-5.4. As an initial step, similarity between the two preprocessed accelerometer data (from the interacting pair) is calculated at every time-step. The similarity at a time-step is computed as the *squared distance* between data-values of the corresponding time-step, resulting in a vector of similarity scores $D = [d_0, d_1, d_2, \ldots, d_n]$. Finally, the correlation coefficient between the similarity scores vector D and the corresponding time-step is

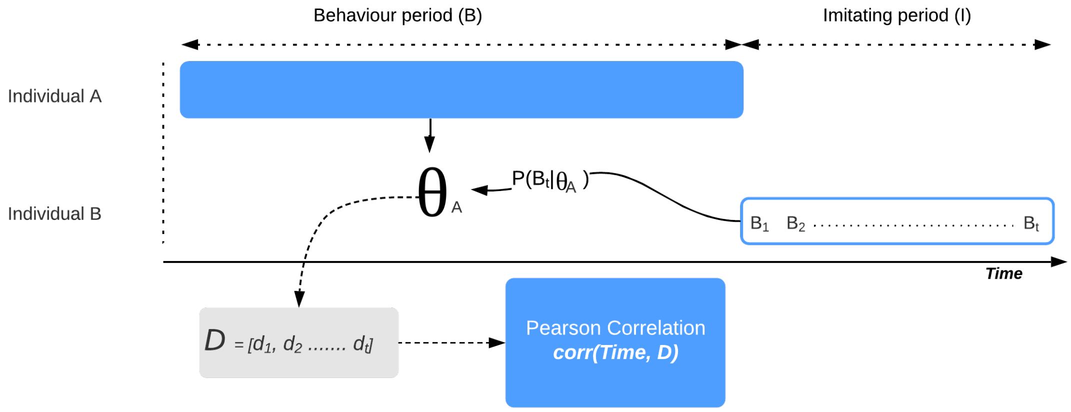


Figure 5.5: An illustration of the asymmetric convergence based feature extraction. The illustration shows the extraction of asymmetric convergence between interacting partners Individual A and B. The θ is the learnt model from A's behaviour period and $B_1, B_2.....B_t$ are the data samples at each time stamp of Individual B. A distance vector D is computed with respected to the probability values $P(B_t|\theta_A)$, which is later used to computer the correlation with *time*, to capture the evolving similarity.

computed, using the pearson correlation. For pearson correlation formula (5.3.1), X becomes the difference scores vector D and Y becomes the corresponding time-step $(1, 2, 3, \ldots, N$, where N is the length of the vector D). By doing this, we intend the capture the increase or decrease of similarity in behaviour of interacting pairs.

With respect to the measure of symmetric convergence between interlocutors, we hypothesis that, larger the resulting correlation coefficient, larger the convergence is, thereby, resulting in a high *Conversation Quality*.

Asymmetric Convergence The measure of asymmetric convergence the same aspect of coordination as the symmetric convergence, but in a time lagged manner. This particular measure was inspired from [88] and [61]. An overview of the methodology implemented to extract asymmetric convergence between an interacting pair can be seen in Figure-5.5. Firstly, similar to the methodology involved in estimating mimicry features, the accelerometer data was split in to two periods - behaviour period (B) and imitating period (I). Using a similar strategy as in mimicry methodology, a GMM was fit on the behaviour periods (B) of both the interacting partners. Then, likelihood scores for each sample in the imitating period (I) of an individual is calculated, with respect the to the GMM model fit on their partner's behaviour period (B), resulting in a vector of similarity scores $D = [d_0, d_1, d_2, \ldots, d_n]$. Finally, the correlation coefficient between the similarity scores vector D and the corresponding time-step is computed, using the pearson correlation. For pearson correlation formula (5.3.1), X becomes the difference scores vector D and Y becomes the corresponding time-step $(1, 2, 3, \ldots, N$, where N is the length of the vector D). By doing this, we intend the capture the increase or decrease of similarity in behaviour of interacting pairs. With that said, it is important to note that, the asymmetric convergence is an asymmetric feature, hence, the same process is performed twice for an interacting pair, once with each individual data as the behaviour period (B).

With respect to the measure of asymmetric convergence between interlocutors, we hypothesis that, larger the resulting correlation coefficient, larger the convergence is, thereby, resulting in a high *Conversation Quality*.

Global Convergence The measure of global convergence the same aspect of coordination as the symmetric convergence, but in a global manner. In other words, for an interacting pair, the global convergence is the decrease or increase in their behavioural similarity between their initial and later parts of their conversation. An overview of the methodology implemented to extract global convergence between an interacting pair can be seen in Figure-5.6. To compute this feature for an interacting pair, as an initial step, the preprocessed accelerometer data of both the individuals is split into two equal halves. The squared distance (euclidean distance) between the two individual's data is calculated for both the halves, resulting in two distance scores of d_1 (inter-individual distance at first half) and d_2 (inter-individual distance at second half). And finally, as

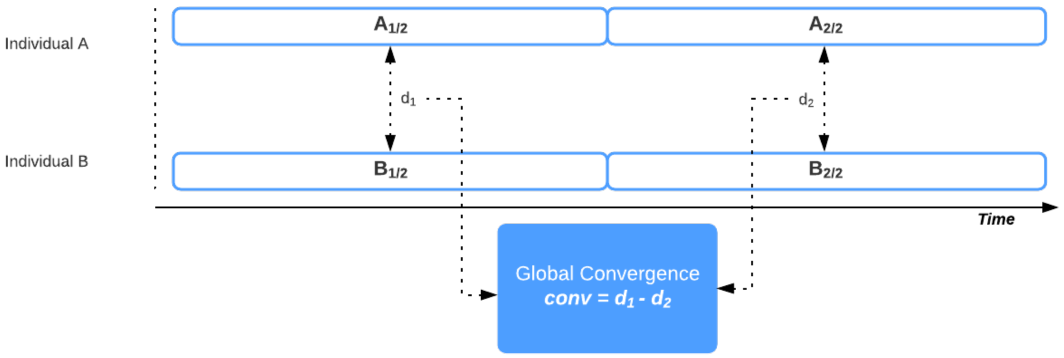


Figure 5.6: An illustration of the global convergence based feature extraction. The illustration shows the extraction of global convergence between interacting partners Individual A and B. Both A and B's accelerometer channels are split into two halves ($A_{1/2}$), $A_{2/2}$), $B_{1/2}$) and $B_{2/2}$)) and squared distances are computed with respect to the two halves for the individuals. Finally, the global convergence is the difference between the squared distances (d_1 and d_2).

the estimate of global convergence, the two scores are subtracted as,

$$c = d_1 - d_2 \tag{5.6}$$

The similarity scores are expected to decrease over time. Hence, a scenario of $c > 0$ ($d_2 \leq d_1$) is expected to have better coordination, that a scenario of $c \leq 0$ ($d_2 \geq d_1$) is expected to have poor coordination. Hence, with respect to the measure of global convergence between interlocutors, we hypothesis that, larger the resulting correlation coefficient, larger the convergence is, thereby, resulting in a high *Conversation Quality*.

An overview of the convergence-based features extracted can be seen in Table-5.3.

	Feature Set Type	Feature Variants	Number of Features
1	Symmetric Convergence	ρ	1
2	Asymmetric Convergence	lag, lead	2
3	Global Convergence	$\mathrm{d}_1 - d_2$	1

Table 5.3: An overview of the 3 sets of convergence based features extracted for an interacting pair.

5.3.4. Group-level Feature

The bodily coordination features extracted (Sections-5.3.1, 5.3.2 and 5.3.3) are pairwise features and are only indicative of dyadic phenomena and constructs. But, the *Conversation Quality* is a group-level construct and group-level features are required to model the construct. To estimate group-level features from previously computed pairwise features, we use a strategy widely used by researchers in modeling group-level constructs. Nanninga et al. (2017) [88] and Zhang et al. (2018) [129], while modeling social and task cohesion, use aggregation strategy to extract group-level features from pairwise features. Such an aggregation strategy considers a group to be a collection of dyads (pairs) and employs several aggregation techniques on the array of pairwise features to extract group-level features. Such a strategy is used with a hypothesis that, the distributions of the group's pairwise features (estimated by the aggregation functions) are indicative of the group-level phenomena like cohesion and conversation quality.

For this research, we extract *six* unique group-level features for each of the pairwise features extracted. That is, the *six* group-level features are calculated for each of the pariwise feature variants seen in Tables-5.1, 5.2 and 5.3 and intended to capture a particular aspect of the distribution of pairwise features in a group. The six group-level features are as follows,

1. *Minimum Pair Value* - For a collection of interacting pairs in a group, this group-level feature captures the interacting pair with the *minimal* value, for a particular pairwise feature type. With respect to this group-level feature, we hypothesise that larger the resulting *minimum pair value* of a group, larger is the group's coordination, thereby, resulting in a high *Conversation Quality*.

2. *Maximum Pair Value* - For a collection of interacting pairs in a group, this group-level feature captures the interacting pair with the *maximum* value, for a particular pairwise feature type. With respect to this group-level feature, we hypothesise that larger the resulting *maximum pair value* of a group, larger is the group's coordination, thereby, resulting in a high *Conversation Quality*.

3. *Mean of Pair Values* - For a collection of interacting pairs in a group, this group-level feature captures the overall *average* of all the pair values, for a particular pairwise feature type. With respect to this group-level feature, we hypothesise that larger the resulting *mean of pair values*, larger is the group's coordination, thereby, resulting in a high *Conversation Quality*.

4. *Variance in Pair Values* - For a collection of interacting pairs in a group, this group-level feature captures the overall *variance* between the pair values, for a particular pairwise feature type. With respect to this group-level feature, we hypothesise that *smaller* the resulting *variance of pair values*, larger is the group's coordination, thereby, resulting in a high *Conversation Quality*.

5. *Median of Pair Values* - For a collection of interacting pairs in a group, this group-level feature captures the *statistical median* of the pair values, for a particular pairwise feature type. With respect to this group-level feature, we hypothesise that closer the resulting *median of pair values* to that of the mean, larger is the group's coordination, thereby, resulting in a high *Conversation Quality*.

6. *Mode of Pair Values* - For a collection of interacting pairs in a group, this group-level feature captures the *statistical mode* of the pair values, for a particular pairwise feature type. With respect to this group-level feature, we hypothesise that closer the resulting *mode of pair values* to that of the mean, larger is the group's coordination, thereby, resulting in a high *Conversation Quality*.

5.3.5. Turn-Taking

Though the prime goal of this research is to model *Conversation Quality* with coordination based features, in addition to the synchrony and convergence features extracted, we also extract a set of individual-level and group-level turn-taking features. From literature, we discussed how turn-taking features have been state-of-the-art for modeling several social constructs. Hence, it would be an interesting contribution to study the effect of several turn-taking features on the *Conversation Quality*. These turn-taking features will be used alongside the coordination features for the predictive modeling and correlation analysis experiments. The turn-taking features are extracted under three categories, with respect to the three aspects of the conversation suggested by Lindley and Monk [79] - *Conversation Equality*, *Conversation Fluency* and *Conversation Synchronisation*.

The turn-taking features were extracted by assuming a speaking turn (*spurts*) to be a continuous speaking activity segment separated by at least *500ms* silence periods. This methodology is similar to that employed in [71] to model participant affect in meetings and [79] to model social experience. In this research, under the category of *Conversation Equality*, the following turn-taking features were used,

1. *Degree of Equality* - As a group-level feature, we calculate this feature using the equation - 2.1. As a individual-level feature, we calculate this feature with a slight modification on the equation - 2.1. The modified equation is as follows,

$$P_{eq} = \frac{T_i - T}{T} \tag{5.7}$$

where T_i is the total turn-duration for individual i and $T = (\sum_i^N (T_i)/N)$. With respect to the conversation equality feature, we hypothesise that larger the resulting *conversation equality*, larger is the group's equality in terms of speaking time, thereby, resulting in a high *Conversation Quality*.

Under the category of *Conversation Fluency*, the following turn-taking features were used,

1. *Percentage of Silence* - As an individual-level feature, the percentage of silence is simply the the percentage of time the individual in the group has the speaking status to be zero. And, as a group-level feature, the percentage of silence is simply the the percentage of time all the individuals (whole group) has their speaking status to be zero. It is calculated as follows,

$$P_s = \frac{T_{len} - \sum_i^{T_{len}} t_s}{T_{len}} \tag{5.8}$$

where t_s is the speaking status vector for the individual/group and T_{len} is the total talk-duration of the group's conversation or length of t_s. With respect to the percentage silence feature, we hypothesise that, smaller the resulting *percentage of silence*, larger is the group's conversation fluency, thereby, resulting in a high *Conversation Quality*.

2. *Number of Back-channels* - In this research, we consider a back-channel to be a Very Short Utterances (VSUs) [70] with continuous speaking status durations *between 500 ms* and *1 sec*. As an individual-level feature, we simply count the number of back-channels used by the individual. As an group-level feature, we simply count the number of back-channels exchanged in the whole group conversation. With respect to the number of back-channels feature, we hypothesise that larger the resulting *number of back-channels*, larger is the group's conversation fluency, thereby, resulting in a high *Conversation Quality*.

Under the category of *Conversation Synchronisation*, the following turn-taking features were used,

1. *Percentage of Overlap* - As an individual-level feature, we calculate the percentage of overlap as follows,

$$P_o = \frac{\sum_{x=1}^{T_{len}} [i_x = g_x]}{T_{len}} \tag{5.9}$$

where i_x and g_x are the speaking status of an individual and the rest of the group-members at time-step x and T_{len} is the total talk-duration of the group's conversation. As a group-level feature, we calculate the percentage of overlap as follows,

$$P_o = \frac{\sum_{x=1}^{T_{len}} [c_x > 1]}{T_{len}} \tag{5.10}$$

where c_x is the number of current speakers in the group conversation at time-step x. With respect to the percentage of overlap feature, we hypothesise that smaller the resulting *percentage of overlap*, larger is the group's conversation synchronisation, thereby, resulting in a high *Conversation Quality*.

2. *Number of Successful Interruption* - We define a successful interruption similar to that of Hung et al. (2010) [56], seen previously in the literature review. As an individual-level feature, we simply count the occurrences of such interruptions, with respect to the individual's speaking turn. As a group-level feature, we simply sum the number of successful interruptions across individuals in the group. With respect to the successful interruption feature, we hypothesise that smaller the resulting *number of successful interruption*, larger is the group's conversation synchronisation, thereby, resulting in a high *Conversation Quality*.

3. *Number of Unsuccessful Interruption* - We define an unsuccessful interruption similar to that of Hung et al. (2010) [56], seen previously in the literature review. As an individual-level feature, we simply count the occurrences of such interruptions, with respect to the individual's speaking turn. As a group-level feature, we simply sum the number of unsuccessful interruptions across individuals in the group. With respect to the unsuccessful interruption feature, we hypothesise that smaller the resulting *number of successful interruption*, larger is the group's conversation synchronisation, thereby, resulting in a high *Conversation Quality*.

An overview of the turn-taking features extracted can be seen in Table-5.4.

	Feature Set Type	**Feature Variants**	**Number of Features**
1	Conversation Equality	Degree of Equality	1
2	Conversation Fluency	Percentage of Silence, #back-channels.	2
3	Conversation Synchronisation	Percentage of Overlap, # successful interrupt, # unsuccessful interrupt.	3

Table 5.4: An overview of the 3 sets of turn-taking based features extracted for an individual and their group.

5.3.6. Feature Processing

In this section, we discuss the preprocessing technique used on the group-level features extracted, before feeding them into a data model. Preprocessing is commonly used to *remove noise* from the dataset, get rid of the *curse of dimensionality* from the dataset, prevent *overfitting* or to basically *improve the quality* of the training data. Such techniques help in modeling the data better.

In this research, we use four different stages of feature preprocessing, such as,

1. Feature Scaling/Normalisation
2. Feature Extraction
3. Feature Selection
4. Sampling

The stages listed above are performed in the listed order, as above. The preprocessing stages are further explained in the below section.

Feature Scaling Feature scaling is a commonly used technique which is applied on the extracted features to standardize them. Extracted features, across data samples, tend to range widely across scales. This phenomena can result in the widely ranging feature to dominate other features. For example, a classifier fit on such an unscaled dataset, might converge on biased decision boundaries, biased towards the widely ranging feature.

There are several techniques to scale the feature, e.g. z-score scaling and min-max scaling. For our research, we use the z-score scaling technique, which is computed as follows,

$$z = \frac{x - \mu_x}{\sigma_x} \tag{5.11}$$

where x is the feature vector consisting of feature values for all data samples and $\mu_x and \sigma_x$ are the mean and variance of vector x. Calculating the z-score across features, scales all the feature vectors to *zero mean* and *unit variance*. To perform this, we use the *StandardScale* method available in the scikit-learn package [96].

Feature Extraction Post the feature scaling technique, we reduce the dimensionality of the dataset by performing a feature extraction. By reducing the dimensionality optimally, we eliminate trivial features which are less indicative of conversation quality, thereby, helping the classifier in learning the dataset better. For this process, we apply Principal Component Analysis (PCA) on the scaled dataset and select only the features which preserve the top 90% present in the dataset. To perform this, we use the *PCA* method available in the scikit-learn package [96].

Feature Selection For further dimensionality reduction, we use a feature selection technique to select the statistically optimal features which best indicate the conversation quality. There are several techniques to achieve feature selection. In this research, we used an *Analysis of Variance* (ANOVA) based feature selection. In this technique, ANOVA based p-values are used as the metric for feature selection. Firstly, a linear model is fit on the dataset and ANOVA F-values are computed between each of the features and their ground-truth labels of conversation quality. Subsequently, only the highly significant features from the ANOVA test were selected for further experiments, where features of high significance had a p-value of $p \leq 0.05$. To perform this, we use the *f_classif* method available in the scikit-learn package [96].

Feature Sampling As we discussed in the Chapter-4, one of the problems to tackle in our dataset is the class imbalance problem, as seen in Figure-4.11. If we directly feed this imbalanced dataset into the classifier model, we might run the risk of the classifier model directly learning the class priors instead of the features in itself. This might prevent our model to scale and generalise across test datasets. This problem can be tackled in more than one way, for example, usage of sampling techniques or modifying the loss function to account for class imbalance. In this research, we use the oversampling technique of Synthetic Minority Oversampling (SMOTE) [23].

SMOTE is an oversampling technique which solves the problem of class imbalance by generating new samples x_{new} of the minority class. The new synthetic samples are generated at a line connecting two randomly selected minority class samples. In other words, for each minority sample x_i, SMOTE randomly selects one of the K-Nearest Neighbours $\hat{x_i}$ and then multiplies the difference between these points by a variable δ which satisfies the property $0 \leq \delta \leq 1$, and the resulting value is considered as the new synthetic sample for the minority class. This process is repeated until the class priors one equal to one another. This process can be mathematically formulated as follows,

$$x_{new} = x_i + (\hat{x_i} - x_i) * \delta \tag{5.12}$$

SMOTE is the last block in our feature processing pipeline and the output of the SMOTE is the final dataset for further modeling experiments.

5.4. Modeling

In the previous section, we had presented the features extracted to study the measure of *Conversation Quality*. In this section, we present the techniques used to model the extracted features to predict and study the measure of conversation quality. Specifically, we present the statistical tests and the classifiers used in this research. By the measure of *Conversation Quality*, we refer to both its individual- and group-level measure - Individual's Experience of Conversation Quality (*IndivCQ*) and Group's Conversation Quality (*GroupCQ*) respectively. The modeling techniques are intended to perform the predictive modeling of conversation quality (performed by classifiers) and also study the conversation quality's properties (performed by statistical tests). Firstly, we will present the statistical tests performed to validate some key hypotheses with respect to the features extracted and the *Conversation Quality*. Subsequently, we will present different classifiers experimented with to predict the *Conversation Quality* of a group using the group-level features extracted.

5.4.1. Statistical Tests

In this section, we explain the experimental setup involved in the statistical testing of *Conversation Quality* with respect to the the *Turn-Taking* and *Coordination* features extracted. Before the predictive modeling of *Conversation Quality*, statistical tests with conversation quality as the dependent variable, helps us understand the properties of measure in itself. In this research, we extract two categories of non-verbal cues to study conversation quality, therefore, the statistical tests are done for both these categories of non-verbal cues. Existing research shows us that several social constructs are largely dependent on the cardinality of the group (the number of individuals in the group) [39][100], hence, in addition to the two categories of non-verbal cues, statistical tests are also performed with respect to the cardinality of the group. Hence, we broadly categorise our statistical tests performed into three unique sections,

1. *Group Cardinality and Conversation Quality* - In this test, the group cardinality variable is considered as the independent variable and its respective conversation quality score is considered as the dependent variable.

2. *Turn-Taking features and Conversation Quality* - In this test, the set of turn-taking features based variables are considered as the independent variables and its respective conversation quality score is considered as the dependent variable.

3. *Coordination features and Conversation Quality* - In this test, the set of coordination features (synchrony, convergence and causality) based variables are considered as the independent variables and its respective conversation quality score is considered as the dependent variable.

Statistical Model In this research, the statistical tests and statistical data exploration are performed using *linear least squares* based data-models, which use the least-squares technique to choose optimal parameters of a linear function involving the set of respective explanatory variables. The least-squares method achieves this by minimizing the sum of the squares of differences between the observed dependent variable and those predicted by the linear function. The smaller the differences of the model, the more robust the model is. Such linear statistical models come in different forms depending on the way in which the model makes assumptions and handles the distribution of the given dependent and independent variables. These linear models by handling uniquely the dependent and independent variables, capture different relationships amongst the variables. With that in mind, in this research, we use two different statistical models to study the properties of Conversation Quality. The two statistical models used are as follows,

1. *Quantile Least Squares* model

2. *Joint LASSO* model

Quantile Least Squares One of the most widely used linear statistical model is the *Ordinary Least Squares* (OLS) model, which is robust and optimal when the independent and dependent variables are exogenous [1], and homoscedastic [2]. That is, the OLS model makes several assumption on the dependent and independent variables, such as,

1. Strict Exogeneity - The errors in the regression should have conditional mean zero. i.e. $\mathrm{E}[\varepsilon \mid X] = 0$. [49].
2. No linear dependence - The independent variables (X) are all linearly independent. Mathematically, this means that the matrix X must have full column rank almost surely, i.e. $\Pr\big[\ \mathrm{rank}(X) = p\ \big] = 1$. [49].
3. Homoscedasticity - $E[\epsilon_i^2|X] = \sigma^2$, which means that the error term has the same variance σ^2 in each observation [49].
4. No autocorrelation - the errors are uncorrelated between observations: $E[\epsilon_i \epsilon_j|X] = 0$, for $i \neq j$ [49].
5. Strict Normality - It is sometimes additionally assumed that the errors have normal distribution conditional on the dependent and independent variables, i.e. $\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 I_n)$ [49].

Hence, it is important to validate these assumptions on our dataset before we choose the statistical model.

A variant of the OLS is the *Quantile Least Squares* (QLS). While the OLS estimates the conditional *mean* of the dependent variable with respect to the independent variable, the QLS estimates the conditional *median* (or other quantiles) of the dependent variable with respect to the independent variables. Such estimates do not require the dependent and independent variables to abide by the assumptions of *Exogeneity*, *Homoscedasticity* and *Normality*. The basic intuition behind the model to estimate the conditional *median* is to perform robustly against outliers in the dependent variables, where the conditional variance and mean of the dependent variable with respect to the predictor variables vary largely between each of the quantiles in the dataset.

	W	p-value
Shapiro–Wilk test	0.8393	6.2677e-08

(a) Normality Test on the independent variable of GroupCQ.

	W	p-value
Shapiro–Wilk test	0.7518	5.6015e-20

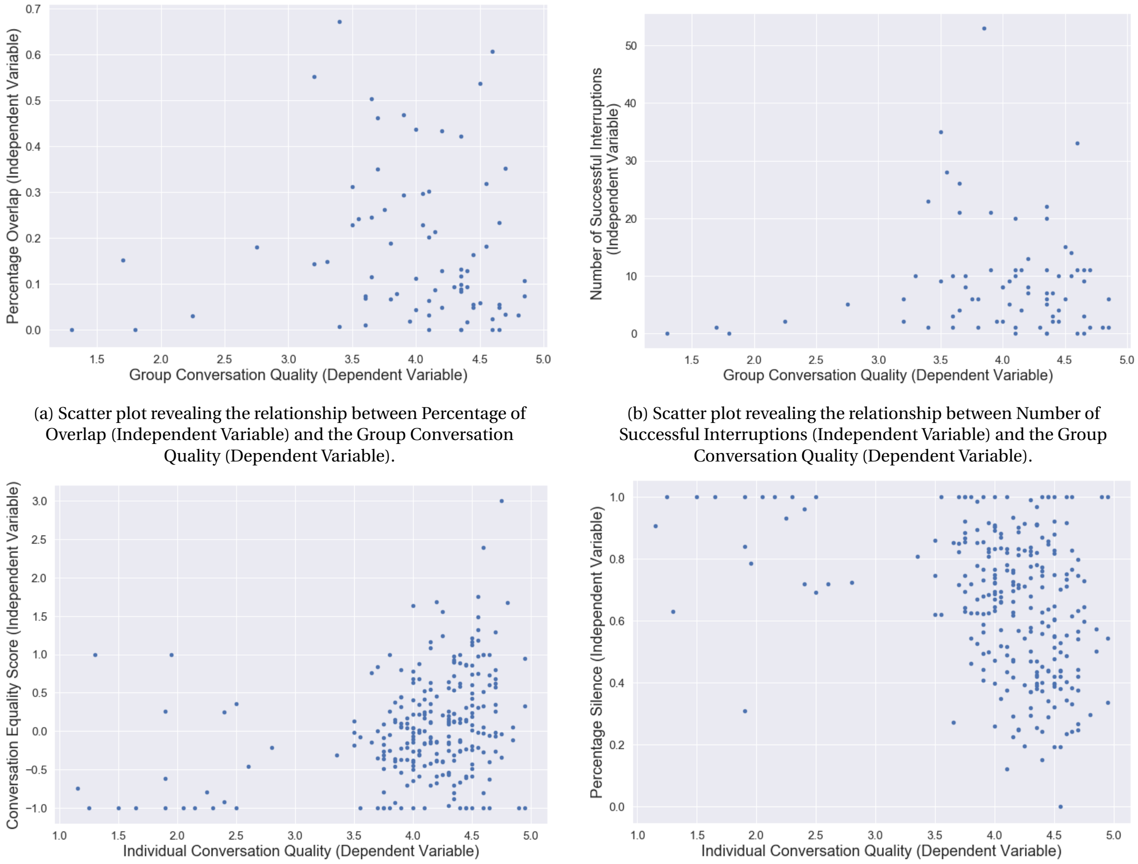
(b) Normality Test on the independent variable of IndivCQ.

Table 5.5: Normality Tests for the independent variable of Conversation Quality.

In our dataset of *Conversation Quality*, the dependent variable (conversation quality) scores for both the individual- and group- level are not normally distributed. Normality was tested using a *Shapiro-Wilk* test, whose results can be seen in Table-5.5. From the results, we can confirm that the distribution of both the dependent variables - *IndivCQ* and *GroupCQ* both do not follow a normal distribution, with a p-value from the *Shapiro-Wilk* test lesser than 0.01 ($p \leq 0.01$). To further explore the distribution of the dependent and the independent variables, we decided to scatter plot the relationship between each of the independent variables and the respective *Conversation Quality* scores. Such a plot helps us to qualitatively evaluate the properties such as *Exogeneity* and *Homoscedasticity*. The explained plots can be seen in Figure-5.7. Here, only few plots of the independent variables are presented, rest of the plots can be found in the Appendix-B. From these plots, we see that the variance (σ^2) varies largely across different quantiles. In other words, the variances of the independent variables along the *lower scales* of IndivCQ and GroupCQ are different from the variances of the independent variables along the *higher scales* of IndivCQ and GroupCQ. Such a behaviour violates the properties of *Exogeneity* and *Homoscedasticity*. With respect to the tests of *Normality*, *Exogeneity* and *Homoscedasticity*, we conclude that our dataset violates the assumptions made by the OLS model and it is not statistically correct to draw inferences on the dataset using a OLS model. To overcome this, we decided to use the *QLS* model which does not make similar assumptions on the *Normality*, *Exogeneity* and *Homoscedasticity* of the dataset. We observe a similar behaviour in all three sections of statistical tests explained above. Hence, the QLS model is used as the statistical model for hypothesis testing in all the three test cases. The QLS model was implemented using the *quantile_regression.QuantReg* method available in the *statsmodel* package [109].

[1] In statistics, an exogenous variable is assumed to be fixed in repeated sampling, which means it is a non-stochastic variable. An implication of this assumption is that the error term in the statistical model is independent of the exogenous variable. [124, p. 49]

[2] In statistics, a sequence of random variables is homoscedastic if all its random variables have the same finite variance. This is also known as homogeneity of variance.



(a) Scatter plot revealing the relationship between Percentage of Overlap (Independent Variable) and the Group Conversation Quality (Dependent Variable).

(b) Scatter plot revealing the relationship between Number of Successful Interruptions (Independent Variable) and the Group Conversation Quality (Dependent Variable).

(c) Scatter plot revealing the relationship between Conversation Equality (Independent Variable) and the Individual Conversation Quality (Dependent Variable).

(d) Scatter plot revealing the relationship between Percentage of Silence (Independent Variable) and the Individual Conversation Quality (Dependent Variable).

Figure 5.7: Scatter plots with respect to few Independent Variables and the Dependent Variables of Conversation Quality. The scatter plots are a qualitative analysis of the independent variable's variance (σ^2) conditioned to the dependent variable of Conversation Quality, and thus examine the Exogeneity and Homoscedasticity of the dataset.

Joint LASSO Though the QLS model does not make assumptions on *Normality*, *Exogeneity* and *Homoscedasticity* of the dependent and independent variables, the QLS model assumes the dependent variables to be uncorrelated with one another (multicollinearity). Because of this, the experiments with the QLS model are performed independently with respect to the different sets of behavioural features. Performing the tests independently might result in the model not accounting for the combined effect of these behavioural features. Hence, to also account for this combined effect, as a second level of complementary analysis, we perform a *joint regression* over all features using a *LASSO model* with L1 prior as the regularizer.

The LASSO (least absolute shrinkage and selection operator) model is a shrinkage and selection method for linear regression. The model is a variant of the basic least square model which uses the L1 based prior regularizer to deal with multicollinearity within the dependent variables. The LASSO model optimises the following loss function (ℓ),

$$\ell(\beta;\lambda) = \sum_{i=1}^{n}(y_i - x_i * \beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \tag{5.13}$$

where x_i and y_i are the column vector and dependent variable of the i^{th} sample, β is the correlation coefficient vector optimised for and the term ($\lambda_1 \sum_{j=1}^{p} |\beta_j|$) is the lasso penalty which handles multicollinearity and induces sparsity for feature selection. The sparsity induced facilitates feature selection and allows us to detect independent variables which are significantly correlated with the dependent variable. In this research, the LASSO model is used to filter out the insignificant features, after performing a joint regression over all features. The LASSO model after inducing sparsity, like any other linear regression model, associates each significant feature with a correlation coefficient (β) which explains the strength and direction of the linear relationship between the feature and the dependent variable. While the β only explains a linear relationship, it might be appropriate to analyse whether the feature actually correlates with the dependent variable in perhaps a non-linear way. To achieve this, a simple Spearman's Rank Correlation is calculated on the LASSO filtered features for drawing further inferences.

The Spearman's Rank correlation measures the strength and direction of monotonic association between two variables, where the monotonic associations include both the linear and non-linear associations between variables. Technically, the Spearman's Rank Correlation is nothing but the pearson correlation between the rank variables. Intuitively, the rank correlation will have a higher correlation coefficient when observations have a *similar rank* across the two variables, and tends to have lower correlation coefficient when observations have a *dissimilar rank* across the two variables. The rank of the variables is simply the relative position label of the observations within the variable: 1^{st}, 2^{nd}, 3^{rd}, etc. The Spearman's Rank Correlation (r_s) for two variables, X_i and Y_i with a sample size of n, can be mathematically framed as follows,

$$r_s = \rho_{\mathrm{rg}_X,\mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X,\mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}}, \tag{5.14}$$

where, $\mathrm{rg}_{X_i},\mathrm{rg}_{Y_i}$ are the relative position label (rank) of the observations X_i and Y_i, ρ denotes the pearson correlation coefficient applied to the rank variables, $\mathrm{cov}(\mathrm{rg}_X,\mathrm{rg}_Y)$ is the covariance of the rank variables, and σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables. The rank correlations have to be performed independently for each independent variables available for the study.

In this research we used the the *LASSO* and the *spearmanr* modules available in the *scikit-learn* package [96], to implement the LASSO and the Spearman Rank correlations respectively.

Evaluation Similar to any linear least squares model, the final parameters of the QLS and LASSO models consists of an intercept and a set of coefficients for each of its independent variables along with their respective standard errors and p-values. That is, in a linear regression model, the dependent variable, y_i, is a linear function of the independent variables,

$$y_i = \beta_1\, x_{i1} + \beta_2\, x_{i2} + \cdots + \beta_p\, x_{ip} + \varepsilon_i \tag{5.15}$$

or in a vector form of,

$$y_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i \tag{5.16}$$

where x_i is a column vector of the i^{th} observations of all the independent variables. (β) is a $p * 1$ vector of unknown parameters and the scalars ϵ_i represents the unobserved random variables (errors), which account for influences upon the responses y_i from sources other than the independent variables x_i.

Additionally, the robustness and the fit of the final QLS model can be analyse using several other metrics such as the r-squared and sparsity. Theses parameters and metrics are vital in inferring the properties of the dependent and independent variables. For the final evaluation and interpretation of the results of the QLS statistical model, we use three different metrics/measures,

1. *R-squared* - or (R^2) is the statistical metric which denotes the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Hence, larger the R^2 more robust is the fit of the model and better the model. We use this metric to explain the robustness of the final statistical model.

2. *p-values* - In statistical hypothesis testing, the p-value is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct. Hence, a lower p-value motivates to rejecting the null hypothesis in itself.

3. *Intercepts and Coefficients* - The intercept term or the constant is the expected mean value of the dependent variable when all the independent variables are zero. This parameter refers the overall average of the dependent value and explains the overall behaviour of the dependent variable. At the same time, the coefficients (β) are the parameters which explain the change in the value of dependent variable with respect to a unit change in the independent variable. Such a parameter explain the correlation between the respective independent variable and its corresponding dependent variable.

5.4.2. Predictive Modeling

In this section, we present the data-driven classification techniques used to predict the measures of *Conversation Quality*, both the *IndivCQ* and *GroupCQ*. By predicting the individual- and group- level measures of *Conversation Quality*, we will be able to contribute towards the development of feedback tools that provide constructive feedbacks on an individual's social interaction. The predictive modeling is performed on the *final dataset* described in the Section-4.3. The final statistics of the final dataset, with respect to the class distribution can be seen in Table-4.5. The first task of the predictive modeling is to predict the *IndivCQ* and also study the predictive capabilities of different feature sets, and, the second task is to perform the same study above with the *GroupCQ*. The rest of this section, presents an overview of the classifier and the evaluation methods used in the process of predictive modeling.

In this research, we performed experiments using a linear white-box model. Specifically, we use a simple *Logistic Regression*.

Logistic Regression with SGD optimizer and Elastic loss The *Logistic Regression* classifier uses a linear logistic function to parameterise and model the relationship between the dependent variable and the set of independent variables. The logistic regression classifier optimises the loss function,

$$min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} log(exp(-y_i(X_i^T w + c)) + 1) + \lambda \parallel w \parallel^2 \tag{5.17}$$

where w are the feature weights, λ is a regularization parameter of the *L2* regularisation term, y_i are the binary ground-truths, X_i is the feature vector and c is a bias term.

In this research, we used a tweaked version of the regular logistic regression. Specifically, we implemented the logistic regression using a *Stochastic Gradient Descent* based optimizer with an *Elastic Loss* based regularisation. The Stochastic gradient descent (SGD) is an iterative method for optimizing an objective function with suitable smoothness properties (e.g. differentiable or subdifferentiable). It can be regarded as a stochastic approximation of gradient descent optimization, since it replaces the actual gradient, calculated from the entire data set, by an estimate thereof calculated from a randomly selected subset of the data [13]. The iterative optimization technique based on the smoothness properties can be mathematically stated as follows,

$$w^{new} = w^{old} - \eta \nabla Q_i(w^{new}) \tag{5.18}$$

where w_{new} and w_{old} are the weights (parameters) for the next iteration and the current iteration respectively, η is the learning rate and $Q_i(w)$ is the value of the loss function at i^{th} sample. The logistic regression model used in this research deploys a loss function with an elastic net based regularisation term. In statistics and, in particular, in the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the *L1* and *L2* penalties of the lasso and ridge methods. The *L1* and *L2* penalties combined regularisation term can be mathematically defined as follow,

$$\hat{R} = \lambda_2 \parallel w \parallel^2 + \lambda_1 \parallel w \parallel_1 \tag{5.19}$$

where λ_1 and λ_2 are the regularisation parameters of the regularisation parameter $\hat{R}$. By using the elastic net regularisation term (5.4.2) in the vanilla logistic regression loss function (5.4.2), the final loss function is obtained as,

$$min_{w \in \mathbb{R}^d} \sum_{i=1}^{N} log(exp(-y_i(X_i^T w + c)) + 1) + \lambda_2 \parallel w \parallel^2 + \lambda_1 \parallel w \parallel_1 \tag{5.20}$$

The above defined logistic regression model is implemented using the *SGDClassifier* method available in the *scikit-learn* package [96].

Evaluation With respect to the class distribution of our dataset (Figure-4.11), we use the *Confusion Matrix*, the *ROC curve* and the *AUC score* as the metrics to evaluate the predictive performance of classifiers, The metrics are calculated as the average across 5-folds in a 5-fold cross-validation process. By using the cross-validation, we also account for the possible overfitting of the classifiers. For each of the fold in the cross-validation process, the optimal hyper-parameters are empirically optimised on the training set before testing on the testing dataset. More details on the hyper-parameters will be presented along with the experiments in the results section.

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The confusion matrix, for a two-class classification problem, is a table with two rows and two columns that reports the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). This allows more detailed class-wise analysis than mere proportion of correct classifications (accuracy), especially important in a class imbalanced dataset like ours.

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters of True Positive Rate and False Positive Rate, calculated from the confusion matrix. The True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows,

$$TPR = \frac{TP}{TP + FN} \tag{5.21}$$

The False Positive Rate (FPR) is defined as follows,

$$FPR = \frac{FP}{FP + TN} \tag{5.22}$$

An ROC curve plots the TPR versus the FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The ROC curve facilitates the qualitative analysis on the classifiers balance between the TPR and the FPR, especially important in a class imbalanced dataset like ours.

The AUC score stands for *Area under the ROC Curve*. In other words, the AUC score measures the entire two-dimensional area underneath the entire ROC curve. When using normalized units, the AUC score is equal to the probability that a classifier will rank a randomly chosen *positive* instance higher than a randomly chosen *negative* one (assuming *positive* ranks higher than *negative*). AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0, one whose predictions are 100% correct has an AUC of 1.0, and the predictions from a random classifier has an AUC estimate of 0.5. The AUC score an evaluation metric is widely used by researchers to evaluate a classifier's performance in a class imbalanced dataset.

6 Experiments & Results

In this chapter, the results of the experiments performed to study the measure of *Conversation Quality* are presented. Specifically, results of the *statistical tests* and the *predictive modeling* of the measure of *Conversation Quality* are presented, for both its manifestations - the *Perceived Group's Conversation Quality* (*GroupCQ*) and the *Perceived Individual's Experience of Conversation Quality* (*IndivCQ*). Firstly, in the Section-6.1, the statistical tests performed with different independent variables and the dependent variable of *Conversation Quality* are presented. Specifically, the three experiments with respect to three set of independent feature variables - *Group Cardinality*, *Turn-Taking* and *Bodily Coordination*, are presented. Secondly, in the Section-6.2, the results of the predictive modeling of *Conversation Quality* are presented. In the same section, we also present the results of several experiments performed to study different sets of features and their predictive capabilities.

6.1. Statistical Analysis

In the Chapter-3, we defined the measure of *Conversation Quality* and its manifestation. When a novel measure is defined and introduced, before attempting to predictively model the social phenomena, it is important to directly study its properties. These set of experiments are intended to answer the second research question of this research (presented in Section-1.3). For this purpose, we perform several statistical test on the dependent variable of *Conversation Quality*, thereby, studying the effects of independent variables such as *Group Cardinality*, *Turn-Taking* and *Bodily Coordination* features.

Firstly, we present the statistical tests performed on the individual-level manifestation of *Conversation Quality* - *Perceived Individual's Experience of Conversation Quality* (*IndivCQ*). Subsequently, we present the statistical tests performed on the group-level manifestation of *Conversation Quality* - *Perceived Group's Conversation Quality* (*GroupCQ*). Under each of these two sections, we present the results of three specific experiments,

1. *Effect of Group Cardinality on Conversation Quality* - This particular experiment statistically tests whether the factor of *Group Cardinality* has a significant effect on the dependent variable of *Conversation Quality*.

2. *Effect of Turn-Taking on Conversation Quality* - This particular experiment statistically tests whether *Turn-Taking* based features have a significant effect on the dependent variable of *Conversation Quality*.

3. *Effect of Bodily Coordination on Conversation Quality* - This particular experiment statistically tests whether *Bodily Coordination* based features have a significant effect on the dependent variable of *Conversation Quality*.

For all the three experiments which are to be performed in this section, we perform two levels of statistical tests, as explained in Section-5.7. Firstly, we perform a statistical test using the Quantile Least Squares (QLS) model, and subsequently, we compare the results of the QLS with that of the Joint LASSO model, as a complementary analysis.

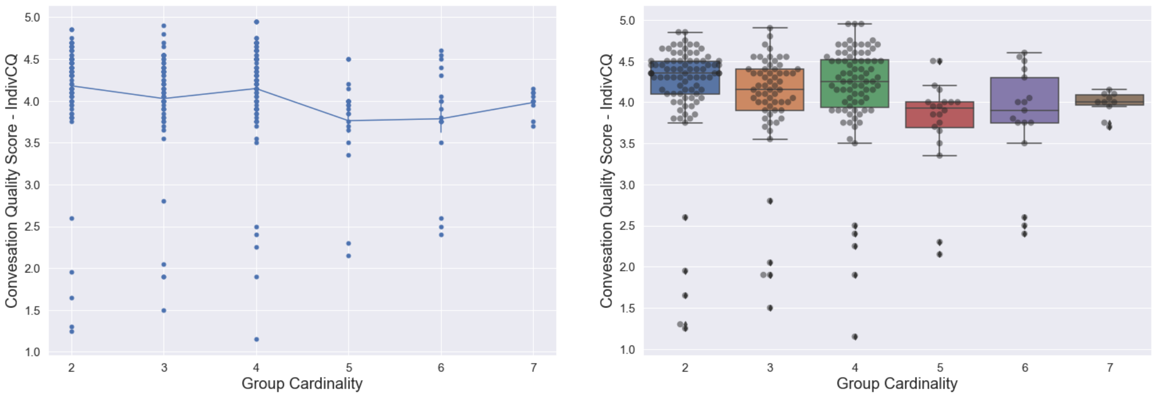
6.1.1. Analysis of Perceived Individual's Experience of Conversation Quality

In this section, we present the results of the statistical tests with respect to the individual-level manifestation of *Conversation Quality* - the Perceived Individual's Experience of Conversation Quality (*IndivCQ*). Additionally, we also draw inferences from theses results.

Experiment - 1: Effect of Group Cardinality on IndivCQ Existing research works such as [39], [105] and [100] shows that social behaviour in group interactions varies with respect to size of the group (the number of individuals in the group interaction, often referred to as the *Group Cardinality*). Hence, it would be an interesting contribution to study the effect of the independent variable - *Group Cardinality* on the dependent variable - *IndivCQ*. In this experiment, we try to study answer the question,

> In an FCG, does the perceived Individual's Experience of Conversation Quality significantly differ with change in the number of individuals in the conversation group (the group cardinality)?

As an initial step, we decided to qualitatively study this effect by plotting the *IndivCQ* scores across different group cardinality, available in our dataset (4.3). This plot can be seen in Figure-6.1.



(a) *IndivCQ* scores across different *Group Cardinality*, in terms of a scatter-plot

(b) *IndivCQ* scores across different *Group Cardinality*, in terms of a swarm-plot

Figure 6.1: Qualitative Analysis of the effect of *Group Cardinality* on the individual-level measure of Conversation Quality - *IndivCQ*. The plots represent the independent variable - *Group Cardinality* along the x-axis and the dependent variable - *IndivCQ* along the y-axis.

From the plots (6.1), we see that the distribution of IndivCQ scores are different across Group Cardinalities, while the sample size also varies across Group Cardinalities. With just a qualitative inference, we can say that the means of IndivCQ are different across Group Cardinalities, especially, the IndivCQ means in Group Cardinalities of 2, 3 and 4 are higher than that of 5, 6, 7. The same can be said with the respect to the variance of IndivCQ across Group Cardinalities, the IndivCQ variance in Group Cardinalities of 2, 3 and 4 are higher than that of 5, 6, 7. With that said, it is also important to quantitatively test this effect as the mean and variances changes across Group Cardinalities are very subtle.

Quantile Least Squares As a first step to quantitatively test this effect, we fit a QLS model on the dataset, with the Group Cardinality as the independent variable and IndivCQ as the dependent variable. The results of the QLS regression test are provided in Table-6.1. The results tests for the null hypothesis which states that the IndivCQ are statistically same across groups of different cardinality.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.5000	0.078	57.795	0.000	*
group_cardinality	**-0.0833**	**0.021**	**-3.978**	**0.000**	*

Table 6.1: Quantile Regression Results - Experiment 1. Test with Cardinality as the independent variable and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.04787, explaining nearly 5% of the total variance in the dataset with 273 observations. * denotes p-value significance at a threshold of 0.05.

From the results, we can conclude that we have statistical evidence that IndivCQ is significantly different across groups of different cardinality, with a p-value ≤ 0.05 which motivates us to reject the null hypothesis. We also see that the β coefficient for the independent variable of Cardinality is negative, suggesting that the IndivCQ is inversely proportional to the Group Cardinality, that is, the perceived Individual's Experience of Conversation Quality decreases with increase in the number of individuals in the group. This is in line with the expectation we had after the qualitative analysis of the plots.

While the previous tests gives us proof to explain the effect of Cardinality on IndivCQ, we still need to perform post-hoc comparisons to ascertain the differences between the cardinalities. To achieve this, we fit multiple QLS models to each possible pair of cardinalities. As we are performing several sub-tests, we are required to correct the significance threshold with respect to the Bonferroni correction for multiple testing. With that said, it is also important to note the imbalance in the number of F-formations of different cardinalities, which is a limitation of this experiment. With sample sizes smaller for cardinalities 5, 6 and 7 when compared to that of cardinalities 2, 3 and 4, we may not be able to draw strong conclusion regarding the cardinalities 5, 6 and 7. Nevertheless, we can draw conclusions strictly with respect to our dataset. The results of the post-hoc comparisons with the Bonferroni corrected significance thresholds can be seen in Table-6.2.

Cardinality Pairs	Intercept	Cardinality (β)	P>|t|	Significance
[2, 3]	4.7500	-0.2000	0.002	*
[2, 4]	4.4500	-0.0500	0.189	
[2, 5]	**4.6497**	**-0.1499**	**0.000**	*
[2, 6]	**4.5750**	**-0.1125**	**0.000**	*
[2, 7]	**4.4900**	**-0.0700**	**0.003**	*
[3, 4]	3.8500	0.1000	0.286	
[3, 5]	4.5244	-0.1248	0.060	
[3, 6]	4.4000	-0.0833	0.097	
[3, 7]	4.2625	-0.0375	0.354	
[4, 5]	4.6484	-0.3496	0.013	
[4, 6]	4.9500	-0.1750	0.033	
[4, 7]	4.5833	-0.0833	0.167	
[5, 6]	3.9024	-0.0004	0.998	
[5, 7]	3.6514	0.0498	0.497	
[6, 7]	3.3000	0.1000	0.569	

Table 6.2: Fifteen Post-Hoc Quantile Regression Comparison results - Experiment 1. Test with Cardinality as the independent variable and IndivCQ as the dependent variable, for each possible pairs of group cardinality. * denotes p-value significance at a threshold of 0.003.

From the last column of the Table-6.2, we see that group cardinality has a significant effect on IndivCQ for cardinality pairs - [2, 3], [2, 5], [2, 6] and [2, 7] at a significance level of 0.003. And, all the significantly different pairs have the β coefficient to be negative. We also see that the significant pairs always has a group cardinality of 2 in it, which suggests that the effect of Group Cardinality on IndivCQ is particularly significant between dyadic and other conversations. Hence, from the post-hoc analysis, we can conclude that, in our dataset [17], the perceived Individual's Experience of Conversation Quality is significantly different (significantly higher) in dyadic group interactions when compared to that of interactions in larger groups (cardinality ≥ 3).

Joint LASSO To further compliment the results of the QLS model, we compare its results with that of the Joint LASSO model. The results of the Joint LASSO model and its respective Rank Correlation, with respect to the group cardinality feature, can be seen in Table-6.3.

Feature Name	LASSO Correlation (β)	Rank Correlation
group_cardinality	-0.0796	-0.24363

Table 6.3: The correlation coefficients of the group cardinality feature with respect to the Joint Correlation and the Rank Correlation models. The complete result of the models can be found in Appendix-D, in Tables-D.2 and D.1 respectively.

The results seen in Table-6.3 supports the results of the QLS model. We see that the Joint LASSO model associates the group cardinality feature with a non-zero correlation coefficient of -0.0796, and at the same time, the rank correlation also gives similar results with a correlation coefficient of -0.24363. This further supports our conclusion that the perceived Individual's Experience of Conversation Quality decreases with increase in the number of individuals in the group.

Experiment - 2: Effect of Turn-Taking on IndivCQ Turn-taking features have been studied extensively by researchers and have shown to be indicative of several subtle social constructs. More recently turn-taking features have been used as the baseline and the state-of-the-art to validate novel methodologies. Hence, it will be an interesting to study the novel measure of IndivCQ with respect to the basic turn-taking features. This particular experiment tests the effect of several aspect of a turn-taking system with respect to the measure of IndivCQ. Specifically, we experiment with three unique aspects of a turn-taking system, as presented by Lindley and Monk [79] - *Conversation Equality*, *Conversation Fluency* and *Conversation Synchronisation* (Explained in Section-5.3.5). In this experiment, we try to study answer the question,

> In an FCG, does *individual-level* turn-taking based features, broadly as Conversation Equality, Conversation Fluency and Conversation Synchronisation, have a significant effect on the perceived Individual's Experience of Conversation Quality?

Quantile Least Squares As a first step to quantitatively test this effect, we fit multiple QLS models with respect to the three aspects of turn-taking features (Section-5.3.5). This helps us study the three unique aspects separately and also account for the assumption of multicollinearity. To account for the multiple tests, the significance threshold is corrected with respect to the Bonferroni correction for multiple testing. From the previous experiment, we saw that the factor of group cardinality has a significant effect on the conversation quality, hence we include the factor of group cardinality as one of the main effects in the following tests. The three separate tests are defined as follows,

1. *Conversation Equality* - In this test, we consider the measure of conversation equality (equation-5.7) as the independent variable and the measure of IndivCQ as the dependent variable. The model is fit on the following relationship - QLS(convq = intercept + group_cardinality + conv_equality)

2. *Conversation Fluency* - In this test, we consider the several *individual-level* measures of conversation fluency (percentage of silence and number of back-channels of an individual (5.3.5)) as the independent variables and the measure of IndivCQ as the dependent variable. The model is fit on the following relationship - QLS(convq intercept + group_cardinality + %silence + #back_channels)

3. *Conversation Synchronisation* - In this test, we consider the several *individual-level* measures of conversation synchronisation (percentage of talk-overlap, number of successful interruptions and number of unsuccessful interruptions of an individual (5.3.5)) as the independent variables and the measure of IndivCQ as the dependent variable. The model is fit on the following relationship - QLS(convq = intercept + group_cardinality + %overlap + #suc_interupt + #unsuc_interupt)

The results with respect to *Conversation Equality* can be seen in Table-6.4, *Conversation Fluency* can be seen in Table-6.5 and *Conversation Synchronisation* can be seen in Table-6.6.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.4465	0.075	59.574	0.000	*
group_cardinality	-0.0723	0.020	-3.555	0.000	*
conv_equality	**0.2136**	**0.040**	**5.302**	**0.000**	*****

Table 6.4: Quantile Regression Results - Experiment 2 with respect to individual-level measures of Conversation Equality. Test with group_cardinality and conv_equality as the independent variables and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.08769, explaining nearly 9% of the total variance in the dataset with 259 observations. * denotes p-value significance at a bonferroni corrected threshold of 0.02. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Conversation Equality in Table-6.4, we see that the measure of conv_equality has significant effect on the IndivCQ. The conv_equality variable also has a positive β coefficient, suggesting that the conv_equality is directly proportional to the perceived IndivCQ. In other words, the larger the equality in talk time between individuals in the interaction, larger the IndivCQ score. The fit QLS model explains nearly 9% of the total variance in the dataset with 259 observations. With the largest r-squared score, amongst

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.6875	0.105	44.686	0.000	*
group_cardinality	-0.0487	0.023	-2.123	0.002	*
%**silence**	**-0.5094**	**0.132**	**-3.860**	**0.000**	*
#back_channels	0.0151	0.013	1.158	0.248	

Table 6.5: Quantile Regression Results - Experiment 2 with respect to individual-level measures of Conversation Fluency. Test with group_cardinality, %silence and #back_channels as the independent variables and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.07452, explaining nearly 7% of the total variance in the dataset with 259 observations. * denotes p-value significance at a bonferroni corrected threshold of 0.02. For the full form of the features listed in the table check Table-D.9.

the other aspects of turn-taking studied, Conversation Equality has the largest statistical significance over the perceived IndivCQ. This also makes sense intuitively as *free-for-all* was one of the aspects measured using the perceived *Conversation Quality* questionnaire 3.4.2.

From the results with respect to Conversation Fluency in Table-6.5, we see that the measure of %silence has a significant effect on the IndivCQ. The %silence variable also has a negative β coefficient, suggesting that the %silence is indirectly proportional to the perceived IndivCQ. In other words, the lesser the percentage of silence in an individual's, larger the IndivCQ score. This is in line with the expectation we had while extracting this particular feature. On the other hand, we also see that the measure of #back_channels does not have a significant effect on the IndivCQ. Nevertheless, the measure of #back_channels has a positive β coefficient, suggesting that the #back_channels is directly proportional to the perceived IndivCQ, though statistically insignificant. The fit QLS model explains nearly 7% of the total variance in the dataset with 259 observations, proving to be one of the largest statistical significance over the perceived IndivCQ.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.5212	0.082	54.903	0.000	*
group_cardinality	-0.0755	0.025	-3.070	0.002	*
%overlap	-0.0820	0.103	-0.794	0.428	
#suc_interupt	-0.0060	0.004	1.489	0.138	
#unsuc_interupt	0.0112	0.010	-1.136	0.257	

Table 6.6: Quantile Regression Results - Experiment 2 with respect to individual-level measures of Conversation Synchronisation. Test with group_cardinality, %overlap, #suc_interupt and #unsuc_interupt as the independent variables and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.04511, explaining nearly 5% of the total variance in the dataset with 259 observations. * denotes p-value significance at a bonferroni corrected threshold of 0.02. For the full form of the features listed in the table check Table-D.9. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Conversation Synchronisation in Table-6.6, we see that the none of the measures of Conversation Synchronisation have a significant effect on the IndivCQ. This is not the same as we expected earlier while extracting the feature. Social constructs such as cohesion and enjoyment has shown be explained well by Turn-Taking Synchronisation features such as overlap and interruptions [56][79].

Nevertheless, at the same time, we see that all the %overlap and #suc_interupt have a negative β coefficient, suggesting that the measures are indirectly proportional to the perceived IndivCQ, though statistically insignificant. Also, we see that #unsuc_interupt is positively correlated with IndivCQ while #suc_interupt has a negative β coefficient. This is an interesting finding, as we see that *unsuccessful* interruptions tend to have a positive effect on IndivCQ while a *successful* interruption tend to have a negative effect on IndivCQ. The fit QLS model only explains nearly 5% of the total variance in the dataset with 259 observations. With the smallest r-squared score, amongst the other aspects of turn-taking studied, Conversation Synchronisation has the least statistical significance over the perceived IndivCQ.

Joint LASSO To further compliment the results of the QLS model, we compare its results with that of the Joint LASSO model. The results of the Joint LASSO model and its respective Rank Correlation, with respect to the turn-taking features, can be seen in Table-6.7.

Feature Name	LASSO Correlation (β)	Rank Correlation
conv_equality	0.17731	0.28856
%silence	-0.02549	-0.22378
%overlap	-0.06046	-0.1141
#suc_interupt	-0.05971	-0.00737
#un_interupt	0.08032	0.14446

Table 6.7: The correlation coefficients of the significant turn-taking features obtained from the Joint Correlation and the Rank Correlation models, with IndivCQ as the dependent variable. Significance of the feature is assumed when the Joint LASSO model associates the feature with a non-zero correlation coefficient (sparsity not induced). The complete result of the models can be found in Appendix-D, in Tables-D.3 and D.1 respectively. For the full form of the features listed in the table check Table-D.9.

The results seen in Table-6.7 further supports our conclusions regarding the conv_equality and %silence features, that the features are positively and negatively correlated respectively, with IndivCQ. At the same time, from the Joint LASSO results, we see that the Conversation Synchronisation features (overlap and interruption features) also have a significant effect on IndivCQ, while the QLS model consider them as insignificant effects. It is also interesting to note that, though the QLS model considered the Conversation Synchronisation to have an insignificant effect, their correlation coefficients are similar to that of the Joint LASSO and Rank Correlation models. For example, the interruption features, successful and unsuccessful interruptions, have a negative and positive correlation coefficients respectively, further supporting our conclusions that *unsuccessful* interruptions tend to have a positive effect on IndivCQ while a *successful* interruption tend to have a negative effect on IndivCQ. An overview on the results of the QLS and Joint LASSO models, with respect to this experiment, suggests that both these models handle the independent variable in a similar fashion, with the QLS model being more stricter on determining variable significance.

Experiment - 3: Effect of Bodily Coordination on IndivCQ Coordination features, across modalities like bodily movements [61] and paralinguistic [88], has been shown to be descriptive of several complex social constructs. In fact, the performance of such features have shown to outperform that of the state-of-the-art turn-taking features. In this research, we consider bodily coordination features measured using tri-axial accelerometers to be the key feature in describing conversation quality. In this experiment, we study the effect of such bodily coordination features on the individual-level conversation quality, IndivCQ. In this experiment, we try to study answer the question,

> In an FCG, does group-level bodily coordination based features, broadly as Synchrony, Mimicry, Convergence and Causality, have a significant effect on the perceived Individual's Experience of Conversation Quality?

Quantile Least Squares As a first step to quantitatively test this effect, using the QLS based regression model, we perform five different hypothesis tests, with only the *magnitude channel* from the accelerometer reading (5.4). Such a setup is used in order to overcome the problem of *multicollinearity* (the linear association between two independent variables) faced by least square regression models. This setup also helps us study the different aspects of coordination separately. From the previous experiments, we saw that the factor of group cardinality has a significant effect on the conversation quality, hence we include the factor of group cardinality as one of the main effects in the following tests. The three separate tests are defined as follows,

1. *Synchrony* - In this test, we consider 22 synchrony based features (e.g. lagged correlation and mutual information) as the independent variable and the measure of IndivCQ as the dependent variable.

2. *Convergence* - In this test, we consider 18 convergence based features (e.g. symmetric, asymmetric and global) as the independent variable and the measure of IndivCQ as the dependent variable.

3. *Lead Mimicry* - In this test, we consider 18 lead mimicry based features as the independent variable and the measure of IndivCQ as the dependent variable. The *Lead Mimicry* features with respect to an individual, captures the event when the individual is leading the conversation and their partner intends to mimic the individual. Technically, the individual's data learnt in the B (Behaviour period 5.3.1) and their partner's data is used as the testing data in period I (Imitation period 5.3.1).

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.2756	0.104	41.261	0.000	*
group_cardinality	-0.0553	0.029	-1.900	0.059	*
mag-min_lagcorr-min	-0.0781	0.194	-0.403	0.688	
mag-max_lagcorr-min	0.0936	0.153	0.611	0.542	
mag-argmin_lagcorr-min	0.0275	0.111	0.248	0.804	
mag-argmax_lagcorr-min	**-5.4044**	**1.265**	**-4.271**	**0.000**	*
mag-mi-min	0.0454	0.092	0.496	0.621	
mag-min_lagcorr-max	-0.5574	0.259	-2.150	0.033	
mag-max_lagcorr-max	0.1744	0.153	1.138	0.257	
mag-argmin_lagcorr-max	**-0.3392**	**0.105**	**-3.226**	**0.002**	*
mag-argmax_lagcorr-max	**-2.8550**	**1.278**	**-2.235**	**0.007**	*
mag-mi-max	-0.0305	0.174	-0.175	0.861	
mag-min_lagcorr-mean	0.6380	0.393	1.623	0.107	
mag-max_lagcorr-mean	-0.2341	0.223	-1.051	0.295	
mag-argmin_lagcorr-mean	**0.2663**	**0.127**	**2.094**	**0.011**	*
mag-argmax_lagcorr-mean	**8.3090**	**2.297**	**3.617**	**0.000**	*
mag-mi-mean	-0.1105	0.188	-0.588	0.557	
mag-min_lagcorr-var	0.0426	0.050	0.845	0.399	
mag-max_lagcorr-var	-0.0010	0.064	-0.016	0.987	
mag-argmin_lagcorr-var	0.1294	0.064	2.025	0.045	
mag-argmax_lagcorr-var	-0.0329	0.043	-0.763	0.447	
mag-mi-var	0.0142	0.064	0.222	0.824	
mag-max_lagcorr-min	0.0936	0.153	0.611	0.542	

Table 6.8: Quantile Regression Results - Experiment 3. Test with **Synchrony based Coordination features as the independent variable** and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.1499, explaining 15% of the total variance in the dataset with 179 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

4. *Lagged Mimicry* - In this test, we consider the 18 lagged mimicry based features (explained in Section-5.3.1) as the independent variable and the measure of IndivCQ as the dependent variable. The *Lagged Mimicry* features with respect to an individual, captures the event when the individual is lagging (or mimicking) and their partner intends to leads the conversation. Technically, the individual's data learnt in the I (Imitating period 5.3.1) and their partner's data is used as the testing data in period B (Behaviour period 5.3.1).

5. *Causality* - In this test, we consider 14 causality based features (e.g. coherence and granger's causality) as the independent variable and the measure of IndivCQ as the dependent variable.

The results with respect to Synchrony can be seen in Table-6.8, Convergence can be seen in Table-6.9, Lead Mimicry can be seen in Table-6.10, Lagged Mimicry can be seen in Table-6.11 and Causality can be seen in Table-6.12.

From the results with respect to Synchrony based Coordination features in Table-6.8, we can infer the following,

- All the statistically significant independent variables are related to the *argmax* and *argmin* variants of lagged correlation features. This suggests us that the time taken to achieve maximum or minimum synchronous coordination has a statistically significant effect on the IndivCQ.

- An interesting insight is with respect to the *min* and *max* based aggregations of these features. The features are negatively correlated with that of the IndivCQ, suggesting that, as hypothesised earlier during the feature extraction, the longer interacting partners take to achieve maximum correlation negatively affects the conversation quality.

- The fit QLS model explains nearly 15% of the total variance in the dataset with 179 observations. With the largest r-squared score, amongst the other aspects of coordination features studied, Synchrony based Coordination features has the largest statistically significant effect over the perceived IndivCQ.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.5811	0.088	52.313	0.000	*
group_cardinality	-0.1253	0.024	-5.122	0.000	*
mag-symconv-min	**0.3303**	**0.122**	**2.703**	**0.008**	*****
mag-lead_asymconv-min	-0.0826	0.084	-0.985	0.326	
mag-lag_asymconv-min	-0.1564	0.120	-1.303	0.194	
mag-globconv-min	**0.7057**	**0.250**	**-2.819**	**0.005**	*****
mag-symconv-max	-0.1615	0.139	-1.165	0.246	
mag-lead_asymconv-max	-0.1885	0.111	-1.706	0.090	
mag-lag_asymconv-max	-0.1932	0.119	-1.618	0.108	
mag-globconv-max	0.2485	0.162	1.531	0.128	
mag-symconv-mean	-0.1910	0.185	-1.032	0.304	
mag-lead_asymconv-mean	0.2354	0.132	1.786	0.076	
mag-lag_asymconv-mean	0.3520	0.197	1.789	0.075	
mag-globconv-mean	0.4277	0.332	1.290	0.199	
mag-symconv-var	**-0.1978**	**0.061**	**3.259**	**0.001**	*****
mag-lead_asymconv-var	0.0779	0.048	1.609	0.110	
mag-lag_asymconv-var	0.0285	0.039	0.732	0.465	
mag-globconv-var	**-0.2380**	**0.068**	**-3.519**	**0.001**	*****

Table 6.9: Quantile Regression Results - Experiment 3. Test with **Convergence based Coordination features as the independent variable** and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.1196, explaining 12% of the total variance in the dataset with 179 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.4635	0.092	48.333	0.000	*
group_cardinality	-0.1011	0.026	-3.923	0.000	*
mag-min_lead_mimicry-min	-0.1514	0.154	-0.981	0.328	
mag-max_lead_mimicry-min	-0.8002	0.528	-1.516	0.132	
mag-mean_lead_mimicry-min	0.0258	0.082	0.316	0.752	
mag-var_lead_mimicry-min	2.5127	1.246	2.017	0.045	
mag-min_lead_mimicry-max	0.2504	0.179	1.399	0.164	
mag-max_lead_mimicry-max	-0.9949	0.559	-1.780	0.077	
mag-mean_lead_mimicry-max	0.1058	0.116	0.908	0.365	
mag-var_lead_mimicry-max	2.1815	1.302	1.676	0.096	
mag-min_lead_mimicry-mean	-0.0045	0.240	-0.019	0.985	
mag-max_lead_mimicry-mean	1.8568	0.957	1.941	0.054	
mag-mean_lead_mimicry-mean	-0.1152	0.152	-0.758	0.449	
mag-var_lead_mimicry-mean	-4.6915	2.371	-1.979	0.050	
mag-min_lead_mimicry-var	**-0.2127**	**0.093**	**-2.289**	**0.010**	*****
mag-max_lead_mimicry-var	1.0717	1.294	0.828	0.409	
mag-mean_lead_mimicry-var	-0.0322	0.050	-0.645	0.520	
mag-var_lead_mimicry-var	-0.6470	1.242	-0.521	0.603	

Table 6.10: Quantile Regression Results - Experiment 3. Test with **Lead Mimicry based Coordination features as the independent variable** and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.1100, explaining 11% of the total variance in the dataset with 179 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Convergence based Coordination features in Table-6.9, we can infer the following,

- All the statistically significant features are related to the *min* and *var* based aggregation features. At the same time, the significant features are related to the *global* and *symmetric* variants of convergence. This suggests that features capturing the least converging interacting pairs (min, variance) in a group has a significant effect on IndivCQ.

- With respect to the correlation coefficients of the *global* and *symmetric* variants of convergence, we see that *min* based aggregation features are positively correlated with IndivCQ, while the *var* based aggregation features are negatively correlated with IndivCQ. This results is inline with the prior expectations made on these group-level aggregation features. The results suggests that the IndivCQ tends to increase when the minimum convergence in a group increases, and also that, the IndivCQ tends to increase when the variance of convergence in a group decreases.

- The fit QLS model explains nearly 12% of the total variance in the dataset with 179 observations. With one of the largest r-squared score, amongst the other aspects of coordination features studied, Convergence based Coordination features has a significant effect over the perceived IndivCQ.

From the results with respect to Lead Mimicry based Coordination features in Table-6.10, we see that only one feature, which captures the variance of the lead mimicry amongst interacting pairs, has a significant effect on IndivCQ. The variance based mimicry feature is negatively correlated with IndivCQ, suggesting that the smaller the variance of mimicry amongst interacting pairs, larger is the conversation quality. Overall, the Lead Mimicry based QLS model is capable of explaining only %11 of the total variance in the dataset and does have a significant effect in IndivCQ.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.4662	0.090	49.641	0.000	*
group_cardinality	-0.1009	0.025	-4.023	0.000	*
mag-min_lag_mimicry-min	-0.1479	0.156	-0.950	0.343	
mag-max_lag_mimicry-min	**-1.2762**	**0.538**	**-2.374**	**0.010**	*
mag-mean_lag_mimicry-min	0.0100	0.068	0.147	0.884	
mag-var_lag_mimicry-min	2.0254	1.132	1.790	0.075	
mag-min_lag_mimicry-max	0.3624	0.180	2.008	0.046	
mag-max_lag_mimicry-max	**1.5949**	**0.551**	**-2.895**	**0.004**	*
mag-mean_lag_mimicry-max	0.0432	0.118	0.367	0.714	
mag-var_lag_mimicry-max	1.2472	1.151	1.084	0.280	
mag-min_lag_mimicry-mean	-0.0866	0.244	-0.355	0.723	
mag-max_lag_mimicry-mean	**2.8599**	**0.974**	**2.936**	**0.004**	*
mag-mean_lag_mimicry-mean	-0.0533	0.141	-0.379	0.706	
mag-var_lag_mimicry-mean	-3.4078	2.114	-1.612	0.109	
mag-min_lag_mimicry-var	**-0.2287**	**0.094**	**-2.439**	**0.011**	*
mag-max_lag_mimicry-var	0.2006	1.200	0.167	0.867	
mag-mean_lag_mimicry-var	-0.0082	0.044	-0.185	0.854	
mag-var_lag_mimicry-var	0.3012	1.183	0.255	0.799	

Table 6.11: Quantile Regression Results - Experiment 3. Test with **Lagged Mimicry based Coordination features as the independent variable** and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.1098, explaining 11% of the total variance in the dataset with 179 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Lagged Mimicry based Coordination features in Table-6.11, we can infer the following,

- We notice that all the four significant features are related to the *max* and *min* variants of the mimicry. This suggests that the features capturing the best and the worst mimicking interacting pairs has a significant effect on the IndivCQ.

- We also see that the *mean* and *max* based aggregation features have a positive correlation with the IndivCQ, while the the *min* and *var* features have a negative correlation with the IndivCQ. The results here are inline with the prior expectation we had over the group-level aggregating features.

- Unlike the Lead Mimicry based model, in the Lagged Mimicry based model, we see more number of statistically significant variables. The fit QLS model explains nearly 12% of the total variance in the dataset with 179 observations. With one of the largest r-squared score, amongst the other aspects of coordination features studied, Lagged Mimicry based coordination features has a significant effect over the perceived IndivCQ, a larger significance that the Lead Mimicry.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.6868	0.099	47.190	0.000	*
group_cardinality	-0.1594	0.027	-5.801	0.000	*
mag-min_coherence-min	-0.1577	0.105	-1.509	0.133	
mag-max_coherence-min	0.2046	0.183	1.119	0.265	
mag-granger-min	-0.0107	0.126	-0.084	0.933	
mag-min_coherence-max	-0.1293	0.199	-0.650	0.517	
mag-max_coherence-max	**0.4708**	**0.177**	**2.655**	**0.009**	*
mag-granger-max	0.1436	0.165	0.870	0.386	
mag-min_coherence-mean	0.2543	0.250	1.017	0.311	
mag-max_coherence-mean	-0.5848	0.268	-2.181	0.031	
mag-granger-mean	-0.0881	0.227	-0.387	0.699	
mag-min_coherence-var	-0.0549	0.053	-1.034	0.302	
mag-granger-var	-0.0861	0.056	-1.550	0.123	

Table 6.12: Quantile Regression Results - Experiment 2. Test with **Causality based Coordination features as the independent variable** and IndivCQ as the dependent variable. The R-squared score of the QLS fit model is 0.0742, explaining 7% of the total variance in the dataset with 179 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Coherence based Coordination features in Table-6.12, we can infer the following,

- Only one feature, which captures the maximum coherence amongst interacting pairs, has a statistically significant effect on the IndivCQ. This suggests that while granger's causality based features do not have any effect on IndivCQ, the *max* based aggregation of coherence feature has a significant effect.

- With respect to correlations, the max_coherence feature is positively correlated with the IndivCQ, suggesting the larger the coherence between the best coordinating interacting pair, larger is the IndivCQ.

- The fit QLS model explains only around 7% of the total variance in the dataset with 179 observations. With the smallest r-squared score, amongst the other aspects of coordination features studied, Causality based coordination features has no significant effect over the perceived IndivCQ.

Joint LASSO To further compliment the results of the QLS model, we compare its results with that of the Joint LASSO model. The results of the Joint LASSO model and its respective Rank Correlation, with respect to the body coordination based features, can be seen in Table-6.13.

From the results of the Joint LASSO model, presented in Table-6.13, we see that the results are very different from that of the QLS model. The QLS results suggested that *argmin* and *argmax* variants of lagged correlation, lagged mimicry and global and symmetric variants of convergence are the only features with the most significant effect on IndivCQ, but, the Joint LASSO results suggest that several other feature sets also have a significant effect on IndivCQ. For example, with respect to the *correlation* based features, the Joint LASSO results reveals that, along with the *argmin* and *argmax* variants of lagged correlation, the basic non-lagged correlation and the *min* and *max* variants of lagged correlation also have a significant effect on IndivCQ. Another difference between the models is that, with respect to the *convergence* based features, while the QLS model revealed that only the *global* and *symmetric* convergence features have a significant effect on IndivCQ, the Joint LASSO results reveals that, along with those features, the *asymmetric* variant of convergence also has a significant effect on IndivCQ. One point of commonality between the two models is that, both the models consider the *lagged* variant of mimicry features to be of more significance that the *lead* variant of mimicry features. Another interesting insight, with respect to the correlation coefficients seen in Table-6.13, is that there exists several significant variables which are treated differently by the Joint LASSO and Rank Correlation models. For example, the lagged mimicry features are given a negative correlation coefficients by the Rank Correlation models (as we had hypothesised/expected during the feature extraction) while they are given positive correlations by the Joint LASSO model. This suggests that there exists a *non-linear* monotonic relationship between the variables and the IndivCQ, rather than a simple linear correlation, hence the LASSO model fails to explain this relationship, giving it a positive correlation value very close to zero. An overview of the results of the models, with respect to the IndivCQ, suggests that the combined effect of body coor-

Feature Name	LASSO Correlation (β)	Rank Correlation
mag-corr–min	0.05409	0.15692
mag-min_lagcorr–min	0.03869	0.02515
mag-corr–max	0.07232	-0.09091
mag-max_lagcorr–mean	-0.05416	-0.11298
mag-argmin_lagcorr–mean	0.00214	-0.02461
mag-argmax_lagcorr–mean	0.01661	0.18065
mag-mi–mean	-0.15075	-0.3407
mag-max_lagcorr–var	0.00078	-0.08692
mag-max_lag_mimicry–min	0.00854	-0.26587
mag-mean_lag_mimicry–max	0.03724	-0.02381
mag-min_lead_mimicry–mean	0.01195	0.10773
mag-mean_lag_mimicry–var	0.01801	-0.09348
mag-globconv–min	-0.05441	0.16752
mag-symconv–max	-0.09201	-0.2088
mag-granger–min	0.07558	0.06138
mag-granger–max	-0.06227	-0.02881
mag-lag_asymconv–max	0.00507	0.03987
mag-symconv–mean	-0.05818	-0.12756
mag-lead_asymconv–mean	0.01143	0.01554
mag-symconv–var	0.02805	-0.18223
mag-lead_asymconv–var	0.04025	-0.15829
mag-lag_asymconv–var	0.02604	-0.09583

Table 6.13: The correlation coefficients of the significant body coordination based features obtained from the Joint Correlation and the Rank Correlation models, with IndivCQ as the dependent variable. Significance of the feature is assumed when the Joint LASSO model associates the feature with a non-zero correlation coefficient (sparsity not induced). The complete result of the models can be found in Appendix-D, in Tables-D.4 and D.1 respectively. For the full form of the features listed in the table check Table-D.9.

dination features (studied using the Joint LASSO) are different from that of the independent effect of body coordination features (studied using the QLS).

6.1.2. Analysis of Perceived Group's Conversation Quality

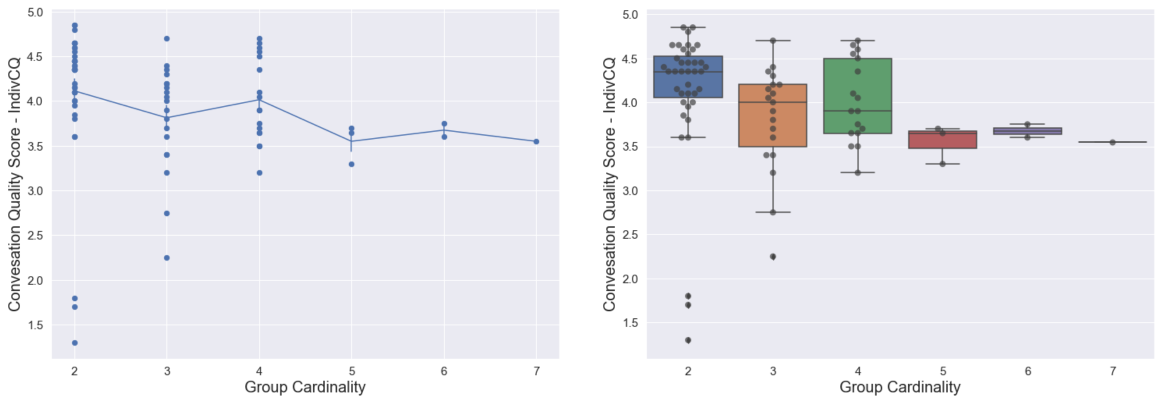
In this section, we present and infer from the results of the statistical tests with respect to the group-level manifestation of *Conversation Quality* - the Perceived Group's Conversation Quality (*GroupCQ*).

Experiment - 1: Effect of Group Cardinality on GroupCQ In the Section-6.1.1, we presented an experiment which studied the effect of Group Cardinality on the IndivCQ. Similarly, here, we present a similar experiment where we intend to study the effect of Group Cardinality on the perceived Group's Conversation Quality (*GroupCQ*). It would be an interesting contribution to check whether similar behaviour is also noticeable in the group-level manifestation of conversation quality.In this experiment, we try to study answer the question,

> In an FCG, does the perceived Group's Conversation Quality significantly differ with change in the number of individuals in the conversation group (the group cardinality)?

As an initial step, we decided to qualitatively study this effect by plotting the *GroupCQ* scores across different group cardinality. This plot can be seen in Figure-6.2.

From the plots (6.2), we see that the distribution of GroupCQ scores are different across Group Cardinalities, while the sample size also varies across Group Cardinalities. With just a qualitative inference, we can say that the means of GroupCQ are different across Group Cardinalities, especially, the GroupCQ means in Group Cardinalities of 2, 3 and 4 are higher than that of 5, 6, 7. The same can be said with the respect to the variance of GroupCQ across Group Cardinalities, the GroupCQ variance in Group Cardinalities of 2, 3 and 4 are higher than that of 5, 6, 7. Such a similar behaviour was also previously seen in the similar experiment with individual-level manifestation (IndivCQ) in Section-6.1.1. With that said, it is also important to quantitatively test this effect.



(a) *GroupCQ* scores across different *Group Cardinality*, in terms of a scatter-plot

(b) *GroupCQ* scores across different *Group Cardinality*, in terms of a swarm-plot

Figure 6.2: Qualitative Analysis of the effect of *Group Cardinality* on the group-level measure of Conversation Quality - *GroupCQ*. The plots represent the independent variable - *Group Cardinality* along the x-axis and the dependent variable - *GroupCQ* along the y-axis.

Quantile Least Squares As a first step to quantitatively test this effect, , we fit a Quantile Least Squares (QLS) model on the dataset, with the Group Cardinality as the independent variable and GroupCQ as the dependent variable. The assumptions made with respect to the variables and the rationale behind choosing QLS as the statistical model for the test was discussed earlier in Section-5.4.1. The results of the QLS regression test are provided in Table-6.14. The results tests for the null hypothesis which states that the GroupCQ are statistically same across groups of different cardinality.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.7833	0.179	26.676	0.000	*
group_cardinality	**-0.2167**	**0.057**	**-3.785**	**0.000**	*

Table 6.14: Quantile Regression Results - Experiment 1. Test with Cardinality as the independent variable and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.1105, explaining 11% of the total variance in the dataset with 81 observations. * denotes p-value significance at a threshold of 0.05.

Cardinality Pairs	Intercept	Cardinality (β)	P>|t|	Significance
[2, 3]	5.0000	-0.4000	0.034	
[2, 4]	4.6500	-0.2250	0.014	
[2, 5]	4.5667	-0.1833	0.105	
[2, 6]	4.5000	-0.1500	0.160	
[2, 7]	4.4600	-0.1300	0.179	
[3, 4]	3.9500	0.0500	0.855	
[3, 5]	4.0250	-0.0750	0.785	
[3, 6]	4.0000	-0.0667	0.759	
[3, 7]	3.9875	-0.0625	0.748	
[4, 5]	4.1500	-0.1000	0.851	
[4, 6]	4.0500	-0.0750	0.826	
[4, 7]	4.0167	-0.0667	0.797	
[5, 6]	3.9000	-0.0500	nan	
[5, 7]	3.9000	-0.0500	nan	
[6, 7]	3.9000	-0.0500	nan	

Table 6.15: Fifteen Post-Hoc Quantile Regression Comparison results - Experiment 1. Test with Cardinality as the independent variable and GroupCQ as the dependent variable, for each possible pairs of group cardinality. * denotes p-value significance at a threshold of 0.003.

From the results, we can conclude that we have statistical evidence that GroupCQ is significantly different across groups of different cardinality, with a p-value ≤ 0.05 which motivates us to reject the null hypothesis.

We also see that the β coefficient for the independent variable of Cardinality is negative, suggesting that the GroupCQ is inversely proportional to the Group Cardinality, that is the perceived Group's Conversation Quality decreases with increase in the number of individuals in the group. This is in line with the expectation we had after the qualitative analysis of the plots.

Similar to the experiment with respect to the IndivCQ and Cardinality, we also perform a pariwise post-hoc test along with the above test to further analyse the effect of cardinality, to ascertain the differences between the cardinalities. To achieve this, we fit multiple QLS models to each possible pair of cardinalities. Similar to IndivCQ, there exists imbalance in the number of F-formations of different cardinalities, which is a limitation of this experiment. With sample sizes smaller for cardinalities 5, 6 and 7 when compared to that of cardinalities 2, 3 and 4, we may not be able to draw strong conclusion regarding the cardinalities 5, 6 and 7. Nevertheless, we can draw conclusions strictly with respect to our dataset. The results of the post-hoc comparisons with the Bonferroni corrected significance thresholds can be seen in Table-6.15.

From the last column of the Table-6.15, we see that none of the group cardinality pairs are significant. Hence, from the above post-hoc analysis on *GroupCQ*, we can conclude that, in our dataset [17], the Group Cardinality does not have a significant effect on the perceived Individual's Experience of Conversation Quality, when the comparison is performed between pairs of Group Cardinality, at a bonferroni corrected significance of 0.003. When drawing inference with both the tests (Table-6.14 and 6.15), we can conclude that while group cardinality has a significant effect on the *GroupCQ* when compared over all samples, but looks to have no significant effect when compared with pairwise of cardinalities.

Joint LASSO To further compliment the results of the QLS model, we compare its results with that of the Joint LASSO model. The results of the Joint LASSO model and its respective Rank Correlation, with respect to the group cardinality feature, can be seen in Table-6.16.

Feature Name	LASSO Correlation (β)	Rank Correlation
group_cardinality	-0.16155	-0.35229

Table 6.16: The correlation coefficients of the group cardinality feature with respect to the Joint Correlation and the Rank Correlation models. The complete result of the models can be found in Appendix-D, in Tables-D.6 and D.5 respectively.

The results seen in Table-6.16 supports the results of the QLS model. We see that the Joint LASSO model associates the group cardinality feature with a non-zero correlation coefficient of -0.16155, and at the same time, the rank correlation also gives similar results with a correlation coefficient of -0.35229. This further supports our conclusion that the perceived Group's Conversation Quality decreases with increase in the number of individuals in the group.

Experiment - 2: Effect of Turn-Taking on GroupCQ In the Section-6.1.1, we presented an experiment which studied the effect of individual-level Turn-Taking features on the IndivCQ. Similarly, here, we present a similar experiment where we intend to study the effect of *group-level* Turn-Taking features on the perceived Group's Conversation Quality (*GroupCQ*). It would be an interesting contribution to check whether similar behaviour is also noticeable in the group-level manifestation of conversation quality. In this experiment, we try to study answer the question,

> In an FCG, does *group-level* turn-taking based features, broadly as Conversation Equality, Conversation Fluency and Conversation Synchronisation, have a significant effect on the perceived Group's Conversation Quality?

Quantile Least Squares As a first step to quantitatively test this effect, we fit multiple QLS models with respect to the three aspects of turn-taking features (Section-5.3.5), similar to Experiment-6.1.1 corresponding to the IndivCQ. The three separate tests are defined as follows,

1. *Conversation Equality* - In this test, we consider the measure of conversation equality (equation-2.1) as the independent variable and the measure of GroupCQ as the dependent variable. The model is fit on the following relationship - QLS(convq = intercept + group_cardinality + conv_equality)

2. *Conversation Fluency* - In this test, we consider the several *group-level* measures of conversation fluency (percentage of silence and number of back-channels, in the whole group (5.3.5)) as the indepen-

dent variables and the measure of GroupCQ as the dependent variable. The model is fit on the following relationship - QLS(convq   intercept + group_cardinality + %silence + #back_channels)

3. *Conversation Synchronisation* - In this test, we consider the several *group-level* measures of conversation synchronisation (percentage of talk-overlap, number of successful interruptions and number of unsuccessful interruptions, in the whole group (5.3.5)) as the independent variables and the measure of GroupCQ as the dependent variable. The model is fit on the following relationship - QLS(convq = intercept + group_cardinality + %overlap + #suc_interupt + #unsuc_interupt)

The results with respect to *Conversation Equality* can be seen in Table-6.17, *Conversation Fluency* can be seen in Table-6.18 and *Conversation Synchronisation* can be seen in Table-6.19.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.4230	0.384	11.521	0.000	*
group_cardinality	-0.1663	0.070	-2.376	0.020	*
conv_equality	0.0631	0.400	0.158	0.875	

Table 6.17: Quantile Regression Results - Experiment 2 with respect to group-level measures of Conversation Equality. Test with group_cardinality and conv_equality as the independent variables and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.08788, explaining nearly 9% of the total variance in the dataset with 74 observations. * denotes p-value significance at a bonferroni corrected threshold of 0.015. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Conversation Equality in Table-6.17, we see that the measure of conv_equality is positively correlated with GroupCQ (positive β coefficients), though statistically *insignificant* with a p-value ≥ 0.02. This result is different from that of similar with respect to IndivCQ, which proves a significant effect between IndivCQ and conv_equality. The fit QLS model explains nearly 9% of the total variance. With the smallest r-squared score, amongst the other aspects of turn-taking studied, Conversation Equality has the least statistical significance over the perceived GroupCQ. This explains that, at a group-level, the Conversation Equality does not have a significant effect on the GroupCQ.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.5068	0.209	21.600	0.000	*
group_cardinality	-0.0487	0.023	-2.123	0.002	*
%silence	-0.0002	0.0002	-0.782	0.437	
#back_channels	0.0267	0.023	1.140	0.258	

Table 6.18: Quantile Regression Results - Experiment 2 with respect to group-level measures of Conversation Fluency. Test with group_cardinality, %silence and #back_channels as the independent variables and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.1060, explaining nearly 10% of the total variance in the dataset with 74 observations. * denotes p-value significance at a bonferroni corrected threshold of 0.02. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Conversation Fluency in Table-6.18, we see that none of the measures of Conversation Fluency has a significant effect on the GroupCQ. That is, both the group-level turn-taking features of %silence and #back_channels does not have a significant effect on the GroupCQ. We also see that none of the β coefficients of the %silence and #back_channels have a large absolute value and looks to have no major effect on the GroupCQ. The fit QLS model explains nearly 10% of the total variance. With one of the smallest r-squared score, amongst the other aspects of turn-taking studied, Conversation Fluency has one of the least statistical significance over the perceived GroupCQ. This explains that, at a group-level, the Conversation Fluency does not have a significant effect on the GroupCQ.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.9031	0.211	23.232	0.000	*
group_cardinality	-0.2403	0.078	-3.071	0.003	*
%overlap	-0.7486	0.584	-1.283	0.204	
#suc_interupt	**-0.0859**	**0.037**	**2.324**	**0.013**	*
#unsuc_interupt	**0.0956**	**0.040**	**-2.385**	**0.020**	*

Table 6.19: Quantile Regression Results - Experiment 2 with respect to individual-level measures of Conversation Synchronisation. Test with group_cardinality, %overlap, #suc_interupt and #unsuc_interupt as the independent variables and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.1326, explaining nearly 13% of the total variance in the dataset with 74 observations. * denotes p-value significance at a bonferroni corrected threshold of 0.02. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Conversation Synchronisation in Table-6.19, we see that the two interruption based measures of Conversation Synchronisation (#suc_interupt and #unsuc_interupt) has a statistically significant effect on the GroupCQ. With respect to the β coefficients of the significant features, we see that, the #suc_interupt variable is negatively correlated with GroupCQ abd the #unsuc_interupt variable is positively correlated with GroupCQ. Such a results was also noted during similar experiments with IndivCQ. At the same time, from the results of the rank correlation (seen in Table-D.7), we see similar results where the #suc_interupt and #un_interupt have a correlation coefficient of -0.0911 and 0.0205 respectively. The fit QLS model only explains nearly 13% of the total variance in the dataset with 74 observations. With the largest r-squared score, amongst the other aspects of turn-taking studied, Conversation Synchronisation has the highest statistical significance over the perceived IndivCQ. This explains that, at a group-level, the Conversation Synchronisation has a significant and a negative effect on the GroupCQ.

Joint LASSO To further compliment the results of the QLS model, we compare its results with that of the Joint LASSO model. The results of the Joint LASSO model and its respective Rank Correlation, with respect to the turn-taking features, can be seen in Table-6.20.

Feature Name	LASSO Correlation (β)	Rank Correlation
conv_equality	0.06478	0.10158
%silence	-0.20162	-0.06321
#back_channels	0.0189	0.08746

Table 6.20: The correlation coefficients of the significant turn-taking features obtained from the Joint Correlation and the Rank Correlation models, with IndivCQ as the dependent variable. Significance of the feature is assumed when the Joint LASSO model associates the feature with a non-zero correlation coefficient (sparsity not induced). The complete result of the models can be found in Appendix-D, in Tables-D.3 and D.1 respectively. For the full form of the features listed in the table check Table-D.9.

From the results seen in Table-6.20, we see that the results of the Joint LASSO model are completely different from that of the QLS model. That is, while the QLS model considered the conversation synchronisation based features (successful and unsuccessful interruptions) to have a significant effect on IndivCQ, the Joint LASSO model considers the conversation equality and fluency (%silence and #back_channels) features to have a significant effect on IndivCQ. At the same time, it is also important to note that, irrespective of the significance of the variables, the independent variables have all been given similar correlation coefficients by both the QLS and Joint LASSO models. This suggests that the QLS and Joint LASSO models, generally, tend to fit the independent variables in a similar fashion, but differ in determining the significance of the variables.

Experiment - 3: Effect of Bodily Coordination on GroupCQ In this research, we consider bodily coordination features measured using tri-axial accelerometers to be the key feature in describing conversation quality. In this experiment, we study the effect of such bodily coordination features on the individual-level conversation quality, GroupCQ. In this experiment, we try to study answer the question,

> In an FCG, does group-level bodily coordination based features, broadly as Synchrony, Mimicry, Convergence and Causality, have a significant effect on the perceived Group's Conversation Quality?

Quantile Least Squares As a first step to quantitatively test this effect, using the QLS based regression model, we perform a similar set of tests as done in Experiment-3 (6.1.1) for IndivCQ. Such a setup is used in order to overcome the problem of *multicollinearity* issues faced by least square regression models. From the previous experiments, we saw that the factor of group cardinality has a significant effect on the conversation quality, hence we include the factor of group cardinality as one of the main effects in the following tests.

The five separate tests are defined as follows,

1. *Synchrony* - In this test, we consider 22 synchrony based features (e.g. lagged correlation and mutual information) as the independent variable and the measure of GroupCQ as the dependent variable.

2. *Convergence* - In this test, we consider 18 convergence based features (e.g. symmetric, asymmetric and global) as the independent variable and the measure of GroupCQ as the dependent variable.

3. *Lead Mimicry* - In this test, we consider 18 lead mimicry based features as the independent variable and the measure of GroupCQ as the dependent variable. The *Lead Mimicry* features with respect to an individual, captures the event when the individual is leading the conversation and their partner intends to mimic the individual. Technically, the individual's data learnt in the B (Behaviour period 5.3.1) and their partner's data is used as the testing data in period I (Imitation period 5.3.1).

4. *Lagged Mimicry* - In this test, we consider 18 lagged mimicry based features as the independent variable and the measure of GroupCQ as the dependent variable. The *Lagged Mimicry* features with respect to an individual, captures the event when the individual is lagging (or mimicking) and their partner intends to leads the conversation. Technically, the individual's data learnt in the I (Imitating period 5.3.1) and their partner's data is used as the testing data in period B (Behaviour period 5.3.1).

5. *Causality* - In this test, we consider 14 causality based features (e.g. coherence and granger's causality) as the independent variable and the measure of GroupCQ as the dependent variable.

The results with respect to Synchrony can be seen in Table-6.21, Convergence can be seen in Table-6.22, Lead Mimicry can be seen in Table-6.23, Lagged Mimicry can be seen in Table-6.24 and Causality can be seen in Table-6.25.

From the results with respect to Synchrony based Coordination features in Table-6.21, we can infer the following,

- Very similar to the same experiment (Experiment-3) with IndivCQ, we see that most of the statistically significant independent variables are related to the *argmax* and *argmin* variants of lagged correlation features. This suggests us that the time taken to achieve maximum or minimum synchronous coordination has a statistically significant effect on the GroupCQ.

- Additionally, we also see that maximum non-lagged correlation and mean non-lagged correlation are also statistically significant. This suggests that the correlations calculated with no time lag also has a statistically significant effect on the GroupCQ.

- With respect to correlation coefficients of the *argmax* and *argmin* variants of lagged correlations, we see that the *max* and *mean* aggregations of the feature are positively correlated, while the *variance* aggregation of the feature are negatively correlated. This suggests that an increase in the max and mean aggregations of lagged correlation result in a higher GroupCQ, and, the decrease in the variance of lagged correlations amongst interacting pairs in group result in a higher GroupCQ.

- With respect to correlation coefficients of non-lagged correlations, we see a similar effect as the previous inference. That is, the *max* and *mean* aggregations of the feature are positively correlated, while the *variance* aggregation of the feature are negatively correlated. That, is the maximum and the mean correlation amongst interacting groups is positively correlated with GroupCQ, while the variance correlation amongst interacting groups is negatively correlated with GroupCQ.

- The fit QLS model explains nearly 50% of the total variance in the dataset with 58 observations. With the largest r-squared score, amongst the other aspects of coordination features studied, Synchrony based Coordination features has the largest statistically significant effect over the perceived GroupCQ.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	5.0339	0.507	9.938	0.000	*
group_cardinality	-0.3666	0.166	-2.215	0.033	
mag-corr-min	-0.1816	0.300	-0.605	0.549	
mag-min_lagcorr-min	2.6954	1.441	1.870	0.070	
mag-max_lagcorr-min	-0.9965	1.189	-0.838	0.407	
mag-argmin_lagcorr-min	7.849e-16	8.91e-16	0.881	0.384	
mag-argmax_lagcorr-min	0.8809	1.068	0.824	0.415	
mag-mi-min	-0.2857	0.707	-0.404	0.689	
mag-corr-max	**2.0609**	**0.708**	**2.909**	**0.006**	*
mag-min_lagcorr-max	0.3161	0.705	0.448	0.657	
mag-max_lagcorr-max	-1.0725	1.119	-0.958	0.344	
mag-argmin_lagcorr-max	**0.2844**	**0.091**	**3.136**	**0.003**	*
mag-argmax_lagcorr-max	-0.3084	1.069	-0.288	0.775	
mag-mi-max	0.1908	0.819	0.233	0.817	
mag-corr-mean	**-1.1928**	**0.498**	**-2.395**	**0.011**	*
mag-min_lagcorr-mean	-3.2111	2.016	-1.593	0.120	
mag-max_lagcorr-mean	1.6796	1.709	0.983	0.332	
mag-argmin_lagcorr-mean	**0.2844**	**0.091**	**3.136**	**0.003**	*
mag-argmax_lagcorr-mean	**0.2844**	**0.091**	**3.136**	**0.003**	*
mag-mi-mean	-0.0820	0.675	-0.121	0.904	
mag-corr-var	**-0.7302**	**0.294**	**-2.486**	**0.011**	*
mag-min_lagcorr-var	1.0006	0.438	2.283	0.028	
mag-max_lagcorr-var	-0.0597	0.325	-0.184	0.855	
mag-argmin_lagcorr-var	**-1.3494**	**0.401**	**-3.366**	**0.002**	*
mag-argmax_lagcorr-var	0.1303	0.226	0.577	0.568	
mag-mi-var	-0.1973	0.421	-0.468	0.643	

Table 6.21: Quantile Regression Results - Experiment 3. Test with **Synchrony based Coordination features as the independent variable** and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.5010, explaining 50% of the total variance in the dataset with 58 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.8853	0.467	10.457	0.000	*
group_sizes	-0.3176	0.152	-2.091	0.042	
mag-symconv-min	-0.1858	0.590	-0.315	0.754	
mag-lead_asymconv-min	-0.0388	0.245	-0.158	0.875	
mag-lag_asymconv-min	-0.0388	0.245	-0.158	0.875	
mag-globconv-min	1.4264	1.466	0.973	0.336	
mag-symconv-max	0.6607	0.579	1.141	0.260	
mag-lead_asymconv-max	-0.0964	0.243	-0.397	0.693	
mag-lag_asymconv-max	-0.0964	0.243	-0.397	0.693	
mag-globconv-max	1.1391	1.251	0.910	0.368	
mag-symconv-mean	-0.4885	0.676	-0.723	0.473	
mag-lead_asymconv-mean	0.0939	0.271	0.347	0.730	
mag-lag_asymconv-mean	0.0939	0.271	0.347	0.730	
mag-globconv-mean	-2.3808	2.072	-1.149	0.257	
mag-symconv-var	-0.3089	0.303	-1.019	0.314	
mag-lead_asymconv-var	0.0617	0.161	0.384	0.120	
mag-lag_asymconv-var	0.0617	0.161	0.384	0.332	
mag-globconv-var	0.1544	0.378	0.408	0.685	

Table 6.22: Quantile Regression Results - Experiment 3. Test with **Convergence based Coordination features as the independent variable** and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.1840, explaining 18% of the total variance in the dataset with 58 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Convergence based Coordination features in Table-6.22, we see that none of the features in the test have a significant effect on GroupCQ. The Convergence based QLS model, fit on the dependent variable of GroupCQ is only capable of explaining %18 of the total variance in the dataset. This is a very less r-squared score when compared to that of the Synchrony based QLS model, and the least amongst the other aspects of coordination features studied. This suggests that Convergence based Coordination features do not have a significant effect on GroupCQ.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.7082	0.451	10.434	0.000	*
group_sizes	-0.2592	0.147	-1.767	0.085	
mag-min_lead_mimicry-min	1.5205	1.232	1.234	0.224	
mag-max_lead_mimicry-min	0.9498	2.012	0.472	0.639	
mag-mean_lead_mimicry-min	-0.3531	0.391	-0.902	0.372	
mag-var_lead_mimicry-min	9.5951	9.905	0.969	0.339	
mag-min_lead_mimicry-max	0.9112	0.949	0.960	0.343	
mag-max_lead_mimicry-max	4.3626	4.948	0.882	0.693	
mag-mean_lead_mimicry-max	0.6338	0.588	1.078	0.288	
mag-var_lead_mimicry-max	17.1296	12.629	1.356	0.183	
mag-min_lead_mimicry-mean	-2.1857	1.876	-1.165	0.251	
mag-max_lead_mimicry-mean	-4.7633	8.025	-0.594	0.556	
mag-mean_lead_mimicry-mean	0.0312	0.558	0.056	0.956	
mag-var_lead_mimicry-mean	-25.9536	21.917	-1.184	0.243	
mag-min_lead_mimicry-var	0.1554	0.660	0.235	0.815	
mag-max_lead_mimicry-var	7.7527	10.863	0.714	0.120	
mag-mean_lead_mimicry-var	-0.4139	0.303	-1.365	0.332	
mag-var_lead_mimicry-var	-3.8942	12.029	-0.324	0.748	

Table 6.23: Quantile Regression Results - Experiment 3. Test with **Lead Mimicry based Coordination features as the independent variable** and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.3240, explaining 32% of the total variance in the dataset with 58 observations. * denotes p-value significance at a threshold of 0.01.

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.7082	0.451	10.434	0.000	*
group_sizes	-0.2592	0.147	-1.767	0.085	
mag-min_lag_mimicry-min	1.5205	1.232	1.234	0.224	
mag-max_lag_mimicry-min	0.9498	2.012	0.472	0.639	
mag-mean_lag_mimicry-min	-0.3531	0.391	-0.902	0.372	
mag-var_lag_mimicry-min	9.5951	9.905	0.969	0.339	
mag-min_lag_mimicry-max	0.9112	0.949	0.960	0.343	
mag-max_lag_mimicry-max	4.3626	4.948	0.882	0.693	
mag-mean_lag_mimicry-max	0.6338	0.588	1.078	0.288	
mag-var_lag_mimicry-max	17.1296	12.629	1.356	0.183	
mag-min_lag_mimicry-mean	-2.1857	1.876	-1.165	0.251	
mag-max_lag_mimicry-mean	-4.7633	8.025	-0.594	0.556	
mag-mean_lag_mimicry-mean	0.0312	0.558	0.056	0.956	
mag-var_lag_mimicry-mean	-25.9536	21.917	-1.184	0.243	
mag-min_lag_mimicry-var	0.1554	0.660	0.235	0.815	
mag-max_lag_mimicry-var	7.7527	10.863	0.714	0.120	
mag-mean_lag_mimicry-var	-0.4139	0.303	-1.365	0.332	
mag-var_lag_mimicry-var	-3.8942	12.029	-0.324	0.748	

Table 6.24: Quantile Regression Results - Experiment 3. Test with **Lagged Mimicry based Coordination features as the independent variable** and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.3240, explaining 32% of the total variance in the dataset with 58 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

From the results with respect to Lead Mimicry based Coordination features in Table-6.23, we see that

	Coef (β)	STD Err	t	P>|t|	Significance
Intercept	4.7045	0.352	13.375	0.000	*
group_sizes	-0.2580	0.113	-2.276	0.028	
mag-min_coherence-min	-0.0431	0.381	-0.113	0.910	
mag-max_coherence-min	-0.1282	0.801	-0.160	0.874	
mag-granger-min	-0.0390	0.131	-0.299	0.766	
mag-min_coherence-max	0.1323	0.512	0.258	0.797	
mag-max_coherence-max	0.1597	0.736	0.217	0.829	
mag-granger-max	0.1163	0.368	0.316	0.754	
mag-min_coherence-mean	-0.0384	0.868	-0.044	0.965	
mag-max_coherence-mean	0.0259	1.250	0.021	0.984	
mag-granger-mean	0.2500	0.228	1.095	0.251	
mag-min_coherence-var	-0.1371	0.423	-0.324	0.556	
mag-max_coherence-var	-0.0441	0.205	-0.216	0.830	
mag-granger-var	-0.3006	0.257	-1.171	0.248	

Table 6.25: Quantile Regression Results - Experiment 3. Test with **Causality based Coordination features as the independent variable** and GroupCQ as the dependent variable. The R-squared score of the QLS fit model is 0.2540, explaining 25% of the total variance in the dataset with 58 observations. * denotes p-value significance at a threshold of 0.01. For the full form of the features listed in the table check Table-D.9.

none of the features in the test have a significant effect on GroupCQ. The Lead Mimicry based QLS model, fit on the dependent variable of GroupCQ is only capable of explaining %32 of the total variance in the dataset. This is a very less r-squared score when compared to that of the Synchrony based QLS model. This suggests that Lead Mimicry based Coordination features do not have a significant effect on GroupCQ.

From the results with respect to Lagged Mimicry based Coordination features in Table-6.24, we see that none of the features in the test have a significant effect on GroupCQ. The Lagged Mimicry based QLS model, fit on the dependent variable of GroupCQ is only capable of explaining %32 of the total variance in the dataset. This is a very less r-squared score when compared to that of the Synchrony based QLS model, and one of the least amongst the other aspects of coordination features studied. This suggests that Lagged Mimicry based Coordination features do not have a significant effect on GroupCQ.

From the results with respect to Causality based Coordination features in Table-6.25, we see that none of the features in the test have a significant effect on GroupCQ. The Causality based QLS model, fit on the dependent variable of GroupCQ is only capable of explaining %25 of the total variance in the dataset. This is a very less r-squared score when compared to that of the Synchrony based QLS model, suggesting that Causality based Coordination features do not have a significant effect on GroupCQ.

Joint LASSO To further compliment the results of the QLS model, we compare its results with that of the Joint LASSO model. The results of the Joint LASSO model and its respective Rank Correlation, with respect to the body coordination based features, can be seen in Table-6.26.

Table 6.26: The correlation coefficients of the significant body coordination based features obtained from the Joint Correlation and the Rank Correlation models, with GroupCQ as the dependent variable. Significance of the feature is assumed when the Joint LASSO model associates the feature with a non-zero correlation coefficient (sparsity not induced). The complete result of the models can be found in Appendix-D, in Tables-D.8 and D.5 respectively. For the full form of the features listed in the table check Table-D.9.

Feature Name	**LASSO Correlation (β)**	**Rank Correlation**
mag-min_lagcorr–min	0.21518	0.2168
mag-corr–mean	0.07573	0.02057
mag-mi–mean	-0.29478	-0.40306
mag-min_lagcorr–var	0.07114	-0.28439
mag-max_lagcorr–var	0.04005	-0.2202
mag-argmin_lagcorr–var	-0.05375	-0.22028
mag-mi–var	0.07855	-0.27232
mag-symconv–min	-0.03703	0.03501
mag-globconv–min	0.00338	0.30425

Continued on next page

Table 6.26 - Continued from previous page

Feature Name	LASSO Correlation (β)	Rank Correlation
mag-lead_asymconv–max	0.16465	-0.04024
mag-lead_asymconv–var	-0.01023	0.02057
mag-granger–min	-0.01411	0.2649
mag-min_coherence–mean	-0.09974	-0.22275
mag-max_coherence–mean	-0.12309	0.32436
mag-granger–mean	0.17962	0.32932
mag-min_coherence–var	-0.06206	0.28057
mag-granger–var	-0.03295	-0.0707
mag-min_lag_mimicry–min	0.02784	0.08746
mag-max_lag_mimicry–min	0.02767	0.10158
mag-var_lag_mimicry–min	-0.03532	0.06321
mag-max_lag_mimicry–mean	0.00229	0.33778
mag-mean_lag_mimicry–mean	0.00012	-0.1712
mag-min_lag_mimicry–var	-0.00019	-0.18453
mag-mean_lag_mimicry–var	-0.00025	-0.13817
mag-min_lead_mimicry–min	0.13314	0.32436
mag-max_lead_mimicry–min	0.05444	0.32932
mag-var_lead_mimicry–min	-0.07414	0.28057
mag-max_lead_mimicry–mean	0.12599	0.33778
mag-mean_lead_mimicry–mean	0.04637	-0.17391
mag-min_lead_mimicry–var	-0.17399	-0.18453
mag-mean_lead_mimicry–var	-0.03462	-0.13994

From the results of the Joint LASSO model, presented in Table-6.26, we see that the results are very different from that of the QLS model. In a comparatively smaller and a class imbalanced dataset for the GroupCQ study, the QLS model had earlier suggested that the *argmin* and *argmax* variants of lagged correlation features are the only features with the most significant effect on GroupCQ, but the results of the Joint LASSO model suggests that there exists more sets of features with such significant effect on the GroupCQ. For example, lead and lagged mimicry features were considered to have a significant effect on GroupCQ by the QLS, but the results of the Joint LASSO model suggest that these features have a significant effect on GroupCQ. Similar inferences can also be drawn with respect to the coherence, granger's causality and convergence features. This suggests that, with respect to the relationship between GroupCQ and the body coordination features, the combined effect of body coordination features (studied using the Joint LASSO) are different from that of the independent effect of body coordination features (studied using the QLS), and the combined effect might reveal the significance of more features, as seen in the results of Joint Lasso.

6.2. Predictive Modeling

In the previous section (6.1), we presented some interesting results with respect to the measure of *Conversation Quality* and different sets of independent variables. In this section, we use the different sets of independent variables, namely, *Turn-Taking* and *Bodily Coordination* features to predicitvely model the two manifestation forms of the social construct of *Conversation Quality*. The predictive modeling is performed as a binary classification task using the final dataset explained in the section-4.3. The class distributions of the dataset can be seen in Table-4.5. These set of experiments are intended to answer the third research question of this research (presented in Section-1.3). As for the modeling technique used, we use a logistic regression classifier regularised with an Elastic Net regularisation term, preceded by a PCA and ANOVA based feature selection (explained in detail in section-5.4.2).

Firstly, we present the results of the predictive modeling of the individual-level manifestation of *Conversation Quality - Perceived Individual's Experience of Conversation Quality* (*IndivCQ*). And, secondly, we present the results of the predictive modeling of the group-level manifestation of *Conversation Quality - Perceived Group's Conversation Quality* (*GroupCQ*). Under each of these two sections, we present the results of three specific experiments and also describe the results of the best performing model,

1. *Influence of Window-Sizes* - This particular experiment is performed to study the influence of the different *window-sizes* (described in section-5.4) on the predictive capability of the model, while modeling

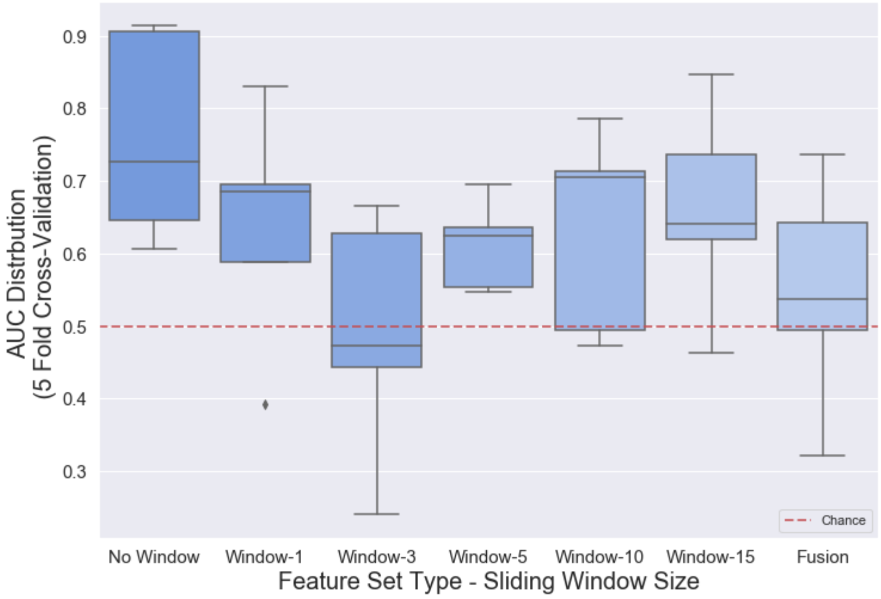


Figure 6.3: Classification performance (Mean AUC) for predicting IndivCQ using different **window-sizes (1, 3, 5, 10, 15)** for feature extraction, using **no sliding-window** for feature extraction and a **feature-level fusion** approach. Different colored lines indicate different approaches with respect to window-sizes.

the measure of *Conversation Quality*.

2. *Influence of Behavioural Feature Sets* - This particular experiment is performed to study the influence of the different *non-verbal feature sets* (described in section-5.3), namely, turn-taking features, synchrony features, convergence features and the fusion of all features, on the predictive capability of the model, while modeling the measure of *Conversation Quality*.

3. *Influence of Group-Level Features* - This particular experiment is performed to study the influence of the different *group-level features* (described in section-5.3.4) on the predictive capability of the model, while modeling the measure of *Conversation Quality*.

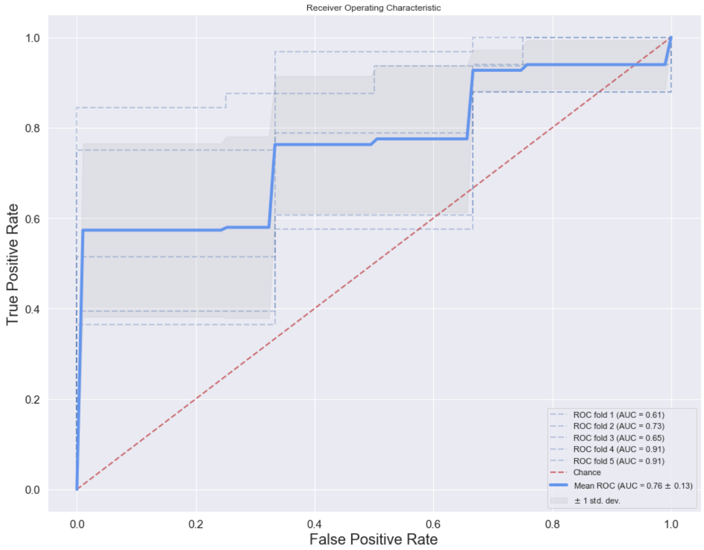
The results of these experiments will be further discussed in the coming sections.

6.2.1. Modeling of Perceived Individual's Experience of Conversation Quality

In this section, we present the experiments and their results with respect to the predictive modeling of Perceived Individual's Experience of Conversation Quality (IndivCQ). Specifically, we perform three experiments related to the prediction of IndivCQ using different feature sets, as explained earlier. This helps us study the influence of different feature extraction and processing techniques on the modeling capabilities of the predictive model. Under each of these experiments, we present the performance of different models under different scenarios and also discuss the best performing model under each of theses experiments.

Experiment - 1: Influence of Window-Sizes In the preprocessing methodology technique (Section-5.4), we had implemented a sliding window approach on the raw accelerometer channels to extract low-level statistical and spectral features, with varying window sizes. Each of the window sizes used captures uniquely the behavioural coordination amongst interacting pairs. This makes an interesting experiment to further study the influence of the window-sizes on the performance capabilities of the perceived Individual's Experience of Conversation Quality, the IndivCQ. Figure-6.3 presents the results of the experiment with respect to the 5 window-sizes used - *1, 3, 5, 10, 15*. In this experiment, we also include the scenario when *no sliding-window* approach was used as a preprocessing step. Additionally, we also include a *fusion of all* above scenarios for further comparison. The results of this experiment can be seen in Figure-6.3.

From the results presented in Figure-6.3, we see that the best performing features are the ones where no sliding-window technique was used. This suggests that the smoothing of accelerometer readings using the sliding-window approach results in loss of information, which might have affected the predictive capabilities of the model. The results might also indicate that bodily coordination between interacting pairs occur in a more subtle level which can be captured directly without the sliding-window approach. The model with no sliding-window based features is capable of predicting the IndivCQ with a mean AUC of 0.76 (±0.13). The ROC curve of this particular model can be seen in Figure-6.4a, along with the respective confusion matrix in Table-6.4b.



(a) Receiver Operating Characteristic curve between the True Positive rate and the False Positive rate, across 5-fold cross validation. The AUC across the folds are also presented with plot.

	Low CQ	High CQ
Low CQ	9	7
High CQ	32	131

(b) Confusion Matrix on the prediction of the model. Low CQ and High CQ represent the samples belonging to classes of Low Conversation Quality (Low IndivCQ) and High Conversation Quality (High IndivCQ) respectively.

Figure 6.4: Predictive performance of the model which uses features extracted **without a sliding-window** based approach. Predictive performance presented in terms of the ROC curve and the Confusion Matrix. The model predicts IndivCQ with a Mean AUC of 0.76 (± 0.13).

From the above results (Figure-6.4), we see that the model has learnt to predict both the classes of Low and High IndivCQ, given the imbalanced dataset. Also, we see that the model's AUC across cross validation folds consistently perform above 0.5 AUC (the AUC of a Random Classifier), denoting the level of sensitivity of the model. This motivates us to use only the features that were extracted without the sliding-window technique for further experiments. This also helps in reducing the dimensionality of the model, thereby might overcome possibilities of overfitting.

Experiment - 2: Influence of Behavioural Feature Sets In the feature extraction module (Section-5.3), we extracted several sets of individual-level and pairwise features like turn-taking, synchrony based body coordination and convergence based body coordination. Each of these features uniquely capture different aspects of group-level behaviour. Hence, it would be interesting to study the predictive capabilities of such features, with respect to the modeling of IndivCQ. In this experiment, we compare performance of several sets of behavioural features, along with feature-level fusion of feature sets. Specifically, we experiment with *eight* different data models trained on different sets of behavioural features,

- *Turn-Taking Features* - In this model, only turn-taking features (Section-5.3.5) are used to train the model. This feature set includes features type like *conversation equality*, *conversation fluency* and *conversation synchronisation*. The significance of such turn-taking features were earlier studied and to study its predictive performance would be interesting.

- *Synchrony Features* - In this model, only synchrony based features (Section-5.3.1) are used to train the model. This feature set includes features type like *correlation*, *lagged correlation*, *mimicry* and *mutual information*. From the statistical tests, the synchrony based features were the found to have the most significant effect on IndivCQ, making it an interesting model to include.

- *Convergence Features* - In this model, only convergence based features (Section-5.3.3) are used to train the model. This feature set includes features type like *symmetric*, *asymmetric* and *global* convergence.

From the statistical tests, the convergence based features were the found to have one of the most significant effect on IndivCQ, making it an interesting model to include.

- *Synchrony + Convergence Features* - In this model, the feature sets of *synchrony* and *convergence* are fused at feature-level and used to train the model. Such a fused feature set has the potential to improve performance, making it an interesting model to include.
- *Causality Features* - In this model, only causality based features (Section-5.3.1) are used to train the model. This feature set includes features type like *coherence* and *granger's causality*. From the statistical tests, the causality based features were the found to have one of the most significant effect on IndivCQ, making it an interesting model to include.
- *Synchrony + Convergence + Causality Features* - In this model, the feature sets of *synchrony*, *convergence* and *causality* are fused at feature-level and used to train the model. This particular feature set basically consists of all features related to bodily coordination based pairwise features. Such a fused feature set has the potential to improve performance, making it an interesting model to include.
- *Turn-Taking + Body Coordination Features* - In this model, the feature sets of *synchrony*, *convergence*, *causality* and *turn-taking* are fused at feature-level and used to train the model. This particular feature set basically consists of all features related to both the pairwise bodily coordination and individual-level turn-taking features. Such a fused feature set has the potential to improve performance, making it an interesting model to include.
- *Synchrony + Convergence + Turn-Taking Features* - In this model, the feature sets of *synchrony*, *convergence* and *turn-taking* are fused at feature-level and used to train the model. Such a fused feature set has the potential to improve performance, making it an interesting model to include.

The list of feature sets framed above are not the complete list with respect to the all possible combinations of the feature sets. But, we believe the above feature sets reveal several interesting properties of IndivCQ. In this experiment, it is also important to note that after the inferences from previous experiment, we decided to perform this experiment without using the features extracted using a sliding-window.

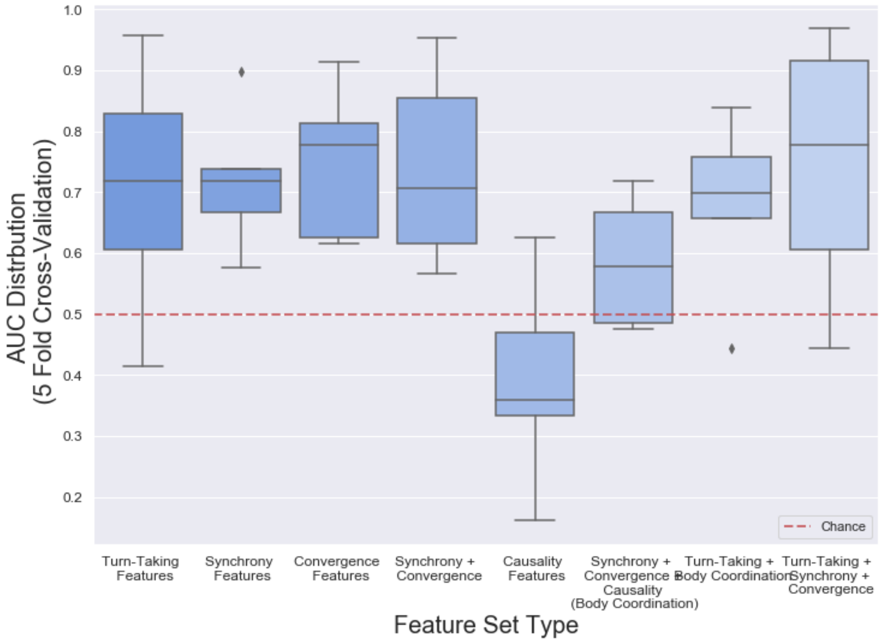


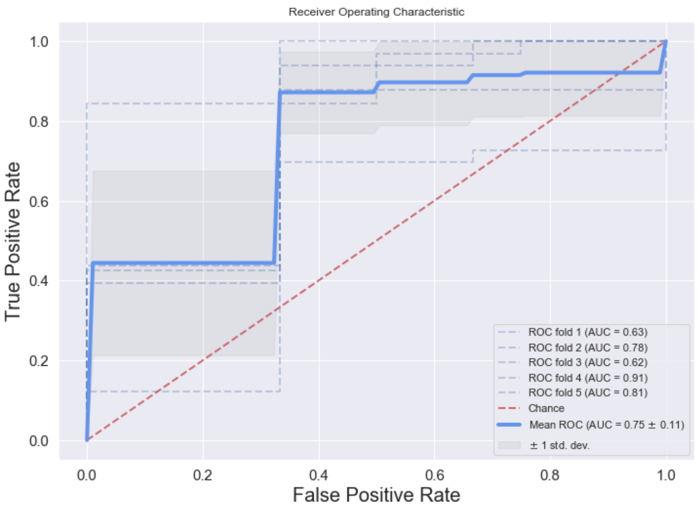
Figure 6.5: Classification performance (Mean AUC) for predicting IndivCQ using different **behavioural feature sets** and also **including fused versions** of feature sets. Different colored lines indicate different approaches with respect to feature sets.

From Figure-6.5, we see that there are several feature sets which closely challenge each other as the best performing features. At the same time, we see that the *Causality* based features perform bad themselves and also decreases the performance of feature sets when fused. Notable are the feature sets seen in Table-6.27, in the order of decreasing performance (with respect to mean and variance of AUC).

	Mean AUC	Variance AUC
Convergence Features	**0.75**	0.11
Synchrony + Convergence Features	**0.75**	0.14
Turn-Taking + Synchrony + Convergence Features	0.74	0.18
Synchrony Features	0.72	**0.10**
Turn-Taking Features	0.71	0.20

Table 6.27: Performance of different **behavioural feature sets** in terms of mean AUC and variance AUC across the 5-folds cross validation.

From Table-6.27, we see that convergence and synchrony based features perform well both by themselves and after feature-level fusion. It is also interesting to note that the two features, by themselves, have a lesser variance AUC, than when they are fused together or with other features, suggesting a high sensitivity of the feature fusion technique. This reveals that the feature fusion tends to increase the complexity of the classification problem, and a more complex classifier model is required to handle these feature, in order to improve performance. At the same time, we also see that though turn-taking features are one of the best performing feature sets by themselves, bodily coordination based features are better predictors of IndivCQ. And also, turn-taking features by themselves have a very high variance AUC of 0.20 with a moderately good mean AUC of 0.71. By fusing turn-taking together with the synchrony and convergence features also increases the variance AUC to 0.18. Hence, considering all the above inferences, the results of this particular experiments suggests that bodily coordination features, especially synchrony and convergence are better predictors of IndivCQ.



(a) Receiver Operating Characteristic curve between the True Positive rate and the False Positive rate, across 5-fold cross validation. The AUC across the folds are also presented with plot.

	Low CQ	High CQ
Low CQ	10	6
High CQ	40	123

(b) Confusion Matrix on the prediction of the model. Low CQ and High CQ represent the samples belonging to classes of Low Conversation Quality (Low IndivCQ) and High Conversation Quality (High IndivCQ) respectively.

Figure 6.6: Predictive performance of the model which uses only **Convergence** based features. Predictive performance presented in terms of the ROC curve and the Confusion Matrix. The model predicts IndivCQ with a Mean AUC of 0.76 (± 0.13).

The Figure-6.6 shows the performance of the best performing feature set of *Convergence*, in terms of ROC curve and Confusion Matrix. From the above results (Figure-6.6), we see that the model has learnt to predict both the classes of Low and High IndivCQ, given the imbalanced dataset. Also, we see that the model's AUC

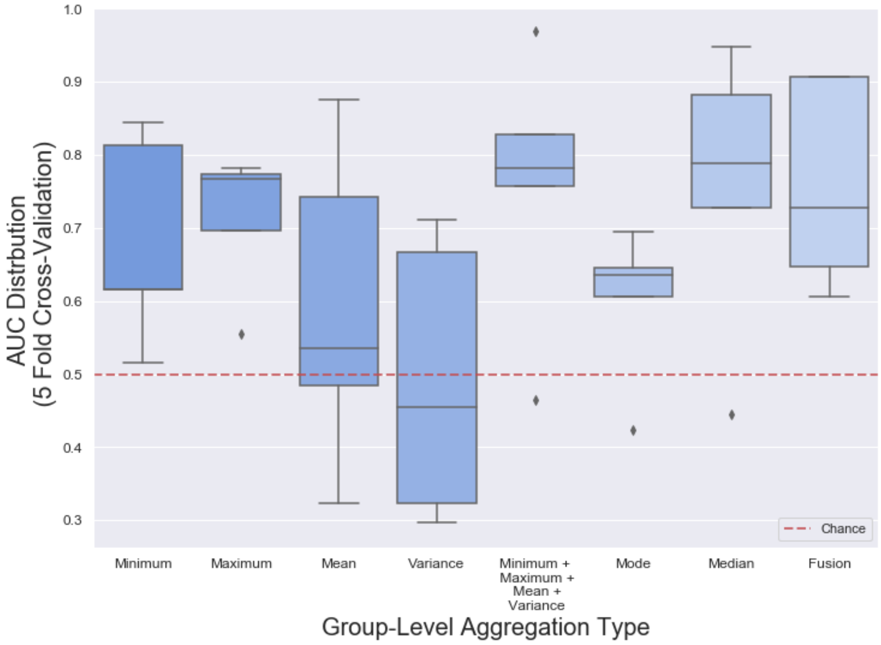


Figure 6.7: Classification performance (Mean AUC) for predicting IndivCQ using different feature sets, with respect to the different **group-level aggregates**, and **including fused versions** of feature sets. Different colored lines indicate different approaches with respect to feature sets.

across cross validation folds consistently perform above 0.5 AUC (the AUC of a Random Classifier), denoting the level of sensitivity of the model.

Experiment - 3: Influence of Group-Level Features The last step of the feature extraction step was to aggregate the individual-level and pairwise features into group-level features, using several aggregators namely *min, max, mean, variance, mode* and *median*, explained in detail in Section-5.3.4. The influence of such aggregators in the prediction capabilities of a model is an interesting topic of discussion. A similar experiment was performed by Nanninga et al. [88] while studying cohesion in meetings. The results presented by the authors show that the *median* based aggregator is the best performing features.

In our experiment, we consider *eight* sets of feature sets with respect to the aggregators, they are,

- *Minimum* - In this model, body coordination features extracted using a *min* aggregator is used to train it. Basically, this feature set captures the least coordinating interacting pair in the group.
- *Maximum* - In this model, body coordination features extracted using a *max* aggregator is used to train it. Basically, this feature set captures the best coordinating interacting pair in the group.
- *Mean* - In this model, body coordination features extracted using a *mean* aggregator is used to train it. Basically, this feature set captures the overall average coordination in the group.
- *Variance* - In this model, body coordination features extracted using a *variance* aggregator is used to train it. Basically, this feature set captures the deviation of coordinations in interacting pairs from the group mean.
- *Minimum + Maximum + Mean + Variance* - In this model, the four feature sets of *Minimum*, *Maximum*, *Mean* and *Variance* are fused at a feature-level and used to train it.
- *Mode* - In this model, body coordination features extracted using a *mode* aggregator is used to train it. Basically, this feature set captures the local maxima in coordination amongst interacting pairs in the group.
- *Median* - In this model, body coordination features extracted using a *median* aggregator is used to train it. Basically, this feature set captures the exact middle point in a vector of interacting pairs and their

	Mean AUC	Variance AUC
Fusion	**0.76**	**0.13**
Median	**0.76**	0.17
Minimum + Maximum + Mean + Variance	**0.76**	0.16
Maximum	0.72	**0.09**
Minimum	0.68	0.13

Table 6.28: Performance of different **group-level aggregators** in terms of mean AUC and variance AUC across the 5-folds cross validation.

coordination. The basic advantage of the median in describing data compared to the mean is that it is not skewed, and so it may give a better idea of a "typical" value.

- *Fusion* - In this model, the features extracted using all the above mentioned unique group-level aggregators are used to train it.

From the results presented in Figure-6.5, we see that five feature sets namely, *Minimum, Maximum, Minimum + Maximum + Mean + Variance, Median* and *Fusion* are the best performing features with a negligible margin in performance between each other. At the same time, we see that the performances of feature sets such as *Mean, Variance* and *Mode* have a poor predictive capability when compared to the rest of the feature sets. It is also interesting to note that both the fusion based approaches used in the experiment result in good predictive performance. The top five performing feature sets and their predictive capabilities in terms of mean AUC and variance can be seen in Table-6.28.

From the results in Table-6.28, we see that the *Fusion* based model is the best performing model with the maximum mean AUC and one of the lowest variance AUC. Similarly, the *Minimum + Maximum + Mean + Variance* based model is one of the best performing models. This suggests that group-level aggregators are optimal when all such aggregators are used and features are fused. This reveals that the combination of such aggreagtors is more indicative of IndivCQ rather than by themselves. As non-fused individual feature sets, *Median* based aggregated feature sets are the best performing with the maximum mean AUC and but with a slightly larger variance AUC. This suggests that feature-level fusion, in terms of group-level aggregators, results in a stable and generalised data model.

6.2.2. Modeling of Perceived Group's Conversation Quality

In this section, we present the experiments and their results with respect to the predictive modeling of Perceived Group's Conversation Quality (GroupCQ). Similar to Section-6.2.1 where we had presented three experiments related to the predictive modeling of IndivCQ using different feature sets, in this section, we present the same three experiments and their results, but with respect to the GroupCQ. This helps us in studying the influence of different feature extraction and preprocessing techniques on the predictive modeling of GroupCQ, and also compare the the same with that of the results from IndivCQ. Under each of these experiments, we present the performance of different models under different scenarios and also discuss the best performing model under each of theses experiments.

Experiment - 1: Influence of Window-Sizes In the preprocessing methodology technique (Section-5.4), we had implemented a sliding window approach on the raw accelerometer channels to extract low-level statistical and spectral features, with varying window sizes. This experiment, presented below, is intended to study the predictive capabilities of different window sizes. This experiment is similar to the Experiment-6.2.1 but uses the GroupCQ labels as the ground-truth. Figure-6.8 presents the results of the experiment with respect to the 5 window-sizes used - *1, 3, 5, 10, 15.* In this experiment, we also include the scenario when *no sliding-window* approach was used as a preprocessing step. Additionally, we also include a *fusion of all* above scenarios for further comparison. The results of this experiment can be seen in Figure-6.8.

From the results presented in Figure-6.8, we see that the best performing features are the ones where no sliding-window technique was used. This is a similar results as seen in the sliding-window based experiment with IndivCQ (6.2.1). This result here adds evidence to the findings that, in the modeling of *Conversation Quality,*sliding-window approach results in loss of information, which might have affected the predictive capabilities of the model. The model with no sliding-window based features is capable of predicting the

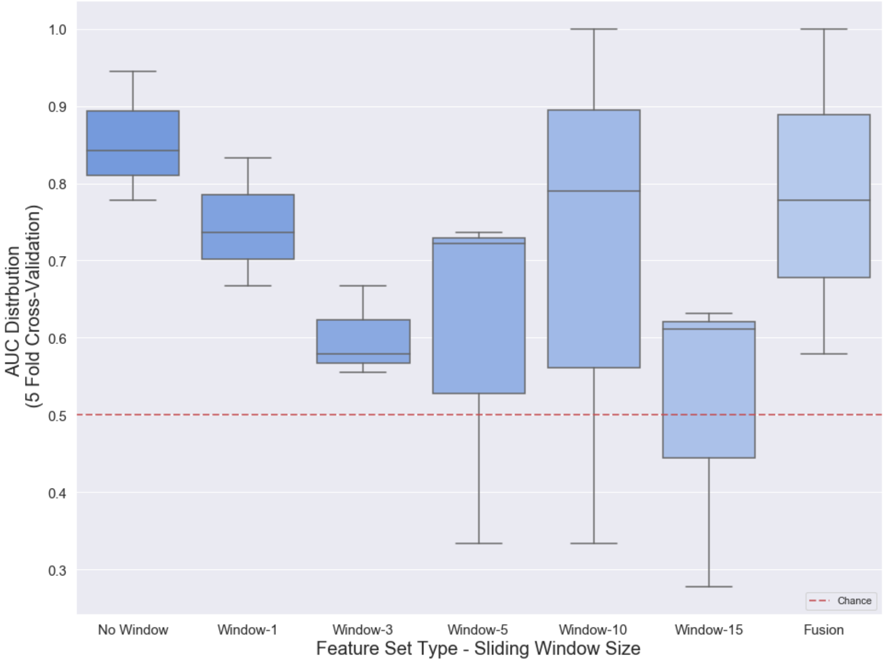
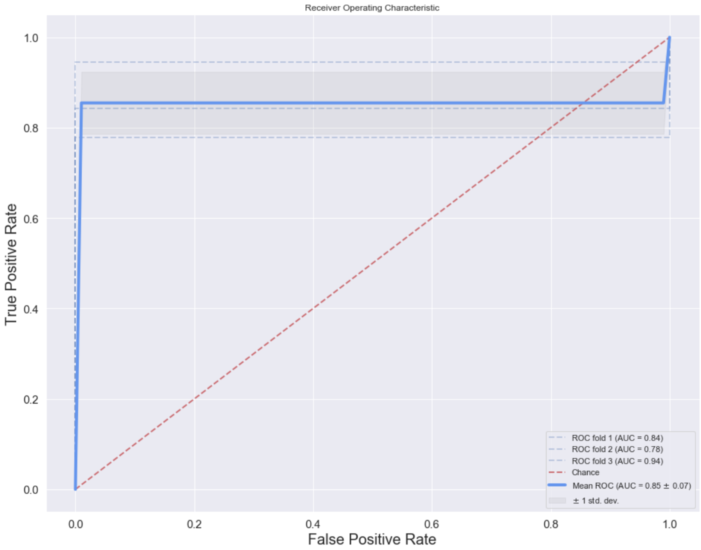


Figure 6.8: Classification performance (Mean AUC) for predicting GroupCQ using different **window-sizes (1, 3, 5, 10, 15)** for feature extraction, using **no sliding-window** for feature extraction and a **feature-level fusion** approach. Different colored lines indicate different approaches with respect to window-sizes.



(a) Receiver Operating Characteristic curve between the True Positive rate and the False Positive rate, across 3-fold cross validation. The AUC across the folds are also presented with plot.

	Low CQ	High CQ
Low CQ	2	1
High CQ	7	48

(b) Confusion Matrix on the prediction of the model. Low CQ and High CQ represent the samples belonging to classes of Low Conversation Quality (Low GroupCQ) and High Conversation Quality (High GroupCQ) respectively.

Figure 6.9: Predictive performance of the model which uses features extracted **without a sliding-window** based approach. Predictive performance presented in terms of the ROC curve and the Confusion Matrix. The model predicts GroupCQ with a Mean AUC of 0.85 (± 0.07).

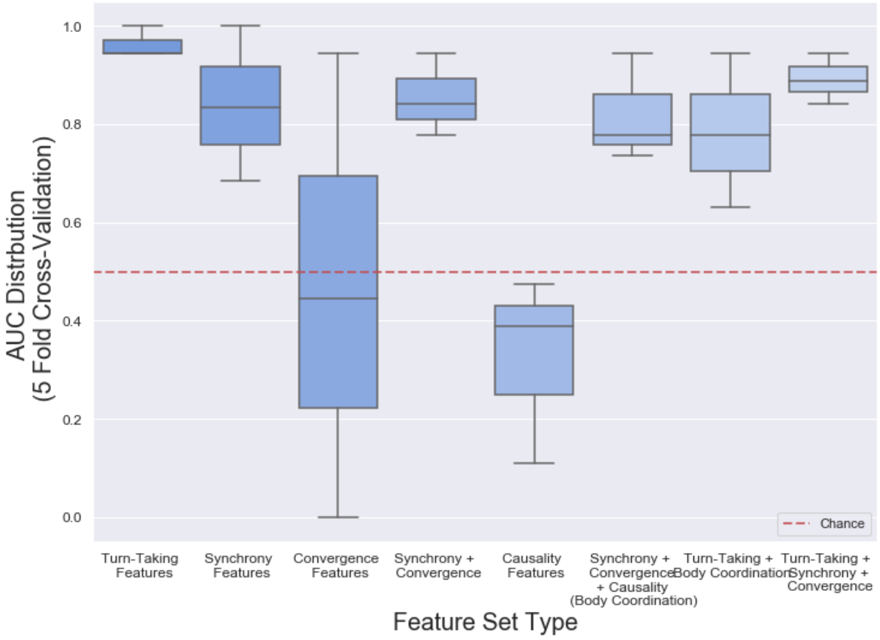


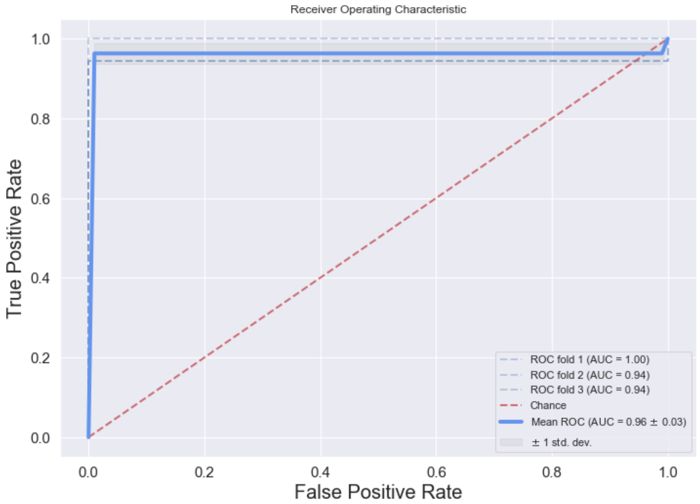
Figure 6.10: Classification performance (Mean AUC) for predicting GroupCQ using different **behavioural feature sets** and also **including fused versions** of feature sets. Different colored lines indicate different approaches with respect to feature sets.

GroupCQ with a mean AUC of 0.85 (±0.07). The ROC curve of this particular model can be seen in Figure-6.9a, along with the respective confusion matrix in Table-6.9b.

From the ROC curve and Confusion Matrix in Figure-6.4, we see that the model (no sliding window based model) has learnt to predict both the classes of Low and High GroupCQ, given the imbalanced dataset. Also, we see that the model's AUC across cross validation folds consistently perform above 0.75 AUC, revealing the stability of the model. From the confusion matrix in Figure-6.9b, we see that the model has misclassified *one* Low GroupCQ sample (False Positive) and *seven* High GroupCQ sample (False Negative). Being a small dataset with an imbalanced class distribution, we cannot draw strong conclusion on the predictive modeling of group-level *Conversation Quality*, nevertheless, the no-sliding window based bodily coordination features have preliminary evidences to be a good predictor of group-level *Conversation Quality* with a mean AUC of 0.85 (±0.07).

Experiment - 2: Influence of Feature Sets In the feature extraction module (Section-5.3), we extracted several sets of behavioural features like turn-taking, synchrony based body coordination and convergence based body coordination. Using these features, in Experiment-6.2.1 we studied the predictive modeling capabilities of these feature sets with respect to the ground-truth labels of IndivCQ. In this section, we perform a similar experiment but with respect to the ground-truth labels of GroupCQ. This experiment will be an interesting study to compare the performance of behavioural features across the two forms of *Conversation Quality* (the individual-level IndivCQ and group-level GroupCQ). For this experiment, we use the same set of eight feature sets as used in Experiment-6.2.1. In this experiment, it is also important to note that after the inferences from previous experiment, we decided to perform this experiment without using the features extracted using a sliding-window.

From Figure-6.10, we see that there are several feature sets which closely challenge each other as the best performing features. Particularly notable are the two feature sets of *Turn-Taking* and *Turn-Taking + Synchrony + Convergence* (fused), with a mean AUC of 0.96 (±0.03) and 0.89 (±0.04) respectively. From the results, we see that Turn-Taking based features are more informative in predicting GroupCQ than the bodily coordination based features. This result is in contract with the results obtained from similar experiments with IndivCQ, where bodily coordination based features were more informative of IndivCQ than Turn-Taking based features. In Table-6.29, we present the top four feature sets in the order of decreasing performance (with respect to mean and variance of AUC).



(a) Receiver Operating Characteristic curve between the True Positive rate and the False Positive rate, across 3-fold cross validation. The AUC across the folds are also presented with plot.

	Low CQ	High CQ
Low CQ	2	1
High CQ	8	47

(b) Confusion Matrix on the prediction of the model. Low CQ and High CQ represent the samples belonging to classes of Low Conversation Quality (Low GroupCQ) and High Conversation Quality (High GroupCQ) respectively.

Figure 6.11: Predictive performance of the model which uses only **Turn-Taking** based features. Predictive performance presented in terms of the ROC curve and the Confusion Matrix. The model predicts GroupCQ with a Mean AUC of 0.96 (± 0.03).

	Mean AUC	Variance AUC
Turn-Taking Features	**0.96**	**0.03**
Turn-Taking + Synchrony + Convergence Features	0.89	0.04
Synchrony + Convergence Features	0.85	0.07
Synchrony Features	0.84	0.13

Table 6.29: Performance of different **behavioural feature sets** in terms of mean AUC and variance AUC across the 3-folds cross validation.

From Table-6.29, we see that Turn-Taking features perform the best in terms of both mean AUC and variance AUC. This reveals that group-level *Conversation Quality*, in our dataset, can best predicted by Turn-Taking features. We also see that Synchrony based feature sets are one of the top four best performing features set and they perform better when feature-level fused with Turn-Taking and/or Convergence features. Another interesting inference is that, to predict GroupCQ, Convergence based feature sets are one of the worst performing feature sets (mean AUC of 0.46 ±0.04), but, when fused with Synchrony and/or Turn-Taking features tend to improve the respective feature's performances. This suggests that the combination in information from convergence in body coordination and synchrony in body coordination or turn-taking features are discriminative of levels of *Conversation Quality*. Another inference from both Table-6.29 and Figure-6.10 is that, in bodily coordination based features, fusion based approaches tend to reduce the variance AUC (≤ 0.08) thereby increasing the generalisability and stability of the predictive model.

The Figure-6.11 shows the performance of the best performing feature set of Turn-Taking, in terms of ROC curve and Confusion Matrix. From the above results, we see that the model has learnt to predict both the classes of Low and High GroupCQ, given the imbalanced dataset. Also, we see that the model's AUC across cross validation folds consistently perform above 0.5 AUC (the AUC of a Random Classifier), denoting the level of sensitivity of the model. In-fact, the best performing Turn-Taking features consistently perform above mean AUC 0.90 across the cross validation folds which reveals the efficiency of the Turn-Taking feature set in predicting GroupCQ.

Experiment - 3: Influence of Group-Level Features The last step of the feature extraction step was to aggregate the individual-level and pairwise features into group-level features, using several aggregators namely min, max, mean, variance, mode and median, explained in detail in Section-5.3.4. In Experiment-6.2.1 we studied the predictive modeling capabilities of group-level aggregation features with respect to the ground-truth labels of IndivCQ. In this section, we perform a similar experiment but with respect to the ground-truth labels of GroupCQ. We use the same eight group-level aggregation feature sets as used in Experiment-6.2.1.

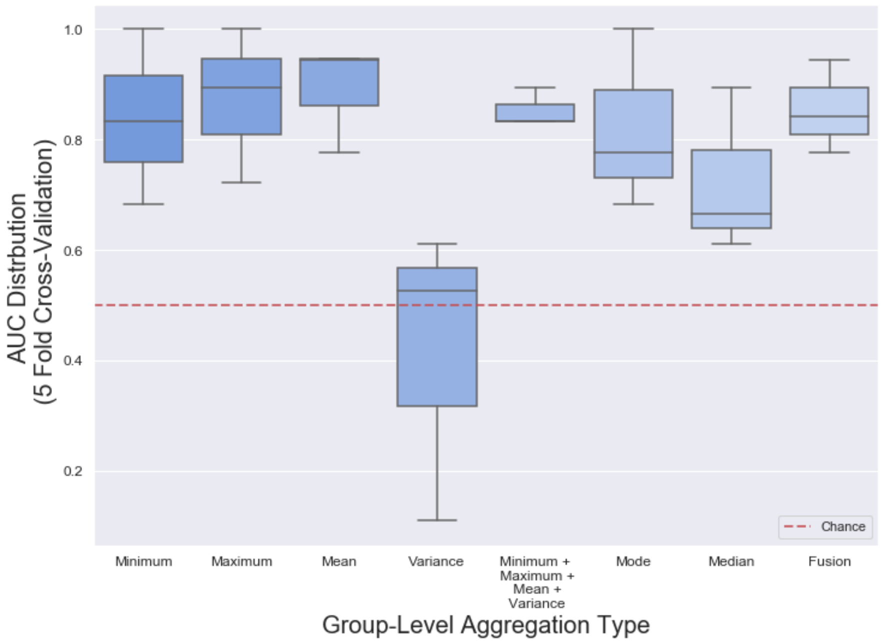


Figure 6.12: Classification performance (Mean AUC) for predicting GroupCQ using different feature sets, with respect to the different **group-level aggregates**, and **including fused versions** of feature sets. Different colored lines indicate different approaches with respect to feature sets.

From the Figure-6.12, we see that *Mean* based group-level aggregation features are the best performing model with a mean AUC of 0.89 (± 0.08), and Variance based features perform the worst with a mean AUC of 0.42 (± 0.22). This is different from what we saw with the similar experiment with respect to IndivCQ, as while predicting IndivCQ, *Median* based group-level features were the best performing feature set. Here, it is more interesting to note that while predicting IndivCQ the non-skewed version of the group-level aggregators, the Median, is the most informative, at the same time, while predicting GroupCQ the skewed version of the group-level aggregators, the Mean, is the most informative. This suggests that the individual-level IndivCQ and the group-level GroupCQ are different social constructs in themselves. At the same time, we also see that there exists more set of features which challenge the *Mean* based aggregation features to be the best features. The top four performing feature sets and their predictive capabilities in terms of mean AUC and variance can be seen in Table-6.30.

	Mean AUC	Variance AUC
Mean	**0.89**	0.08
Maximum	0.87	0.11
Minimum + Maximum + Mean + Variance	0.85	**0.03**
Fusion	0.85	0.07

Table 6.30: Performance of different **group-level aggregators** in terms of mean AUC and variance AUC across the 5-folds cross validation.

From Table-6.30, we see that *Mean* and *Maximum* based group-level aggregators are the best performing features with 0.89 (± 0.08) and 0.87 (± 0.11) respectively. At the same time, the feature-level fusion based approaches *Minimum* + *Maximum* + *Mean* + *Variance* and *Fusion* are one of the best performing features with

0.85 (± 0.03) and 0.85 (± 0.07) respectively. Similar to the previous experiment, we see here that though *Mean* and *Maximum* based features are the best performing, fusion based approaches tend to reduce the variance AUC. This suggests that feature-level fusion based features sets, in modeling Conversation Quality, tend to decrease the variance AUC, thereby, increasing the generalisability and stability of the predictive models.

7 Discussions

In chapter, we discuss, with a retrospective view on this research work, the main contributions of this thesis and the results of the experiments performed, with an intention to answer the thesis' research questions (Section-1.3). While discussing the results, we also highlight some key findings during the research. Rest of this chapter will be categorised with respect to the three research questions of this thesis.

Research Question 1: Quantifying spontaneous conversations Building on Edelsky's work on cooperative floors [31], in this research, we successfully operationalised the comprehensive measure of Conversation Quality. In order to comprehensively quantify spontaneous interactions with respect to individual experiences, the Conversation Quality measure captures four rich aspects of an interaction, namely *Quality of Interaction, Degree of Rapport, Degree of Likeness* and *Free-for-All.* To measure Conversation Quality in spontaneous interactions, we rely on the perceived form of the measure. Social interactions being multi-level systems [67], we measure the perceived Conversation Quality at two different levels of perceptions, namely the individual- (Perceived Individual's Experience of Conversation Quality) and group- level (Perceived Group's Conversation Quality). With this setup, drawing inspirations from Cuperman and Ickes' (PES) [28] and Lindley and Monk [79], we framed two questionnaires (QPCQ) which capture the Conversation Quality at their respective levels of perception. The scientifically backed QPCQ questionnaire relied on external naive annotators to annotate for Conversation Quality by only using video clips of spontaneous interactions and hence, the QPCQ is a time efficient and easy-to-use questionnaire to comprehensively measure Conversation Quality.

In this research, the QPCQ questionnaire was annotated by external annotators to collect the ground-truth labels of perceived Conversation Quality, in the MatchNMingle [17] dataset. The Conversation Quality annotations received for the 115 spontaneous interactions was already an interesting dataset with several interesting patterns. One important finding was that the inter-annotator agreement was higher in the individual-level Conversation Quality annotations than that of the group-level Conversation Quality annotations. This suggests that a higher degree of subjectivity is involved in the perception group-level Conversation Quality than that of the individual-level Conversation Quality. The subjectivity involved may be directly influenced by how external annotators aggregate perceived individual-level experiences to determine the group-level Conversation Quality. The lack of agreeability in group-level Conversation Quality annotations demanded us to ignore a large portion of data samples for further experiments. Probably, such a lack of agreeability amongst annotators can solved by replacing naive annotators with trained annotators. Another interesting pattern, observed in the annotations of both the individual- and group- level Conversation Quality is that, the inter-rater agreeability is higher along the extremes of the Conversation Quality scale and lower between the extremes of the scale, a similar result was obtained by Hung et al. [56], while studying cohesion in meetings.

With the above contributions, we believe this research has successfully devised a novel measure to comprehensively quantify spontaneous interactions. Such a measure, the *Conversation Quality*, has wide range of applications in building social intelligent systems from Computer Supported Cooperative work to development of Feedback Systems.

Research Question 2: Properties of spontaneous conversations To understand the properties of the measure of *Conversation Quality*, we performed several statistical tests, with the *Conversation Quality* annotations as the dependent variable and the behavioural features extracted (Section-5.3) as the independent variable. The statistical tests involved two-levels of complementary analysis, the QLS based statistical test followed by the Joint LASSO test. While the QLS model studied the independent effect of different feature sets on the respective *Conversation Quality* variables, the Joint LASSO studied the combined effect of all the features on the respective *Conversation Quality* variables. To achieve this, a robust Quantile Least Squares (QLS) model, which is free of the assumptions made by an OLS model was used with a Bonferroni Correction for significance, and, a scalable Joint LASSO model coupled with a rank correlation test was used with significance determined based on the sparsity induced by the LASSO model.

The quantitative analysis on the individual-level Conversation Quality, the Perceived Individual's Experience of Conversation Quality (IndivCQ), revealed several interesting insights. With respect to the group cardinality, the results suggest that the individual-level conversation quality decreases with the increase in number of participants in the group interaction. Such a result is in-line with the works on schisms [105] and conversational floors [100]. This effect of group cardinality on the individual-level Conversation Quality was further statistically backed by the results of the Joint LASSO model's results. With respect to turn-taking based features, we see that the equality in talk duration and the percentage of an individual's silence are positively and negatively correlated with IndivCQ and these inferences are statistically backed by both the QLS and Joint LASSO models. Interestingly, the Joint LASSO model reveals that the conversation synchronisation based features also have a significant effect on individual-level Conversation Quality, with the number of successful and unsuccessful interruptions that an individual experiences in a conversation having a negative and a positive effect respectively on the perceived individual-level Conversation Quality. With respect to the body coordination features, we see that the results of the QLS and the Joint LASSO model are different from one another, with little commonality in their results. The results of the two models commonly agree on the results that *argmin* and *argmax* variants of lagged correlation, lagged mimicry and the symmetric convergence features have a statistically significant effect on the individual-level Conversation Quality. This suggests that the synchrony based bodily coordination have the most significant effect on the perceived individual-level Conversation Quality.

The results of the quantitative analysis on the group-level Conversation Quality, the Perceived Group's Conversation Quality (GroupCQ), was different from that of the individual-level Conversation Quality, suggesting that the properties of Conversation Quality differs significantly with respect to their forms of perception. With respect to the group cardinality, the results of both the QLS and the Joint LASSO models, suggest that the group-level conversation quality decreases with the increase in number of participants in the group interaction. This reveals that the number of participants in the conversation has a negative effect on both the individual- and group- level Conversation Quality. With respect to turn-taking based features, the results of the QLS model and the Joint LASSO model are very different from one another and do not complement each other's results. While the QLS model considers that conversation synchronisation features have the most significant effect on GroupCQ, the Joint LASSO model considers that the conversation equality and fluency features have the most significant effect on GroupCQ. An interesting finding is that the interruption based features, successful and unsuccessful interruptions have a similar effect on both the individual- and group-levels of Conversation Quality with a negative and a positive effect respectively. With respect to the relationship between the body coordination features and the GroupCQ, the QLS model is stricter in determining statistical significance and only considers the *argmin* and *argmax* variants of lagged correlation as the only significant features. The significance of the *argmin* and *argmax* variants of lagged correlation features are also statistically backed by the results of the Joint LASSO model. In addition to the *argmin* and *argmax* variants of lagged correlation features, the results of the Joint LASSO model suggests that the lagged mimicry and causality based features also have a significant effect on the group-level Conversation Quality. It also important to note the *argmin* and *argmax* variants of lagged correlation features have a significant effect on both individual- and group- levels of Conversation Quality.

With two levels of complementary statistical tests, the properties of the both the individual- and group-level Conversation Quality were studied. From the statistical tests, an important takeaway is that, at the group-level, the study of independent effect of different feature sets (studied using the QLS) and the study of their combined effect (studied using the Joint LASSO) reveal different properties of the Conversation Quality. This result could be due to the fact that group-level Conversation Quality dataset was comparatively a smaller dataset with a high class imbalance and hence the QLS model was stricter while revealing significant effects. The same two studies, at the individual-level, backed each other with their results to an extent, with the Joint

LASSO model revealing additional properties of the Conversation Quality.

Research Question 3: Predictive modeling in spontaneous conversations We used a simple Logistic Regression based model, optimised using the Stochastic Gradient algorithm, to answer this particular research question. The predictive modeling of Conversation Quality, both the individual- and the group- level, was performed using the same modeling technique.

With respect to the individual-level Conversation Quality, the Perceived Individual's Experience of Conversation Quality (IndivCQ), our modeling technique was able to predict the construct with a Mean AUC of 0.76 (± 0.13). An interesting finding is that the pre-processing of accelerometer signals using a sliding-window approach tends to reduce the model performance substantially, and the model which uses features without any sliding-window based pre-processing is the best performing model, contradicting Kapcak et al.'s [61] study on estimating romantic interest. With respect to the the best performing feature sets, our results indicate that the Synchrony and Convergence based features are the best performing feature sets, outperforming the Turn-Taking based features. At the same time, amongst the group-level aggregation features, the Median based feature set performs the best, in-par with the model which uses a feature-level fusion of all the group-level feature aggregators. The performance of the Median based feature set is in line with the inferences drawn by Nanninga et al.'s [88] while studying cohesion in meetings. This is an interesting finding for the fact that, the non-skewed version of the group-level aggregators, the *Median*, is the most informative of IndivCQ, suggesting that the skewness in distribution of body coordination amongst the interacting pairs in the group is highly descriptive of the Conversation Quality.

With respect to the group-level Conversation Quality, the Perceived Group's Conversation Quality (GroupCQ), our modeling technique was able to predict the construct with a Mean AUC of 0.96 (± 0.03). At this moment, it is also important to note that the dataset used to predictively model GroupCQ was relatively small (58 samples), and with a large class imbalance (only 3 samples of the negative class). Nevertheless, several interesting insights were found during the modeling of GroupCQ. Surprisingly, the Turn-Taking based features are the best performing feature set in predicting GroupCQ, outperforming the bodily coordination based features. This is in contrast to the modeling of IndivCQ, where the bodily coordination features outperformed the Turn-Taking based features by a large margin. This suggests that while perceiving GroupCQ, annotators tend to rely more on the turn-taking based cues, rather than other social cues displayed by the group. A similar contrasting results was obtained during the group-level features experiment as well, where the Mean based group-level aggregation features outperformed both the Median and Fusion based features, while predicting GroupCQ. Considering both the above contrasting results, an interesting conclusion we can draw is that the two levels of *Conversation Quality*, as a social construct, are fundamentally very different from one another, and hence, require separate attention while modeling the respective constructs.

With the predictive modeling methodology, using behavioural features such as turn-taking and bodily coordination, we were able to successfully predict both the individual- and group- level forms of *Conversation Quality*, and also quantitatively detect the best defining features.

8 Conclusion

In this chapter, we present the concluding remarks on our study of *Conversation Quality* in spontaneous interactions. Additionally, we also discuss the limitations of this study and some potential future research questions with respect to *Conversation Quality* in spontaneous interactions.

8.1. Main Conclusion

In this research, we designed a novel measure, the perceived *Conversation Quality*, which describes the perceived quality of spontaneous social interactions, with respect to the individual experiences. Such a perceived measure of spontaneous interactions are a vital contribution towards the development of social robots and feedback systems. To quantitatively study the perceived Conversation Quality, we devised a questionnaire which measures, at the individual- and at the group- level, the perceived Conversation Quality in spontaneous interactions. To the best of our knowledge, there is no existing work in the literature which has attempted to comprehensively quantify spontaneous interactions.

The MatchNMingle dataset [17], a multi-modal dataset consisting of in-the-wild spontaneous interactions, was used in this research. Inspired from existing literature, we extracted two categories of behavioural features, Turn-Taking and Bodily Coordination features, to further model the perceived *Conversation Quality* and study its properties. Using a Logistic Regression, optimised using the Stochastic Gradient Descent algorithm, we were able to predict the individual- and the group- level perceived Conversation Quality with a mean AUC of 0.76 (±0.13) and 0.96 (±0.03) respectively. From the experiments performed, we see that the Synchrony and Convergence based bodily coordination features are the best performing feature sets while predicting the *individual-level* perceived Conversation Quality and, the Turn-Taking based features are the best performing feature sets while predicting the *group-level* perceived Conversation Quality. The results suggest that the two forms of perceived Conversation Quality, the individual- and the group- level, are completely different from one another and demands researchers to handle them with respective considerations.

To further reveal the properties of the perceived Conversation Quality measure, we studied the effect of different social factors on the measure by performing suitable statistical tests. The results show that the perceived Conversation Quality, both the the individual- and the group- level, decreases with the increase in number of participants in the spontaneous interaction. Moreover, the equal distribution of talk time amongst participants and the duration of their silence periods have a significant effect on the perceived Conversation Quality, suggesting that external annotators tend to use these factors as proxy to judge individual experiences. Another interesting finding is that successful and unsuccessful interruptions have a positive effect and a negative effect on perceived Conversation Quality respectively, aligning with the fact that successful interruptions might result because of back-channels and talk overlap which are indicative of an individual's involvement with their interacting partners. Moreover, the results show that the time factor revealing bodily coordination features (lagged correlations and convergence) have a significant effect on both the two levels of Conversation Quality, suggesting that the time factor involved in the bodily coordination is more informative than the degree of coordination in itself.

8.2. Future Research

While we contextualise the results, it is also important to address the short-comings of this research and speculate on potential future research questions. In this section, we present several such potential research questions intended to extend further research on *Conversation Quality* in spontaneous interactions.

Trained external annotators for perceived Conversation Quality As discussed earlier, the lack of agreeability in group-level Conversation Quality annotations demanded us to ignore a large portion of data samples for further experiments, resulting in a comparatively smaller dataset with high class imbalance. This prevented us from drawing strong conclusions with respect to the group-level form of Conversation Quality. The lack of agreeability can be owed to the subjectivity involved in group-level annotations, the fact that different annotators use a different aggregation of individual experiences strategy to annotate for group-level Conversation Quality. The subjectivity involved is especially large in case of naive external annotators, as used in this research. For this reason, for future research, annotations of perceived Conversation Quality, both the individual- and group- level, can be collected using trained external annotators. Using trained external annotators, in place of naive annotators and with the same questionnaire, might provide us with a richer dataset for study on Conversation Quality.

Dynamics in Conversation Quality In the current study, we had described a spontaneous interaction using one perceived Conversation Quality, with an assumption that there exists one perceived conversation quality score which is stable throughout the interaction. But, social interactions and individual experiences are dynamic in nature and requires a more fine grained approach. Several researchers have handled this using a thin-slice based annotations. Such a thin-slice based approach can help us further study the Conversation Quality along with its dynamics and such a study of dynamics in individual experiences can be a substantial contribution towards building social robots.

Individual-level Conversation Quality's influence on group-level Conversation Quality In this study, we had modeled the two levels of perceived Conversation Quality, the individual- and the group- level, as two independent phenomena in spontaneous interaction. But literature works in Emergent Systems [107] and Multi-level Systems [41] consider lower level (e.g. individual-level) constructs to have significant influence on higher level (e.g. group-level) constructs and vice versa, and consider the higher level constructs are more than just the aggregation of lower level constructs. With respect to these literature works, it would be an interesting research topic to study the dynamic influence of the two levels of Conversation Quality on each other. Such a study will be strong contribution towards development of Feedback Systems, where the system's understanding of the influence across levels can help in providing interventions and constructive feedbacks.

Data-driven aggregation for group-level features Due to the fact that synchrony is a window into pair-wise and dyadic alignment, synchrony as a group-level feature has been understudied and less understood. Synchrony has been extensively studied in dyadic settings and, in a small group setting the pair-wise synchrony measures are generally aggregated to form group-level features (socio-evolutionary perspective which treats groups as aggregate of individuals). In this research, we extracted group-level features with respect to the socio-evolutionary perspective, using minimum, maximum, mean, variance, mode and median based aggregations. From the results, we see that the data models fit the dataset with respect to the type of aggregation used for the particular feature, suggesting that group-level feature aggregation is very important in modeling group behaviour. Hence, it would be an interesting perform more data-driven based techniques to extract group-level features. For example, an end-to-end neural network based netVLAD architecture [4] can be used to extract group-level features in a data-driven manner. Such a modeling technique could help in better predicting the perceived Conversation Quality.

Bibliography

[1] John Aldrich et al. Correlations genuine and spurious in pearson and yule. *Statistical science*, 10(4): 364–376, 1995.

[2] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[3] John H Antil. Conceptualization and operationalization of involvement. *Advances in consumer research*, 11(1):203–209, 1984.

[4] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2015.

[5] Michael C Ashton and Kibeom Lee. Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and social psychology review*, 11(2):150–166, 2007.

[6] U. Avci and O. Aran. Predicting the performance in decision-making tasks: From individual cues to group interaction. *IEEE Transactions on Multimedia*, 18(4):643–658, April 2016. ISSN 1941-0077. doi: 10.1109/TMM.2016.2521348.

[7] Roman Bednarik, Shahram Eivazi, and Michal Hradis. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*, pages 1–6, 2012.

[8] Frank J Bernieri, John S Gillis, Janet M Davis, and Jon E Grahe. Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, 71(1): 110, 1996.

[9] Diane S Berry and Jane Sherman Hansen. Personality, nonverbal behavior, and interaction quality in female dyads. *Personality and Social Psychology Bulletin*, 26(3):278–292, 2000.

[10] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. *IEEE Transactions on Multimedia*, 20(2):441–456, 2017.

[11] Niall Bolger and Jean-Philippe Laurenceau. *Intensive longitudinal methods: An introduction to diary and experience sampling research.* Guilford Press, 2013.

[12] Joseph A Bonito and Joann Keyton. Multilevel measurement models for group collective constructs. *Group Dynamics: Theory, Research, and Practice*, 23(1):1, 2019.

[13] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 351–368, 2008.

[14] Mckenzie Braley and Gabriel Murray. The group affect and performance (gap) corpus. In *Proceedings of the Group Interaction Frontiers in Technology*, pages 1–9. 2018.

[15] Oliver Brdiczka, Jérôme Maisonnasse, and Patrick Reignier. Automatic detection of interaction groups. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 32–36, 2005.

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

[17] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. 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*, PP(c):1, 2018. ISSN 19493045. doi: 10.1109/TAFFC.2018.2848914.

[18] Nick Campbell and Stefan Scherer. Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[19] Jean Carletta, Simon Garrod, and Heidi Fraser-Krauss. Placement of authority and communication patterns in workplace groups: The consequences for innovation. *Small Group Research*, 29(5):531–559, 1998.

[20] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer, 2005.

[21] Milly Casey-Campbell and Martin Martens. Sticking it all together: A critical assessment of the group cohesion–performance literature. *International Journal of Management Reviews*, 11, 05 2009. doi: 10.1111/j.1468-2370.2008.00239.x.

[22] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8749:1–15, 2014. ISSN 16113349. doi: 10.1007/978-3-319-11839-0_1.

[23] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[24] Shammur Absar Chowdhury, Evgeny A. Stepanov, and Giuseppe Riccardi. Predicting user satisfaction from turn-taking in spoken conversations. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 08-12-Sept:2910–2914, 2016. ISSN 19909772. doi: 10.21437/Interspeech.2016-859.

[25] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[26] Jason A. Cohen. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, 70 4:213–20, 1968.

[27] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. pages 23.1–23.12, 01 2011. ISBN 1-901725-43-X. doi: 10.5244/C.25.23.

[28] Ronen Cuperman and William Ickes. Big Five Predictors of Behavior and Perceptions in Initial Dyadic Interactions: Personality Similarity Helps Extraverts and Introverts, but Hurts "Disagreeables". *Journal of Personality and Social Psychology*, 97(4):667–684, 2009. ISSN 00223514. doi: 10.1037/a0015741.

[29] Owen Daly-Jones, Andrew Monk, and Leon Watts. Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus. *International Journal of Human-Computer Studies*, 49(1):21–58, 1998.

[30] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012. ISSN 19493045. doi: 10.1109/T-AFFC.2012.12.

[31] Carole Edelsky. Who's got the floor? *Language in Society*, 10(3):383–421, 1981. ISSN 00474045, 14698013. URL `http://www.jstor.org/stable/4167262`.

[32] Jens Edlund, Julia Bell Hirschberg, and Mattias Heldner. Pause and gap length in face-to-face interaction. 2009.

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

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

[35] John Garcia and Andrew R Gustavson. The science of self-report. *APS Observer*, 10(1), 1997.

[36] Daniel Gatica-perez. Automatic nonverbal analysis of social interaction in small groups : A review. *Image and Vision Computing*, 27(12):1775–1787, 2009. ISSN 0262-8856. doi: 10.1016/j.imavis.2009.01.004. URL `http://dx.doi.org/10.1016/j.imavis.2009.01.004`.

[37] Daniel Gatica-Perez, Iain McCowan, Dong Zhang, and Samy Bengio. Detecting group interest-level in meetings. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, pages 489–492, 2005. doi: 10.1109/ICASSP.2005.1415157. URL `https://doi.org/10.1109/ICASSP.2005.1415157`.

[38] Ekin Gedik and Hayley Hung. Speaking status detection from body movements using transductive parameter transfer. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: adjunct*, pages 69–72, 2016.

[39] Ekin Gedik and Hayley Hung. Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):163, 2018.

[40] Ekin Gedik, Laura Cabrera-quiros, Claudio Martella, and Gwenn Englebienne. Towards Analyzing and Predicting the Experience of Live Performances with Wearable Sensing. *IEEE Transactions on Affective Computing*, PP(8):1, 2018. doi: 10.1109/TAFFC.2018.2875987.

[41] Riemannian Geometry and Geometric Analysis. *Advancing Multilevel Research Design: Capturing the Dynamics of Emergence - Steve*. Number Cdm. ISBN 9783540773405.

[42] Erving Goffman. *Encounters: Two studies in the sociology of interaction*. Ravenio Books, 1961.

[43] Erving Goffman. *Behavior in public places*. Simon and Schuster, 2008.

[44] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[45] Hatice Gunes and Björn Schuller. *Automatic Analysis of Social Emotions*, pages 213 – 224. 05 2017. doi: 10.1017/9781316676202.016.

[46] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez. Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 613–616, Oct 2011. doi: 10.1109/PASSAT/SocialCom.2011.143.

[47] Juan Lorenzo Hagad, Roberto Legaspi, Masayuki Numao, and Merlin Suarez. Predicting Levels of Rapport in Dyadic Interactions Through Automatic Detection of Posture and Posture Congruence. *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 613–616, 2011. doi: 10.1109/PASSAT/SocialCom.2011.143.

[48] Edward Twitchell Hall and T Hall. *The silent language*, volume 948. Anchor books, 1959.

[49] Fumio Hayashi. Econometrics. page 10 and 34, 2000.

[50] Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[51] JM Helmhout, Henk WM Gazendam, and René J Jorna. Emergence of social constructs and organizational behavior. In *21st EGOS colloquium, Berlin*, 2005.

[52] Lotta Hirvenkari, Johanna Ruusuvuori, Veli-Matti Saarinen, Maari Kivioja, Anssi Peräkylä, and Riitta Hari. Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PloS one*, 8(8), 2013.

[53] J Hofmann, F Stoffel, A Weber, and T Platt. The 16 enjoyable emotions induction task (16-eeit). *Unpublished Research instrument, Department of Psychology, University of Zurich, Switzerland*, 2012.

[54] Leaetta M. Hough. The 'big five' personality variables–construct confusion: Description versus prediction. 1992.

[55] Chiao-Yin Hsiao, Wan-Rong Jih, and Jane Hsu. Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns. 07 2012.

[56] Hayley Hung and Daniel Gatica-perez. Estimating Cohesion in Small Groups Using. *IEEE Transactions on Multimedia*, 12(6):563–575, 2010. doi: 10.1109/TMM.2010.2055233.

[57] Hayley Hung, Dinesh Babu, Sileye Ba, Jean-marc Odobez, and Daniel Gatica-perez. Investigating Automatic Dominance Estimation in Groups From Visual Attention and Speaking Activity. pages 2–5.

[58] Hayley Hung, Gwenn Englebienne, and Jeroen Kools. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 207–210, 2013.

[59] Joseph Jaffe, Beatrice Beebe, Stanley Feldstein, Cynthia L Crown, Michael D Jasnow, Philippe Rochat, and Daniel N Stern. Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the society for research in child development*, pages i–149, 2001.

[60] Natasha Jaques, Daniel McDuff, Yoo Lim Kim, and Rosalind Picard. Understanding and predicting bonding in conversations using thin slices of facial expressions and body language. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10011 LNAI:64–74, 2016. ISSN 16113349. doi: 10.1007/978-3-319-47665-0_6.

[61] Ö. Kapcak, J. Vargas-Quiros, and H. Hung. Estimating romantic, social, and sexual attraction by quantifying bodily coordination using wearable sensors. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 154–160, Sep. 2019. doi: 10.1109/ACIIW.2019.8925137.

[62] Adam Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63, 02 1967. doi: 10.1016/0001-6918(67)90005-4.

[63] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[64] Adam Kendon, Richard M Harris, and Mary R Key. *Organization of behavior in face-to-face interaction*. Walter de Gruyter, 2011.

[65] Katherine J Klein, Amy Buhl Conn, D Brent Smith, and Joann Speer Sorra. Is everyone in agreement? an exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, 86(1):3, 2001.

[66] Steve W J Kozlowski and Katherine J Klein. A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. *Multilevel theory, research and methods in organizations: Foundations, extensions, and new directions*, (October 2012):3–90, 2000.

[67] Steve W.J. Kozlowski, Georgia T. Chao, James A. Grand, Michael T. Braun, and Goran Kuljanin. *Capturing the multilevel dynamics of emergence: Computational modeling, simulation, and virtual experimentation*, volume 6. 2016. ISBN 2041386614. doi: 10.1177/2041386614547955.

[68] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[69] Uliyana Kubasova, Gabriel Murray, and McKenzie Braley. Analyzing verbal and nonverbal features for predicting group performance. *arXiv preprint arXiv:1907.01369*, 2019.

[70] Catherine Lai and Gabriel Murray. Predicting group satisfaction in meeting discussions. *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, MCPMD 2018*, 2018. doi: 10.1145/3279810.3279840.

[71] Catherine Lai, Jean Carletta, and Steve Renals. Modelling participant affect in meetings with turn-taking features. In *Proc. Workshop of Affective Social Speech Signals*, 2013.

[72] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/2529310`.

[73] Sylvia Tamara Lenz. Alan agresti (2013): Categorical data analysis. *Statistical Papers*, 57(3):849, 2016.

[74] Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731, 2015.

[75] Rivka Levitan, Agustín Gravano, and Julia Hirschberg. Entrainment in speech preceding backchannels. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2:113–117, 2011.

[76] Rivka Levitan, Agustín Gravano, and Julia Hirschberg. Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 113–117, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P11-2020`.

[77] Rivka Levitan, Stefan Benus, Agustin Gravano, and Julia Hirschberg. Entrainment and turn-taking in human-human dialogue. In *2015 AAAI spring symposium series*, 2015.

[78] Siân E Lindley and Andrew F Monk. Social enjoyment with electronic photograph displays: Awareness and control. *International Journal of Human-Computer Studies*, 66(8):587–604, 2008.

[79] Siân E. Lindley and Andrew F. Monk. Measuring social behaviour as an indicator of experience. *Behaviour & Information Technology*, 32(10):968–985, oct 2013. ISSN 0144-929X. doi: 10.1080/0144929X.2011.582148. URL `http://www.tandfonline.com/doi/abs/10.1080/0144929X.2011.582148`.

[80] Florian Lingenfelser, Johannes Wagner, Elisabeth André, Gary McKeown, and Will Curran. An event driven fusion approach for enjoyment recognition in real-time. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 377–386, 2014.

[81] Claudio Martella, Ekin Gedik, Laura Cabrera-quiros, Gwenn Englebienne, Hayley Hung, Instituto Tecnolgico, De Costa Rica, and Costa Rica. How Was It ? Exploiting Smartphone Sensing to Measure Implicit Audience Responses to Live Performances Categories and Subject Descriptors. pages 201–210.

[82] Ross Mead, Amin Atrash, and Maja Matarić. Automated proxemic feature extraction and behavior recognition: Applications in human-robot interaction. *International Journal of Social Robotics*, 5, 08 2013. doi: 10.1007/s12369-013-0189-8.

[83] Jan Michalsky and Heike Schoormann. Pitch convergence as an effect of perceived attractiveness and likability. In *INTERSPEECH*, pages 2253–2256, 2017.

[84] Kevin W Mossholder and Arthur G Bedeian. Cross-level inference and organizational research: Perspectives on interpretation and application. *Academy of Management Review*, 8(4):547–558, 1983.

[85] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. *CoRR*, abs/1801.06055, 2018. URL `http://arxiv.org/abs/1801.06055`.

[86] Gabriel Murray and Catharine Oertel. Predicting group performance in task-based interaction. *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, pages 14–20, 2018. doi: 10.1145/3242969.3243027.

[87] Gabriel Murray and Catharine Oertel. Predicting group performance in task-based interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 14–20, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5692-3. doi: 10.1145/3242969.3243027. URL `http://doi.acm.org/10.1145/3242969.3243027`.

[88] Marjolein C. Nanninga, Yanxia Zhang, Nale Lehmann-Willenbrock, Zoltán Szlávik, and Hayley Hung. 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*, ICMI '17, pages 206–215, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5543-8. doi: 10.1145/3136755.3136811. URL `http://doi.acm.org/10.1145/3136755.3136811`.

[89] Lance J. Rips Norman M. Bradburn and Steven K. Shevell. Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. In *New Series 1987*, pages 236(4798):157–167. 1987.

[90] David A Northrup. *The problem of the self-report in survey research*. Institute for Social Research, York University, 1997.

[91] Catharine Oertel and Giampiero Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 99–106, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2129-7. doi: 10.1145/2522848.2522865. URL `http://doi.acm.org/10.1145/2522848.2522865`.

[92] Catharine Oertel, Céline De Looze, Stefan Scherer, Andreas Windmann, Petra Wagner, and Nick Campbell. Towards the automatic detection of involvement in conversation. In Anna Esposito, Alessandro Vinciarelli, Klára Vicsi, Catherine Pelachaud, and Anton Nijholt, editors, *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, pages 163–170, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-25775-9.

[93] Catharine Oertel, Stefan Scherer, and Nick Campbell. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August):1541–1544, 2011. ISSN 19909772.

[94] Catharine Oertel, Joakim Gustafson, Kenneth A.Funes Mora, and Jean Marc Odobez. Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, (November):107–114, 2015. doi: 10.1145/2818346.2820759.

[95] Daniel Olguin Olguin, Peter A. Gloor, and Alex (Sandy) Pentland. Capturing Individual and Group Behavior with Wearable Sensors. *Proceedings of the 3d International ICST Conference on Pervasive Computing Technologies for Healthcare*, 2009. doi: 10.4108/ICST.PERVASIVEHEALTH2009.6033. URL `http://eudl.eu/doi/10.4108/ICST.PERVASIVEHEALTH2009.6033`.

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

[97] Lars Penke and Jens B Asendorpf. Beyond global sociosexual orientations: a more differentiated look at sociosexuality and its effects on courtship and romantic relationships. *Journal of personality and social psychology*, 95(5):1113, 2008.

[98] Alex Pentland and Anmol Madan. Perception of social interest. In *Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI)*. Citeseer, 2005.

[99] Marshall Scott Poole, Andrea B. Hollingshead, Joseph E. McGrath, Richard L. Moreland, and John Rohrbaugh. Interdisciplinary perspectives on small groups. *Small Group Research*, 35(1):3–16, 2004. doi: 10.1177/1046496403259753. URL `https://doi.org/10.1177/1046496403259753`.

[100] Chirag Raman and Hayley Hung. Towards automatic estimation of conversation floors within F-formations. *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019*, pages 175–181, 2019. doi: 10.1109/ACIIW.2019.8925065.

[101] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. *Journal of consulting and clinical psychology*, 79 (3):284, 2011.

[102] David Reitter, Johanna D Moore, and Frank Keller. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. 2010.

[103] Daniel C Richardson and Rick Dale. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science*, 29(6):1045–1060, 2005.

[104] Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

[105] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

[106] Emanuel Schegloff. *Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences*, pages 71–93. 01 1982.

[107] Klaus R Scherer. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3459–3474, 2009.

[108] Alexander Schmitt and Stefan Ultes. Interaction quality. *Speech Commun.*, 74(C):12–36, November 2015. ISSN 0167-6393. doi: 10.1016/j.specom.2015.06.003. URL `https://doi.org/10.1016/j.specom.2015.06.003`.

[109] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[110] Candace L Sidner, Christopher Lee, Cory Kidd, Neal Lesh, and Charles Rich. Explorations in engagement for humans and robots. *arXiv preprint cs/0507056*, 2005.

[111] Stanley Smith Stevens et al. On the theory of scales of measurement. 1946.

[112] June P Tangney, Roy F Baumeister, and Angie Luzio Boone. High selfcontrol predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of personality*, 72(2):271–324, 2004.

[113] Deborah Tannen. Interpreting interruption in conversation. *Gender and discourse*, pages 53–83, 1989.

[114] Debra Umberson and Jennifer Karas Montez. Social relationships and health: A flashpoint for health policy. *Journal of health and social behavior*, 51(1_suppl):S54–S66, 2010.

[115] Heleen Van Mierlo, Jeroen K Vermunt, and Christel G Rutte. Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, 12(2):368–392, 2009.

[116] Giovanna Varni, Marie Avril, Adem Usta, and Mohamed Chetouani. Syncpy: a unified open-source analytic library for synchrony. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, pages 41–47, 2015.

[117] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[118] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: https://doi.org/10.1038/s41592-019-0686-2.

[119] Marynel Vázquez, Aaron Steinfeld, and Scott Hudson. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. pages 3010–3017, 09 2015. doi: 10.1109/IROS.2015.7353792.

[120] MA Walker, DJ Litman, CA Kamm, and A Abella. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech Language*, 12(4):317 – 347, 1998. ISSN 0885-2308. doi: https://doi.org/10.1006/csla.1998.0110. URL `http://www.sciencedirect.com/science/article/pii/S0885230898901103`.

[121] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[122] Leon Watts, Andrew Monk, and Owen Daly-Jones. Inter-personal awareness and synchronization: assessing the value of communication technologies. *International Journal of Human-Computer Studies*, 44(6):849–873, 1996.

[123] Candace West and Don H Zimmerman. Small insults: A study of interruptions in cross-sex conversations between unacquainted persons. In *American Sociological Association's Annual Meetings, Sep, 1978, San Francisco, CA, US; This is a revised version of a paper presented at the aforementioned conference.* Routledge/Taylor & Francis Group, 2015.

[124] Jeffrey M Wooldridge. *Introductory econometrics: A modern approach.* Nelson Education, 2016.

[125] D. Wyatt, T. Choudhury, and H. Kautz. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–213–IV–216, 2007.

[126] Danny Wyatt, Tanzeem Choudhury, Jeff Bilmes, and James A Kitts. Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):1–41, 2011.

[127] Victor H Yngve. On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*, pages 567–578, 1970.

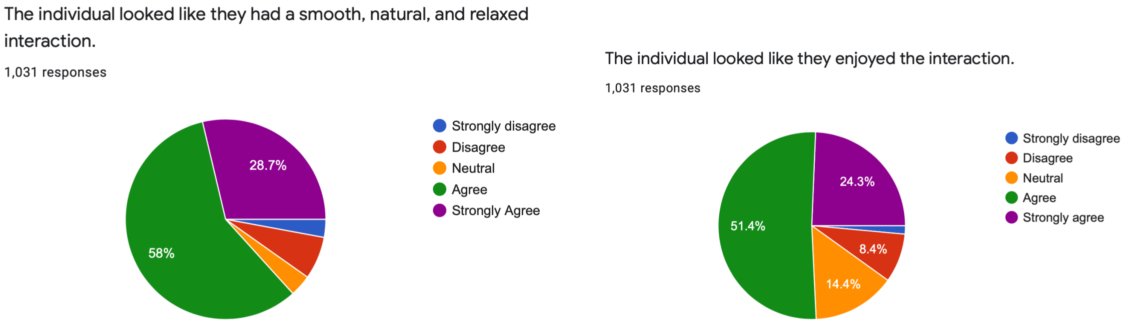
[128] Lu Zhang and Hayley Hung. Beyond F-formations: Determining social involvement in free standing conversing groups from static images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:1086–1095, 2016. ISSN 10636919. doi: 10.1109/CVPR.2016.123.

[129] Yanxia Zhang, FX Palo Alto Laboratory, Usa Jeffrey Olenick, Steve W J Kozlowski, Jeffrey Olenick, Chu-Hsiang Chang, and Hayley Hung. TeamSense: Assessing Personal Affect and Group Cohesion in Small Teams through Dyadic Interaction and Behavior Analysis with Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*, 2(150):150, 2018. doi: 10.1145/3264960. URL `https://doi.org/10.1145/3264960`.

A

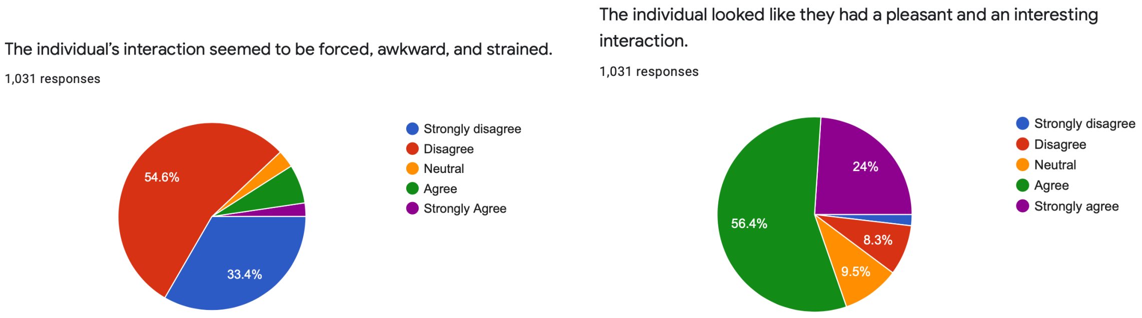
Appendix A: QPCQ Annotation Analysis

A.1. Annotations Distribution: Perceived Individual's Experience of Conversation Quality (IndivCQ)



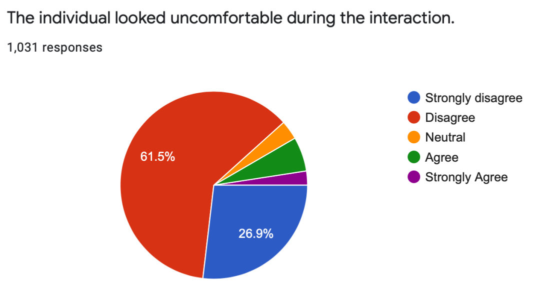
(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 1**.

(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 2**.

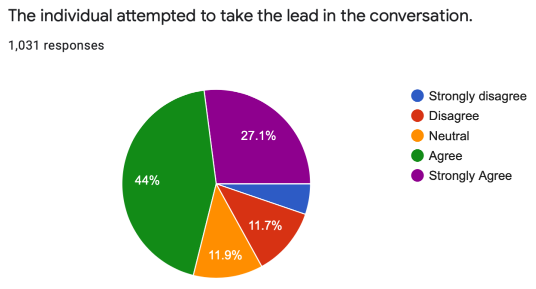


(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 3**.

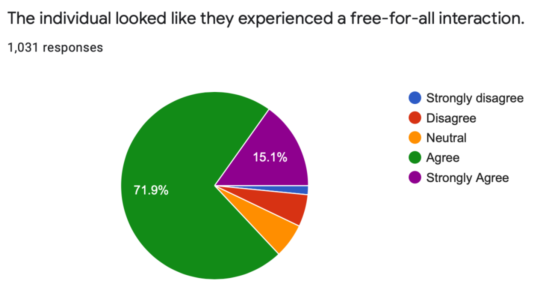
(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 4**.



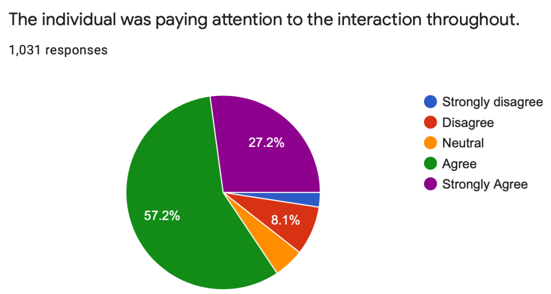
(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 5**.



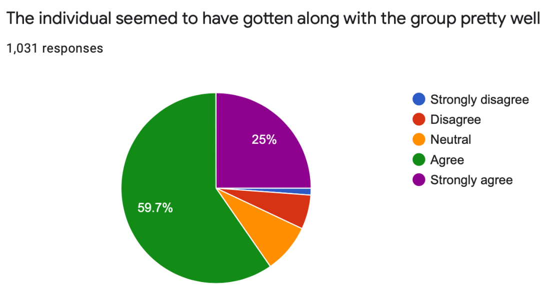
(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 6**.



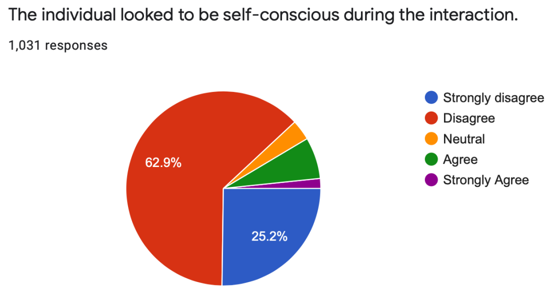
(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 7**.



(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 8**.

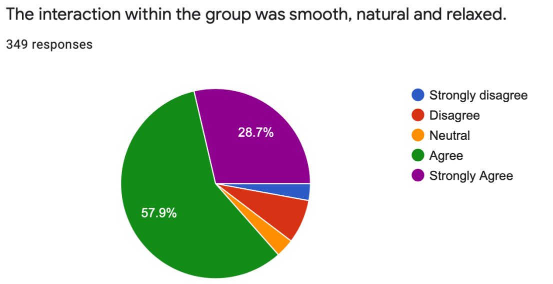


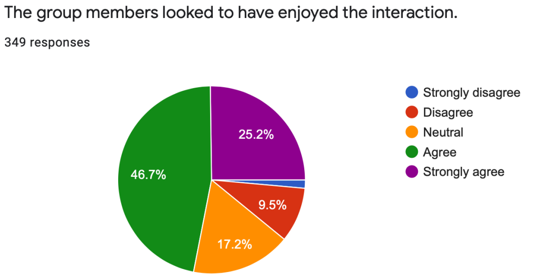
(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 9**.



(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 10**.

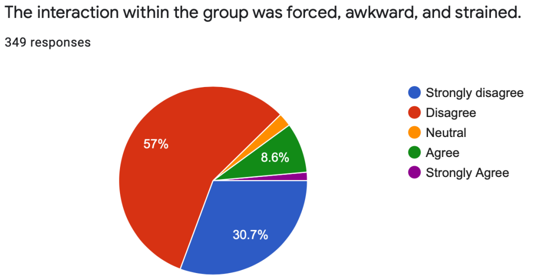
A.2. Annotations Distribution: Perceived Group's Conversation Quality (GroupCQ)

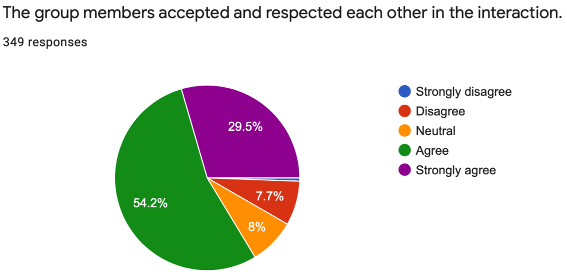




(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 1**.

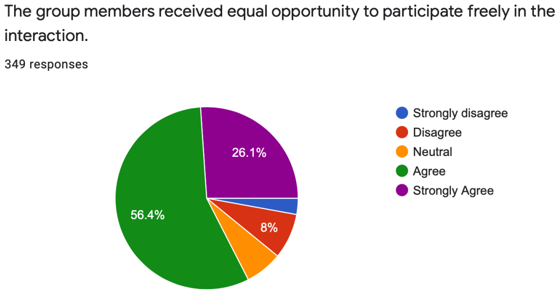
(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 2**.

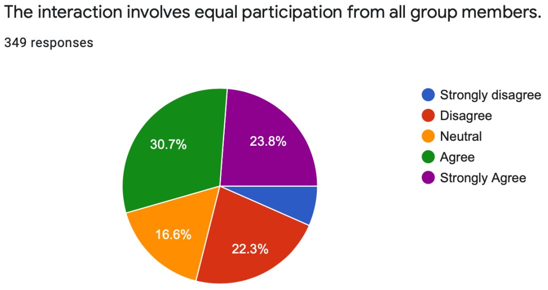




(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 3**.

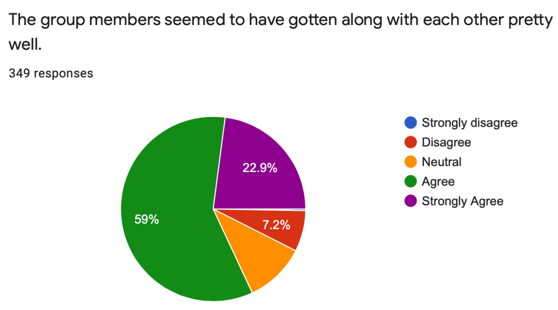
(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 4**.



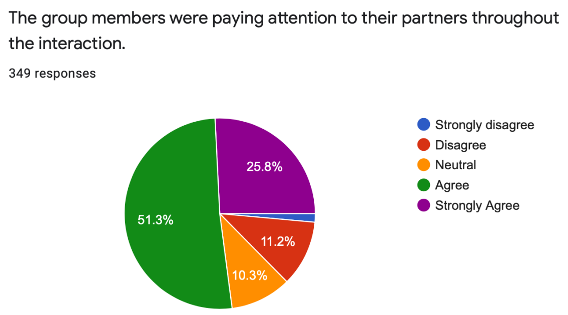


(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 5**.

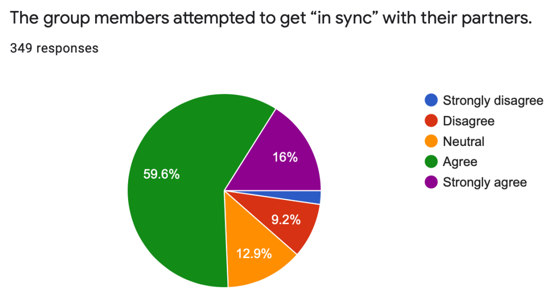
(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 6**.



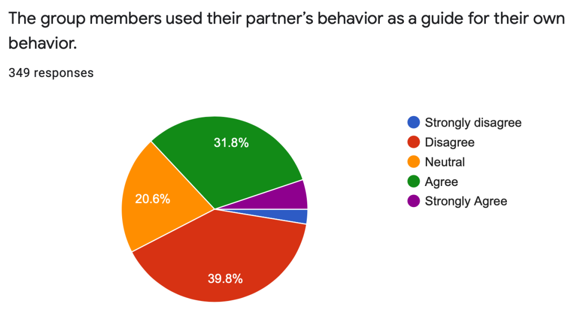
(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 7**.



(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 8**.



(a) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 9**.

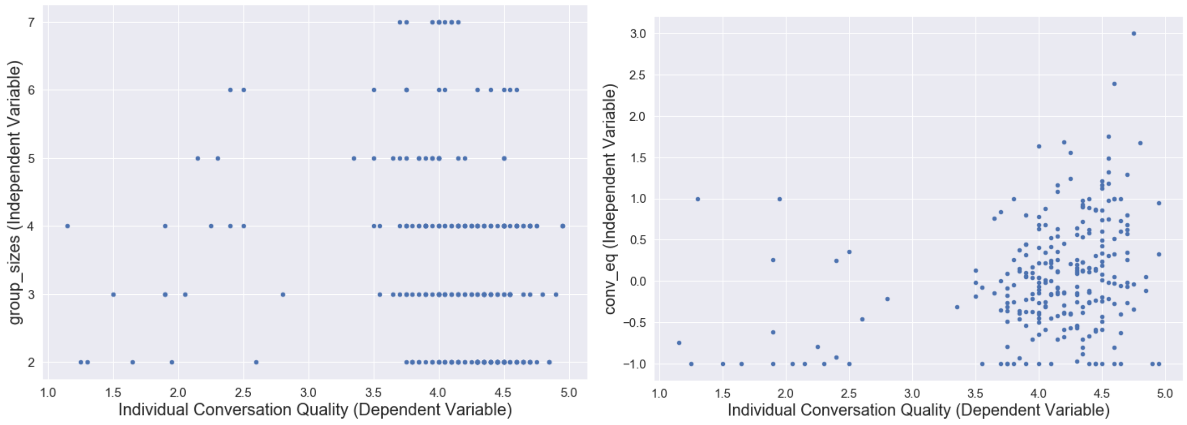


(b) Annotation Responses Distribution (collected from 3 annotators) for **Questionnaire item 10**.

B

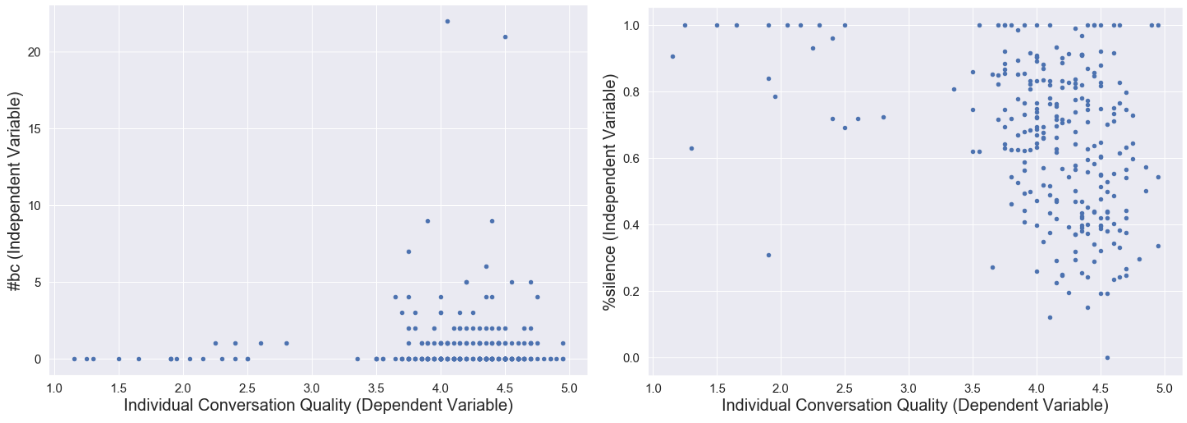
Appendix B: Independent Variable and Dependent Variable plots

B.1. Analysis on Perceived Individual's Experience of Conversation Quality (IndivCQ)



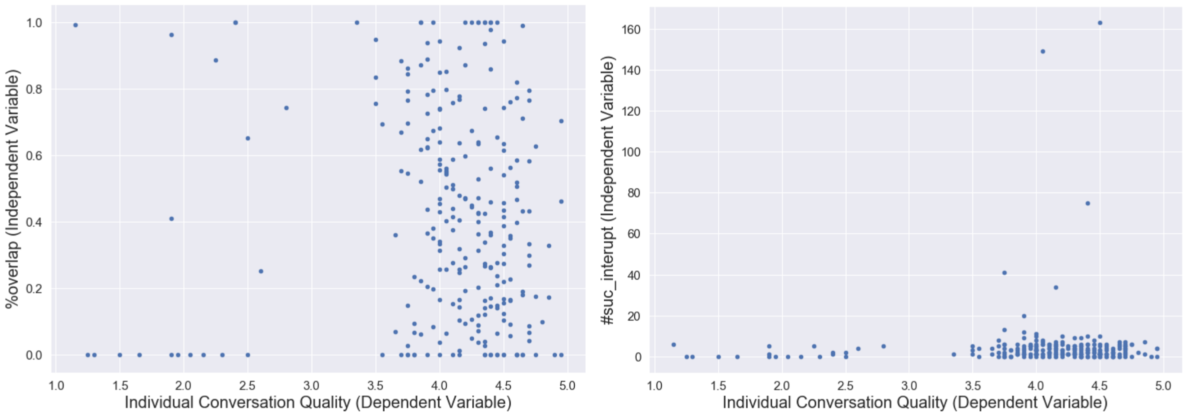
(a) Scatter plot revealing the relationship between Group Cardinality (Independent Variable) and the Individual Conversation Quality (Dependent Variable).

(b) Scatter plot revealing the relationship between Conversation Equality (Independent Variable) and the Individual Conversation Quality (Dependent Variable).



(a) Scatter plot revealing the relationship between Number of Backchannels (Independent Variable) and the Individual Conversation Quality (Dependent Variable).

(b) Scatter plot revealing the relationship between Percentage Silence (Independent Variable) and the Individual Conversation Quality (Dependent Variable).



(a) Scatter plot revealing the relationship between Percentage Overlap (Independent Variable) and the Individual Conversation Quality (Dependent Variable).

(b) Scatter plot revealing the relationship between Number of Successful Interruptions (Independent Variable) and the Individual Conversation Quality (Dependent Variable).

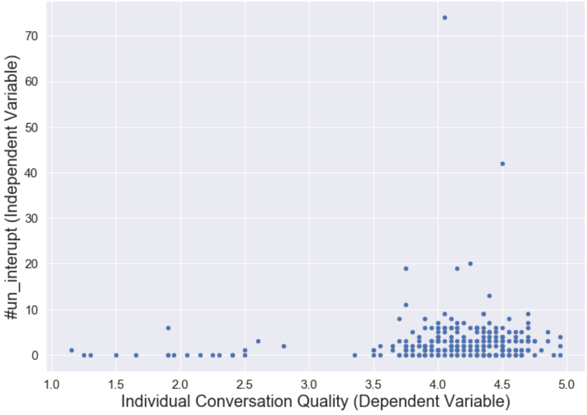
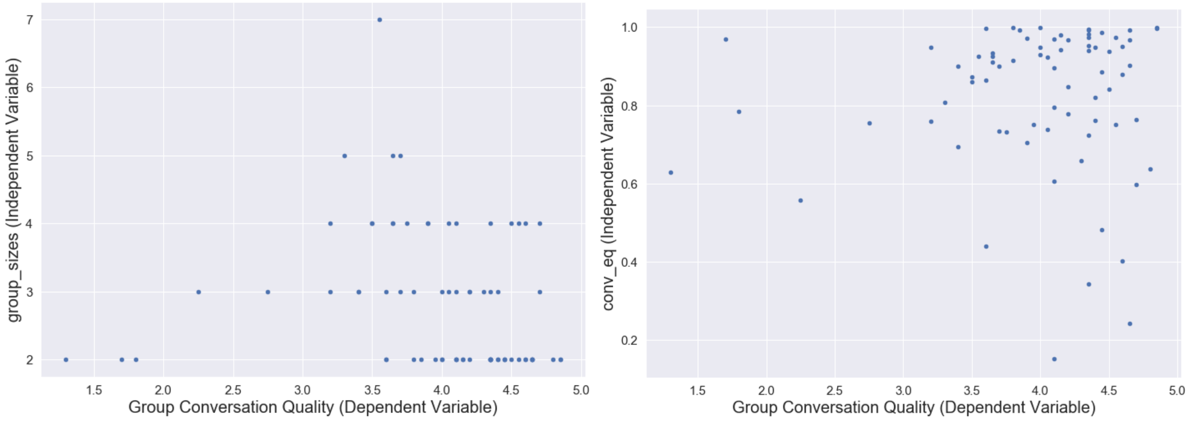


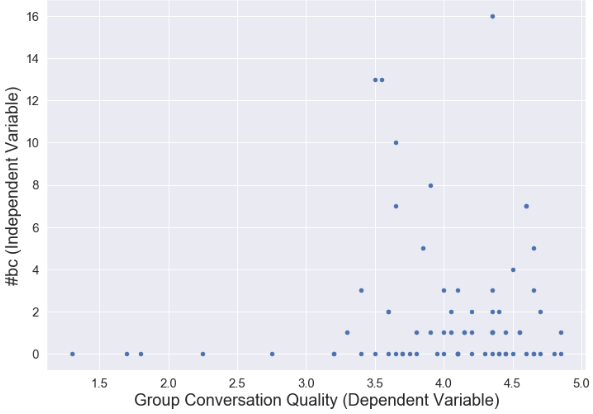
Figure B.4: Scatter plot revealing the relationship between Number of Unsuccessful Interruptions (Independent Variable) and the Individual Conversation Quality (Dependent Variable).

B.2. Analysis on Perceived Group's Conversation Quality (GroupCQ)

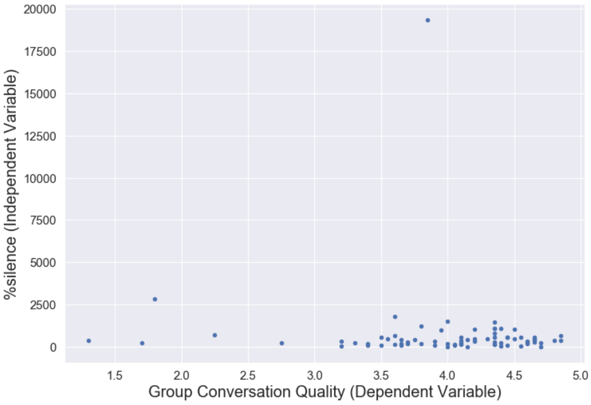


(a) Scatter plot revealing the relationship between Group Cardinality (Independent Variable) and the Group Conversation Quality (Dependent Variable).

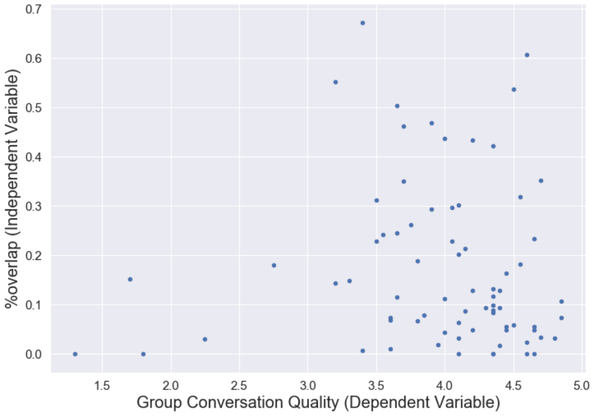
(b) Scatter plot revealing the relationship between Conversation Equality (Independent Variable) and the Group Conversation Quality (Dependent Variable).



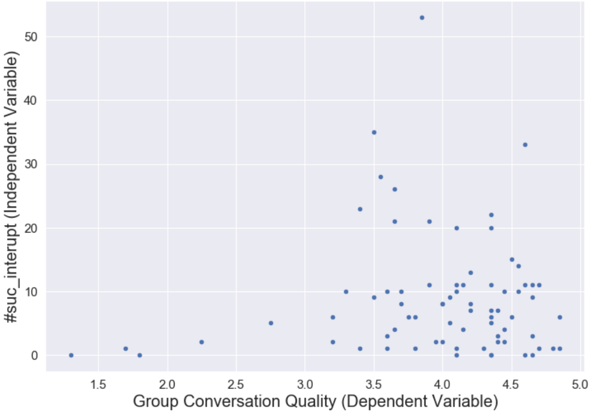
(a) Scatter plot revealing the relationship between Number of Backchannels (Independent Variable) and the Group Conversation Quality (Dependent Variable).



(b) Scatter plot revealing the relationship between Percentage Silence (Independent Variable) and the Group Conversation Quality (Dependent Variable).



(a) Scatter plot revealing the relationship between Percentage Overlap (Independent Variable) and the Group Conversation Quality (Dependent Variable).



(b) Scatter plot revealing the relationship between Number of Successful Interruptions (Independent Variable) and the Group Conversation Quality (Dependent Variable).

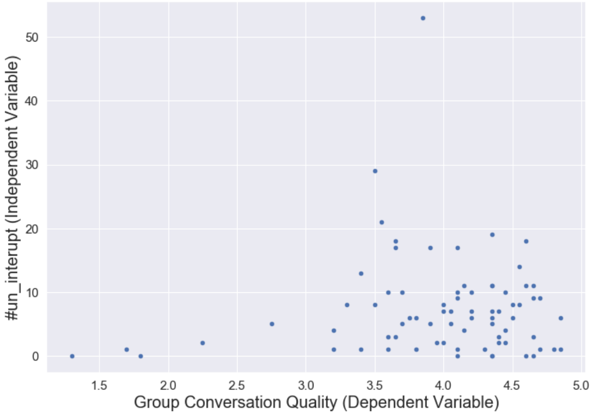
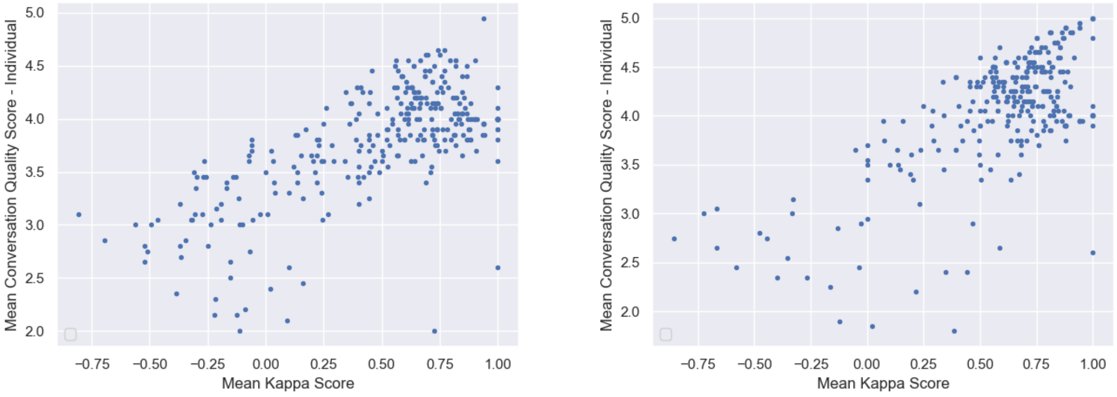


Figure B.8: Scatter plot revealing the relationship between Number of Unsuccessful Interruptions (Independent Variable) and the Group Conversation Quality (Dependent Variable).

C

Appendix C: Pairwise Mean Conversation Quality Vs Mean Kappa score

C.1. Analysis on Perceived Individual's Experience of Conversation Quality (IndivCQ)



(a) Individual-level Conversation Quality: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses between Annotator 1 and Annotator 2.

(b) Individual-level Conversation Quality: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses between Annotator 2 and Annotator 3.

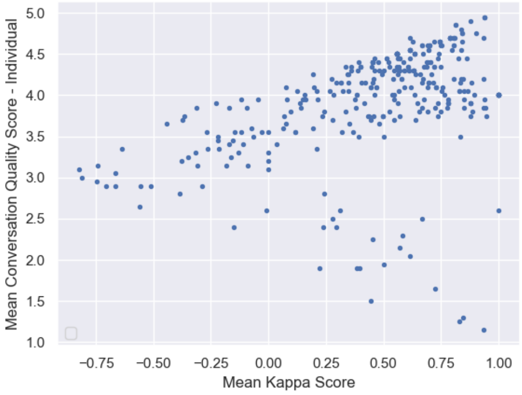
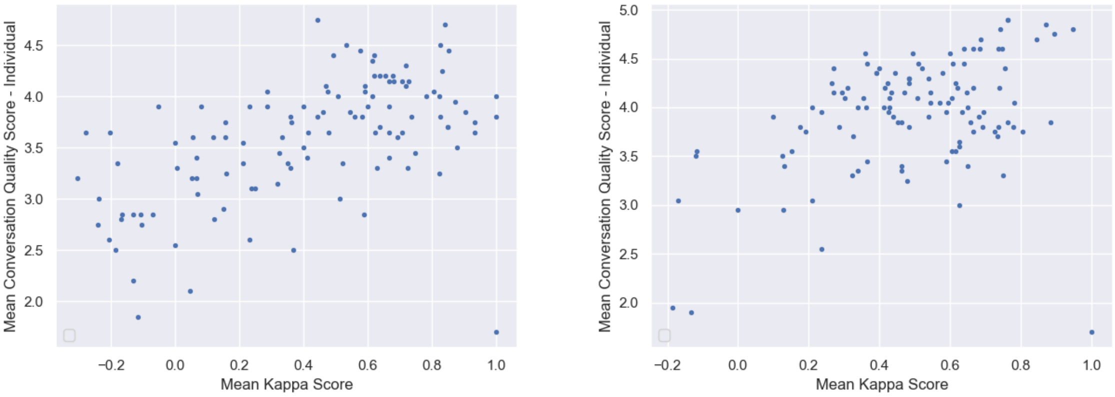


Figure C.2: Individual-level Conversation Quality: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses between Annotator 1 and Annotator 3.

C.2. Analysis on Perceived Group's Conversation Quality (GroupCQ)



(a) Group-level Conversation Quality: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses between Annotator 1 and Annotator 2.

(b) Group-level Conversation Quality: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses between Annotator 2 and Annotator 3.

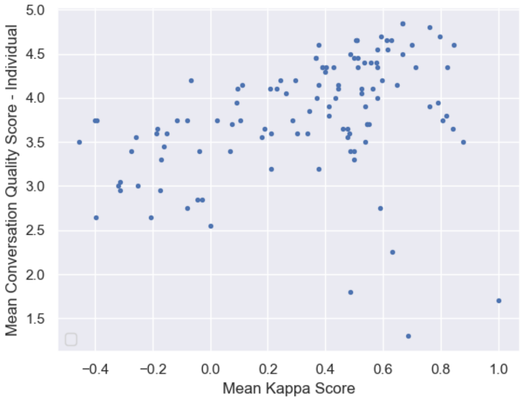


Figure C.4: Group-level Conversation Quality: Mean Conversation Quality score Vs Mean Kappa score - Inter-annotator agreeability of annotation responses between Annotator 1 and Annotator 3.

D

Appendix D: Joint LASSO and Rank Correlation Results

D.1. Analysis on Perceived Individual's Experience of Conversation Quality (IndivCQ)

Table D.1: LASSO Correlation (One model for all features): **All the features extracted** (i.e. Group Cardinality, Turn-Taking and Bodily Coordination feature, Section-5.3) as the independent variable and IndivCQ as the dependent variable. For the full form of the features listed in the table read Table-D.9.

	Feature Name	Correlation Coefficient
0	group_cardinality	-0.0796
1	tt-conv_eq	0.17731
2	tt-%silence	-0.02549
3	tt-#back_channels	0.0
4	tt-%overlap	-0.06046
5	tt-#suc_interupt	-0.05971
6	tt-#un_interupt	0.08032
7	mag-corr–min	0.05409
8	mag-min_lagcorr–min	0.03869
9	mag-max_lagcorr–min	-0.0
10	mag-argmin_lagcorr–min	0.0
11	mag-argmax_lagcorr–min	0.0
12	mag-mi–min	-0.0
13	mag-corr–max	0.07232
14	mag-min_lagcorr–max	-0.0
15	mag-max_lagcorr–max	-0.0
16	mag-argmin_lagcorr–max	-0.0
17	mag-argmax_lagcorr–max	0.0
18	mag-mi–max	-0.0
19	mag-corr–mean	0.0
20	mag-min_lagcorr–mean	-0.0
21	mag-max_lagcorr–mean	-0.05416
22	mag-argmin_lagcorr–mean	0.00214
23	mag-argmax_lagcorr–mean	0.01661
24	mag-mi–mean	-0.15075
25	mag-corr–var	-0.0
26	mag-min_lagcorr–var	0.0
27	mag-max_lagcorr–var	0.00078

Continued on next page

Table D.1 - Continued from previous page

	Feature Name	Correlation Coefficient
28	mag-argmin_lagcorr–var	0.0
29	mag-argmax_lagcorr–var	-0.0
30	mag-mi–var	-0.0
31	mag-min_lead_mimicry–min	0.0
32	mag-max_lead_mimicry–min	0.0
33	mag-mean_lead_mimicry–min	0.0
34	mag-var_lead_mimicry–min	-0.0
35	mag-min_lag_mimicry–min	0.0
36	mag-max_lag_mimicry–min	0.00854
37	mag-mean_lag_mimicry–min	-0.0
38	mag-var_lag_mimicry–min	-0.0
39	mag-min_lead_mimicry–max	0.0
40	mag-max_lead_mimicry–max	0.0
41	mag-mean_lead_mimicry–max	0.0
42	mag-var_lead_mimicry–max	-0.0
43	mag-min_lag_mimicry–max	0.0
44	mag-max_lag_mimicry–max	0.0
45	mag-mean_lag_mimicry–max	0.03724
46	mag-var_lag_mimicry–max	-0.0
47	mag-min_lead_mimicry–mean	0.01195
48	mag-max_lead_mimicry–mean	0.0
49	mag-mean_lead_mimicry–mean	-0.0
50	mag-var_lead_mimicry–mean	-0.0
51	mag-min_lag_mimicry–mean	0.0
52	mag-max_lag_mimicry–mean	0.0
53	mag-mean_lag_mimicry–mean	0.0
54	mag-var_lag_mimicry–mean	-0.0
55	mag-min_lead_mimicry–var	0.0
56	mag-max_lead_mimicry–var	0.0
57	mag-mean_lead_mimicry–var	0.0
58	mag-var_lead_mimicry–var	0.0
59	mag-min_lag_mimicry–var	0.0
60	mag-max_lag_mimicry–var	0.0
61	mag-mean_lag_mimicry–var	0.01801
62	mag-var_lag_mimicry–var	0.0
63	mag-symconv–min	-0.0
64	mag-lead_asymconv–min	0.0
65	mag-lag_asymconv–min	-0.0
66	mag-globconv–min	-0.05441
67	mag-symconv–max	-0.09201
68	mag-lead_asymconv–max	0.0
69	mag-lag_asymconv–max	0.00507
70	mag-globconv–max	0.0
71	mag-symconv–mean	-0.05818
72	mag-lead_asymconv–mean	0.01143
73	mag-lag_asymconv–mean	0.0
74	mag-globconv–mean	-0.0
75	mag-symconv–var	0.02805
76	mag-lead_asymconv–var	0.04025
77	mag-lag_asymconv–var	0.02604
78	mag-globconv–var	0.0
79	mag-min_coherence–min	0.0
80	mag-max_coherence–min	0.0

Continued on next page

Table D.1 - Continued from previous page

	Feature Name	Correlation Coefficient
81	mag-granger–min	0.07558
82	mag-min_coherence–max	0.0
83	mag-max_coherence–max	-0.0
84	mag-granger–max	-0.06227
85	mag-min_coherence–mean	0.0
86	mag-max_coherence–mean	-0.0
87	mag-granger–mean	-0.0
88	mag-min_coherence–var	0.0
89	mag-max_coherence–var	0.0
90	mag-granger–var	-0.0

	Feature Name	Spearman Correlation	p-value
0	**group_cardinality**	**-0.24363**	5e-05

Table D.2: Spearman Rank Correlation (Independent model for each feature): **Group Cardinality** as the independent variable and IndivCQ as the dependent variable. Highlighted features are the ones significant from the Joint LASSO test (Table-D.1).

	Feature Name	Spearman Correlation	p-value
0	**tt-conv_eq**	**0.28856**	0.0
1	**tt-%_silence**	**-0.22378**	0.00028
2	tt-#back_channels	0.03029	0.62754
3	**tt-%overlap**	**-0.1141**	0.06675
4	**tt-#suc_interupt**	**-0.00737**	0.90603
5	**tt-#un_interupt**	**0.14446**	0.02003

Table D.3: Spearman Rank Correlation (Independent model for each feature): Individual-level **Turn-Taking features** as the independent variable and IndivCQ as the dependent variable. Highlighted features are the ones significant from the Joint LASSO test (Table-D.1). For the full form of the features listed in the table read Table-D.9.

Table D.4: Spearman Rank Correlation (Independent model for each feature): **Bodily Coordination features** as the independent variable and IndivCQ as the dependent variable. Highlighted features are the ones significant from the Joint LASSO test (Table-D.1). For the full form of the features listed in the table read Table-D.9.

	Feature Name	Spearman Correlation	p-value
0	**mag-corr–min**	**0.15692**	0.03592
1	**mag-min_lagcorr–min**	**0.02515**	0.73825
2	mag-max_lagcorr–min	-0.04477	0.55182
3	mag-argmin_lagcorr–min	0.00059	0.9938
4	mag-argmax_lagcorr–min	0.18872	0.01141
5	mag-mi–min	-0.12928	0.08458
6	**mag-corr–max**	**-0.09091**	0.22619
7	mag-min_lagcorr–max	-0.11267	0.13319
8	mag-max_lagcorr–max	-0.14076	0.06019
9	**mag-argmin_lagcorr–max**	**-0.03962**	0.5985
10	mag-argmax_lagcorr–max	0.17695	0.01781
11	**mag-mi–max**	**-0.3407**	0.0
12	mag-corr–mean	0.03703	0.62259
13	mag-min_lagcorr–mean	-0.06488	0.38818
14	**mag-max_lagcorr–mean**	**-0.11298**	0.13211
15	**mag-argmin_lagcorr–mean**	**-0.02461**	0.74371
16	**mag-argmax_lagcorr–mean**	**0.18065**	0.01552
17	mag-mi–mean	-0.30212	4e-05

Continued on next page

Table D.4 - Continued from previous page

	Feature Name	Spearman Correlation	p-value
18	mag-corr–var	-0.20133	0.00688
19	mag-min_lagcorr–var	-0.17997	0.01593
20	mag-max_lagcorr–var	-0.12025	0.10885
21	**mag-argmin_lagcorr–var**	**-0.08692**	0.24732
22	mag-argmax_lagcorr–var	-0.10729	0.15285
23	mag-mi–var	-0.26053	0.00043
24	mag-symconv–min	0.00885	0.09069
25	mag-lead_asymconv–min	0.09229	0.21917
26	mag-lag_asymconv–min	0.07497	0.31858
27	**mag-globconv–min**	**0.16752**	0.025
28	**mag-symconv–max**	**-0.2088**	0.00503
29	mag-lead_asymconv–max	-0.05894	0.43324
30	**mag-lag_asymconv–max**	**0.03987**	0.59615
31	mag-globconv–max	0.03338	0.6573
32	**mag-symconv–mean**	**-0.12756**	0.08882
33	**mag-lead_asymconv–mean**	**0.01554**	0.83647
34	mag-lag_asymconv–mean	0.06738	0.37017
35	mag-globconv–mean	0.10985	0.14322
36	**mag-symconv–var**	**-0.18223**	0.01463
37	**mag-lead_asymconv–var**	**-0.15829**	0.03432
38	**mag-lag_asymconv–var**	**-0.09583**	0.20193
39	mag-globconv–var	-0.08492	0.2584
40	mag-min_coherence–min	-0.04322	0.56566
41	mag-max_coherence–min	0.02537	0.73603
42	**mag-granger–min**	**0.06138**	0.41439
43	mag-min_coherence–max	-0.1279	0.08798
44	mag-max_coherence–max	0.04705	0.0517
45	**mag-granger–max**	**-0.02881**	0.70187
46	mag-min_coherence–mean	-0.12083	0.10715
47	mag-max_coherence–mean	0.0093	0.90165
48	mag-granger–mean	0.02128	0.77735
49	mag-min_coherence–var	-0.09577	0.2022
50	mag-max_coherence–var	-0.11263	0.13334
51	mag-granger–var	-0.13775	0.06595
52	mag-min_lag_mimicry–min	0.19733	0.0081
53	**mag-max_lag_mimicry–min**	**-0.26587**	0.00032
54	mag-mean_lag_mimicry–min	0.16299	0.02926
55	mag-var_lag_mimicry–min	0.26784	0.00029
56	mag-min_lag_mimicry–max	0.03394	0.65195
57	mag-max_lag_mimicry–max	0.12887	0.08557
58	**mag-mean_lag_mimicry–max**	**-0.02381**	0.75168
59	mag-var_lag_mimicry–max	0.00527	0.94416
60	mag-min_lag_mimicry–mean	0.1034	0.16837
61	mag-max_lag_mimicry–mean	0.21197	0.00439
62	mag-mean_lag_mimicry–mean	0.007	0.92592
63	mag-var_lag_mimicry–mean	0.10909	0.14605
64	mag-min_lag_mimicry–var	-0.10139	0.17685
65	mag-max_lag_mimicry–var	-0.15566	0.03746
66	**mag-mean_lag_mimicry–var**	**-0.09348**	0.21325
67	mag-var_lag_mimicry–var	-0.1239	0.09845
68	mag-min_lead_mimicry–min	0.20545	0.0058
69	mag-max_lead_mimicry–min	0.25661	0.00053
70	mag-mean_lead_mimicry–min	0.14945	0.04586

Continued on next page

Table D.4 - Continued from previous page

	Feature Name	Spearman Correlation	p-value
71	mag-var_lead_mimicry–min	0.26983	0.00026
72	mag-min_lead_mimicry–max	0.03131	0.67735
73	mag-max_lead_mimicry–max	0.15546	0.03772
74	mag-mean_lead_mimicry–max	-0.04682	0.5337
75	mag-var_lead_mimicry–max	0.00911	0.90368
76	**mag-min_lead_mimicry–mean**	**0.10773**	0.15118
77	mag-max_lead_mimicry–mean	0.22662	0.00228
78	mag-mean_lead_mimicry–mean	-0.00628	0.9335
79	mag-var_lead_mimicry–mean	0.11193	0.13579
80	mag-min_lead_mimicry–var	-0.10695	0.15417
81	mag-max_lead_mimicry–var	-0.13694	0.06756
82	mag-mean_lead_mimicry–var	-0.11145	0.13748
83	mag-var_lead_mimicry–var	-0.12849	0.08651

D.2. Analysis on Perceived Group's Conversation Quality (GroupCQ)

Table D.5: LASSO Correlation (One model for all features): **All the features extracted** (i.e. Group Cardinality, Turn-Taking and Bodily Coordination feature, Section-5.3) as the independent variable and GroupCQ as the dependent variable. For the full form of the features listed in the table read Table-D.9.

	Feature Name	Correlation Coefficient
0	group_cardinality	-0.16155
1	tt-conv_eq	0.06478
2	tt-%silence	-0.20162
3	tt-#back_channels	0.0189
4	tt-%overlap	0.0
5	tt-#suc_interupt	0.0
6	tt-#un_interupt	0.0
7	mag-corr–min	0.0
8	mag-min_lagcorr–min	0.21518
9	mag-max_lagcorr–min	-0.0
10	mag-argmin_lagcorr–min	0.0
11	mag-argmax_lagcorr–min	0.0
12	mag-mi–min	-0.0
13	mag-corr–max	0.0
14	mag-min_lagcorr–max	0.0
15	mag-max_lagcorr–max	0.0
16	mag-argmin_lagcorr–max	0.0
17	mag-argmax_lagcorr–max	0.0
18	mag-mi–max	0.0
19	mag-corr–mean	0.07573
20	mag-min_lagcorr–mean	0.0
21	mag-max_lagcorr–mean	-0.0
22	mag-argmin_lagcorr–mean	0.0
23	mag-argmax_lagcorr–mean	0.0
24	mag-mi–mean	-0.29478
25	mag-corr–var	-0.0
26	mag-min_lagcorr–var	0.07114
27	mag-max_lagcorr–var	0.04005
28	mag-argmin_lagcorr–var	-0.05375
29	mag-argmax_lagcorr–var	0.0
30	mag-mi–var	0.07855
31	mag-min_lead_mimicry–min	0.13314

Continued on next page

Table D.5 - Continued from previous page

	Feature Name	Correlation Coefficient
32	mag-max_lead_mimicry–min	0.05444
33	mag-mean_lead_mimicry–min	0.0
34	mag-var_lead_mimicry–min	-0.07414
35	mag-min_lag_mimicry–min	0.02784
36	mag-max_lag_mimicry–min	0.02767
37	mag-mean_lag_mimicry–min	0.0
38	mag-var_lag_mimicry–min	-0.03532
39	mag-min_lead_mimicry–max	0.0
40	mag-max_lead_mimicry–max	0.0
41	mag-mean_lead_mimicry–max	0.0
42	mag-var_lead_mimicry–max	-0.0
43	mag-min_lag_mimicry–max	0.0
44	mag-max_lag_mimicry–max	0.0
45	mag-mean_lag_mimicry–max	0.0
46	mag-var_lag_mimicry–max	-0.0
47	mag-min_lead_mimicry–mean	0.0
48	mag-max_lead_mimicry–mean	0.12599
49	mag-mean_lead_mimicry–mean	0.04637
50	mag-var_lead_mimicry–mean	-0.0
51	mag-min_lag_mimicry–mean	0.0
52	mag-max_lag_mimicry–mean	0.00229
53	mag-mean_lag_mimicry–mean	0.00012
54	mag-var_lag_mimicry–mean	-0.0
55	mag-min_lead_mimicry–var	-0.17399
56	mag-max_lead_mimicry–var	0.0
57	mag-mean_lead_mimicry–var	-0.03462
58	mag-var_lead_mimicry–var	0.0
59	mag-min_lag_mimicry–var	-0.00019
60	mag-max_lag_mimicry–var	0.0
61	mag-mean_lag_mimicry–var	-0.00025
62	mag-var_lag_mimicry–var	0.0
63	mag-symconv–min	-0.03703
64	mag-lead_asymconv–min	0.0
65	mag-lag_asymconv–min	0.0
66	mag-globconv–min	0.00338
67	mag-symconv–max	0.0
68	mag-lead_asymconv–max	0.16465
69	mag-lag_asymconv–max	0.0
70	mag-globconv–max	0.0
71	mag-symconv–mean	-0.0
72	mag-lead_asymconv–mean	-0.0
73	mag-lag_asymconv–mean	-0.0
74	mag-globconv–mean	0.0
75	mag-symconv–var	0.0
76	mag-lead_asymconv–var	-0.01023
77	mag-lag_asymconv–var	-0.0
78	mag-globconv–var	-0.0
79	mag-min_coherence–min	0.0
80	mag-max_coherence–min	-0.04336
81	mag-granger–min	-0.01411
82	mag-min_coherence–max	-0.0
83	mag-max_coherence–max	-0.0
84	mag-granger–max	-0.0

Continued on next page

Table D.5 - Continued from previous page

	Feature Name	Correlation Coefficient
85	mag-min_coherence–mean	-0.09974
86	mag-max_coherence–mean	-0.12309
87	mag-granger–mean	0.17962
88	mag-min_coherence–var	-0.06206
89	mag-max_coherence–var	0.0
90	mag-granger–var	-0.03295

	Feature Name	Spearman Correlation	p-value
0	**group_cardinality**	**-0.35229**	0.00208

Table D.6: Spearman Rank Correlation (Independent model for each feature): **Group Cardinality** as the independent variable and GroupCQ as the dependent variable. Highlighted features are the ones significant from the Joint LASSO test (Table-D.5).

	Feature Name	Spearman Correlation	p-value
0	**tt-conv_eq**	**0.10158**	0.38912
1	**tt-%silence**	**-0.06321**	0.59263
2	**tt-#back_channels**	**0.08746**	0.4587
3	tt-%overlap	-0.15949	0.17468
4	tt-#suc_interupt	-0.0911	0.93863
5	tt-#un_interupt	0.0205	0.86238

Table D.7: Spearman Rank Correlation (Independent model for each feature): Individual-level **Turn-Taking features** as the independent variable and GroupCQ as the dependent variable. Highlighted features are the ones significant from the Joint LASSO test (Table-D.5). For the full form of the features listed in the table read Table-D.9.

Table D.8: Spearman Rank Correlation (Independent model for each feature): **Bodily Coordination features** as the independent variable and GroupCQ as the dependent variable. Highlighted features are the ones significant from the Joint LASSO test (Table-D.5). For the full form of the features listed in the table read Table-D.9.

	Feature Name	Spearman Correlation	p-value
0	mag-corr–min	0.17794	0.18143
1	**mag-min_lagcorr–min**	**0.2168**	0.10211
2	mag-max_lagcorr–min	0.08483	0.52665
3	mag-argmin_lagcorr–min		
4	mag-argmax_lagcorr–min	0.25622	0.05221
5	mag-mi–min	-0.20045	0.13137
6	mag-corr–max	0.1603	0.02935
7	mag-min_lagcorr–max	-0.07606	0.57042
8	mag-max_lagcorr–max	-0.14226	0.28677
9	mag-argmin_lagcorr–max	0.22012	0.09684
10	mag-argmax_lagcorr–max	0.14368	0.28194
11	mag-mi–max	-0.50119	6e-05
12	**mag-corr–mean**	**0.02057**	0.8782
13	mag-min_lagcorr–mean	0.04188	0.75495
14	mag-max_lagcorr–mean	-0.04625	0.73029
15	mag-argmin_lagcorr–mean	0.22012	0.09684
16	mag-argmax_lagcorr–mean	0.22012	0.09684
17	**mag-mi–mean**	**-0.40306**	0.00171
18	mag-corr–var	-0.23805	0.07195
19	**mag-min_lagcorr–var**	**-0.28439**	0.0305
20	**mag-max_lagcorr–var**	**-0.2202**	0.09673
21	**mag-argmin_lagcorr–var**	**-0.22028**	0.09661

Continued on next page

Table D.8 - Continued from previous page

	Feature Name	Spearman Correlation	p-value
22	mag-argmax_lagcorr–var	-0.11212	0.40208
23	**mag-mi–var**	**-0.27232**	0.03864
24	**mag-symconv–min**	**0.03501**	0.79417
25	mag-lead_asymconv–min	0.1854	0.16353
26	mag-lag_asymconv–min	0.1854	0.16353
27	**mag-globconv–min**	**0.30425**	0.02023
28	mag-symconv–max	-0.30718	0.019
29	**mag-lead_asymconv–max**	**-0.04024**	0.76422
30	mag-lag_asymconv–max	-0.04024	0.76422
31	mag-globconv–max	0.07738	0.56371
32	mag-symconv–mean	-0.17443	0.19033
33	mag-lead_asymconv–mean	0.02728	0.83891
34	mag-lag_asymconv–mean	0.02728	0.83891
35	mag-globconv–mean	0.23081	0.0813
36	mag-symconv–var	-0.27216	0.03876
37	**mag-lead_asymconv–var**	**-0.09342**	0.48548
38	mag-lag_asymconv–var	-0.09342	0.48548
39	mag-globconv–var	-0.23694	0.07332
40	mag-min_coherence–min	-0.03018	0.82209
41	mag-max_coherence–min	0.13191	0.32362
42	**mag-granger–min**	**0.2649**	0.04448
43	mag-min_coherence–max	-0.27894	0.03397
44	mag-max_coherence–max	-0.12501	0.34978
45	mag-granger–max	-0.00924	0.94513
46	**mag-min_coherence–mean**	**-0.22275**	0.09283
47	**mag-max_coherence–mean**	**0.0376**	0.77933
48	**mag-granger–mean**	**0.16156**	0.22566
49	**mag-min_coherence–var**	**-0.24607**	0.0626
50	mag-max_coherence–var	-0.24473	0.0641
51	**mag-granger–var**	**-0.0707**	0.59795
52	**mag-min_lag_mimicry–min**	**0.32436**	0.01299
53	**mag-max_lag_mimicry–min**	**0.32932**	0.0116
54	mag-mean_lag_mimicry–min	0.1005	0.45286
55	**mag-var_lag_mimicry–min**	**0.28057**	0.0329
56	mag-min_lag_mimicry–max	-0.03473	0.79576
57	mag-max_lag_mimicry–max	0.27312	0.03805
58	mag-mean_lag_mimicry–max	-0.19768	0.1369
59	mag-var_lag_mimicry–max	-0.03587	0.78921
60	mag-min_lag_mimicry–mean	0.07017	0.60067
61	**mag-max_lag_mimicry–mean**	**0.33778**	0.00951
62	**mag-mean_lag_mimicry–mean**	**-0.1712**	0.19881
63	mag-var_lag_mimicry–mean	0.05967	0.65634
64	**mag-min_lag_mimicry–var**	**-0.18453**	0.16553
65	mag-max_lag_mimicry–var	-0.12063	0.36708
66	**mag-mean_lag_mimicry–var**	**-0.13817**	0.30099
67	mag-var_lag_mimicry–var	-0.21757	0.10087
68	**mag-min_lead_mimicry–min**	**0.32436**	0.01299
69	**mag-max_lead_mimicry–min**	**0.32932**	0.0116
70	mag-mean_lead_mimicry–min	0.09835	0.46266
71	**mag-var_lead_mimicry–min**	**0.28057**	0.0329
72	mag-min_lead_mimicry–max	-0.03473	0.79576
73	mag-max_lead_mimicry–max	0.27312	0.03805
74	mag-mean_lead_mimicry–max	-0.20245	0.12748

Continued on next page

Table D.8 - Continued from previous page

	Feature Name	Spearman Correlation	p-value
75	mag-var_lead_mimicry–max	-0.03587	0.78921
76	mag-min_lead_mimicry–mean	0.07017	0.60067
77	**mag-max_lead_mimicry–mean**	**0.33778**	0.00951
78	**mag-mean_lead_mimicry–mean**	**-0.17391**	0.19169
79	mag-var_lead_mimicry–mean	0.05967	0.65634
80	**mag-min_lead_mimicry–var**	**-0.18453**	0.16553
81	mag-max_lead_mimicry–var	-0.12063	0.36708
82	**mag-mean_lead_mimicry–var**	**-0.13994**	0.29477
83	mag-var_lead_mimicry–var	-0.21757	0.10087

	Abbreviation	Full Form
1	tt	Turn-Taking
2	conv	Conversation
3	eq	Equality
4	%silence	Percentage of Silence
5	#back_channels	Number of Back-channels
6	%overlap	Percentage of Overlap
7	mag	Magnitude
8	mi	Mutual Information
9	min	Minimum
10	max	Maximum
11	corr	Correlation
12	lagcorr	Lagged correaltion
13	argmin	Index of Minimum
14	argmax	Index of Maximum
15	mean	Mean
16	var	Variance
17	mimicry	Mimicry
18	lead	Lead (in asymmetric features)
19	lag	Lagged (in asymmetric features)
20	symconv	Symmetric Convergence
21	asymconv	Asymmetric Convergence
22	globconv	Global Convergence
23	coherence	Coherence Causality
24	granger	Granger's Causality

Table D.9: Legend: Key to read abbreviations of feature names, seen in the results tables. The bodily coordination based features are named as abbreviations of the form **CH-BFV-GA**, where *CH* is the **Channel Name** (e.g. x(xaxis), y(yaxis) or mag(magnitude of tri-axial data)), *BFV* is the **Behavioural Feature Variant** (e.g. correlation, argmin of lagged correlation, minimum of lead mimicry etc..,) and *GA* is the **Group-level Aggregation** features (e.g. min, max, mean, variance).