# REAL-TIME HUMAN/GROUP INTERACTION MONITORING PLATFORM INTEGRATING SENSOR FUSION AND MACHINE LEARNING APPROACHES

By

Sylmarie Dávila-Montero

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Electrical and Computer Engineering - Doctor of Philosophy

2022

## ABSTRACT

A person's social intelligence impacts their physical and mental health, and the productivity levels of the individuals involved, for example, in workplace interactions. To promote successful social interactions, this dissertation explores the use of sensor technology and machine learning algorithms to monitor and quantify nonverbal behavior indicators in real time. This dissertation conducts extensive convergence research between psychology, communication science, and engineering and establishes a new real-time human/group interaction monitoring platform. From sensor selection to data collection and algorithm design, existing human behavior monitoring systems vary widely in the type of methods employed for their design. Many of these systems were trained with data collected in controlled environments, making them not practical for real-life scenarios. Moreover, existing systems lack the capabilities needed to recognize behaviors in a manner that could support machine-augmented social intelligence. To address these issues, the developed human/group interaction monitoring platform combines a real-time enabled multi-sensor system with a machine learning framework that establishes training and algorithm design methods for behavior recognition. Methods for the execution of human studies, collection of natural human behavior data, and data annotation procedures were also established to train machines to recognize human behaviors impacting the quality of social interactions. The contributions of this dissertation, which can be universally applied to other behavior studies, will advance the design of human behavior monitoring systems for group interactions and facilitate future real-time feedback to increase self-awareness and promote successful social interactions.

To my dearest family, both on earth and in heaven…

## ACKNOWLEDGEMENTS

I would like to thank Michigan State University Graduate School, the National Science Foundation, the GEM Consortium, and the SLOAN program for their financial support.

I also thank my supervisory committee: Dr. Selin Aviyente, Dr. Erin Purcell, Dr. Angela Hall, Dr. Gary Bente, and Dr. Andrew J. Mason for their support, guidance, and ideas throughout this process. Special thanks to Dr. Bente for his continuous advice on the psychophysiological aspects of this work and the social aspects of experimental design; and Dr. Hall for her economic support to perform human studies and provide assistance with experimental design and human resources to help me with data labeling. My immense gratitude goes to my supervisory committee chair and advisor, Dr. Andrew J. Mason, for giving me the opportunity to work with him, first as an undergraduate summer research intern and then as part of his research group as I worked towards my doctoral degree. It has been an honor to learn from him and grow as a professional under his guidance. I am grateful for his support, advice, mentorship, trust, and patience.

I would also like to thank the MSU SLOAN and the MSU AGEP communities. More specifically, I would like to thank Dr. Percy Pierre, Dr. Nelson Sepulveda, and Steven Thomas for their academic guidance and financial support during this process. To Dr. Nelson Sepulveda, I will be forever grateful for his constant support and advice.

Thanks to my lab mates, past and current, for their constant feedback, collaboration, and support during this process. Special thanks to my friends, here in Michigan and far away. Keilyn Vale, Yeilyn Vale, Alex Román, Dr. Adrian Ildefonso, Dr. Lisaura Maldonado, and my writing group (Mara Cuebas, Lizbeth Dávila, Nichole Montero, and Gabriela Ortiz), thanks for all the love and non-stop support. Very special thanks to Dr. Keisha Castillo-Torres for your unconditional

support, love, feedback, and motivation every step of the way, even when miles away. I could not have asked for a better friend and colleague for this journey. We did it!

Thanks to my family. To my parents, Marilyn Montero-Caro and José L. Dávila-Estrada, from which I learned the value of education and hard work, thanks for your unconditional love, support, visits to help me cope with stressful moments, and words of wisdom. Thanks to my sisters, Bianca, Genna, and Josnaly, and my cousin Eidy Montero for their constant support and admiration, which has motivated me to be the best version of myself. I would also like to thank my grandparents whose support has meant the world. To Miriam Montero, thanks for taking the time to learn about my research, your interest and feedback in my work have helped me become a better communicator. To my life partner, Ajay Delarosa, thanks for supporting me in all the ways that you did during my writing process. Your words of admiration, motivation, and respect were fuel to finish my work. I am blessed to have you by my side.

Thanks to my previous mentors for motivating me to pursue this adventure: Baldomero Llorens, Dr. Domingo Rodriguez, Dr. Nestor Rodriguez, and Dr. Nayda Santiago. Special thanks to Dr. Santiago for always believing in me and being one of the best role models any women engineer could ask for.

Thanks to everyone that has not been mentioned but that has walked a piece of this adventure with me by providing me feedback, editing my essays and papers, and/or supporting me in any possible way. But most importantly, my biggest gratitude goes to God because his presence in my life has kept me focused on my purpose: to use my talents to contribute positively to the life of those I get to cross paths with.

## TABLE OF CONTENTS

# 1. INTRODUCTION

## 1.1.Social Awareness: Challenges and Opportunities

Teamwork and social interactions are at the core of new discoveries, product developments, and general successful organizational outcomes. In the U.S. alone, 55M team meetings are estimated to be carried out per day, costing organizations ~$1.4T/year [1]. However, of those, ~$250B/year is wasted on team meetings that have poor outcomes [1]. Research has shown that the quality of social interactions within a group can either foster team effectiveness where individuals work well together or can encourage teams to fall apart [2]. Therefore, ineffective social interactions are one of the principal causes of poor team productivity.

Social interactions are constructed and influenced by the human behaviors of two or more individuals. An aspect of human behaviors that can degrade the quality of social interactions is unconscious biases. Throughout a team meeting, unconscious biases can cause behavioral events such as interruptions or ostracization (exclusion or the act of ignoring) towards another team member. To a certain extent, we have grown accustomed to these types of social behaviors and the biases that cause them. However, they can have a negative impact on the individuals experiencing these behavioral events eventually affecting, not just the outcomes of an organization but also, the individuals' health. In fact, the quantity and quality of social interactions influence a range of health conditions including cardiovascular diseases, compromised immunity, and depression [3]–[5].

Unconscious biases, also known as implicit biases, are defined as social stereotypes or attitudes held subconsciously about certain groups of people that affect the way individuals behave around them. Unconscious biases are more prevalent than conscious prejudice, which is bias people know they have and intentionally act upon. The actions resulting from unconscious biases

can lead to microinequity or microaggressions [6]. Microinequity refers to demeaning or marginalizing someone, whereas microaggression is the act of expressing prejudice against a marginalized group or person. Many of our behaviors, including unconscious bias behaviors, are motivated by unconscious triggers and emotions [7]. Hence, research has suggested that unconscious biases can be prevented by increasing our social awareness, which includes being self-aware of our emotions, intentions, and ways in which we communicate with each other.

The simplest way of assuring effective social interactions is by individuals becoming more self-aware of their behaviors and environmental stimuli, i.e., by improving their situational awareness. Our perception and awareness of our behaviors and the behaviors of others play an important role in our daily lives and the quality of social interactions [7], [8]. However, it is known that as humans, our awareness and perception of our environment and the behaviors of others and ourselves can be limited by a variety of factors. For instance, various studies have shown that even when we are able to perceive intentions or environmental stimuli, we may not always be processing them in our conscious mind, making us unaware of the event [9]. In fact, psychologists and social and behavioral scientists agree in that much of what we do on a daily basis is unconscious [10], [11]. Thus, a step toward improving an individual's social awareness is to apply technology to study and monitor in real time their human behaviors and the behaviors of those around them.

### 1.2.Social Behavior Monitoring Technologies: Challenges and Opportunities

Modern sensor technologies have permitted the objective assessment of behaviors that influence the well-being of humans, such as physical activity [12], sleep patterns [13], stress levels [14], food intake patterns [15], and social interactions [16]–[20]. As the understanding increases of how social interactions influence human well-being and productivity, a variety of sensor technologies and computational methods have been applied for the study and recognition of human

behaviors that influence the quality of group interactions, both for in-person and virtual interaction environments.

In general, rudimentary technologies exist for the monitoring of individual-level behaviors and aspects of group-level behaviors [21]. Individual-level monitoring technologies focus on the recognition of emotions; their purpose is to increase emotional awareness to enhance aspects of inter-personal attraction, physical presence, and social presence [22]–[24]. On the other hand, group-level monitoring technologies focus on the recognition of conversation dynamics and attention with the end-goal of increasing balance participation and improving collaboration and group performance [25], [26]. Still, the real-time monitoring of complex social interaction dynamics, such as unconscious biases, that have a major impact on the well-being of humans, requires the integration of both individual-level and group-level behaviors. In addition, most of the existing technologies lack real-time capabilities and system features, such as a feedback framework or mechanism, that will permit the enhancement of social interactions in real time. Furthermore, many such systems are not configurable for the monitoring of complex human behaviors that could lead to identifying complex group dynamics and lack of awareness. In order to achieve a combination of individual-level and group-level behavior recognition, and create a system that will allow real-time feedback, a platform capable of collecting data from natural human interactions and processing in real time behavioral cues from multiple individuals is necessary.

**Figure 1** illustrates the concept of real-time group behavior monitoring technologies improving awareness during social interactions. Here, the technology captures human behavior data during interactions and provides informative real-time feedback to everyone regarding the social ecosystem to improve each user's awareness of individual, dyadic, and group behaviors. Feedback messages could relate to conversation dynamics (e.g., who is dominating the

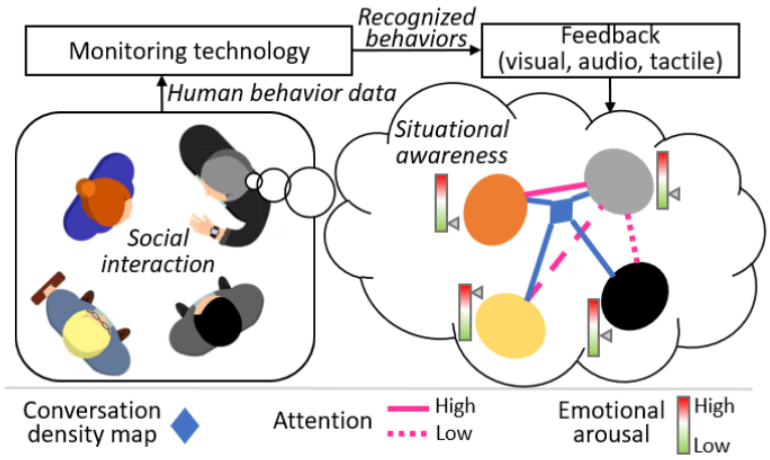


***Figure 1****. Technology to monitor individual and group behaviors during a social interaction can measure aspects of conversation dynamics, levels of attention, and levels of emotional arousal. Being aware of our behaviors and the behaviors of others has been shown to help improve social interactions, positively impacting individuals' wellbeing, organizational outcomes, etc. © 2021, IEEE.*

conversation and number of interruptions), levels of attention, or even the levels of emotional arousal affecting group dynamics or that can be related to implicit bias behaviors. Still, the information that could be fed back to the individuals involved in the interaction greatly depends on the capabilities of the social behavior monitoring technologies.

### 1.3.Requirements and Challenges in the Design of Real-time Monitoring Technologies

The design and implementation of a group behavior monitoring system to improve social interactions face theoretical and technical challenges. Humans communicate consciously and unconsciously using multiple channels, i.e., gestures, movements, tone of voice, etc. Therefore, the effective monitoring of complex group interaction dynamics requires the recognition of human behaviors through various sensing modalities that allow the integration of communication patterns' information and emotional states. Furthermore, data from natural human interactions should be utilized to design computational models embedded in behavior monitoring systems. However, existing behavior monitoring systems lack sensing modalities, frameworks for data collection, or

computational capabilities needed to recognize in real time complex social dynamics, potentially because the design of these systems presents numerous challenges.

The challenges that need to be overcome to achieve a functional group behavior monitoring system to increase self-awareness and enhance social interactions can be defined as follows:

- Group interactions are complex and include a combination of individual behaviors and dyadic behaviors. Thus, monitoring group interactions using sensor technology requires the understanding of psychological and communication theories at the individual, dyadic, and group levels and their application to the engineering design of a system. Methods that could quantify the quality of a group interaction using technology are still an area to investigate and no well-established methods exist.
- Real-time monitoring of individual behaviors and group interactions requires coordination of hardware and software components to perform automatic synchronization and processing of data collected across individuals in the interaction. Challenges in this area include the combination of sensing modalities, data synchronization across modalities and sensor nodes, and the selection of optimal signal processing parameters and computational models for the recognition of individual human behaviors and group interactions.
- Effective recognition of behaviors requires computational models trained with data from natural human behavior interactions. Challenges in this area include the design of data processing modules with variations in data sources, lack of guidelines and/or infrastructure for data collection that informs methods for performing human studies, and management of data preparation/annotation.

### 1.4.Goals

The goal of this project is to establish a platform through which the challenges described in Section 1.3 can be solved to bridge the gap between psychology, communication science, and engineering. Such a platform will be bringing individual, dyadic, and group-level information to the design of group behavior monitoring systems. The achievement of this goal will allow the implementation of a multimodal system to monitor, in real time, non-verbal and physiological behavior indicators; it will also facilitate real-time feedback to promote successful social interactions by bringing awareness to our unconscious behaviors.

### 1.5.Outline

This thesis is organized as follows: Chapter 2 presents the psychological theories and concepts that underpin the analysis of social interactions and methods to monitor them; Chapter 3 presents the design of a framework for the data collection of nonverbal indicators of individual behavior and group interaction using sensor technology and a framework for the processing of collected data; Chapter 4 presents initial data collection studies, processes to prepare the collected data for future processing, and base signal characteristics and algorithms for the recognition of nonverbal indicators of human behavior found in speech and body motion signals; Chapter 5 presents the design and execution of a social interaction study and relationship of base nonverbal behaviors with reported rapport experienced, and Chapter 6 presents contributions of this dissertation and future work.

## 2. BACKGROUND

The multidisciplinary nature of the goals of this dissertation work spans from social sciences to engineering technology. Therefore, this chapter is designed to give a sense of the social science theories and concepts that have guided the psychological study of human behavior and social interactions. In addition, this chapter also describes the methods that have been employed to monitor human behaviors and the elements involved in using technology for such end.

### 2.1.Human Behavior and its Effectors

Humans are a highly social species expressing individual behaviors that, when accumulated, create the social environment in which society operates. In general, human behavior is driven by personal factors such as thoughts and emotions that are influenced by our environment and social interactions. Social interactions, constructed by the behaviors of two or more individuals, are highly complex and play an important role in our health and survival [3]. Behavior is generally defined as the "observable consequences of the choices a living entity makes in response to external or internal stimuli" [27]. Internal stimuli could be a person's thoughts, memories, perceptions, or attitudes, while external stimuli come from the environment the person interacts with, including social interactions. In humans, depending on the level of situational and personal awareness that they possess, responses to external and internal stimuli (effectors of human behaviors) can be voluntary or involuntary. **Figure 2** shows the dynamics of the effectors of human behavior, which can include personal factors and components of social interactions.

As illustrated in **Figure 2**, personal factors are inside the person. They can come from a person's biology or psychology. Personal factors that come from a person's psychology are in the mind and are not externally observable; however, the behaviors a person expresses because of the influences of their psychological personal factors are directly observable. Social behaviors, a

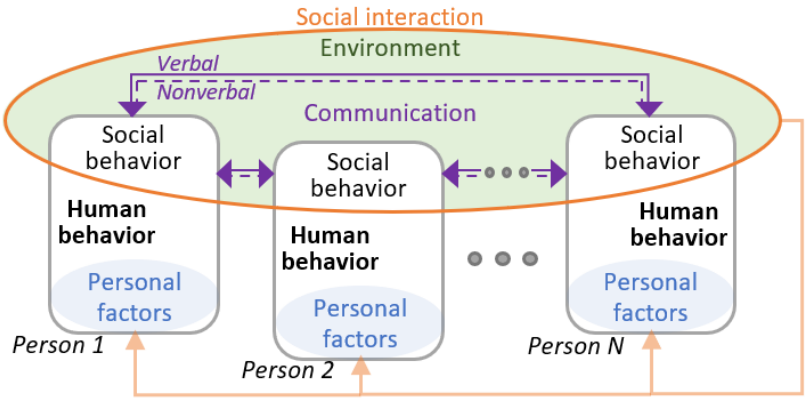


***Figure 2.*** *Diagram describing the effectors of human behavior and their dynamics. In short, given an environment, personal factors influence human behaviors, which influence our social behaviors affecting how we communicate during social interactions. In a reciprocal loop, the elements involved in social interactions influence back our personal factors, which influence our human behaviors and so on. © 2021, IEEE.*

subset of human behaviors that are specifically directed at other people or that involve social action, are directly observable. Communication, both verbal and nonverbal, is a vital aspect of social interactions and is directly observable. As illustrated in **Figure 2**, social behaviors strongly influence communication dynamics during a social interaction while, in a reciprocal loop, our social behaviors get influenced by our social interactions. Part of this idea is captured by the well-established *Social Cognitive Theory (SCT)*, which contends that individuals' perceptions of their environment can influence their emotional, physiological, and behavioral reactions [28], [29], subsequently influencing future behaviors in a reciprocal loop.

To properly understand the technology developed to monitor human behavior, one must first understand the personal factors that underpin human behaviors and the theories and concepts that have guided the psychological study of social interactions. These two topics are briefly summarized below to provide a scholarly foundation for the research described in this work.

## 2.2.Theories and Concepts of Human Behavior

### 2.2.1. Personal Factors

The psychological factors that have been commonly studied that contribute to human behavior are affect and dispositions. Affect generally means anything related to a person's emotions or moods, and it can be divided into two categories: states and traits. State affect is an emotion or mood that is experienced in a certain moment, whereas trait affect is a more enduring part of one's personality. Emotions are mental and physiological experiences of feeling that are acutely experienced (intense) and discrete in that they have a beginning and an end point, while moods refer to the positive or negative feelings that are in the background of our everyday experiences; these are diffuse (not acutely experienced) and longer-lasting states than emotions; however, they are not as enduring as trait affect. Trait affect is part of one's personality – it is a tendency to experience certain emotions and moods in general. For example, someone might have a negative affectivity trait, which is the tendency to experience negative moods and emotions more often than others. Together, these states of emotional experiences and traits constitute affect.

A disposition in the social sciences is thought of as a natural proclivity (biological or psychological) to respond to situations in a particular way. Because dispositions are "natural" and inherent in the person, they are thought to be the most stable and enduring phenomenon studied in the psychological sciences that are discussed in this chapter (i.e., more enduring across time than state affect, attitudes, and behaviors). However, despite their stability across time, dispositions do not relate to behavior with perfect consistency because there are environmental factors that also influence behavior. For example, a person might have a biological disposition to develop a psychological disorder, but through certain training environments like therapy, they are able to override their disposition to develop the disorder. For another example, someone may be

genetically predisposed to have a reserved personality, but they are put in social environments that constantly require them to talk to others, so they override their genetic predisposition. Dispositions influence human behavior more when the situation is weak, like when a person is in a casual social interaction. On the other hand, dispositions influence human behavior less when the social situation is strong and enforces certain norms, such as a professional environment in which all individuals are expected to behave in a certain way regardless of their personalities [30]. This Section will focus on personality traits, which are influenced by dispositions as well as by the environment [31].

Lastly, another important personal factor that impacts behavior is attitudes. An attitude is a psychological tendency to evaluate a particular target with some degree of favor or disfavor [32]. The "target" could be another person or a non-living thing such as a food, brand, or idea. An attitude, at its core, is an evaluation. Thus, it differs from affect and dispositions. Whereas affect could include an emotion that arises in response to a target, an attitude is a feeling towards the target and a set of judgments about the target. Attitudes are more enduring than a state but less enduring than a trait or disposition. **Figure 3** shows the relationship between these personal factors and time. The duration of these factors and the interactions between them play an important role in understanding how technology can be used to understand human behaviors.

It was contended that there is a lack of research using behavior monitoring technologies to study the role of attitudes in human behavior during social interactions. Thus, next, a review of psychological theories that delve specifically into the explanation of state affect and personality traits is presented.

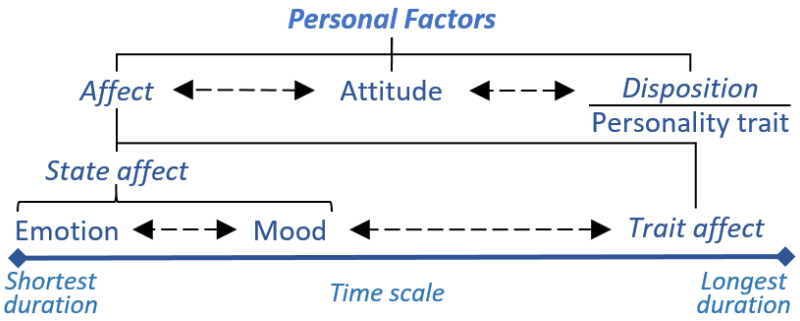


***Figure 3***. *Diagram describing the personal factors that influence human behavior and how they manifest through time. Personal factors include affect, attitudes, and dispositions, of which, affect and dispositions are the most studied. Affect is divided into states and traits. State affect is related to acute emotions and mood, in contrast to trait affect which is related to a human's disposition to experience positive or negative emotions and is a more enduring part of human personality. © 2021, IEEE.*

*2.2.1.1.State-Affect Related Theories*

Discrete, acute emotions often provoke a person to mentally narrow in on a specific action or set of actions. For example, the experience of fear leads to the activation of thoughts in the mind about defending oneself or running away (also known as "fight-or-flight" response), and the experience of interest can activate a person's thoughts aimed at exploring and taking in new information [33]. "Activations of thought" driven by emotions can occur subconsciously. Indeed, the body mobilizes physiological resources to complete these actions without the person's conscious awareness.

Based on the explored idea that emotions reflect responses of the sympathetic nervous system [34], the *Polyvagal Theory* explains how state affect alters brain processes and biological processes that occur in the rest of the body [35], [36]. In addition, this theory provides insights into the relationship between measurable physiological states, linked to the autonomic and central nervous systems, and the resulting human behavior, suggesting a bidirectional relationship between the brain and the body. It also suggests that the environment affects behaviors that consequently alter physiological states. Thus, monitoring changes in the physiological states of the human body, such

as respiration rate, heart rate, and perspiration rate, among others, can provide insights into the affective state of an individual [37]. Likewise, monitoring environmental conditions can provide information on how the environment influences emotional states and other factors.

Emotions can be-understood to fall somewhere along two orthogonal dimensions: (1) of how pleasurable the emotion is, (2) and of how much arousal or activation the emotion involves. As shown in **Figure 4**, emotions are commonly arranged in a *circumplex model of affect* [38], according to where they fall on both dimensions. For example, excitement is an emotion that is pleasurable and high on arousal, whereas calmness is an emotion that is pleasurable and low on arousal. Anger and fear are unpleasant, high on arousal emotions close together on the circumplex, whereas boredom is a low-arousal unpleasant emotion. The circumplex model of affect is a mainstream and well-established theory. However, other dimensional models of emotions have also been used to study emotional states, such as the *Pleasant, Arousal, and Dominance (PAD) emotional state model* [39] that, in addition to modeling emotions in a valence-arousal scale,

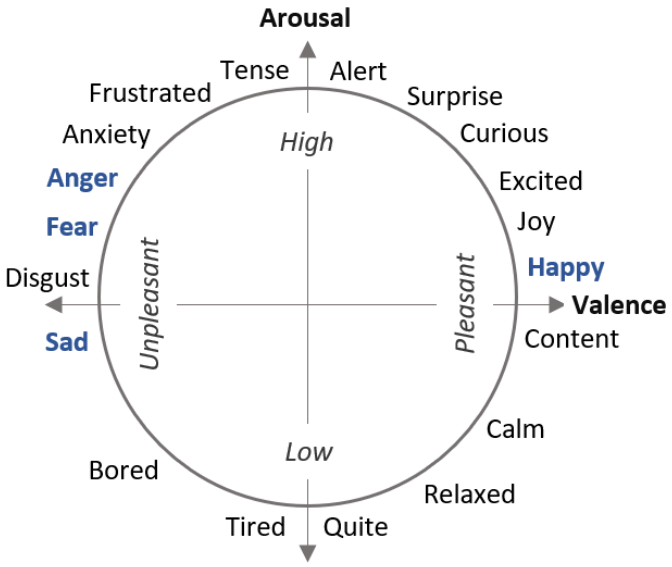


***Figure 4****. A typical circumplex model of affect that describes affective states using two fundamental neurophysiological systems: valence and arousal. Valence describes the level of pleasure or displeasure of an emotion, while arousal describes its level of activation. Emotions in blue color represent the four most commonly studied emotions in affective computing. © 2021, IEEE.*

contains a dominance dimension representing the controlling nature of an emotion. The *Plutchik's model* [40] is another dimensional model of emotions that organizes discrete emotions from the most basic to the most complex ones. In the field of affective computing, happiness, sadness, anger, and fear are the four most studied emotions.

A discrete emotion can affect someone's response to a social interaction whether the emotion was caused by that interaction or not. The well-established framework of *Emotions As Social Information (EASI) model* [41], [42] asserts that emotions serve a social function by relaying information when they are expressed. For example, if a person is late to a meeting with a coworker and the coworker appears to be angry, this provides information that leads to certain inferences, such as the inference that the person was late, the inference that the behavior of being tardy was inappropriate, and the inference that the person should strive to arrive earlier in the future [41]. The information that emotions relay to others in social interactions is valuable for adjusting future behavior.

*2.2.1.2.Personality Traits Related Theories*

The dominant theory in the organizational sciences used to taxonomize personality is the Five-Factor Model (FFM) of personality, also known as the "Big Five." The "Big Five" factors of personality can be abbreviated with the acronym "OCEAN": Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Openness involves being open to new experiences, unconventional, nonconforming, creative, and imaginative, while conscientiousness is the "tendency toward being dependable, disciplined, purposeful, organized, and achievement-oriented" [43]. Extraversion, in the sense of the FFM, is the "tendency to be social, talkative, energetic, and active" [43]. It was found that among the personality factors, extraversion has the strongest relationship with leadership (both being recognized as a leader by others and being

effective as a leader) [44]. On the other hand, agreeableness tends to be a catchall factor related to aspects of personality that are likable and harmonious with others, such as being trusting of others, polite, empathetic, and compliant [45]. The last one of the Big Five is Neuroticism, which is often labeled as its opposite instead, emotional stability. Those who are high on neuroticism are more likely to experience negative emotions like anxiety, anger, irritation, frustration, and jealousy. On the other hand, having low neuroticism or high emotional stability means that a person tends to be more even-keeled, calm, and unwavering (not necessarily positive or enthusiastic).

There are many other taxonomies of personality, such as the HEXACO model [45] which breaks the agreeableness factor of the FFM into agreeableness and humility. The Dark Triad is another taxonomy that has only three undesirable personality traits: narcissism, Machiavellianism, and psychopathy [46]. Another trait is *locus of control*, which describes the extent to which individuals believe that they control their own outcomes as opposed to having their successes and failures determined by external forces [47]. So far, the traits covered by the FFM and the *locus of control* have been studied using human behavior monitoring technologies. In general, the activation of these traits during social interactions contributes to observable social behaviors that make up the social environment that people operate in.

#### 2.2.2. Communication

One of the most important factors influencing our human behavior is the social behavior of our interaction partners. We might think of it as a situational factor, but this would ignore the dynamic nature of mutual adaption within the communication process. By definition, "communication is a transactional process in which people generate meaning through the exchange of verbal and nonverbal messages in specific contexts, influences by individual and societal forces and embedded in culture" [48]. Verbal communication refers to the use of spoken and written

language (words). It is usually organized in distinct on-off patterns of messages or utterances with iterating sender and receiver (speaker, listener) roles. The use of words requires a shared explicit code usually to be found in a dictionary. Spoken language, however, can also carry implicit, so-called paraverbal, information, for instance, encoded in the floor possession and pausing or in prosodic features such as pitch, speed, and volume of the vocal output.

Nonverbal communication comprises all aspects of communication that are not encoded into words. In stark contrast to verbal communication, nonverbal communication is continuous, i.e., always on, and largely implicit, i.e., it lacks a dictionary and is produced and processed widely automatically and unconsciously. Therefore, it is hard to control, and its effects impose on the observer with an irrefutable force. Even a lack of nonverbal expressions is interpreted by the observer, for instance as disinterest. In this sense, it has been said that "we cannot not communicate" [49]. It has been argued that our social perception and impression formation is much more dependent on nonverbal cues than on verbal behavior and that nonverbal communication can be conceived as meta-communicative [49], in the sense that it even largely defines how we understand and interpret the spoken words. Thus, even in the presence of verbal communication, successful communication largely depends on the efficient use of nonverbal communication channels [50].

As nonverbal communication largely withdraws deliberate manipulation, it is supposed to provide information about unobservable processes such as the individual's emotional state, intentions, personality traits, etc. [51]. The view that nonverbal communication is a reliable source of true information, although still under debate [52], has made it the focus of study of many areas dedicated to understanding social behavior and the human mind. For example, the social signal processing [53], [54] and affective computing [55] literature, areas of engineering and computer

science that study social interactions and human emotions, respectively, focus on the messages produced by nonverbal communication channels. They are better known in those areas as “social signals”; a notion that was first introduced in the field of computational social science and organization engineering [56]. Thus, here the focus is on reviewing the most used nonverbal communication channels.

Due to its unique level of complexity, the analysis of nonverbal communication poses considerable methodological challenges [57]. Nonverbal communication implies multiple channels and serves various functions. As illustrated in **Figure 5**, nonverbal communication includes gestures, body movements, and postures [58]–[61], facial expressions [62], [63], and eye gaze [60], among others. This work treats paraverbal communication, such as prosody, pitch, volume, and intonation [64], [65], under the broader construct of nonverbal communication.

Nonverbal communication comprises attentional functions, interpretations, and most importantly the regulation of interpersonal relations. We distinguish three, distinct, yet interdependent functions of nonverbal communication [66]: (1) discourse functions, (2) dialog functions, and (3) socio-emotional functions that influence our social behaviors.

Discourse functions are closely related to speech production and understanding. Emblems, pointing gestures, illustrative gestures, and beat gestures belong to this functional category [67]. But also, prosodic aspects, such as pausing and variations in voice pitch and volume. In general, they influence aspects of interpersonal communication and engagement that includes listener attention, interest, understanding, and interpretation.

Dialogue functions include turn-taking signals (e.g., eye contact, raise of voice, pausing) and back-channel signals (e.g., head nods, ‘uh-huh’, etc.), which serve to smooth the flow of interaction when exchanging speaker and listener roles [68]. In addition, dialogue functions influence aspects

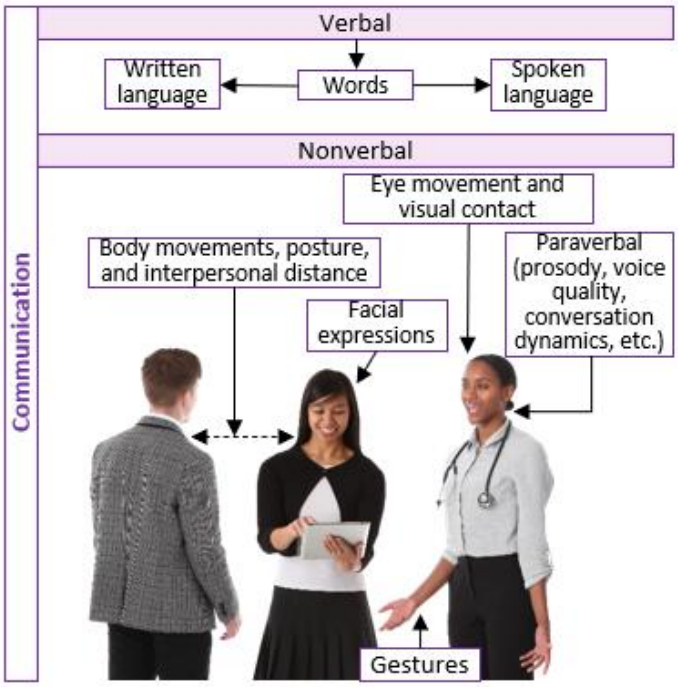


***Figure 5.*** *Elements of communication relevant to social behaviors during interactions modeled by an exchange of verbal and nonverbal messages. Verbal communication involves the use of words through written or spoken language. Nonverbal communication involves the use of gestures, facial expressions, paraverbal communication, eye movements, visual contact, body movements, posture, and interpersonal distance. © 2021, IEEE.*

of social interactions that include communication patterns, conversation dynamics, and the level of interaction between individuals. Dialogue functions also influence aspects of collaboration such as cooperation, in addition to aspects of dominance, and leadership roles.

Socio-emotional functions of nonverbal behavior include the communication of emotions and interpersonal attitudes and their regulation, which are crucial for establishing rapport. Whether we harmonize in an interaction, take others' perspectives or are capable of establishing a smooth flow of interaction very much depends on the exchange of those socio-emotional cues. Socioemotional functions are not independent from dialogue and discourse functions of nonverbal behavior. A smooth flow of the conversation will most likely influence positively the interaction climate. Power relations are evident in body postures, eye contact, voice amplitudes and more [69]. Harmony or interpersonal rapport shows in expressiveness or responsiveness [70] as well as in

mutual attentiveness (body orientation, eye contact), reciprocal positivity (smiles, interpersonal distance, body lean, and orientation), and behavioral coordination (motor mimicry, posture sharing and synchrony, and activity entrainment). Thus, monitoring nonverbal messages provides insights into the human and social behaviors being displayed given an environment [16], [71].

*2.2.2.1.In-Person and Virtual Communication*

Verbal and nonverbal communication are essential in both in-person and virtual interactions. By definition, in-person interactions are synchronistic, in that it occurs when two individuals are in the same place at the same time. This is the traditional and the richest media of all communication forms because it allows the individuals involved in the interaction to observe nonverbal cues such as facial expressions and body language [72]. On the other hand, virtual interactions come in various forms including email, telephone, instant messaging, and video calls, among others. This form of communication can be found to be asynchronous or synchronous. Of interest are video calls that, similar to in-person interactions, allow for real-time (synchronistic) communication, feedback, and transmission of important nonverbal cues [73]. Virtual interactions in the form of video calls can achieve the sense of "same space" inherent in in-person interactions.

Virtual interactions are comprised of multiple nonverbal communication cues (embedded in audio, video, and text forms) that happen simultaneously and differently than they do when in-person [74]. Research has found that voice, including paraverbal, is the most important communicative cue of meetings [74]–[76]. Nevertheless, video is essential in terms of "social presence," defined as the sense of intimacy and immediacy with others [74]. During the COVID-19 pandemic, Microsoft surveyed its employees to collect their experiences in virtual meetings, finding that participants reported that small group meetings with video turned on can be engaging and interactive [74]. Additionally, it was also found that to maintain a similar experience to in-

person interactions, nonverbal cues (derived, for example, from facial expressions and body movements) are essential to provide information about engagement, attention, and focus [77].

## 2.3.Methods for Monitoring Human Behaviors

### 2.3.1. Ethnographic Methods

Traditionally, scientists interested in studying social interactions have made use of ethnographic research methods, such as observations (including experiments) and surveys. Expert observation is probably the most common method for studying social interactions. Data collected through expert observation is a method used in all sciences and is independent of people's willingness to provide verbal information about their behaviors and feelings. One of the greatest advantages of employing expert observation is the depth of the collected data, which can be very detailed to "explain behavior and communication patterns in ways that a survey, interview, or experimental design cannot" [78]. On the other hand, self-reported data through methods such as surveys present the advantage that a wide range of information can be collected. Methods such as surveys make it possible to study very large populations, their attitudes, values, beliefs, and past behaviors [79]. However, even when human behaviors have been studied using expert observations and/or through surveys, these methods are not suitable for the interpretation of human and social behaviors in settings that would benefit from real-time feedback to improve on the observed behaviors.

### 2.3.2. Biometric Technology

In addition to expert observations and surveys to monitor human behaviors, biometric methods have been employed to measure psychophysiological processes related to human behavior. Biometric methods for studying human behavior include the use of different technologies, including sensors and algorithms, to monitor the activation of personal factors

(emotions and personality traits) and aspects of social behaviors during social interactions that manifest in the form of physiological processes and nonverbal messages. Some of the advantages of using biometric methods include the potential for unbiased and consistent measurements. Moreover, using the right experimental setup and biometric technologies, such as wearable sensor platforms, real-time data collection and analysis of human behavior can be achieved. Once "in-the-wild" human behavior analysis is achieved, real-time feedback can be provided to create behavioral awareness in the individuals from which the behavior was detected. To this end, a wide range of biometric technologies have been developed, and a variety of sensor modalities and algorithms employed to facilitate the realization of studies in real-life scenarios.

### 2.4.Biometric Technologies and its Components

Technologies for real-time monitoring of human behavior require the employment of a variety of components. The three main components of these types of technologies are sensors, signal features, and computational models. This section provides a glance at the literature available in this area, which is thoroughly revised and analyzed in Chapter 3. Note that this section, and this work in general, omits the review of works that employ cameras for the monitoring of human behaviors. The primary reasons to exclude cameras from this work are because most of the reported use of video cameras for the monitoring of human behavior has been for offline applications and because its use increases the computational load and power consumption of the system [80]. In addition, it has been a topic of debate that the use of cameras to monitor human behavior presents a concern for user privacy. Thus, as video and image modalities present limitations for real-time and wearable applications, we consider them out of scope.

### 2.4.1. Wearable Sensors for Collecting Physiological Signals and Nonverbal Messages

In the 1990s, the early development of wearables to study human behaviors was focused on identifying aspects of social interactions. These initial systems, still used nowadays, employed InfraRed (IR) and/or quasipassive radio frequency (RF) sensor modules to track position and proximity among individuals wearing these devices [81]–[83]. In an effort to create wearable systems with the ability to capture more informative data about human behaviors, research groups started working on the integration of multiple sensing modalities. One of the earliest initiatives was the MIThril project pioneered by A. Pentland [84]. The MIThril project focused on developing a "practical, modular system of hardware and software for research in wearable sensing and context-aware interaction" [84]. With the introduction in the early 2000s of the personal digital assistant (PDA) devices, the MIThril project first developed a modular wearable system comprised of a variety of sensors such as accelerometers, InfraRed (IR) active tag readers, GPS units, analog microphones, 2-channel electromyography (EMG) sensors, 2-channel electrodermal activity (EDA) sensors, and skin temperature monitors [85]. The sensors were wired to a PDA intended to perform real-time processing and communicate with other units of the same kind through Wi-Fi. However, there seem to be no reports of data collected using this system. In an effort to study communication patterns of groups of people during meetings in real time, Eagle and Pentland [86], designed a wearable system employing a headset microphone connected to individuals' PDAs to allow streaming of high-quality audio signals over a network, also with the choice of storing the audio locally on the device for post-processing. Conversations were detected in all streamed audio signals and conversation features extracted, including inferring the proximity among participants. Later, the same research group made use of the UbER-Badge [87], a device with a microphone, a two-axis accelerometer, and a forward-oriented IR transceiver, to measure human interest levels

during interactions in a conference meeting, all among dyads [88], [89]. With a similar system called the Sociometer, people involved in an interaction were identified, and through audio signals, conversation dynamics studied [90]. This version of the Sociometer was later optimized. Modified versions of the Sociometer have been known as the Sociometric badge [91]–[93], Open badge [94], and Rhythm badge [95]; all these sensor platforms have been used in the study of social interactions.

Mobile phones have also been used as a platform for monitoring social interactions. In [96], the Bluetooth and microphone units from a mobile phone were used to detect proximity and conversation dynamics in real time to infer levels of interest in a social interaction. Moreover, they have also been used to connect with badges to display feedback information useful to improve social interactions[93], [95]. A review of additional wearable sensors used for social interaction recognition can be found in [97].

Besides social interactions, wearables have also been used for real-time emotion recognition. In [98] and [99], accelerometer, gyroscope, ambient light, temperature, and humidity sensors were integrated into a watch-like device to monitor levels of anxiety in human subjects. In an improved version that includes a MEMS microphone and a skin temperature sensor, Jiang et al. [100] used this wearable system for health monitoring to study the relationship between mental health and physical health. In [22], Breeze, a wearable pendant placed around the neck, with an inertial measurement unit (IMU), was employed to measure breathing patterns as these are closely linked to emotions. The goal of the researchers was to improve the emotional states of the Breeze users by providing real-time feedback on the user's breathing patterns. Also related to the regulation of emotional states, in [101], a wearable system in the form of a glove containing an EDA, a blood volume pulse (BVP), and a skin temperature sensor was designed to continuously monitor changes

in the physiological signals that could relate to emotional mental states. On the other hand, Girardi et al. [102] used commercially available wearable sensors to capture electroencephalography (EEG), EDA, and EMG signals to detect emotions in the arousal-valence dimensions. Also using commercially available sensors, McGinnis et al. [103] employed accelerometers and gyroscopes to diagnose anxiety and depression in young children. A comprehensive list of commercially available wearable physiological sensors, used especially for the monitoring of emotions, can be found in [104].

### 2.4.2. Signal Features

The processing of sensor signals plays a critical role in the design of accurate real-time human behavior monitoring systems. Methods applied for the treatment of sensor signals include digital signal processing and machine learning techniques. The goal of digital signal processing is to apply pre-processing techniques to enhance signal quality and to compute statistically identifiable signal characteristics or measurable signal properties, typically referred to as signal "features", that are informative of human behaviors. Pre-processing techniques include signal filtering, normalization, and standardization which help eliminate signal artifacts and any other unwanted information from the collected signals [105]. On the other hand, extracting features from signals involves finding a variety of mathematical methods that could help identify patterns in the data [106]. Features are extracted/calculated using time-domain feature extraction techniques, frequency-domain feature extraction techniques, and time-frequency-domain feature extraction techniques. Time-domain features include zero crossing rate, slope sign changes, waveform length, statistical values, and Shannon entropy, among others; frequency-domain features include measurements derived from a Discrete Fourier transform and power spectral density analysis, among others; and time-frequency-domain features include short-time Fourier transform, Hilbert transform, Morlet

wavelet, and wavelet transform, among others [105]. After features are extracted, machine learning methods such as feature selection can be used to reduce redundancy in extracted signal characteristics and/or reduce the dimensionality of a given dataset. Selecting a final set of signal features has an important influence on the size of examples needed to create models capable of recognizing human behaviors, on the cost of computation, and on the time needed for recognizing such behaviors [105].

### 2.4.3. Computational Models

Based on the features extracted from sensor signals, computational models are trained and used to predict or classify human behavior. Therefore, the performance of computational models, also referred to in this work as machine learning models, can depend on the provided set of features. Likewise, the effectiveness of signal features can also depend, in part, on the type of computational method used to evaluate the feature's contribution.

The two principal types of machine learning models employed in the human behavior recognition literature are classification and regression models. Classification models focus on recognizing discrete or categorical classes, while regression models focus on predicting continuous numerical values. The use of a machine learning model is application specific. For example, the problem of emotion recognition can be treated as one with categorical values (e.g., happy, sad, neutral) or as one with continues numerical values (i.e., reflecting levels of arousal and valence based on a numerical scale). More details and analysis of signal features related to human behavior and computational models will be given throughout Chapter 3.

### 2.4.4. Training Frameworks

The design of technologies for real-time monitoring of human behavior is not complete without the collection of datasets to support the evaluation of signal features and the training of

machine learning models. Datasets to train machine learning models for human behavior recognition are application specific and can be classified as acted/evoked datasets and natural datasets. Acted/evoke datasets refer to data collected by requesting a subject to behave in a certain way (e.g., actors) or by controlling the environment to elicit the behavior of interest (e.g., watching images or movies to elicit specific emotions). On the other hand, natural datasets refer to data collected during spontaneous/natural interactions where behaviors cannot be controlled. Natural datasets are the hardest to construct as it involves designing the scenario to encourage spontaneous interactions and preparing annotation schemes. According to Cognilytica, an analyst firm, 80% of the time spent in machine learning and artificial intelligence projects goes into data collection, organization, and annotation [107], [108].

The scientific community has worked collaboratively to create publicly available datasets to advance the design of algorithms. For the design of human behavior monitoring, a variety of datasets have been created [109]–[117]. The creation of datasets has been mostly performed for applications in the area of emotion recognition, however, in the last 10 years, more databases have surged reflecting aspects of social interactions.

For emotion recognition applications, a combination of acted/evoke and natural datasets exist, in addition to datasets containing multimodal data. The HUMAINE database contains a collection of 48 audiovisual acted/evoked and naturalistic clips, some of them also containing physiological data, with labels describing affective responses. Clips were obtained mainly from TV shows/interviews and human-computer conversations, but clips from other sources were also obtained. This has been one of the most comprehensive databases, also providing a labeling scheme to identify emotional responses in audiovisual data [117]. DEAP is an evoked multimodal database containing frontal face video and physiological signals obtained while participants were

watching music videos. Collected data was labeled using self-reported ratings for arousal, valence, and like/dislike [112]. The MAHNOB-HCI is also an evoked multimodal database containing audio, visual, eye gaze, and physiological data. Videos and images were used to evoke emotions in 27 participants [113]. Other databases, such as BioVid Emo DB [109], also contain multimodal data collected while individuals were matching videos. Most of the data in these databases was collected with sensors that restricted the natural movement of participants. In addition, because their focus was on emotion recognition, their experimental design and collected data do not reflect the reactions of natural social/group interactions.

To provide more naturalistic data on person-to-person interaction environments, other databases have been made available. For example, the RECOLA database contains data from natural remote collaborative environments [110], [111]. Multimodal data, including audio, video, and physiological signals were collected from interactions between dyads (two individuals). This dataset was labeled by external annotators using a continuous arousal and valence scale and social behavior dimensions. External annotators looked at the following social behaviors: agreement, dominance, engagement, performance, and rapport. A more recent database, SEWA DB, contains audiovisual data from individuals watching adverts and then dyads having a conversation about such adverts [115]. This dataset includes facial landmarks, facial action units, vocalizations, mirroring, affective state (valence, arousal), and social behavior (liking, agreement) annotations. However, none of these databases capture the dynamics of groups.

There still exists a need for databases containing data from group interaction environments with their respective annotation schemes. For the design of human and group behavior monitoring systems, data capturing a combination of emotional reactions with social behaviors at the individual, dyadic, and group level are needed.

## 2.5.Summary

Reviewing the personal factors that underpin human behavior and the theories and concepts that have guided the psychological study of social interactions provides a scholarly foundation for understanding the methods that have been employed for monitoring human behavior. Methods employing wearable sensors for the monitoring of human behaviors have been focused on recognizing emotions or aspects of group behaviors, separately. However, for those works that have focused on group behaviors, only specific aspects of communication have been implemented in those systems, which do not capture the entirety of the social interaction complexity to identify disruptive behaviors and bring awareness. Even when many efforts have been done to study and design technologies to monitor individual emotions and aspects of group behaviors, very little has been done in designing robust systems that could measure a higher number of elements influencing social interactions within groups of people. In addition, none of those works have been focused on identifying aspects of the interaction that could be potentially useful to bring awareness to members of a group. It is also important to note that, currently there are no standard technologies, methods, and/or processes to study and design group behavior monitoring systems.

# 3. DESIGN OF A MULTI-SENSOR SYSTEM WITH A MACHINE LEARNING FRAMEWORK TO MONITOR GROUP INTERACTIONS IN REAL TIME

The reviewed literature provided a unique social science perspective with a focus on identifying critical elements to consider for the design of social behavior monitoring systems. The literature surrounding technology for human behavior monitoring is vast and varied. Focusing only on technologies with the potential to advance automatic and/or real-time monitoring of human behaviors, as was chosen for this work, motivated the creation of a classification system that could synthesized reported technologies to enable an analytical perspective. This classification system or taxonomy would relate to behavioral elements that helped define the individual, dyadic, and group metrics involved in group interactions. Of particular interest was a behavioral element of group interactions called rapport, which helps define the quality of social interaction between dyads. We hypothesized that by improving real-time self-awareness, rapport levels between individuals (dyads) could be increased, possibly affecting the overall group interaction. Here, technologies used for the monitoring of human behavior and rapport are presented. With the goal of establishing a framework for the design of group interaction monitoring systems with the capability of providing feedback that can improve the quality of social interactions, this work leverages on existing theories to monitor dyadic interactions and presents efforts in the design of a multi-sensor monitoring system for the real-time detection of group consonance. The term *group consonance* is introduced in this work to define the subset of rapport composed of *monitorable* behavioral components that contribute to establishing good rapport between dyads and its effect on the overall group interaction. As part of the design efforts, a comprehensive review and analysis of sensor technologies used for the study of human behaviors are presented and discussed.

***Table 1**. Taxonomy summarizing human behavior elements monitored using sensor technologies. © 2021, IEEE.*

| *Effector classes (complexes)* | *Elements (aspects/components/dimensions)* | *References* |
|---|---|---|
| Emotions | Dimensional: valence, arousal, potency<br>Categorical (basic emotions):<br>happy, angry, sad, quiet, disgust, anxiety, surprise<br>Others: curiosity, boredom, uncertainty, puzzlement | [22], [24], [98]–[103], [109]–[113], [132], [141]–[146], [182], [184]–[190], [193], [194], [237]–[257] |
| Personality factors | Personality traits: leadership emergence, openness, conscientiousness, extraversion, agreeableness, and neuroticism<br>Person Perception Dimensions: valence, dominance, activity<br>Others: empathy, honesty | [53], [86], [93], [111], [147], [150]–[154], [258]–[266] |
| Social Interactions | Cooperation or collaboration, agreement and disagreements, attraction, interest, attention, emphasis, vigilance, group performance, cohesion, communication patterns and dynamics, level of interaction, rapport | [17], [24]–[26], [83], [86], [88]–[96], [111], [123]–[126], [155]–[158], [162]–[165], [183], [191], [267]–[278] |

### 3.1.Measuring the Quality of Group Interactions

#### 3.1.1. Taxonomy for Monitoring Elements of Human Behavior

An underlying goal of this work, and indeed of most of the previous efforts in the literature, is to enhance the potential for technologies that augment human capability, toward a future of increasingly effective human-machine interactions. To further promote this human-centered approach, this work established a taxonomy for behavior-sensing technology that is based on the relevant psychological theory summarized in Chapter 2. Specifically, this taxonomy assigns technologies to the human behavior effectors that they target, and it defines three effector classes that encompass the reviewed literature, as shown in **Table 1**. The defined effector classes cover personal factors (i.e., emotions and personality traits) as well as social interaction factors observed through nonverbal communication channels, all of which influence human behavior.

In brief, **Table 1** assigns the emotions effector class to works that concentrated on recognizing categorical and dimensional emotional structures, most of which focus on understanding an

individual's emotional state, rather than the dynamics of emotional expression and exchange during a social interaction. The personality factors class was allocated to works related to personality traits and person perception dimensions as well as to works centered on the detection of empathy and honesty. Finally, the social interactions class was assigned to works covering aspects of interpersonal communication and engagement such as levels of interest, level of cohesion, communication dynamics, and rapport.

This taxonomy will be maintained throughout this chapter. The taxonomy will be used to discern similarities and differences in the various sensors, signal features, and computational models employed for monitoring within these prescribed human behavior effector classes.

### 3.1.2. Individual, Dyadic, and Group Nonverbal Behaviors

As described in Chapter 2, our social environments are created by the sum of individual behaviors interacting with each other. This work defines individual elements of behavior as the basic unit of all interactions, especially, dyadic interactions. A dyadic interaction describes the interaction between two people, which represents the smallest possible social group. Dyads represent the basic social interaction unit of groups of three or more people. Here, individual, dyadic, and group metrics of nonverbal behavior are explained to guide the design of our group interaction monitoring system. As highlighted in Chapter 2, the focus of this work is on the use of nonverbal behavior indicators because of their influence and importance on social interaction perception and the advantages of keeping privacy risks at their lowest, as will be further explained in Section 3.2.

#### *3.1.2.1.Individual nonverbal metrics*

As displayed in **Table 1**, emotions and personality factors have been studied and monitored using technology. Those two effector classes group literature that present technological

advancements with a focus on understanding behavioral elements that reside in a single individual, rather than the interaction between two or more individuals. Even when emotions and certain personality factors are influenced by the social environment, these ones have generally been monitored at the individual level. Generally, emotions and personality factors, such as personality traits and person perception dimensions, can be studied and monitored through physiological and paraverbal communication changes in an individual. Emotions and personality factors can also be studied through facial expressions, gestures, and posture. As discussed in Chapter 2, changes in physiological reactions can be driven by changes in emotional states. Likewise, paraverbal communication, which includes intonation, tempo, voice quality, volume, speaking time, turns, and interruptions, can be used to determine levels of individual activity and dominance, in addition to emotional states. However, the works that focus on determining individual nonverbal metrics of behavior differ widely in the approach that they take and the technology they employ, in terms of sensors, features, and computational models.

*3.1.2.2.Dyadic nonverbal metrics*

This work will consider a dyad to be the simplest form of a group and the smallest unit of analysis of an interaction within three or more individuals. Most of the reviewed literature targets dyads to study the elements of social interaction listed in **Table 1**. Generally, aspects of nonverbal communication such as body movements, postures, eye movement, visual contact, facial expressions, gestures, and paraverbal are more widely used to determine dyadic metrics of interaction. However, a small number of works have used synchrony analysis between physiological signals collected from dyads to determine levels of collaboration, synchronicity, and coordination. This demonstrates how individual nonverbal metrics of behavior from different individuals can be combined to monitor social dynamics. Likewise, other individual nonverbal

metrics, such as speaking time, turns, and interruptions can be compared across participants of an interaction to determine levels of overall interaction, cooperation, and cohesion. Also, gestures and postures help determine levels of mimicry or coordination, which is essential to establish rapport.

At the core, rapport has been the primary aspect of social interactions attributed to dyads. Rapport is a complex social behavior mostly correlated with nonverbal communication channels. In fact, research has demonstrated that nonverbal behavioral cues are more indicative of rapport than verbal communication channels [119]. Rapport is a harmonious relationship and connection with someone, where the feelings or ideas of others and ourselves are understood, and communication runs smoothly. People that experience/develop high levels of rapport have a higher quality of social interactions, better team dynamics, and, consequently, are more productive in the workplace.

In 1990, Tickle-Degnan and Rosenthal [120] proposed a theoretical model for rapport that describes three essential components of this complex social behavior: mutual attention, shared positive feeling, and synchrony or coordination. Tickle-Degnan and Rosenthal made various observations about how rapport manifests through time and how it varies depending on the context. Tickle-Degnan and Rosenthal suggested that at the beginning of an interaction, strong feelings of rapport are dominated more by emotional positivity and attentiveness than by coordination. However, in more developed interactions, attentiveness and synchrony or coordination are more dominant. **Figure 6** illustrates the idea of the relative importance of the three essential ingredients for rapport and their relationship through time, presented in [120].

One of the most important observations of Tickle-Degnen and Rosenthal included that the three components defining rapport (i.e., coordination, mutual attentiveness, and positivity) were

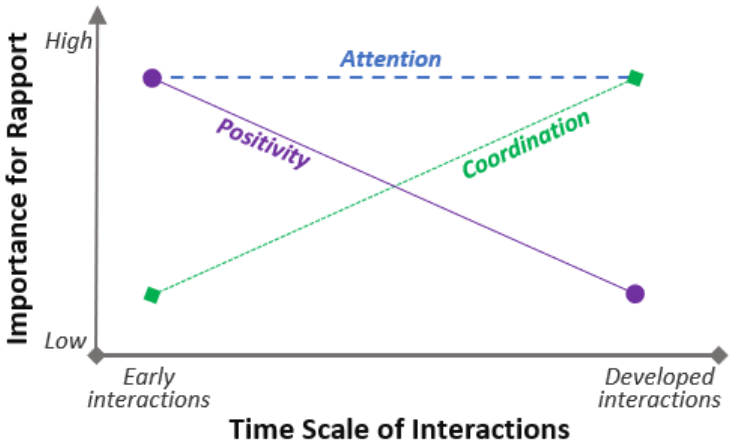


***Figure 6**. Relative importance of the three essential ingredients for rapport and their relationship through time. Figure adapted from Tickle-Degnen and Rosenthal [120].*

encoded in expressed behaviors [121]. The study of rapport and how it might be perceived by others differ from the study of personality perception because the former does not reside on a single individual, instead, it is constructed based on the relationship between two individuals [121].

*3.1.2.3.Group nonverbal metrics*

Similar to the case of dyadic nonverbal metrics, although at a smaller scale, the elements of social interaction listed in **Table 1** have been studied in groups. Similarly, many of the nonverbal metrics used in the study of dyadic interactions have been applied to group interactions. In most works, individual nonverbal metrics combined with dyadic nonverbal metrics have been used to determine low and high levels of rapport [122] and cohesion [123] in meetings. Other metrics such as overall group speaking length, speaking turns, and speaking interruptions throughout a meeting have been used to characterize groups as cooperative or competitive [124]. Although high rapport is considered an essential factor in the establishment of quality interactions, none of the works that focus on monitoring this phenomenon provide a framework suitable for behavioral feedback that could resolve complex dynamics. Thus, this lack of information motivated the work in this

dissertation to provide new means of assessing human behaviors, which is further explored in the next Section.

### 3.1.3. Technologies that Measure Rapport in Group Interactions: Challenges and Opportunities

Because rapport is considered essential to effective dyadic and group interactions and encompasses multiple components of human behavior, in this work, it is considered a guiding factor in the design of the group interaction monitoring system. Many works in the literature have used the Tickle-Degnan and Rosenthal model, directly or indirectly, to guide their studies, observations, and automatic analysis of rapport using technology. For example, Hagad et al. [125] pointed out that posture mirroring behavior, which is related to coordination, has been linked to rapport. Thus, using a video camera, the authors extracted signal features describing the individuals' posture during a dyadic interaction, trained posture classification models for each individual in the interaction, and then used the results of these models to determine posture congruence in dyads, achieving a ~71% average classification accuracy when recognizing between low, neutral, and high rapport. In another work by Cerekovic et al. [126], rapport was predicted using 1-minute segments of audio-visual data collected from an individual interacting with a virtual agent. The authors trained binary regression and classification models to predict/recognize between positive and negative rapport, achieving an 87% average accuracy. However, features employed in the prediction/recognition task included verbal audio features, not just nonverbal ones. In general, virtual agents have been commonly employed in the recognition of rapport or its components [127]–[129], however, that has limited the automatic recognition of rapport to just dyads. In an effort to recognize rapport in groups, Muller et al. [122] investigated the automatic prediction of low rapport during natural interactions within small groups. The authors were

particularly interested in recognizing the overall degree to which an individual in an interaction is able to build rapport with others. To do this, they analyzed audio-visual data and extracted features describing nonverbal messages such as facial expressions, hand motion, gaze, speaker turns, and speech. The data labels were obtained by averaging the rapport scores given to an individual by the other members of the group. The authors trained a classification model to recognize low versus medium/high overall group rapport and studied the correlation of the features with the overall level of rapport, achieving up to a 70% average classification precision.

So far, the automatic recognition of rapport in dyads and groups, as defined by Muller et al., has been performed by measuring just a single component of rapport or by training a model that identifies between low and high rapport by taking all extracted features at once. Even when this has contributed to the monitoring of the quality of social interactions, current methods of monitoring group interactions may not provide the necessary information to deliver a feedback message in real time to help individuals improve their quality of interactions. Real-time or near real-time processing is required in group interaction monitoring systems to provide information that can impact human behaviors as they happen. From the works that focus on providing real-time feedback, efforts have been concentrated on improving paraverbal communication patterns or providing individual awareness of emotions, both important aspects of social interaction. Still, the real-time monitoring of complex behavioral dynamics, such as rapport, that have a major impact on the well-being of humans, requires the integration of multiple nonverbal metrics of behavior and respective recognition capabilities. To the best of our knowledge, no work has been focused on establishing a framework to monitor rapport at the level of its components to facilitate feedback in human-to-human interaction to contribute to the improvement of the quality of the interaction.

Monitoring the quality of social interactions by calculating an overall rapport score may not provide enough information to deliver effective user feedback that can enhance the quality of the interaction. This work hypothesizes that monitoring individual components of rapport, i.e., attentiveness, positivity, and coordination, will allow us to extract an overall measure of rapport and identify which component of rapport needs attention when low rapport is detected. This can also be combined with other general group nonverbal metrics. However, the use of rapport as a measure of group interaction quality requires a deep understanding of human behavior dynamics, nonverbal metrics that contribute to each of the rapport components and their interactions over time, the dyadic attributes, and the effective employment of sensors and computational models. A system framework based on the rapport model needs to take into consideration the smallest unit of interaction, individuals, and the basic unit of group interaction, dyads, and build upon that a group model. Because the use of technology could limit the aspects of rapport that can be monitored, this work will refer to the monitoring of rapport using technology as monitoring group consonance.

### 3.2.Deep Analysis of Technology for Behavior Monitoring

To better understand the design space for group interaction monitoring systems and establish a framework that monitors group consonance based on rapport components, a deep analysis of available technologies was performed. Here, the categorization and details of sensors, signal features, and computational models employed in the monitoring of human behaviors are presented.

#### 3.2.1. Categorizing Behavior Monitoring Sensors

In general, machine monitoring of human behavior starts with the appropriate selection of sensors. Commonly used sensors in monitoring human behavior can be grouped as sensors that capture video and images, audio, physiological, movement, orientation, proximity, and environmental signals. The selection of a sensor or multiple sensors is driven by the type of

behavior that intends to be monitored and its associated nonverbal messages and physiological reactions. Based on the analysis of reviewed literature, 21 different sensors were found to have been used to monitor human behaviors. **Table 2** lists the sensors used in the reviewed literature and the nonverbal messages, physiological reactions, and/or environmental conditions that can be captured by them. In addition, **Table 2** summarizes information related to sensor placement and the level of superficial invasiveness to the user. Here, the definition of superficial invasiveness centers on the degree to which the sensor requires to enter into contact with the body and not whether the sensor needs to be implantable in the body. Thus, this work classifies the level of superficial invasiveness to the user in three categories: skin contact (sensor requires direct contact with the skin), body contact (sensor has to be placed on the body but does not require direct contact with the skin), and no contact, with skin contact being the most invasive and no contact completely non-invasive. For sensors that require skin contact, it is indicated if they require a single point of contact or multiple points of contact. This information is useful when assessing the level of obtrusiveness of a given system or evaluating sensors for the design of wearable systems.

From the sensors listed in **Table 2**, the top 11 most frequently used sensors in the literature were studied and their frequency of use was plotted with respect to the effector classes presented in **Table 1**. The relationship between the top 11 most frequently used sensors and the effector classes are summarized in **Figure 7**. In the mentioned figure, it can be observed that the monitoring of emotions has been one of the areas of most interest followed by the monitoring of social interactions, with microphones as one of the most common sensor modalities used for their study. It can be noticed that microphones, cameras, and EDA sensors are the only sensing modalities used in the monitoring of all three effector classes. In the cases of microphones and cameras, it is presumably because of the quality and quantity of the information that they provide, the numerous

advances in the areas of speech and image processing, and advantages in terms of superficial invasiveness and placement. However, the use of cameras (to capture image and/or video) requires

***Table 2****. Categorization of sensor technologies used in the literature* to monitor human behavior, together with its informants, associated effector classes and sensor modality, and the level of invasiveness of the sensors relative to their placement. Abbreviations: Emotions (E), Personality Factors (PF), Social Interactions (SI), Unimodal (Uni), Multimodal (Multi).* 

| *Type* | *Sensor* | *Nonverbal, physiological, & other informants* | *Effector classes* | | | *Sensor modality* | | *Level of superficial invasiveness* | *Placement* |
|---|---|---|---|---|---|---|---|---|---|
| | | | *E* | *PF* | *SI* | *Uni* | *Multi* | | |
| *Audio* | Microphone | Prosody, pitch, speech volume, intonation, turn-taking, pauses, speech duration | ✓ | ✓ | ✓ | ✓ | ✓ | Body contact or no contact | Chest or in front of an individual (on a table) |
| *Video and Image* | Camera | Gestures, body movements, body lean and orientation, postures, facial expressions, eye gaze | ✓ | ✓ | ✓ | ✓ | ✓ | No contact | In front of the individual or room view |
| *Movement, orientation, and proximity* | Accelerometer | Body movements, body lean and orientation, postures, gestures, breathing patterns | ✓ | | ✓ | | ✓ | Body contact | Chest, left wrist, belt, necklace, in the right trouser pocket, shirt pocket, or bag |
| | Gyroscope | | ✓ | | | | ✓ | | |
| | Magnetometer | | ✓ | | | | ✓ | | |
| | InfraRed (IR) sensor | Orientation (face-to-face time), proximity | | | ✓ | | ✓ | | Chest, head |
| | Ultrasonic sensor | | | | ✓ | | ✓ | | Chest |
| | GPS | Proximity | | | ✓ | | ✓ | | Chest, belt, pocket, or bag |
| | Radio Frequency (RF) – Bluetooth included | Proximity, gestures, body movements | ✓ | | ✓ | ✓ | ✓ | Body contact or no contact | Chest, belt, pocket, bag, or room |
| | Eye tracker (optical) | Eye gaze | ✓ | | ✓ | | ✓ | | Face or in front of an individual (on a monitor) |

***Table 2.*** *(cont'd).*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Physiological* | Blood volume pulse (BVP) /Photoplethysmography (PPG) sensor | Blood volume in arteries and capillaries, heart rate | ✓ | | ✓ | | ✓ | Skin contact – single point of contact | Wherever there is an easy access to a pulse. Fingers or earlobes are commonly used |
| | Respiration (RSP) sensor | Respiration rate | ✓ | | ✓ | | ✓ | | Chest |
| | Skin temperature monitor | Skin temperature | ✓ | | ✓ | | ✓ | | Any site on the body with preference in the axilla and forehead |
| | Electrodermal activity (EDA) /Galvanic Skin Response (GSR) sensor | Skin conductivity | ✓ | ✓ | ✓ | ✓ | ✓ | Skin contact – two points of contact | Fingers, palm of the hands, soles of the feet, or wrist |
| | Electrocardiogram (ECG) | Heart rate | ✓ | | ✓ | ✓ | ✓ | Skin contact – multiple points of contact | Chest or limbs |
| | Electroencephalography (EEG) | Brain activity | ✓ | | ✓ | ✓ | ✓ | | Along the scalp |
| | Electroglottography (EGG) | Pitch, turn-taking, pauses, speech duration, utterances | ✓ | | | | ✓ | | Surface of the neck |
| | Electromyography (EMG) | Facial expressions | ✓ | | | | ✓ | | Facial muscles |
| | Electrooculography (EOG) | Eye gaze | ✓ | | ✓ | | ✓ | | Face, around the eyes |
| *Environ -ment* | Ambient temperature sensor | Environmental factors | ✓ | | | | ✓ | Body contact or no contact | Wrist or in a room |
| | Humidity sensor | | ✓ | | | | ✓ | | |
| | Ambient light sensor | | ✓ | | | | ✓ | | |

* Information presented in this table was obtained by analyzing collected information from the articles referenced in Table 1.

a large data bandwidth of communication compared to microphone data. In fact, most of the literature reporting the use of video cameras for the monitoring of human behavior has been for offline applications. It is important to mention that when video and image data are included to analyze human behavior, the computational load and power consumption of the system increase [80]. In addition, it has been a topic of debate that the use of cameras to monitor human behavior

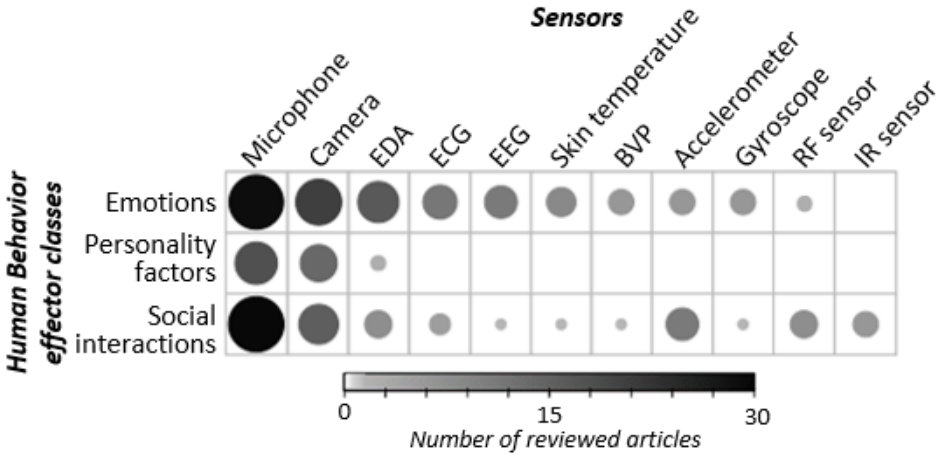


***Figure 7****. Graphic representation of where the work related to human behaviors has been concentrated relative to the 11 most used sensor modalities. The monitoring of emotions has been one of the areas of most interest followed by the monitoring of social interactions, with microphone as one of the most common sensor modalities used for their study. Microphones, cameras, and EDA sensors are the only sensing modalities used in the monitoring of all of the three effector classes. © 2021, IEEE.*

presents a concern for user privacy. Thus, as video and image modalities present limitations for real-time and wearable applications, we exclude them from further analysis in this work. However, information on the use of video and image sensor modalities for the monitoring of human behaviors, including the use of facial expressions for the recognition of the emotion effector, can be found in [130], [131]. On the other hand, although the privacy issue could be argued to also apply to microphone data, in the case of monitoring human behavior as presented in this work, speech recognition is not the goal. In this area, microphones are used mostly to perform speech detection to extract acoustic features in speech (e.g., volume, signal energy, pitch), which can be performed in a local device before any data transmission and at reasonable computational and power consumption rates [100], [132].

Besides cameras and microphones, and in addition to EDA sensors, four physiological sensors that have been commonly used are ECG, EEG, skin temperature, and BVP sensors. The wearability of physiological sensors, which has been possible due to advancements in CMOS and circuits technologies [12], [133], has allowed the study of human behavior effectors in different

scenarios. ECG, EEG, skin temperature, and BVP sensors have helped understand acute and long-term changes in the physiology of the human body that are often altered by internal and external stimuli, but that our conscious mind cannot control. All of those four physiological sensors have been used to monitor emotions and aspects of social interactions, as noted in **Figure 7**. On the other hand, accelerometers, gyroscopes, IR sensors, and RF sensors are among the most frequently used sensors from the movement, orientation, and proximity sensor types. While accelerometers, gyroscopes, and RF sensors have been used in the recognition of emotions, IR sensors have just been used in the monitoring of social interactions to measure the proximity between individuals.

In addition to sensor modalities that are directly related to measurements of an individual, the contextual or environmental information in which signals of an individual are collected could help improve machine understanding of behavior. Although the use of environmental sensors in the human behavior monitoring literature is scarce, it is starting to be used to add to the contextual understanding of behavior. For example, in [98] and [100], environmental sensors such as temperature, humidity, and ambient light were used in a wearable sensing device to help determine moments of personal anxiety.

*3.2.1.1.Analyzing unimodal versus multimodal sensor systems*

While **Table 2** and **Figure 7** illuminate the breadth of sensors employed for behavior monitoring and their relative popularity in the literature, it is also important to consider the number of different sensor modes employed among these studies. To provide some insight into this, **Figure 8** plots the distribution of sensor modalities concerning the identified effector classes across the articles that were analyzed. This plot shows that, of the ~72 reviewed works, around 59% of them rely on unimodal sensing, including all works targeting personality factors. Moreover, these unimodal efforts utilize only five of the sensor types defined in **Table 2**, namely microphones,

EDA, EEG, ECG, and RF sensors. In contrast, roughly 40% of works that were found to use two or more sensor modes, defined as multimodal in **Figure 8**, utilize all sensor types listed in **Table 2** (except for cameras, excluded from this analysis). One might expect that, as sensor technologies advance, a trend toward multimodal sensing would be evident, and the performed analysis supports this, showing that 66% of the multimodal works have been published since 2017, compared to only 14% of unimodal works. Multimodal sensing also makes practical sense considering that, as social individuals, humans often communicate using multimodal signals in a complementary and redundant manner. Thus, our own actions would suggest that multimodal sensor systems would be ideal for the recognition of human behaviors.

In the area of human-computer interaction, specifically in the detection of emotions, it has been recognized that multimodal systems improve the recognition rate of human behaviors when compared to unimodal approaches [55], [134], [135]. **Figure 9** presents the range, where the central red mark indicates the median accuracy, of the reported computational classification performance accuracies for both unimodal and multimodal sensor systems in the reviewed literature. Note that **Figure 9** collects information only from works that performed a classification task and reported their results using a percentage of performance accuracy.

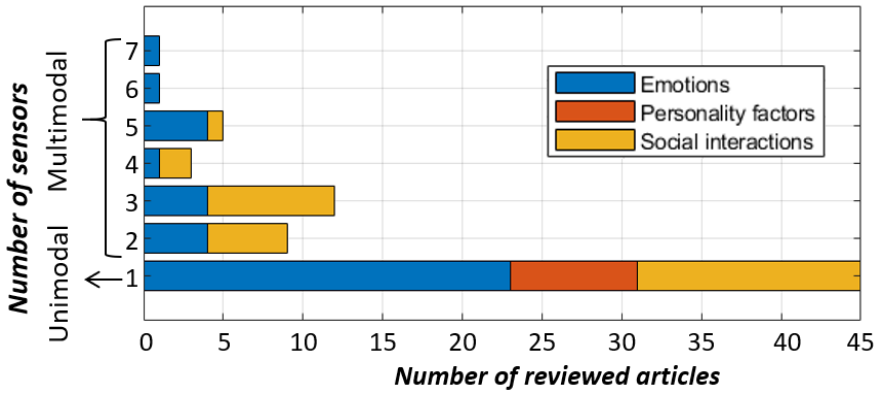


***Figure 8**. Distribution of the use of unimodal and multimodal (excluding video and images) sensor modalities to monitor human behaviors. Of ~74 reviewed works, around 59% of them rely on unimodal sensing, including all works targeting personality factors, and roughly 40% use two or more sensor modes. © 2021, IEEE.*

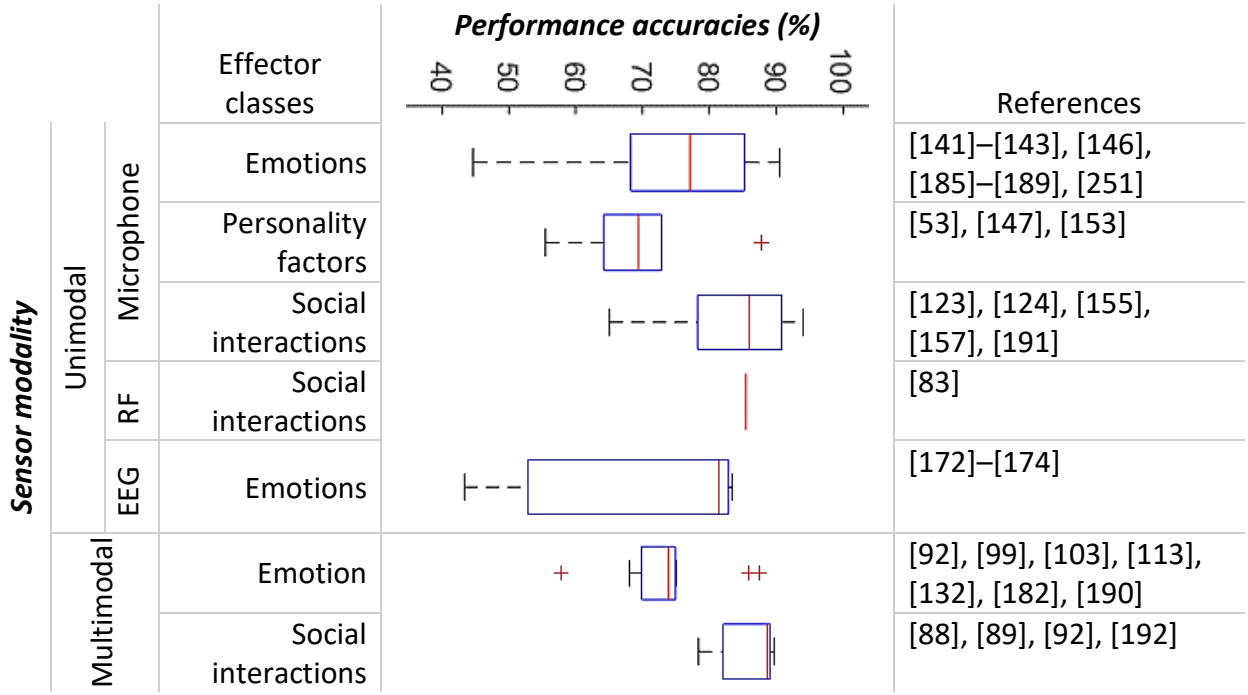


***Figure 9.*** *Summary of reported performance accuracies of unimodal and multimodal (excluding video and images) sensor systems of reviewed literature. The central mark indicates the median accuracy, and the left and right edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme accuracy values not considered outliers, while accuracy values considered outliers are plotted individually using the '+' symbol. © 2021, IEEE.*

From **Figure 8** and **Figure 9**, it can be observed that the effector classes that most utilize multimodal sensing are emotions and social interactions, a fact also noted in behavior monitoring review papers [16], [136]–[138]. On the other hand, the works reviewed in this chapter show a lack of multi-sensor modalities for monitoring personality factors. Although unimodal approaches have helped the scientific community in evaluating how information from a specific sensor contributes to understanding a certain behavior, studying the integration of multi-sensor modalities advances the development of more accurate and robust social sensing systems. Compared to unimodal sensing, multimodal sensing is still in its infancy and encounters new layers of complexity in defining and assessing accuracy. This may explain the lack of multimodal accuracy

improvements observed in **Figure 9**. However, multimodal systems do demonstrate less variability in accuracy, which could indicate advantages in precision and system robustness.

3.2.2. Signal Features Informative of Human Behaviors

The sensor modalities discussed in the previous section are just one of the components necessary to capture the physiological processes and nonverbal messages associated with human behaviors. The processing of sensor signals also plays a critical role in the design of accurate real-time human behavior monitoring systems. The goal of sensor signal processing is to compute statistically identifiable signal characteristics or measurable signal properties, typically referred to as signal "features", that are informative of human behaviors.

To analyze the sensor signal processing reported in the reviewed behavior monitoring literature, works were first grouped based on their use of unimodal sensor signals and multimodal sensor signals. Then, the unimodal works were organized by their sensor modalities and the four most used modes (excluding cameras, for reasons stated earlier), based on data in **Figure 7**, were selected for further analysis. Within each modality, sensor signal processing elements such as signal characteristics, pre-processing approaches, and features were studied and summarized to illuminate the design space employed in the literature. For the analysis of signal features, reported works were grouped by their behavior effector class defined by the taxonomy established in **Table 1**. Then, works, where the contribution of features to the recognition of a particular behavior was reported using correlation analysis or feature selection algorithms, were summarized below. Feature selection has two advantages: it reduces computational costs, and it removes noisy data that otherwise could degrade system performance.

The understanding gained from the analysis of unimodal sensor signal processing elements was then applied to make a qualitative assessment of their utility and design considerations in

multimodal systems. Finally, we attempted to integrate this information with an analysis of the limited works presenting signal processing for multimodal systems. This effort allowed us to make the summary observations presented at the end of this section that may be helpful for the design of real-time human behavior monitoring systems.

*3.2.2.1.Audio signals*

Audio signals collected from microphones are sound waves converted into electrical energy that, when employed in human behavior recognition systems, are typically used to monitor paraverbal communication. Audio signals used to monitor paraverbal communication are usually collected using a minimum sampling rate of 8 kHz, but rates up to 44.1 kHz have also been reported. The use of higher sampling frequencies provides better signal resolution, but it is not necessary for the extraction of the acoustic features of interest. The processing of audio signals is mainly composed of four parts: speech detection, speech segmentation, signal pre-processing, and feature extraction. Thus, identifying levels of noise, periods of silence, and periods of speech becomes a key task to ultimately extract accurate features and associate them with behaviors of interest. In real-time processing, audio signals are processed in frames of ~30ms to ~80ms, often with overlaps between each consecutive frame. These frames of data are used to detect speech. In general, after detecting speech in the audio signal, audio segmentation is performed. Audio segmentation refers to the task of dividing the audio signal into acoustic segments from which acoustic features will be extracted [139].

Typically, in the area of human behavior monitoring, audio segmentation has been done in two ways, through an utterance-based approach or a windowing-based approach. The utterance-based approach includes segments taken based on linguistic units such as vowels, phonemes, words, and phrases. However, when dealing with automatic and real-time processing, an automatic

speech recognizer (ASR) is needed to make use of an utterance-based approach. Although the use of ASR typically does not degrade the performance of a system [140], [141], it does increase the computational complexity of the system and could represent a threat to users' privacy. On the other hand, a windowing-based approach makes use of a window of time (in milliseconds or seconds), windows of speech activity (defined by pauses or silence), and/or windows of voiced or unvoiced signals. Windowing-based approaches are preferred in real-time systems because they are very fast and computationally efficient. However, this efficiency could be compromised when high amounts of memory space are needed to extract features of interest. While very small windows of time may not provide enough information to determine a change in a behavioral state, longer windows of time provide information similar to the one obtained from utterance-based approaches. This is because, in general, an utterance is comprised of pauses or breath segments and voiced-unvoiced speech segments [142]. Thus, accumulating data from audio frames creates a larger window of speech activity with speech and salient segments, similar to the information of utterance-based approaches. A good balance between performance and computational complexity can be found by evaluating different time window sizes as done in [143]. Here, we discuss works that make use of both approaches with the goal of extracting general information about relevant features.

Before extracting acoustic features, it is good practice to pre-process the audio signal using a pre-emphasis filter and a window function (i.e., Hamming window) applied to each frame to reduce signal discontinuity in order to avoid spectral leakage. **Table 3** describes all the identified acoustic features used in the reviewed literature. Acoustic features were grouped by several feature categories: prosodic in speech, conversational characteristics, voice quality characteristics, cepstral coefficients, formant characteristics, frequency spectrum coefficients, and others. The

***Table 3***. *Audio features found in the reviewed literature associated with human behavior effector classes. © 2021, IEEE.*

| *Feature categories* | *E* | *PF* | *SI* |
|---|---|---|---|
| **Prosodic features:** Volume amplitude (statistics*), intensity (statistics*), energy (entropy, RMS, linear regression, statistics*), voice pitch (linear regression, statistics*), autocorrelation (maximum peaks, # of peaks), voiced time | [99], [100], [110], [132], [141]–[145], [182], [184]–[186], [189], [190], [193], [194], [237], [251] | [53], [86], [148], [150], [151], [153], [154] | [86], [88], [89], [92], [123], [124], [126], [155]–[157], [191], [270] |
| **Conversational features:** Turn duration, # of turns, speaking duration (statistics*), speaking rate, overlapping speech duration, interruptions, pause duration (statistics*), # of pauses | [141], [145], [185] | [86], [147], [148], [150]–[154] | [86], [90], [123], [124], [126], [155], [156], [191], [270], [279] |
| **Voice quality features:** Zero-crossing rate, harmonics-to-noise ratio (HNR), jitter, shimmer, glottal features (# of glottal pulses, relaxation coefficient (Rd), functions of phase-distortion (FPD)) | [99], [110], [142]–[144], [184]–[186], [190], [193], [237] | [150] | - |
| **Cepstral features:** Shifted delta cepstrum (SDC), mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) cepstral coefficients, linear prediction-based cepstral coefficients (LPCC), plus their delta and acceleration values | [99], [110], [142]–[144], [182], [184]–[190], [193], [194], [237], [251] | - | - |
| **Formant features:** Formant frequencies (first and second), bandwidths (first and second), statistics* | [100], [110], [132], [141], [142], [190], [193], [251] | [53], [150] | [270] |
| **Frequency spectrum coefficients:** Brightness, center of gravity, distance between the 10 and 90 % frequency quantile, slope between the strongest and the weakest frequency, linear regression, spectral energy (statistics*) | [99], [100], [132], [143], [182], [185], [186], [193], [194], [237], [251] | [150] | [270] |
| **Others:** Wavelet coefficients, air pressure distribution in the vocal tract | [146], [189] | - | - |

Note. * Statistics include mean, std, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

definition of some of the features vary depending on the applied segmentation approach, therefore, we do not define them here, but information can be found in the references listed in **Table 3**.

Because different audio features have been reported to contribute in different ways to the recognition of human behaviors, it is valuable to look more deeply into the level of contribution that various audio features provide toward behavior recognition.

*Emotions* - Lee and Narayanan [141] evaluated, using a feature selection method, a set of prosodic (voice pitch, energy, speech duration, and their statistics) and formant features extracted at an utterance level to improve the recognition of two emotion classes: negative and non-negative emotions (valence dimension). The feature selection method consisted of evaluating classification accuracies using the k-nearest neighborhood classifier with a leave-one-out cross-validation method. While the authors separated speech data by gender (female, male), the ratio of the duration of the voiced and unvoiced region, energy median, and F0 (voice pitch) regression coefficient were included in the five-best features for both genders. In this same line, Tahon et al. [144] employed an ANOVA test and a classifier to study the contribution of prosodic, cepstral, and voice quality features in the detection of positive and negative emotions (valence dimension). They concluded that the mean and std of the relaxation coefficient (a parameter associated with how relaxed is the human voice), the harmonics-to-noise ratio (HNR), and the unvoiced ratio are of interest for valence detection. Also, of interest resulted a combination of features consisting of the functions of phase-distortion (FPD) (a distortion of the phase spectrum around its linear phase component), voice pitch, energy, and shimmer features.

In the recognition of discrete emotions, the recognition of frustration and calmness can be found. Ang et al. [145] showed, through the use of "a brute-force iterative feature selection algorithm", how prosodic features extracted at the utterance level contributed to the recognition of frustration. They concluded that longer durations of vowels or phonemes in an utterance (a word in their case) and slower speaking rates (the number of vowels divided by the duration of the

utterance) were associated with frustration. In addition, high values in voice pitch features such as maximum pitch in the longest vowel, the maximum overall pitch, the times that the maximum and minimum pitch occurred, the maximum speaker-normalized pitch rise, and the distance of various pitch statistics from the speaker baseline were all associated with frustration, representing the highest percentage of the total information used by the classifier. Other features that were associated with frustration were speaker-normalized RMS energy features, the number of dialog exchanges between the user and the system, and raised voice.

In addition to the direct use of extracted features, some works have applied principal component analysis (PCA) to reduce the dimensionality of the feature vector used to perform classification. Sahoo and Routray [146] estimated the pressure distribution in the vocal tract, which often results in a minimum of 40 feature values that increase depending on the number of vowels present in a given utterance or window of time. Thus, the authors applied PCA and made use of the first 6 principal components to classify calm and aggressive speech segments.

*Personality factors* – When monitoring elements of personality factors, and also social interactions, two types of features can be extracted: individual-level features and group-level features. Individual-level features are extracted based on the audio signals of a single individual and could include any of the acoustic features listed in **Table 3**. Group-level features are extracted from individual-level features; they describe the dynamics of a group of people. Thus, they are typically extracted using a window of time with a size in the order of minutes [147].

Related to personality traits, prosodic features have been found to be important for modeling observed extraversion, emotional stability, and openness to experience. Mairesse et al. [148] analyzed how those three aspects of personality were correlated to prosodic features. It was found that the maximum voice pitch, and the mean, std, and maximum values of intensity in dB were

highly correlated with extraversion. Emotional instability was highly correlated with the voiced time and the minimum and mean values of voice pitch, while openness was correlated to the maximum voice pitch values and voiced time. The authors also showed that prosodic features are very good predictors of extraversion in comparison to other types of non-acoustic features. In this sense, analytical studies have shown that extroverts speak more rapidly, with fewer pauses and hesitations than introverts [149]. Extraversion has also been associated with high values of voice pitch and higher variations in fundamental frequency, shorter periods of silence, and higher voice quality and intensity. This was confirmed by Vinciarelli et al. [150] when studying how acoustic features correlated to personality traits. The higher the voice pitch and speaking rate, the higher the perceived extraversion. A higher center of mass in the power spectrum and higher spectral tilt were correlated with perceptions of less agreeableness. Voices for which the power spectrum is peakier and tends to be skewed towards higher frequencies are perceived as more agreeable. These latter cues affect the perception of conscientiousness in the same way, together with the speaking rate (people that talk faster are perceived as more competent). In the case of neuroticism, the higher the voice pitch and first formant mean, the higher the perceived neuroticism. However, no evidence of correlation was found for openness.

Related to the person perception dimensions, Tusing [151] studied how much the amplitude of the speech signals in decibels (dB), the voice pitch, and the speech rate in words per minute (wpm) contribute to the perception of dominance. Through regression models, it was concluded that the mean amplitude, the amplitude standard deviation, the average voice pitch, and speech rate were correlated with aspects of dominance. This is particularly interesting because it has been noted that dominant people tend to be verbally active while non-dominant individuals are less so. One of the greatest advantages of using speaking rate and features like speaking length [152], [153]

to infer dominance revolves around its fast computation and easy use in real-time human behavior monitoring systems. This was employed by Eagle and Pentland [86], who made use of conversation features such as speaking rate, energy, duration of time holding the floor, interruptions, and turn-tacking transition probabilities to build over time profiles of participants' typical social behavior. This allows us to recognize relationships and dominant behaviors. In a work by Jayagopi et al. [153], features such as speaking turn duration histogram, total successful interruptions, total speaking turns, and total speaking energy also proved to be a good combination of features to identify the most and the least dominant individuals in an interaction. Similar features were shown in [154] to help identify emergent leaders.

*Social interactions* – Using individual-level features, Hillard et al. [155] studied the automatic detection of agreements and disagreements using prosodic and linguistic features. There it was found that prosodic features such as the average, maximum, and initial pause duration, the maximum and average voice pitch values, and the average and maximum duration of an utterance are almost as good as linguistic features in identifying segments of agreements. Investigating the automatic detection of the level of interest and involvement of individuals in an interaction, Gatica-Perez et al. [156] found through a feature selection method that speech energy, speaking rate, and voice pitch were the best audio features for the task. Moreover, voice pitch values have also been associated with the detection of emphasis during meetings [157]. On the other hand, Cerekovic et al. [126] studied the correlation of acoustic features with self-reported and judged evaluations of rapport between a subject and a virtual agent. It was found that interactions with fewer and shorter pauses, long speech segments, and louder speech were correlated with high rapport. Turn-taking patterns were also correlated with rapport.

Using group-level features, conversation dynamics have been very well explored. It has been studied that global features, such as group speaking interruption-to-turns ratio and group speaking turns egalitarian measure, have been found to discriminate with high accuracy between a competitive meeting and a cooperative meeting [124]. Features such as turn-taking have also been used to identify conversations between two individuals by calculating the mutual information between the turn-taking features of the individual's audio streams [90] and to detect conflicts [158]. Other features such as the sum of all the individual's pause duration, the maximum speaking rate during overlapping speech among individuals, the minimum average turn length among individuals, the total time that at least two people are speaking at the same time (total overlap time), the average energy that is observed for any participant when they are speaking at the same time as at least one other person, and the speaking rate during overlapping speech were reported to have high values in high-cohesion meetings [123].

*3.2.2.2.Electrodermal activity (EDA) signals*

Electrodermal activity (EDA), also known as galvanic skin response, are signals that represent the flow of current between two points of skin contact at which an electrical potential is applied. EDA signals represent properties of the skin that are regulated by changes in sweat glands' secretion, which are controlled by the sympathetic nervous system; sweat secretion increases with increments in emotional arousal. As a result, EDA is considered a good indicator of emotional arousal [159]. EDA signals can be sampled at a rate as low as 4 Hz.

The EDA signal is a time series signal with two activity components, called phasic and tonic, with frequency components of interest between 0.05 and 3 Hz. The tonic component is a slow-changing signal, on the scale of tens of seconds to minutes, which is also known as the skin conductance level (SCL). On the other hand, the phasic component, also known as the skin

conductance response (SCR), is typically the component considered in human behavior recognition tasks. EDA signals are usually pre-processed to identify and remove movement and respiratory artifacts [160], [161].

Similar to audio signals, in the automatic processing of EDA signals, windows of time are used to extract features of interest. Because EDA signals are slower changing signals than audio signals, the window size used to extract EDA features can vary from 5 seconds to 1 minute. **Table 4** describes all the identified EDA features used in the reviewed literature. We grouped the EDA features per category: raw EDA features, SCR features, SCL features, frequency features, and coupling indexes. General information on their definitions can be found in the references listed in **Table 4**. In unimodal systems specifically, EDA signals have been used in the recognition of personality factors and aspects of social interactions; and have been consistently processed using coupling indexes.

***Table 4.*** *EDA features found in the reviewed literature associated with human behavior effector classes. © 2021, IEEE.*

| *Feature categories* | *E* | *PF* | *SI* |
|---|---|---|---|
| **Raw EDA features:** # of local minima, # of local maxima, derivatives, non-stationary index & statistics* | [110], [113], [237] | - | [163] |
| **SCR features:** # of peaks, peak amplitude, rise time, recovery time, peak duration, zero-crossing rate of slow response (0-2.4Hz), & statistics* | [102], [110], [113], [182], [184], [194], [237], [254] | - | [165] |
| **SCL features:** Zero-crossing of very slow response (0-0.2Hz) & statistics* | [110], [181], [194], [237], [254] | - | - |
| **Frequency features:** Spectral power coefficients & statistics* | [110], [113], [237] | - | - |
| **Coupling indexes:** Pearson's correlation coefficient (PCC), signal matching, instantaneous derivative matching (IDM), directional agreement (DA), Fisher's z-transform of the PCC, single session index (SSI) | - | [162] | [163]–[165], [183], [267] |

Note. * Statistics include mean, std, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

*Personality factors* – Empathy has been one of the personality factors monitored using EDA signals. Slovák et al. [162] studied the monitoring of empathy in dyads. Raw EDA signals were first smoothed using a rectangular smoothing algorithm and then uniformly scaled based on a running minimum and maximum value taken from each participant from which data was collected. Using a 15-second window with a moving rate of 1 second, signals from pairs of individuals were combined using a Pearson correlation algorithm. In addition, the single session index (SSI), which "represents an index of synchrony over a longer period of time and is calculated as the natural logarithm of the ratio of the sum of positive synchrony divided by the sum of negative synchrony over the specified time," [162] was then computed for the entire recording section (4 minutes). It was concluded that high emotional engagement of individuals in the conversation was consistently associated with high EDA synchrony. On the other hand, low emotional engagement was associated with moments of inconsistency or fluctuating EDA synchrony.

*Social interactions* – In addition to Pearson's correlation coefficient (PCC), other physiological coupling indices that have been found in the literature are signal matching, instantaneous derivative matching (IDM), directional agreement (DA), and Fisher's z-transform of the PCC. In the area of collaboration, a regression analysis showed that out of the five coupling indices, IDM and DA were good predictors of collaborative behavior [163]. Haataja et al. [164] presented an analysis of synchronicity that, first, calculates the average slope of an EDA signal in a 5-second window and then calculates the PCC between EDA signals of two individuals using a moving 15-second window. Similar to the case of empathy, the SSI was calculated but using a window of 2 minutes. Results indicated that physiological synchrony does occur during collaborative learning at a statistically significant level. Because the analysis was performed offline, resulting moments of synchrony could not be correlated to specific monitoring instances.

However, results do suggest that physiological synchrony might be a relevant condition when joint understanding is better built within groups. In an effort for studying the dynamics of collaboration related to the degree of physiological activation of triads, Pijeira-Díaz et al. [165] calculated the number of peaks per minute in SCR signals using a moving window with a window width of 1 minute and a moving step of 250ms, and then calculated the arousal DA as a measure of the synchrony degree. Results showed that most of the time participants were at different arousal levels, but when they were in synchrony it was mostly in the low arousal level. Although results were not correlated with specific instances, the authors showed the potential of using arousal DA to characterize collaborative behaviors.

*3.2.2.3.Electroencephalography (EEG) signals*

Electroencephalography (EEG) signals represent the electrical activity of the brain. Systems that record EEG signals can have as few as one electrode channel to as many as 256 channels. The placement of EEG electrodes along the scalp is of great importance. Thus, their placement adheres to international standards such as the 10/20 system (also known as International 10/20 system) [166], 10/10, or 10/5 systems [167], the last two also known as the Modified Combinatorial Nomenclature (MCN). These standards aim to standardize the exact position of each electrode and assign names to each of them to facilitate the identification of the brainwave location that may serve a specific brain function. For example, in the area of emotion recognition, specific electrode positions are of interest. T3 and T4, electrodes placed in the temporal lobe regions, are found to be near emotional processors. P3, P4, and Pz, electrodes placed in the parietal brain region, are located near sources that reflect activities of perception and differentiation. While frontal lobe electrodes (i.e., F3, F4, F7, F8) have proximity to sources of emotional impulses and have been

***Table 5**. EEG features found in the reviewed literature associated with human behavior effector classes. © 2021, IEEE.*

| *Feature categories* | *E* | *SI* |
|---|---|---|
| **Time domain features:**<br>Power, derivatives, Hjorth features (activity, mobility, complexity), non-stationary index, fractal dimension, higher order crossings (HOC), & statistics* | [102], [173], [280] | - |
| **Frequency domain features (per band):**<br>Energy spectrum (ES), power spectrum, power spectral density (PSD), differential entropy (DE), rational asymmetry (RASM) of DE features in a channel pair, differential asymmetry (DASM) of DE features in a channel pair, differential caudality (DCAU) between DE features, higher order spectra (HOS), & statistics* | [113], [172]–[174] | [195] |
| **Time-frequency domain features:**<br>Hilbert-Huang spectrum (HHS), discrete wavelet coefficients (DWC) | [173] | - |

Note. * Statistics include mean, std, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

used for emotion recognition [168], [169]. EEG signals are typically sampled at a rate of ~256Hz but can be sampled at a lower rate depending on the signal components of interest.

As EEG signals have a low signal-to-noise ratio and are prone to muscle movement artifacts [170], pre-processing of these signals includes filtering and signal inspection for artifact removal. Before extracting features, typically a window function (i.e., Hamming window, etc.) is applied to each window of time or frame of data to reduce signal discontinuity in order to avoid spectral leakage. Windows of time are of at least 1 second in size. **Table 5** describes all the identified EEG features used in the monitoring of human behavior. We grouped the EEG features per category: time-domain features, frequency-domain features, and time-frequency domain features. General information on their definitions can be found in the references listed in **Table 5**. Traditionally, EEG signals have been analyzed using event-related potential (ERP) features. However, when EEG signals are analyzed based on identified ERPs, an event (or trigger) needs to be identified and then features describing the response to that event are extracted [171]. This approach is not suitable for real-time implementation since it is unknown when an "event" will happen. On the

other hand, when EEG signals are analyzed using either time, frequency, or time-frequency domain features, EEG signals are first divided into frequency bands containing slow, moderate, and fast brainwaves that are associated with specific brain states (i.e., sleep, relaxed, and alert, among many others). These frequency bands are delta band (1-4Hz), theta band (5-8Hz), alpha band (9-12Hz), beta band (13-25Hz), and gamma band (>25Hz). However, the exact frequency values used to extract the frequency band can vary across researchers by 1 or 2 units of Hz per band. Typically, features are extracted specifically per frequency band. In unimodal systems, EEG signals have been used mostly for the recognition of individual emotions.

Duan et al. [172] extracted frequency domain features in five frequency bands from signals recorded from a 62-channel electrode cap to classify positive or negative emotional states of the individuals participating in their study. All features used were smoothed using a linear dynamic system (LDS) approach. They found that emotional states relate to EEG signals in the gamma band more closely than other frequency bands and that using differential entropy (DE) as a feature provides better results than using more traditional features such as energy spectrum (ES). Likewise, Jenke et al. [173] evaluated different time, frequency, and time-frequency feature sets from signals recorded from a 64-channel electrode cap. Using feature selection methods, it was concluded that features such as power spectrum, higher order spectra (HOS), Hilbert-Huang spectrum (HHS), and discrete wavelet coefficients (DWC) computed from beta and gamma bands were better at classifying emotions. Zheng et al. [174] investigated, not just the frequency domain features and critical frequency bands for the recognition of three emotions (positive, neutral, and negative), but also the performance of a combination of four, six, nine, and 12 channels in the recognition of the three emotions. They concluded that DE performed better as a feature when compared to power spectral density (PSD), differential asymmetry (DASM), rational asymmetry

(RASM), and differential caudality (DCAU). In addition, as noted in previously discussed works, they also confirmed that beta and gamma oscillation of brain activity are more related to emotion processing than other frequency bands. Using a weight distribution of a trained deep belief network (DBN), the 12 channels that collect the most emotional information are FT7, FT8, T7, T8, C5, C6, TP7, TP8, CP5, CP6, P7, and P8 (named based on the MCN system). If reduced to four channels, they found them to be FT7, FT8, T7, T8, wherein the 10/20 system, T7, and T8 are T3 and T4, respectively.

*3.2.2.4.Electrocardiogram (ECG) signals*

Electrocardiogram (ECG) signals represent the electrical activity of the heart. Frequency components of interest in ECG signals are below 20Hz, although a commonly used sampling frequency is of 1kHz. A heartbeat (or cardiac cycle) is associated with ECG signal phases and specific signal characteristics. A complete cardiac cycle is made up of five waves that construct an ECG signal, namely P wave, Q wave, R wave, S wave, and T wave. From those five waves, five signal phases are identified: PR interval, PR segment, QRS complex, ST segment, and QT interval. Each of them is associated with how the electrical signal travels through the heart. For

***Table 6.*** *ECG features found in the reviewed literature associated with human behavior effector classes. © 2021, IEEE.*

| *Feature categories* | *E* | *SI* |
|---|---|---|
| **Time domain features:** Heart rate (HR) (expressed in beats per minute (bpm)), inter-beat interval (IBI) (measured in ms), zero-crossing rate, non-stationary index, heart rate variability (HRV), & statistics* | [110], [113], [180], [184], [194], [237] | - |
| **Frequency domain features:** Spectral power, power spectral density, spectral entropy, derivatives & statistics* | [110], [113], [237] | - |
| **Coupling indexes:** Pearson's correlation coefficient (PCC), Fisher's z-transform of the PCC, weighted coherence | - | [183], [267] |

Note. * Statistics include mean, std, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

the heart rate measurement (or frequency of the cardiac cycle), the QRS complex is the most important signal phase because the instantaneous heart rate is calculated from the time between any two consecutive QRS complexes (R-R interval).

Similar to other physiological signals, ECG signals are prone to noise and artifacts, which are typically tackled at the input of the signal acquisition system [175] or the pre-processing stage. A review of this topic can be found in [176]. Noise and artifact removal of ECG signals is important before feature extraction. **Table 6** describes all the identified ECG features used in the monitoring of human behavior. We grouped the ECG features per category: time-domain features, frequency-domain features, and coupling indexes. General information on their definitions can be found in the references listed in **Table 6**. In unimodal systems, ECG signals are used to monitor an individual's emotional arousal states through parameters such as heart rate (HR) (expressed in beats per minute (bpm)), inter-beat interval (IBI) (measured in ms), and heart rate variability (HRV) [177]–[179]. For example, Quintana et al. [180] used correlation analysis to study how different social conditions affect HRV and its relation to emotional states. They concluded that high levels of HRV during resting state are associated with improved emotion perception, while reduced HRV is associated with impairments in social cognition.

*3.2.2.5.Multi-signal modalities*

Signals from multi-sensor modalities have been used to increase the robustness of human behavior monitoring systems. However, the integration of multiple sensors involves managing inconsistencies in the collected data before feature extraction. Different sensor signals are typically collected using different sampling frequencies, they use different pre-processing methods, and they require different windows of time to extract features. All of these contribute to inconsistencies in the data collected across sensor modalities and present a great challenge for data synchronization,

which is important to achieve robustness in human behavior monitoring systems. Nonetheless, when signal features from two or more sensing modalities are used, the reviewed literature identifies two common methods to combine information: feature-level fusion and decision-level fusion. In feature-level fusion, the features extracted from individual sensors are consolidated into a single feature set. A simple solution to synchronize extracted features at the feature-level fusion is to extract them using the largest window size among the selected sensor modalities and then build a single feature vector. Thus, statistics are commonly employed in the feature extraction process. In decision-level fusion, also called model-level, the decisions from multiple classifiers (usually one classifier per sensor modality) are combined into a common decision. More on the theory of fusion mechanisms can be found in [181]. As performed in the discussion of features from audio, EDA, EEG, and ECG signals, we focus on discussing works performing feature-level fusion and the correlated or best-performing set of features from combined sensor modalities. **Table 7** lists additional sensing modalities used in the reviewed literature together with the type of features that are typically extracted from each of them.

*Emotions* – In [182], a total of five sensors were used for the recognition of four emotions. Features from audio, EDA, EMG, PPG, skin temperature, and RSP signals were extracted. Through a sequential backward selection algorithm, features such as the sub-band spectral entropy from PPG, the number of peaks within 4 seconds in EDA and EMG, and the mean values of the MFCCs in the speech features stood out in the recognition of the four emotions. On the other hand, [99] and [100] made use of audio and movement (from accelerometers and gyroscopes) signals to recognize anxiety levels and other individuals' well-being characteristics, respectively. Both made use of a Pearson product-moment correlation coefficient (PPMCC) analysis to investigate the most relevant features associated with anxiety and well-being. In [99], it was found that at least

***Table** 7. Sensor signal features used in multimodal systems. © 2021, IEEE.*

| *Features per sensor modality* | *E* | *SI* |
|---|---|---|
| **RF sensor:** Raw received signal strength indicator (RSSI) values, duration in time of a RSSI value, mean of measurements from two RSSI RF signals, difference between two RSSI RF signals | [256], [257] | [83], [192] |
| **IR sensor:** Number of detected encounters with another IR sensor, sum of lengths of all encounters, and length of an encounter | - | [88], [89], [92] |
| **Accelerometer, gyroscope, and magnetometer:** Signal energy, energy-entropy, correlation coefficient between axis, pitch, roll, peak value in frequency domain, statistics* | [22], [98]–[100], [103], [132] | [88], [89], [92], [192] |
| **Skin temperature:** Derivatives, spectral power in low frequency bands, PCC, weighted coherence, statistics* | [113], [182], [184], [254] | [183] |
| **Respiration:** Signal energy, derivatives, breathing rhythm, breathing rate, sub-band power spectral, PCC, weighted coherence, statistics* | [113], [182], [184] | [183] |
| **Blood volume pulse:** Mean signal, variance, sub-band power spectral, power spectral density, heart rate, heart rate variability, blood flow, pulse, statistics* | [182], [184], [254] | - |
| **Electromyogram:** Statistics* | [102], [182] | - |
| **Eye-tracker:** Pupil diameter, gaze distance, eye blinking, gaze coordinates, statistics, coupling indexes | [113] | [183] |
| **EOG:** Blink rate, blink amplitude, power of blink amplitude, statistics* | - | [195] |

Note. * Statistics include mean, std, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

brightness and MFCC5 from speech, and std of the axis of gyroscopes and their peak value in the frequency domain were highly correlated with the degree of anxiety of the individuals in the study. Likewise, [100] found that the formants, energy, entropy, and brightness features from audio signals and both time and frequency domain features from accelerometers and gyroscopes were strongly correlated with aspects of mental health.

*Social interactions* – Gips and Pentland [88] and Laibowitz et al. [89] used three sensors for the recognition of interest during a social encounter. Initially, a 15-dimensional feature vector was constructed per dyad encounter with features from accelerometers, microphones, and IR sensors. Based on a correlation analysis, the six highest ranked encounter features for the recognition of interest were: std of accelerometer measurements in the x-axis and y-axis, mean and std of average audio signal amplitude, mean average audio difference between averaged readings, and std of the difference between the average amplitude and the average difference. In the use of a combination of physiological signals, Pun et al. [183] used a total of five sensors for the recognition of collaborative behaviors. Coupling features from EDA, ECG, eye-tracker, skin temperature, and RSP signals were extracted. Through a fast-correlation-based filter with mean squared linear regression, the correlation between the extracted features and the degree of perceived collaboration was determined. Coupling features were calculated using the signals from dyads in an interaction. From the physiological signals, the coherence of the IBI in the very low frequencies (0.003Hz-0.05Hz) and the low frequencies (0.05Hz-0.15Hz) were correlated with aspects of collaboration. While from eye-movement signals, the number of times participants looked at the same place at the same time and the number of times participants looked at the same place within a ±6 second window were correlated with collaborative behaviors. Related to group cohesion, Zhang et al. [92] made use of a wearable sociometer badge with accelerometer, microphone, and an IR sensor to measure cohesion at an individual level and a group level. Using Pearson correlation coefficients, it was found that at the individual level, the mean movement energy was positively correlated with cohesion task. At a dyadic level, the correlation of vocal activities was also positively correlated with cohesion task.

*3.2.2.6.Analysis and Discussion*

To eliminate redundant information and optimize algorithms for real-time implementation, it is important to perform correlation analysis or feature selection to analyze the contribution of signal features in the recognition of a human behavior effector. As noted in **Table 3- Table 7**, a wide range of features from different sensor modalities have been employed for the recognition of human behavior effectors. Although not all works referenced in the tables performed correlation analysis or feature selection on extracted signal features, significant consistency exists among the best-found features to be used in recognizing emotions and those to be used in recognizing personality factors and social interactions.

From the agglomeration of references in **Table 3 - Table 7**, one can observe that the most common features for recognizing emotions are: prosodic, cepstral, voice quality, and frequency spectrum coefficients from audio signals; SRC features from EDA signals; frequency domain features per frequency band from EEG signals; and time domain features from ECG. More specifically, from prosodic features of audio signals, features related to voice pitch appear to greatly contribute to the recognition of positive and negative emotions (i.e., emotional valence levels). From EEG signals, the DE feature extracted from the gamma frequency band has also proven to be effective in the recognition of positive and negative emotions. Moreover, from ECG signals, the HRV, which can also be determined from PPG signals, has been found to be a good indicator of emotional valence, emotional arousal, and emotion perception. On the other hand, features from sensor signals used in multi-signal modalities such as Std Dev of gyroscope's axis values and their peak value in the frequency domain have been found to be correlated with anxiety levels.

In the case of personality factors and social interactions, from audio signals, prosodic and conversational features are the most commonly used. More specifically, from prosodic features, voice pitch has proven to greatly contribute to the recognition of extraversion, dominance, and emphasis during meetings. On the other hand, conversational features such as speaking rate and speaking length have proven to contribute to recognizing cooperative meetings, in addition to extraversion and dominance. In general, speaking length and speaking rate are also attractive for real-time use because of their low computational complexity and fast computation. For social interactions alone, other commonly used features found to be relevant in the recognition of social interaction elements, such as collaboration and cohesion, are coupling indexes from EDA signals; distance between individuals and duration of the encounter obtained from IR sensor signals; eye-movement related features from eye-tracker sensor signal; and Std Dev features of accelerometer measurements in the x-axis and y-axis.

To date, analyses of features' contribution to the recognition of human behavior effectors come from works on unimodal systems and less so from works in multimodal sensor systems. This could, arguably, be due to the large number of works in unimodal sensor systems. Still, from observation, the most common sensor signals' combinations used in multi-sensor modalities include microphones with physiological sensors and/or movement and proximity sensors, and combinations of physiological sensors. However, further research is encouraged in the evaluation of the best feature or features to be used in multi-sensor modalities for the recognition of human behavior effectors. As sensor features are identified as contributing to the recognition of more than one human behavior effector, more optimized and robust systems could be designed. For example, voice pitch, from audio signals, has been observed to be a good contributor to the recognition of

all human behavior effectors. Thus, using voice pitch when designing a system to recognize multiple human behavior effectors could help increase system efficiency.

### 3.2.3. Computational Models for Human Behavior Recognition

Based on the features extracted from sensor signals, computational models are trained and used to predict or classify human behavior. Therefore, the performance of computational models can depend on the set of features provided. Likewise, the effectiveness of signal features can also depend, in part, on the type of computational method used to evaluate the features' contribution.

The two principal types of computational models employed in the human behavior recognition literature are classification and regression models. Classification models focus on recognizing discrete or categorical classes, while regression models focus on predicting continuous numerical values. The use of a computational model is application dependable. For example, the problem of emotion recognition can be treated as one with categorical values (e.g., happy, sad, neutral) or as one with continuous numerical values (i.e., reflecting levels of arousal and valence based on a numerical scale).

An analysis of reported computational methods used in the monitoring of human behaviors was performed as follows. First, reviewed literature was grouped based on their use of classification and regression models. Then, within each of the two model groups, different types of models and the number of predicted or classified classes were summarized to illustrate the design space employed in the literature. This summary analysis allowed us to make observations regarding the most commonly used computational models, which are presented at the end of the section. Specifically related to classification models, we analyzed and compared their accuracy values to define the state-of-the-art system performances that may help drive the future design of real-time human behavior monitoring systems.

***Table 8***. *List of classification models used in the reviewed literature. © 2021, IEEE.*

| *Models* | *E* | *PF* | *SI* |
|---|---|---|---|
| **Support Vector Machine (SVM):** Classic SVM, adaptive SVM, and incremental SVM | [102], [113], [142], [172], [174], [184], [186], [187], [190] | [53], [153] | [123], [124] |
| **k-Nearest Neighbor (k-NN)** | [141], [172], [174], [189], [190] | - | - |
| **Naïve Bayes (NB)** | [102], [173], [185], [190] | - | [123] |
| **Log-likelihood ratio** | - | - | [124] |
| **Logistic regression** | [174] | [53] | [92] |
| **Linear regression** | - | - | [89] |
| **Linear Discriminant Analysis (LDA)** | [141], [182] | - | - |
| **Decision and Regression Tree** | [102] | - | [155], [192] |
| **Random Forest (RF)** | [182] | - | - |
| **Hidden Markov Models (HMMs)** | [146] | - | [156], [191] |
| **Gaussian Mixture Model (GMM)** | [142], [189] | - | - |
| **Neural networks:** Convolutional NN, Multilayer perceptron (MLP), self-organizing map, deep belief networks (DBNs) | [142], [143], [174], [188] | - | - |
| **Partial Least Squares-Discriminatory Analysis (PLS-DA)** | [190] | - | - |
| **Latent Dirichlet Allocation model** | - | | - |
| **Sets of rules:** Rule-based, rank-level fusion, collective classification approach | - | | - |
| **Clustering models:** k-means | [99] | - | - |

*3.2.3.1.Classification models*

In general, based on the reviewed literature, classification models have been widely used in emotion, personality factors, and social interaction recognition tasks. The reviewed literature presents variations in the number and type of classes that classification models are trained to recognize and variations in the classification models being employed. A summary of the classification models employed in the reviewed literature associated with human behavior effector classes can be found in **Table 8**.

*Emotions* – Lee et al. [141] investigated the performance of a k-Nearest Neighbor (k-NN) and a Linear Discriminant Analysis (LDA) classifier to predict two emotion classes (negative and non-negative) when using audio data from males and females separately. While for female data, LDA consistently performed better than k-NN, for male data there were cases in which k-NN performed better than LDA. Gu et al. [99] made use of a K-means classifier to recognize high anxiety and low anxiety using features from audio signals. The authors obtained 72.73% of performance accuracy by using just two features: brightness and MFCC. In this line, Sahoo and Routray [146] trained Hidden Markov Models (HMMs) to detect aggression and calmness also using audio signals. By using pressure distribution features a performance accuracy of 93.5% was achieved. Later, by using the same features, the authors trained an HMM to recognize four emotion classes (anger, boredom, happy, and neutral) achieving an 80% overall recognition accuracy. On the other hand, using EEG signals, Duan et al. [172] evaluated two classifiers, a Support Vector Machine (SVM) and a k-NN to predict two emotion classes (positive and negative emotion). In general, SVM outperformed k-NN achieving a performance accuracy of up to 86.69%. Using a multimodal sensor system, Chanel et al. [184] investigated the performance of Random Forest (RF) and SVM classifiers in predicting emotional and non-emotional moments using audio and physiological (EDA, ECG, BVP, skin temperature, and respiration) signals during social interaction. The authors investigated the performance of decision-level fusion by combining the output scores of classifiers trained on signal features from each individual in the interaction. Regardless of the classifier type (RF or SVM), it was found that by adding emotional information from all individuals in the interaction, the emotional response of one individual can be predicted with higher accuracy than just using the classification model from the individual of interest.

Related to the recognition of three emotion classes, Zheng and Lu [174] investigated the performance of four classifiers in predicting positive, neutral, and negative emotion classes using EEG signals. The four classifiers were deep belief networks (DBNs), SVM, logistic regression, and k-NN with resulting average classification accuracies of 86.08%, 83.99%, 82.70%, and 72.60%. However, the highest reported accuracy of DBNs was by taking EEG features from 62 channels, whereas the highest reported accuracy of SVM was 86.65% when taking EEG features from 12 channels.

Related to the recognition of four emotion classes, Kim [182] trained a LDA classifier in combination with a sequential backward selection to predict low and high arousal and high and low valence using audio and physiological (EDA, ECG, BVP, EMG, skin temperature, and respiration) signals. The author trained a model for each subject (three in total) and a subject-independent model achieving an average accuracy of 78.67% and 55%, respectively. Similarly, Vogt et al. [185] trained a Naïve Bayes (NB) classifier to predict four emotion classes (joy, satisfaction, anger, and frustration) but just using audio signals. The authors trained subject-dependent models for 29 subjects, achieving accuracy values that ranged from 24% to 74%, with an average of 55%. They also trained a subject-independent model using data from 10 subjects achieving a 41% recognition accuracy. Their use of NB was motivated by its fast computation and ability to take high-dimensional feature vectors. However, Vogt et al. suggested that a more accurate classifier would be an SVM and that with a vector size under 100 features, it could be suitable for real-time implementation. In this line, using EEG and eye gaze signals, Soleymani et al. [113] trained SVM subject-dependent models to predict four emotion classes (high and low arousal and high and low valence). Classification accuracies for arousal and valence were 67.7% and 76.1%, respectively. Using audio signals, Abdelwahab and Buso [186] investigated the use of

two modified versions of SVM to classify the same four emotion classes (high and low arousal and high and low valence). They trained an adaptive SVM model and an incremental SVM model, which aims at maintaining or improving their classification performance even under mismatched training and testing conditions. The authors concluded that both methods provide similar performance, but a precise accuracy value was not reported. On the other hand, Wu and Liang [142], also using audio signals, trained three types of models, Gaussian Mixture Model (GMM), SVM, and a multilayer perceptron (MLP) to predict four emotion classes (neutral, happy, angry, and sad). A Meta Decision Tree (MDT) was then used for classifier fusion, achieving an overall performance accuracy of 80%. However, the results from SVM alone were close to the results of MDT fusion classifier because the MDT is a classifier selection approach instead of a combination of all classifiers. Moreover, Cen et al. [187] trained a SVM model for offline and real-time recognition of the same four emotional states (neutral, happy, angry, and sad) also using just audio signals. Their results showed a 90% and 78.78% classification accuracy for automatic offline and real-time emotion recognition, respectively. In addition, Girardi et al. [102] investigated the performance of SVM, J48 (algorithm based on decision trees), and Naïve Bayes (NB) on predicting low and high arousal and high and low valence by using physiological signals such as EDA, EEG, and EMG. Results showed that SVM outperforms the other classifiers and that EEG signal features alone provided the best performance accuracy for valence classification, while EEG+EDA performed the best for arousal classification. On the other hand, using a Convolutional Neural Network (CNN) to predict the same four previously mentioned emotion classes, Rajak and Mall [188] using audio signals, specifically, MFCC features achieved a classification accuracy of 76.2%.

In the recognition of more than four emotion classes, Jenke at al. [173] trained NB subject-dependent models using EEG signals to predict five emotion classes (happy, curious, angry, sad, and quiet) achieving a performance accuracy of 36.80%. Later, Lanjewar et al. [189] made a comparison between the performance of a GMM and a k-NN to predict six emotion categories using audio signals. In general, their results showed that the GMM performed better than the k-NN model with 66% and 52% of classification accuracy, respectively. However, the speed of computation is faster for the k-NN classifier than for GMM, which makes it attractive when time constraints are critical to consider, like for real-time applications. The computational time of GMM increased when the number of features increased in the training phase. However, it was noted that GMM was better at predicting angry and sad emotion classes, while k-NN performed better at predicting happy as well as angry emotion classes. Also using audio signals, Balti and Elmaghraby [143] implemented a self-organizing map with a response integration approach to predict seven emotion classes (anger, boredom, disgust, anxiety/fear, happiness, sadness, and neutral), achieving a 70.86% performance accuracy. Likewise, Jing et al. [190] investigated the performance of SVM, k-NN, NB, and Partial Least Squares-Discriminatory Analysis (PLS-DA) in also predicting seven emotion classes (sad, joy, fear, surprise, neutral, anger, and disgust) by using audio and EGG signals. The authors evaluated the models using acoustic features only and combined feature sets independently for males and females. However, the results consistently showed that SVM got a higher average emotional recognition accuracy for both genders when compared to the other classification models, with a classification accuracy of ~72%.

*Personality factors* – Using audio signals, Jayagopi et al. [153] trained an unsupervised classification model and an SVM model to predict the most-dominant person and the least-dominant person in a group conversation. The unsupervised model computed either the largest or

smallest accumulated value of each extracted feature, depending on whether the goal was to predict the most dominant or the least dominant person. In addition, two SVM models were trained. One to predict the most and the non-most dominant person in the group conversation, and another one to predict the least and the non-least dominant person in the same group conversation. Their results showed that SVM performed better than the unsupervised model in predicting the most-dominant person, being their best performance accuracies of 91.2% and 85.3%, respectively. On the other hand, both models performed the same when predicting the least-dominant person with an 83.9% accuracy. The same author, in [147], also using audio signals, trained a Latent Dirichlet Allocation model to predict three classic leadership styles: autocratic, participative, and free-rein, achieving a 79.20% classification accuracy. Likewise, Sanchez-Cortes et al. [154] evaluated four approaches using audio signals to infer an emergent leader in a group. The four approaches were a rule-based approach (search for the person with the highest feature value in a group and select that as the leader), a rank-level fusion (extension of rule-based that handles fusion of multiple features), SVM, and a collective classification approach. Results showed that the rank-level fusion provided the best performance with 72.5% of accuracy. It also performed the best in identifying perceived dominance with 65% of accuracy. Related to personality traits, Mohammadi and Vinciarelli [53], also using audio signals, evaluated the performance of a logistic regression and an SVM in predicting high and low extraversion, agreeableness, conscientiousness, neuroticism, and openness. Results suggest that logistic regression performs better than SVM in predicting conscientiousness and neuroticism with a 72.55% and 66.10% classification accuracy, respectively. On the other hand, SVM performed better than logistic regression in predicting extraversion, agreeableness, and openness with 73.45%, 63.10%, and 52.75% classification accuracy, respectively.

*Social behaviors* – Similar to the previous sub-sections, most of the literature reported here has trained their models with features from audio signals. Using prosodic and conversational features, Hillard et al. [155] trained a Decision tree (DT) classifier to predict moments of agreement and disagreement during meetings. They achieved an overall performance accuracy of 64%. Similarly, Jayagopi et al. [124] evaluated a log-likelihood ratio model and an SVM model in classifying conversational group dynamics into cooperative-type or competitive-type. Using an SVM with a quadratic kernel, 100% classification accuracy was obtained.

In line with meetings, McCowan et al. [191] trained an HMM to predict eight meeting actions (monologues from individuals (total of 4), note-taking, presentation, discussion, and white-board talk) achieving an 83.9% classification accuracy. Also using HMM, Gatica-Perez et al. [156] predicted two levels of interest, high and low, during a meeting. By training an HMM with a feature vector constructed from calculating the mean of the features from all the subjects in the interaction, 84% recall and 63% precision performance measures were achieved, while by just concatenating the features from all the subjects an 80% recall and 58% precision performances were achieved. Also investigating levels of interest, but during social encounters, Laibowitz et al. [89] trained a Linear Regression model using accelerometer signals, in addition to audio signals. Their model achieved an 86.2% classification accuracy.

Related to cohesion, Hung and Gatica-Perez [123], evaluated the classification performance of an NB model and an SVM model when predicting high and low cohesion using audio signals. However, both classifiers showed similar classification performances, achieving up to 90% accuracy. Moreover, Zhang et al. [92] employed a logistic regression classifier to recognize between task cohesion and social cohesion among dyads by using audio, accelerometer, and IR signals. Their approach achieves 80.30% and 64.62% classification accuracy when predicting task

cohesion and social cohesion, respectively. On the other hand, Katevas et al. [192] used a XGBoost regression tree classifier to detect interactive groups of various sizes (node and group level) by using an accelerometer, gyroscope, and RF signals, achieving a 94% performance accuracy.

*3.2.3.2.Regression models*

In general, works that have made use of regression models are focused on the prediction of emotions and social interactions. Regression models have been found to be particularly attractive when it is of interest to predict or recognize levels of emotional arousal, emotional valence, collaboration, and vigilance on a continuous numerical scale. A summary of the regression models employed in the reviewed literature associated with human behavior effector classes can be found in **Table 9**.

*Emotions* – Wöllmer et al. [193] introduced a framework for continuous monitoring of arousal and valence levels using audio signals. The authors evaluated two regression models: Support Vector Regression (SVR) and a long short-term memory recurrent neural network (LSTM-RNN). Their results showed that LSTM-RNN performed better than SVR at predicting arousal levels with a Mean Squared Error (MSE) performance measurement of 0.08 and 0.10, respectively. On the other hand, both regression models performed the same at predicting valence levels with an MSE

***Table 9.*** *Regression models found in the reviewed literature associated with human behavior effector classes. © 2021, IEEE.*

| *Models* | *E* | *SI* |
|---|---|---|
| **Support Vector Regression (SVR)** | [110], [193], [194] | - |
| **Regression Trees** | - | [183] |
| **Least Squared regression** | - | [183] |
| **Neural networks:**<br>Long short-term memory recurrent neural network (LSTM-RNN), Feed-forward (FF), Bilateral long short-term memory (BLSTM) | [110], [193], [194] | - |
| **Structured regression model:**<br>Continuous conditional neural field (CCNF), continuous conditional random field (CCRF) | - | [195] |

of 0.18. Ringeval et al. [110] used a hybrid decision fusion based on SVR with a lineal kernel and Neural Networks (NN) to recognize arousal and valence emotional levels based on data from audio, EDA, and ECG sensors obtained from the AV+EC 2015 database [110]. For NN, they explored three types of architectures: feed-forward (FF), LSTM, and bilateral long short-term memory (BLSTM). The authors found that SVR performs best on the audio features for valence prediction with a 0.069 Concordance Correlation Coefficient (CCC) and NN performs best on EDA features for arousal with a 0.79 CCC. Moreover, FF provided the best performance for EDA features. Their hybrid decision-fusion method achieved the best arousal prediction with a 0.228 CCC and 0.173 RMSE performance metric using audio features while achieving their second-best valence prediction performance with a 0.195 CCC and 0.119 RMSE using EDA features. However, when the authors employed decision-fusion on their multi-modal data, their results improved achieving 0.444 CCC and 0.164 RMSE for arousal prediction, and 0. 382 CCC and 0.113 RMSE on valence prediction, demonstrating the value of a multi-modal approach. Also using SVM and LSTM models, Brady et al. [194] used a decision-level approach to predict these arousal and valence levels. The authors trained an SVR model for audio signals and an LSTM for physiological signals (EDA and ECG) and combined their decisions using a Kalman filter framework. They found that models for ECG and EDA provided significant performance improvements for valence prediction, obtaining 0.364 CCC and 0.117 RMSE for models trained with HR and HRV data and 0.177 CCC and 0.124 RMSE for EDA data.

*Social interactions* – Contrary to emotion recognition, which mainly focuses on predicting arousal and valence levels, in the area of social interactions, the target classes vary greatly from one work to another. Chanel et al. [183] used Bag of Regression Trees (BRT) and Least Squared regression with a fast-correlation-based filter (FCBF LS) to predict collaborative behaviors (i.e.,

degree of conflict, confrontation, emotional management, etc.) based on data from EDA, ECG, skin temperature, respiration, and eye-tracker. Physiological and eye-tracker data were treated separately, and different regression models performed differently based on the sensor data modality and the targeted collaborative behavior. For example, the FCBF LS model provided the lowest RMSE value, with a 0.44 RMSE performance value, when using eye-tracker data to predict the degree of convergence in a group of people. However, the BRT model performed better at predicting confrontation using physiological signals when compared to the FCBF LS model. On the other hand, Zheng and Lu [195] employed an SVR with a radial basis function to estimate the level of vigilance based on data from EEG and EOG. The authors introduced a continuous conditional neural field (CCNF) and a continuous conditional random field (CCRF) to the design of their vigilance estimation model with the goal of incorporating the temporal dependency present in vigilance. It was demonstrated that the fusion of multimodal sensor features improves model performance, achieving 0.09 RMSE performance value, compared to features from a single modality that achieved 0.12 and 0.13 RMSE performance values for EOG-based and EEG-based methods, respectively. In addition, the temporal dependency-based models demonstrated to also enhance vigilance estimation.

*3.2.3.3.Analysis and Discussion*

A wide range of computational models, as noted in **Table 8** and **Table 9**, have been employed for the recognition of human behavior effectors. To deeply analyze the use of classification models, performance metrics related to the accuracy values reported per classification model are organized by effector class and presented in **Figure 10**. From the agglomeration of references in **Table 8**, it can be observed that SVM has been the most popular classification model used for the recognition of human behavior effectors followed by k-NN and NB. In addition, from **Figure 10**, it can be

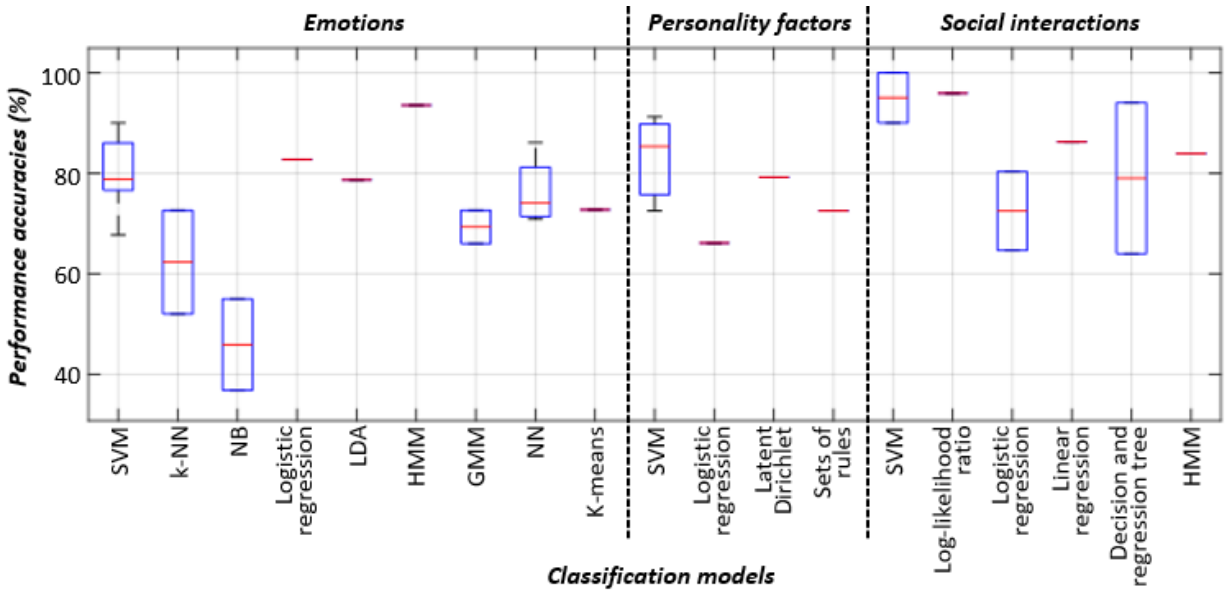


***Figure 10.*** *Summary analysis of reported performance accuracies of classification models per human behavior effector groups. The central mark indicates the median accuracy, and the top and bottom edges of the box indicate the 75$^{th}$ and 25$^{th}$ percentiles, respective. The whiskers extend to the most extreme accuracy values not considered outliers These results were obtained by analyzing data from the references in Table 8. © 2021, IEEE.*

observed that SVM provides one of the highest levels of accuracy across all effector classes. On the other hand, k-NN and NB have been specifically used in emotion recognition, and although they follow SVM in popularity, their levels of accuracy are among the lowest across all other employed classification models. An important factor to consider when evaluating the performance of classification models is the number of classes that they are trained to predict. For example, in **Figure 10** under emotions, HMM reports the highest accuracy but classifies just two classes, whereas the accuracies reported for SVM are for models trained to recognize from two to four classes. Moreover, the accuracy and complexity of these computational models vary depending on 1) the number of classes that they are trained to predict and 2) the quantity of information (number of features) that they take to accurately predict a class. Both of these factors are also critically important when considering real-time implementations. The four classification models that have

been trained to recognize human behaviors in real time are k-means [99], HMM [146], NB [185], and SVM [123], [187].

Our review has identified that classification models have been more widely employed than regression models. This indicates that the problem of identifying human behaviors using sensor technologies has generally been treated as a "discrete problem" rather than a continuous one. However, it has been argued that human behaviors change gradually, on a continuous scale rather than in discrete states [196]. Thus, the use of continuous numerical values for the recognition of such behaviors may be preferred. To date, the use of regression models to treat behavior recognition as a continuous case (i.e., using continuous numerical values to recognize or predict a behavior) has varied with the behavior effector class being monitored; SVR and NN regression models have been common for emotion recognition, while regression trees, least-squared regression, and structured regression models have been used in the prediction of aspects of social interactions. A general observation related to regression models is that, although these models have been employed in the automatic recognition of human behaviors, so far, they do not appear to have been used in real time. However, as regression models are attractive for the prediction of continuous classes, further study of these models for the real-time prediction of human behavior effector classes is highly encouraged. In addition, although there are a limited number of works employing regression models to predict a behavior effector class and different performance metrics (MSE, RMSE, CCC) have been used, hybrid decision-fusion appears to achieve the best prediction performances.

In general, different computational models tend to fit feature sets from different sources in unique ways. Decision-level fusion methods, as described in Section 3.2.2.5, combine decisions from multiple computational models into a common decision, and their use should become more

popular as the number of sensor modalities within systems increases. Decision-level methods such as a set of rules and hybrid decision fusion have started to gain traction in conjunction with classification and regression models, respectively.

Although the number of features is a highly important factor in the training of computational models, nearly half of the reviewed works did not report this value. However, from those works that did report it, the number of features ranges from 1 to ~1000. On average, emotion recognition models tend to be trained with a higher number of features than models for the recognition of personality factors and social interactions, suggesting that emotion recognition systems are more computationally complex. Based on current studies, it is unclear if this computational complexity is linked to the complexity inherent to the personalization of human emotion. Emotion recognition models have also been more widely explored, and their complexity may be an artifact of the relative maturity of those models.

### 3.3.Design of a Multi-Sensor System with a Machine Learning Framework to Monitor Group Consonance using the Rapport Theoretical Model

As mentioned in the introduction of this chapter and Section 3.1.3, this work hypothesizes that (1) in the absence of self-awareness, rapport levels among dyads may decrease, possibly affecting the overall group interaction, and (2) by establishing a framework to monitor components of rapport (i.e., attentiveness, positivity, and coordination), the system could determine both an overall value of group consonance and the component(s) affecting rapport needing attention.

As rapport is established through multiple channels of communication, especially nonverbal, a multi-sensor system with a machine learning framework is needed. Existing social interaction monitoring systems lack accessibility, sensing modalities, or computational capabilities needed to recognize complex social dynamics in real time. Nevertheless, the design of these systems presents

numerous challenges that include sensors' position and wearability, sensors' networking, the integration of information from different sensor modalities, management of different sampling rates and pre-processing methods, use of optimal window length for real-time processing, time-alignment of the collected multimodal sensor signals, and variations in feature formats and extraction, effective feature selection, and computationally efficient but accurate human behavior and social interaction classification models.

### 3.3.1. System Requirements

To design a system that can be used for the study and real-time monitoring of group interactions, in both in-person and virtual environments, the fact that in virtual environments the head area conveys many of the nonverbal messages was considered. This limitation led to evaluating sensor modalities that can be worn on the head while providing human behavior and social interaction insights. In addition, to avoid inducing mobility constraints, the search was limited to wearable and non-invasive sensors.

Because a group is composed of three or more individuals and behavioral information is communicated through various channels, the ability of this system to manage multi-sensor connectivity, data processing, and communication of at least three sensor nodes is required. Each sensor node will be dedicated to collecting behavioral information from a single individual in the interaction. Further, a framework to manage data synchronization and communication across sensor nodes is needed to be able to identify complex social behavior. The multi-sensor system architecture should allow for data recording to permit the design of real-time signal processing and machine learning algorithms. Furthermore, the framework needs to allow the implementation and execution of real-time signal processing and machine learning methods to accomplish (near) real-time monitoring of group interactions.

### 3.3.2. Sensor Selection

To select sensors with the capability of collecting signals reflecting nonverbal messages of interest (including physiological reactions), information from reviewed literature was extracted from **Table 2** and a mapping of nonverbal messages of interest with sensors that can contribute to their detection (as shown in **Table 10**) was created. Cameras were excluded from the mapping because of our interest in designing a real-time human behavior monitoring system that minimizes privacy issues and computational complexity. Based on the potential for wearability on the head and contributions to the detection of most nonverbal messages of interest, the sensors listed in **Table 10** were further analyzed; and a smaller group of sensors were selected as part of the multi-

***Table 10.*** *Mapping of nonverbal messages of interest with sensors that can contribute to their detection. Highlighted in gray are selected sensor modalities for the multi-sensor system of this dissertation project. © 2022, IEEE.*

| | ***Nonverbal messages*** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Sensors of interest*** | Body movement | Body orientation | Back-channel signals | Posture | Facial expressions | Eye gaze | Paraverbal | Interpersonal distance | Gestures | Physiological |
| Microphone | | | | | | | x | | | |
| Accelerometer | x | x | x | x | | | | | x | x |
| Gyroscope | x | x | x | x | | | | | x | x |
| Magnetometer | x | x | x | x | | | | | x | x |
| IR | | x | | | | | | x | | |
| RF | | x | | | | | | x | | |
| Eye tracker | | | | | | x | | | | |
| PPG | | | | | | | | | | x |
| Skin temperature monitor | | | | | | | | | | x |
| EEG | | | | | | | | | | x |
| EGG | | | | | | | x | | | |
| EMG | | | | | x | | | | | |
| EOG | | | | | | x | | | | |

sensor framework. Per the previous review, nonverbal information of interest includes pitch and other prosodic features in speech signals, physiological reactions, and head/body activity.

Accelerometer, gyroscope, and magnetometer were selected for the detection of body movements, orientation, posture, gestures, and back-channel signals (e.g., head nods and headshakes). Accelerometers measure the magnitude and direction of acceleration, gyroscopes measure the angular velocity of rotation, and magnetometers measure the direction and strength of the magnetic field in the local vicinity. Although infrared (IR) and radio frequency (RF) sensors can gather interpersonal distance information, in addition to body orientation, these would not be practical for virtual social environments and thus dropped from further consideration. For the collection of paraverbal communication messages, microphones were selected over electroglottography (EGG) because of their advantages in terms of placement and wearability.

For the detection of physiological responses, photoplethysmography (PPG) and electroencephalography (EEG) were selected because of their information-rich signal content, especially for the recognition of changes in emotional states. EEG-sensor electrodes measure the electrical activity of the brain. The analysis presented in Section 3.2 revealed that electrodes placed in the temporal lobe and the frontal lobe regions of the brain have proximity to sources of emotional impulses. PPG sensors measure the volumetric variations of blood circulation at specific body locations such as the finger, wrist/forearm, forehead, and earlobe [197]. These variations reflect physiological parameters that are linked to the cardiovascular and respiratory systems affected by changes in emotional states. From all those locations, of particular interest are the forehead and the earlobe since a head-mounted sensor system is the aim of this work. Compared to the forehead position, the earlobe is the most frequently used measurement site because this location is not comprised of cartilage, thus they contain large blood supplies [197]. Similar to

ECG, PPG waves can be used to identify regular or irregular heart rate (HR). Using PPG sensors to monitor HR has several advantages when compared to traditional ECG-based systems. PPG sensor systems make use of a simpler hardware architecture, are cost-effective, and only require a single sensor to be in contact with the human body, which simplifies wearability [197]. A typical PPG sensor contains a light source (infrared light emitting diode (LED) or green LED) and a photodetector. PPG uses the photodetector to measure the intensity of reflected light from the tissue, which is then used to calculate blood volume changes. Other sensors listed in **Table 10** with the capability to gather facial expression information and eye gaze were dropped from further consideration as they are mostly used for emotion recognition and attention monitoring, respectively; both human behavior factors could be captured by the six sensor modalities selected.

Commercially available wearable sensor devices containing the selected sensor modalities were researched to be integrated into sensor nodes for the designed multi-sensor system. Wearable sensors needed to contain a long-lasting battery of at least two hours and provide access to an application programming interface (API) to manage sensors' data as needed without proprietary permission from sensor manufacturers. From a variety of available sensor systems (a comprehensive list can be found in [104]), the Shimmer GSR+Unit was selected because of its ability to collect data from an accelerometer, gyroscope, magnetometer, and PPG sensors using an earlobe clip. The Shimmer GSR+Unit also has the capability to collect electrodermal activity (EDA) signals; however, because there was not an optimal way to place the EDA electrodes in the forehead (which is the recommended placement area of the head) [198], this sensor modality was not included in the system design. For EEG, a BrainBit EEG headband was chosen. BrainBit has four EEG dry electrodes, two in the occipital lobe region (O1 and O2) and two in the temporal lobe region (T3 and T4). Compared to other alternatives such as Emotiv [199] and Neuroelectrics

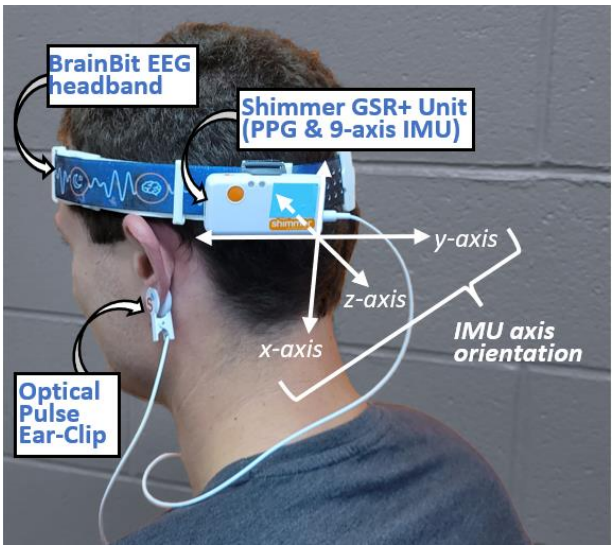


***Figure 11**. Head-mounted wearable sensors selected as part of the multi-sensor framework for the monitoring of social interactions. All selected sensors are shown except for microphone, which taken from the PC used for virtual meetings. © 2022, IEEE.*

Enobio [200] for recording EEG, BrainBit offered electrode positions of interest (Temporal lobe for emotion detection and Occipital lobe for visual attention recognition) and a less obtrusive design. Although Emotiv offers electrodes positioned in the Frontal lobe and integrates inertial movement sensors, combining BrainBit and Shimmer offered a more flexible design in terms of placement. On the other hand, comparing Shimmer with other systems such as Empatica E4 [201] for heart rate, Shimmer offers a higher integration of sensing modalities and flexibility for placement since it can collect PPG signals from the earlobe or the fingers if placed as a wristband. In addition, the Shimmer and BrainBit devices offer a compact design and API support for independent applications. They also use Bluetooth communication which allows the users to move their heads freely. **Figure 11** shows the Shimmer mounted on the BrainBit headband. The multi-sensor framework involves the use of personal computers, which act as system nodes; therefore, it also makes use of the integrated computer microphones for the collection of audio signals.

### 3.3.3. Data Collection Interface and Sensor Data Synchronicity

Data from multiple sensor modalities need to be collected simultaneously and synchronized before further processing. To help with networking and sensor data synchronization, the designed framework makes use of the Lab Streaming Layer (LSL). In addition to handling both the network and the time-synchronization of sensor signals, LSL also allows (near) real-time access to the measured time series as well as optional centralized collection and disk recording of the data. LSL also provides core libraries in various language interfaces and a suite of tools built on top of the libraries [202]. From the already available tools, a recording program and an audio acquisition application were integrated into the system architecture. To allow the collection of data from BrainBit and Shimmer, their respective API tools were combined with the LSL core libraries to build customized data collection applications.

The BrainBit and Shimmer application interfaces were developed using MATLAB 2019b. BrainBit integration into our platform was facilitated by the BrainFlow libraries, designed to obtain, parse, and analyze physiological signals from biosensors such as EEG. The application obtains EEG data from BrainBit at a sampling rate of 250 Hz and a voltage range of ±0.4μV. On the other hand, Shimmer integration was facilitated by the MATLAB API provided by the Shimmer GSR+Unit manufacturer [203]. This custom application obtains Shimmer data at a sampling rate of 128 Hz, pre-filters IMU and PPG sensor data before transmission to the LSL managed network and estimates heart rate based on PPG sensor data. The functionality of the application requires the installation of Realterm Serial Terminal to access the computer terminal to which Shimmer connects. **Figure 12** shows the user interfaces for the custom-built applications. All the resources and MATLAB code needed to establish sensor connection as described in this section are available at https://gitlab.msu.edu/davilasy/sensor-connection-atlas.

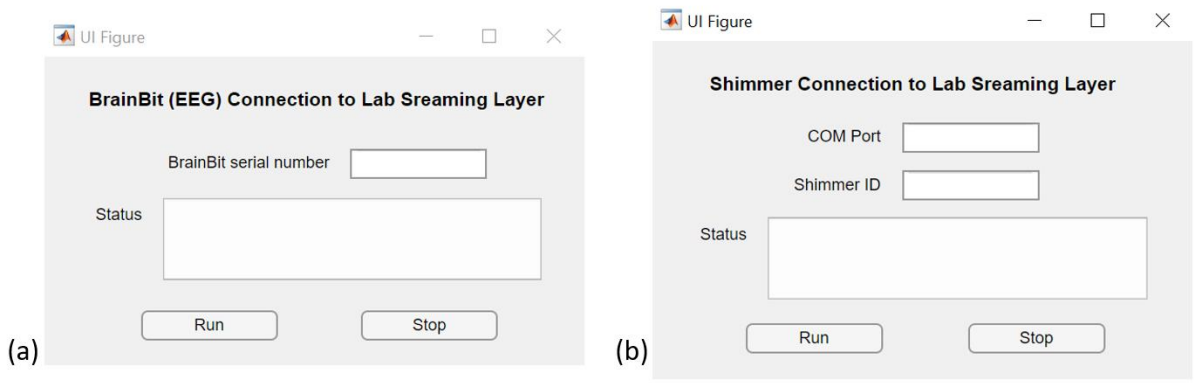


***Figure 12.*** *Graphical user interfaces (GUIs) of the custom-built applications. (a) The BrainBit application takes the serial number of the device to be connected to the corresponding sensor node, use it to establish connection, and assigns collected data to a sensor ID. (b) The Shimmer application takes the Shimmer serial number or ID and the COM port to which it is connected via Bluetooth to the node computer to establish connection and assign ID to the collected data. Both GUIs contain a status window to show connectivity problems or confirm data is being streamed.*

The overall architecture of the multi-sensor system is shown in **Figure 13**. The overall multi-sensor system consists of three sensor nodes, each using the LSL tools and the custom application to collect and synchronize data from all sensing modalities. Sensors are connected to their respective nodes using Bluetooth 5.0. A separate computer acts as a central unit that, when all nodes are connected to the same WIFI network, allows the synchronized collection of data from all nodes. The multi-sensor architecture allows for data storage and the implementation of a machine learning framework, including the extraction of local signal features (related to individual human behaviors) and global signal feature extraction (related to social behaviors) for the classification of group interactions.

### 3.3.4. Proposed Real-time Machine Learning Framework for Sensor Data Processing

An essential part of group interaction monitoring systems is the processing of sensor data to recognize behavior indicators of interest. In this work, a machine learning framework motivated by the goal of monitoring group interactions and informed by the analysis of signal features and computational models presented in Section 3.2 was established. As rapport is considered essential

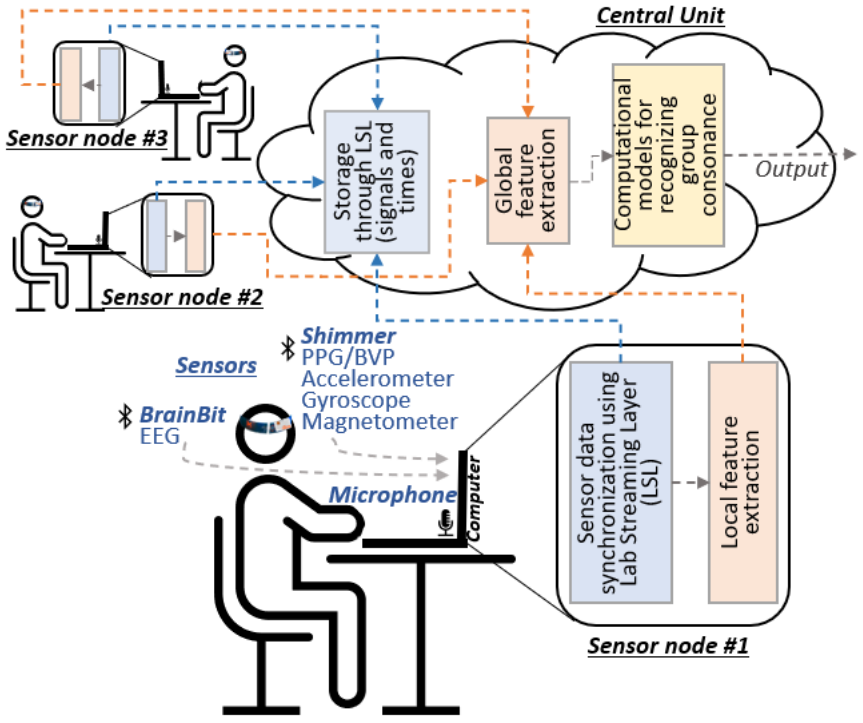


***Figure 13.*** *Overall architecture of the multi-sensor framework for the real-time monitoring of social interactions. The framework consists of three sensor nodes and a central unit. Each sensor node is composed of six sensor modalities. Synchronization, networking, and storage of sensor signals are managed with LSL (https://github.com/sccn/labstreaminglayer). © 2022, IEEE.*

for the quality of group interactions, a machine learning framework was designed considering rapport is modeled as a 3-component paradigm based on the Tickle-Degnan & Rosenthal Theoretical Model [120]. The machine learning framework is primarily composed of three components: (1) signal pre-processing, (2) feature extraction, and (3) training of computational models or, more specifically, classification models.

*3.3.4.1.Signal pre-processing*

Signal pre-processing techniques involve the establishment of adequate data buffer sizes, data window sizes, and filter types for the treatment of signals before and after feature extraction. In this work, data buffers and data windows differ from each other in that data buffers are the chunks of raw sensor data that are taken for extraction of low-level features, wherein data windows are

typically bigger in size than data buffers containing low-level features that will be used to extract higher-level features either at the sensor nodes or the central unit. Sizes of data buffers and data windows vary per sensor modality and require analyzing different sizes for optimal recognition of behavioral cues and real-time processing.

*3.3.4.2.Feature extraction*

As shown in **Figure 13**, the feature extraction process is divided into two layers involving local and global feature extraction. Local feature extraction involves the extraction of features that describe the activity of a single individual and are calculated at the sensor node, whereas global feature extraction involves the extraction of features that describe dyadic or group dynamics and, therefore, requires data from more than one individual. Global features include all features calculated at the central unit.

Local feature extraction is composed of two types of features: (1) low-level features or Type A features and (2) transformed features or Type B features. Type A features are extracted from the data buffers using statistical or signal processing methods. Type B features are extracted from a collection of Type A features being held by a window of data or are higher-level behavior features obtained from a classification model. **Figure 14** illustrates the idea behind the extraction of Type A and B features. Note that global features mainly contain Type B features, which are derived from Type A features calculated from multiple individuals.

At a local level, features will be extracted from all sensor signals using their respective data buffer sizes. Global features will be extracted at a slower rate than local features. The exact time windows to be used for the extraction of features will be determined during the analysis of the collected data. However, due to characteristic frequency components present in each of the sensor signals of interest, it is expected that local features from audio signals will be extracted more

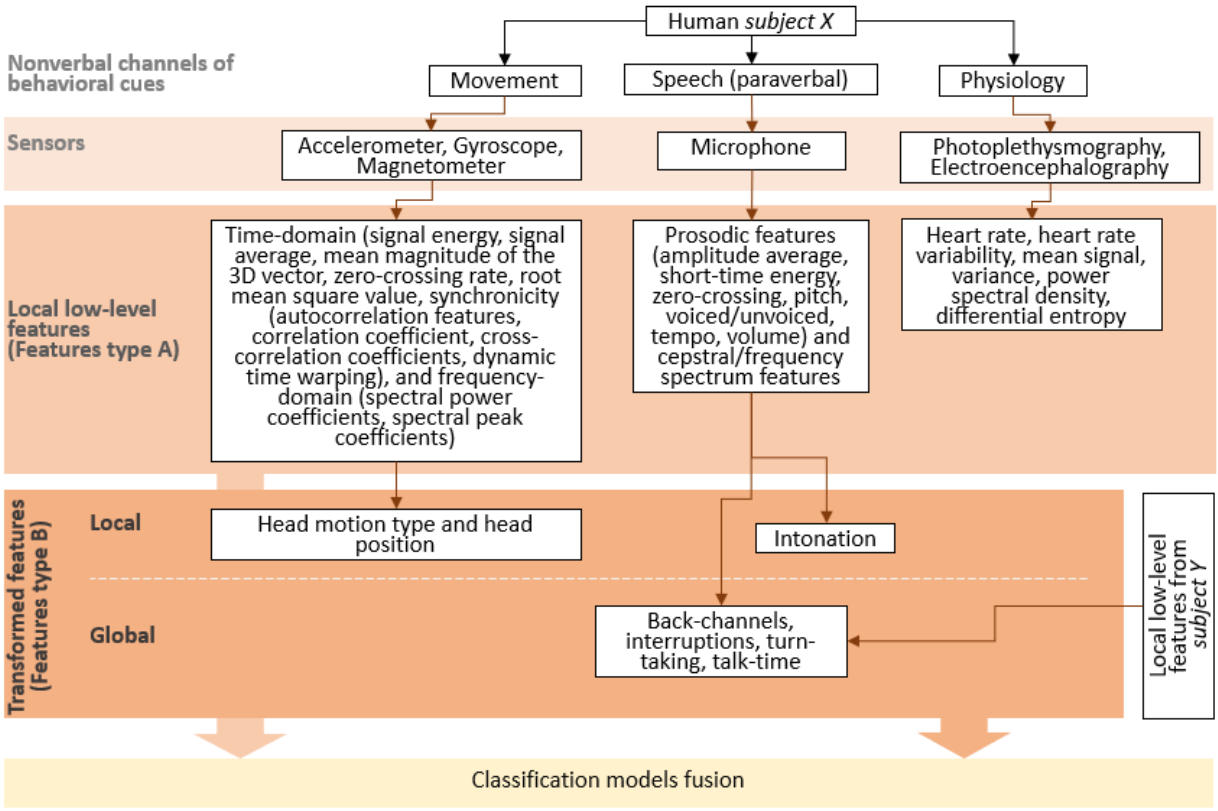


***Figure 14***. *Proposed feature extraction process for the selected sensor modalities, which includes extracting local low-level features or features Type A and transformed features or features Type B.*

rapidly than features from IMU, EEG, or PPG signals. At the global level, because we are interested in identifying coordinated behaviors, the extraction of features may be driven by the rate of change of the accelerometer data because audio is a faster-changing signal.

In microphones, features describing back-channel signals, prosodic, and conversation dynamics, among others, that can provide insights into levels of synchrony between individuals could be extracted to help recognize attentiveness, positivity, and coordination levels. Audio features such as power energy and pitch will be extracted locally to determine intonations, while detected frames of speech signal will be used to extract global features such as talk-time, turn-taking, overlapping talk, and back-channel signals like “uh-huh”. The extraction of global features requires interaction with data from all subjects in the interaction. For IMU data, features describing

back-channel signals such as head nodding and movement signals that can provide insights into levels of attention, positivity, and coordination will be extracted. The proposed machine learning framework considers that the recognition of intonations from speech signals and head motion from IMU signals require the transformation of low-level signals using classification models. The calculation of these type B features requires extensive study and experimentation, which is presented in Chapter 4.

In PPG, features describing the heart rate (HR) and heart rate variability (HRV) will be extracted locally to help identify levels of positivity. Finally, in EEG, features describing the signal power in different frequency bands, especially in the alpha frequency band (~8-12 Hz), in the occipital lobe positioned electrodes will be explored for attention measurement, and in the occipital and temporal lobe for emotional state. A list of features of interest is also presented in **Figure 14**.

*3.3.4.3.Training of computational models to determine group consonance*

This proposed machine learning framework suggests the use of a model fusion approach. **Figure 15** illustrates a 2-Layer recognition framework that combines features and model fusion

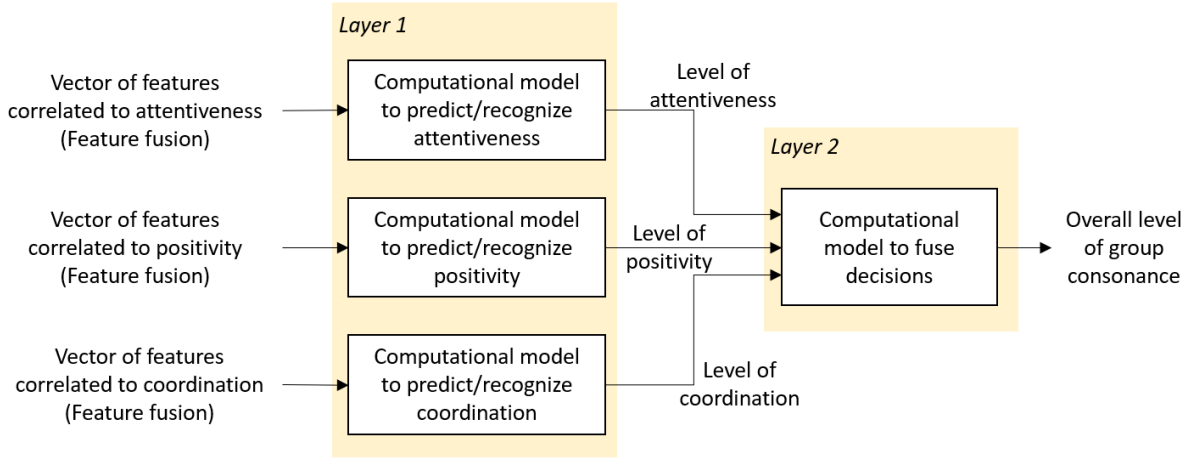


***Figure 15**. Model fusion approach to recognize/characterize components related to rapport that are affecting the overall interaction and consequently the overall calculated level of group consonance. In the first layer, three models are suggested to be trained to predict/recognize the different behavioral components of rapport. In the second layer, a model will combine the output of the models in Layer 1 and provide an overall measurement of dyadic consonance.*

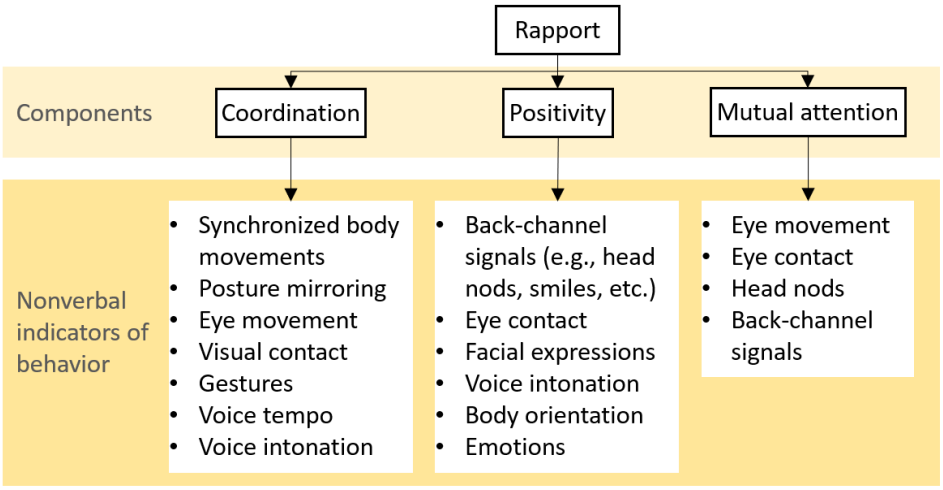


***Figure 16**. Nonverbal behavioral indicators associated with components of rapport that will be used to determine group consonance.*

approaches. In the first layer involving computational models, three models are suggested to be trained to predict/recognize the different behavioral components of rapport. **Figure 16** shows a list of behavioral indicators associated with components of rapport that will guide the fusion of extracted features and training of Layer 1 computational models. Then, in a second layer, a model will combine the output of the models using weighting factors and provide an overall measurement of dyadic consonance. These dyadic consonance values determined by a machine can then be combined to determine the level of group consonance. Before starting the implementation of a machine learning framework, sensor connection and synchronization are validated.

### 3.3.5. Results and Discussion

To test the collection of data and LSL network connection of the implemented multi-sensor framework for the monitoring of social interactions, a short data collection study during a staged meeting was performed. The data collection was approved by the Michigan State University Institutional Review Board (IRB) and conducted under strict physical distance and protection of privacy protocol guidelines. A total of 3 subjects were recruited voluntarily. Subjects were in separate rooms on the same building floor. The wearable sensors (Shimmer and BrainBit) were

attached as shown in **Figure 11**. The meeting of participants was conducted through the Zoom video conferencing program and consisted of a series of questions that the participants answered. For each question and response, sensor data were collected simultaneously for all three individuals,

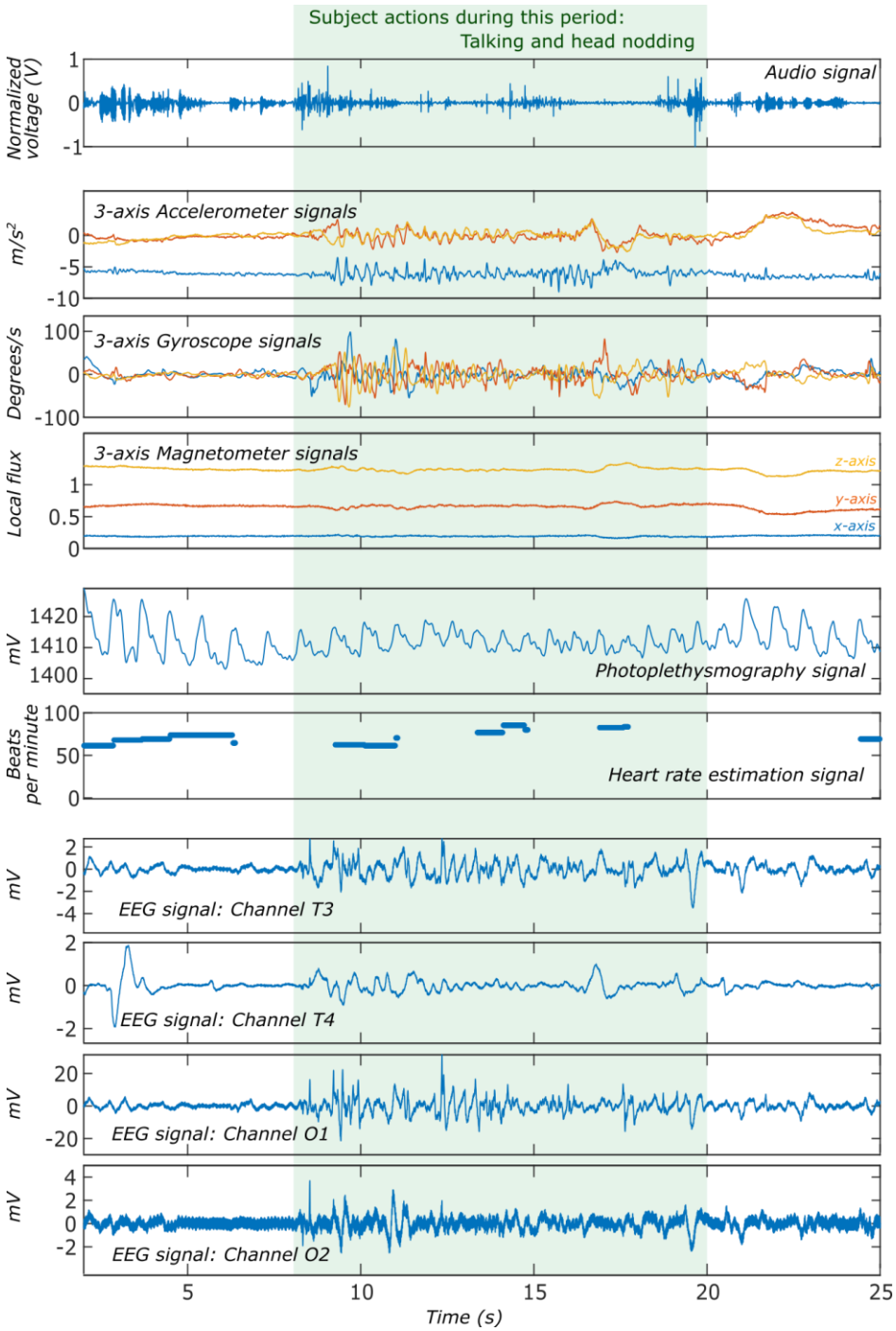


***Figure 17.*** *Synchronized signals from a single subject collected using the presented multi-sensor framework during a portion of a team meeting. © 2022, IEEE.*

representing a total of 3*16 synchronized data streams (16 data streams per sensor node). The meeting was recorded for data annotation purposes.

**Figure 17** shows all sensor signals collected and synchronized during a period of 25s from a single individual (sensor node). The EEG signals were post-processed using a 3rd-order bandpass IIR filter with cut-off frequencies of 1Hz and 50Hz for better visualization. A section of **Figure 17** was highlighted to indicate that during that period the subject was talking and head nodding. Head motion is reflected on the accelerometer and gyroscope signals. Signal synchronization allows observing how PPG signals appear to be degraded with head nodding, which consequently appears to cause discontinuities in the heart rate estimation signal. Therefore, the framework allows studying causes of signal degradation that will permit the design and implementation of real-time pre-processing methods to be implemented per sensor node. On the other hand, **Figure 18** shows one set of sensor signals for each subject and identified nonverbal messages. Synchronization of these signals was performed by the central unit of the multi-sensor framework.

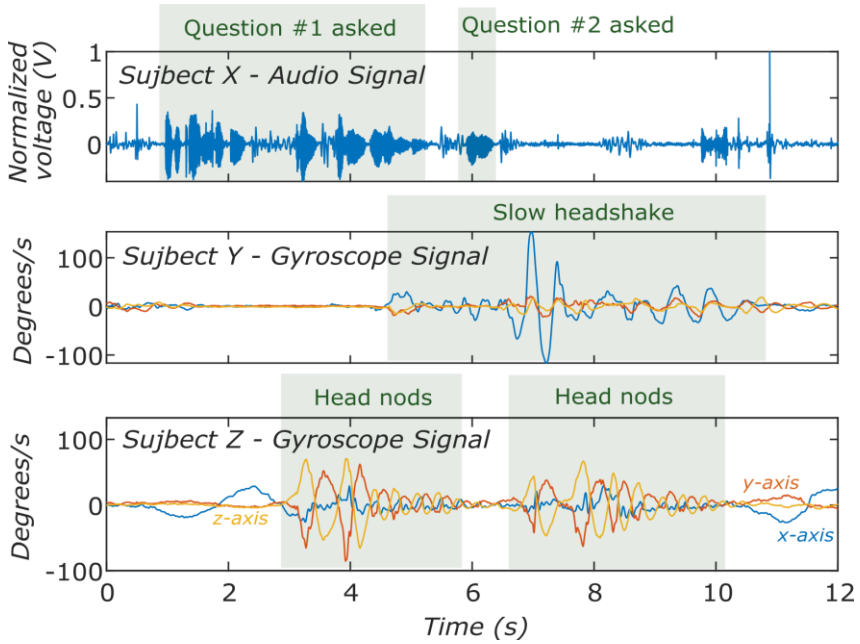


***Figure 18.*** *Synchronized signals from three subjects collected using the presented multi-sensor framework during a portion of a team meeting. © 2022, IEEE.*

Once again, signal synchronization, as shown in **Figure 18**, permits the identification of nonverbal messages' dynamics. In this case, speech from Subject X is followed by head nods and head shakes of Subject Y and Subject Z. Therefore, showing validation of the multi-sensor framework data collection and network connectivity.

### 3.4.Summary

Reviewing the personal factors that underpin human behavior and the theories and concepts that have guided the psychological study of social interactions provided a scholarly foundation for understanding the methods that have been employed for monitoring human behavior. Methods employing wearable sensors for the monitoring of human behaviors have been focused on recognizing emotions and social behaviors separately. To better understand the landscape of human behaviors that have been monitored using sensor technologies, the collected body of literature allowed this work to establish a taxonomy of human behavior monitoring technologies based on the reviewed psychological theories with the purpose of grouping existing human behavior monitoring literature. This helped us see that even when many efforts have been done to study and design technologies to monitor the defined human behavior effectors, very little has been done in designing robust systems that could measure complex social interactions.

Towards the goal of overcoming existing design challenges for the real-time monitoring of complex social interactions, this chapter presented an analysis of theoretical and technical aspects of human behavior monitoring technologies, established rapport as a measure of the quality of dyadic interactions and group consonance, and introduces a new and accessible multi-sensor system that allows the study and real-time analysis of both in-person and virtual interactive environments. The system integrates six sensing modalities, selected based on a deep analysis of technologies for behavior monitoring, and leverages the use of existing commercially available

wearable sensors. The system allows the synchronized collection of sensor data from at least three sensor nodes. Details of sensor integration and networking protocols to manage sensor data synchronicity were presented and a real-time machine learning framework was introduced. Results validate sensor data collection by our system and nonverbal messages that could be identified due to data synchronization. By this means, the physical infrastructure for monitoring individual, dyadic, and group-level behaviors is introduced. The next chapters show insights into the implementation of aspects of the machine learning framework.

*Disclaimer: A substantial portion of this chapter was published in* [118] *(© 2021, IEEE) and* [204] *(© 2022, IEEE).*

# 4. SIGNAL PROCESSING FOR THE RECOGNITION OF LOCAL TRANSFORMED FEATURES: FROM DATA COLLECTION TO ALGORITHM DESIGN

The machine learning framework presented in Chapter 3 proposes the use of local low-level features (Type A features) to extract local transformed features (Type B features). This chapter presents the first efforts in designing and implementing data processing blocks to recognize head activity and intonations using IMUs and audio signals, respectively. The design of these data processing blocks requires the collection of data that contain behavioral targets of interest and approximate real-life scenarios. Currently, no publicly available datasets or processes exist for the collection and preparation of data as is needed for the goals of this work. Therefore, this chapter also presents the study procedures employed for the collection and design of the datasets used for the training of computational models. Evaluations of signal features and training of computational models to identify head actions from IMU data and speech intonation from microphone data are also presented.

## 4.1.Real-Time Detection of Head Actions using IMUs

IMUs have been widely used for the identification of human activity, primarily focusing on physical activity detection, such as the ones seen integrated into smartwatches. IMUs have also been used in wearable devices placed in the head area. Head motion and position reveal a vast amount of information about the quality of social interaction. In general, IMUs placed in the head have been employed to assist individuals with disabilities. For example, the recognition of head gestures has been used as commands to control video players [205], computer cursors [206], wheelchairs, and robot hands [207]. More recently, IMUs have been used for the recognition of head motions associated with human behaviors and social interactions [208]. In [209], a 6-axis IMU was placed on the forehead and used to recognize four movements: pitch, roll, yaw, and

immobility. Classification models were trained with statistical signal features and raw signal data, achieving a 92% and 95% of recognition accuracy, respectively. In [210], a 9-axis IMU was mounted on the front side of a cap and considered the recognition of six types of head gestures (nod, shake, and facing up, down, left, and right), achieving an average classification accuracy of 95%. Generally, head motion recognition systems using IMUs have achieved classification accuracies that range from 72% to 99% [211]. However, the classification performance of these systems is highly dependent on the position of the sensors, the head motions of interest, and the set of signal features used for classification. Currently, no gold standard exists to automatically detect head activity from IMUs [211]. Therefore, this work represents the first effort in establishing methods for the design of head activity models using the sensors selected in Chapter 3 as part of the human behavior monitoring system. Data was collected from people wearing the sensor headset presented in **Figure 11** and performing specific head activities of interest, which included positioning the head at given angles and nodding, shaking, and rolling their head in response to specific questions.

### 4.1.1. Designing a Real-time Head Position and Motion Detection Algorithm

#### *4.1.1.1.Real-time model fusion architecture*

To study the best signal feature sets, reduce the computational complexity of data processing, and reduce data transmission rates for the multi-sensor system presented in Chapter 3 [204], a model fusion architecture was proposed as shown in **Figure 19**. The model fusion architecture is composed of two classification blocks: one for the detection of head position (static stage) and another for head motion (dynamic stage). The static stage allows studying the signal feature set that will best contribute to the classification of basic head positions (center, tilted to the right, tilted

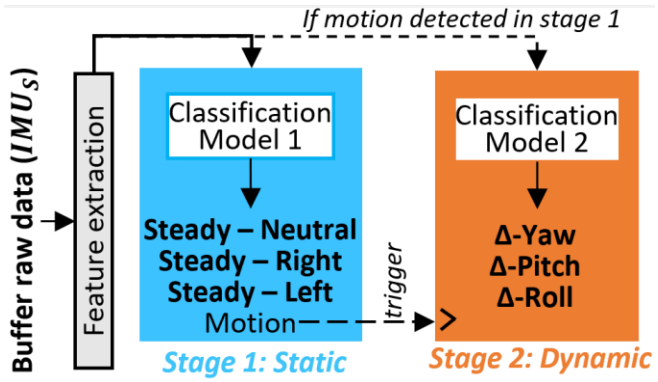


***Figure 19**. Two-stage model fusion architecture for the design and optimization of a head action detection (HAD) processing unit.*

to the left) versus general head motion. Likewise, the dynamic stage allows the study of an optimal set of features to classify Δ-pitch, Δ-yaw, and Δ-roll head motions. The combination of the two stages with their respective parameters creates the head action detection (HAD) processing unit

Moreover, the overall model fusion architecture is designed to, after feature extraction, perform a first classification where it is detected if the head is at one of the three steady positions (neutral, right, left) or if it is in motion. If the classifier detects motion, then a second classification is performed where Δ-yaw, Δ-roll, or Δ-pitch are identified. Because the classification models are recognizing two types of activity, static and dynamic, this model fusion approach provides the opportunity to retrain models separately, accelerating and reducing future re-training time for these classifiers.

*4.1.1.2.Signal segmentation*

The implementation of the fusion model architecture requires the study of optimal model parameters, including data buffer sizes for real-time signal segmentation and processing. Research has demonstrated that evaluating different buffer/window sizes contributes to finding a good balance between system performance and computational complexity [143]. To study the contribution that a buffer size can have in the extraction and processing of signal features, buffer sizes ranging from 1 to 4.5 seconds with 50% overlaps were evaluated. A buffer size of 1 second

was selected as the smallest buffer size because of interest in head motions that could be carrying frequency components around 1 Hz.

To perform this evaluation, pre-recorded signals are first buffered to simulate the real-time acquisition of the collected sensors' data. The buffer is applied per sensor signal type and axis. Resulting in

$$IMU_E = \begin{cases} [x_1, x_2, \dots, x_L] \\ [y_1, y_2, \dots, y_L] \\ [z_1, z_2, \dots, z_L] \end{cases} \quad (1)$$

where *E* represents the sensor type (accelerometer, gyroscope, or magnetometer), *x* the data in the x-axis, *y* the data in the y-axis, *z* the data in the z-axis, and *L* the buffer size.

*4.1.1.3.Pre-processing and feature extraction*

For each $IMU_E$ data buffer, as presented in (1), 70 features including time-domain, frequency-domain, and synchronization features were extracted. A list of features is shown in **Table 11**. Signal features were divided into two groups: extracted before band-pass filtering and extracted after filtering. Features extracted before filtering include signal energy for all three axis components, the average value in a signal buffer for all three axis components, the mean magnitude of the three-dimensional vector, and the zero-crossing rate for all three axis components.

Signal features extracted after filtering include: the root mean square (RMS) value for all three axis components, three autocorrelation features for all three axis components (height of main peak and height and position of the second peak), correlation coefficients across axis components, cross-correlation features across all three axis components (height of main peak and height and position of the second peak), dynamic time wrapping coefficients across axis components, three spectral power features per axis component (in 3 adjacent pre-defined frequency bands ranging from 0.2

***Table 11**. List of features extracted from IMU signals and evaluated to measure their contribution to the recognition of head motion. 70 features were extracted in total per sensor signal modality.*

| *Feature type* | *Feature name* | *Abbreviation with axis subscript* $(i \rightarrow x, y, z)$ |
|---|---|---|
| Time-domain | Signal energy | $SE_i$ |
| | Average value | $Av_i$ |
| | Mean magnitude of the three-dimensional vector | $Mnorm_{xyz}$ |
| | Zero-crossing rate | $ZC_i$ |
| | Root mean square value | $RMS_i$ |
| Synchronicity | Autocorrelation (height of main peak; height and position of second peak) | $AC\_1h_i$, $AC\_2h_i$, $AC\_2p_i$ |
| | Correlation coefficient | $Corr\_coef_{xy}$, $Corr\_coef_{yz}$, $Corr\_coef_{xz}$ |
| | Cross-correlation coefficients | $CCorr\_coef\_1h_i$, $CCorr\_coef\_2h_i$, $CCorr\_coef\_2p_i$ |
| | Dynamic time warping | $DTW_{xy}$, $DTW_{xz}$, $DTW_{yz}$ |
| Frequency-domain | Spectral power coefficients | $SpecP\_1b_i$, $SpecP\_2b_i$, $SpecP\_3b_i$ |
| | Spectral peak coefficients (height and position of first 4 peaks) | $Speak\_1p_i$, $Speak\_2p_i$, $Speak\_3p_i$, $Speak\_4p_i$, $Speak\_1h_i$, $Speak\_2h_i$, $Speak\_3h_i$, $Speak\_4h_i$ |

Hz to 10 Hz), and 8 spectral peak features per axis components (height and position of first 4 peaks).

The band-pass filter was applied to remove gravitational contributions and unwanted fast movements. The filter was a digital infinite impulse response (IIR) Butterworth with a first stopband frequency of 0.05 Hz, a first passband frequency of 0.1 Hz, a second passband frequency of 14 Hz, and a second stopband frequency of 16 Hz. We made use of an IIR filter because of their speed on high throughput applications, filter resolution, and consideration of memory consumption.

*4.1.1.4.Feature selection*

A feature selection method was applied to reduce the dimensionality of the extracted feature vector by studying feature contribution, removing redundant or noisy data that could degrade the

classification performance of head motions, and decreasing computational costs involved in the real-time feature extraction process to be ultimately implemented. A decision tree (DT) classifier and a function that computes estimates of predictor importance for the classification tree were used for this evaluation. The function that computes estimates of feature importance adds changes in the risk due to splits on every feature and divides the sum by the number of branch nodes. Because the two blocks of the fusion model architecture are classifying different types of events, feature selection was applied individually to both blocks.

*4.1.1.5.Classification model*

Subject-independent binary DT classifiers were trained for each stage in the fusion model architecture. For stage 1, based on the results of the predictor importance function, feature sets were further reduced and used to train classifiers for final evaluation and implementation. For stage 2, four types of feature sets derived from the feature importance analysis were used for the same end. Feature sets for stage 2 included the use of all extracted features, the use of the most important features based on the predictor importance function, and two sets of features engineered based on results from previously trained classifiers.

DT classifiers were selected because of their computational efficiency when paired with optimized feature sets. The Gini's diversity index was used as the split criterion with 20 as the maximum number of splits. Of the total dataset, 70% was used for training and 30% for testing. K-fold cross-validation of 10 was used to estimate the performance of the classifier on unseen data. Classification performance was measured using accuracy, defined as

$$Accuracy = (1 - \frac{C_e}{N_b}) \times 100 \quad (2)$$

where $C_e$ is the number of signal buffers misclassified and $N_b$ is the total number of buffers in the training/testing set. The complexity of the model was evaluated based on the number of features used by the model, the depth of the model, and the number of nodes in the tree.

4.1.2. Study and Data Collection Procedure

To design and optimize the HAD unit based on the described fusion model architecture, a validation study was performed and approved by the Michigan State University Institutional Review Board. A total of 3 subjects were recruited voluntarily. Subjects were located in separate rooms and the wearable sensor headset, as shown in **Figure 11**, was attached to their heads. Sensor data was collected at a rate of 128 Hz. The participants were then given access to a computer and connected to the Zoom video conferencing program to virtually interact with a study administrator. Sensor data from all participants were collected simultaneously using the system infrastructure introduced in Chapter 3. In addition, the Zoom interaction was recorded for data annotation purposes.

The study consisted of instructing participants to perform specific head movements at different motion rates and/or inclinations for 30 seconds. Head actions, inclinations, and motion rates are described in **Table 12**. Immobility refers to the absence of motion; Δ-pitch to a downward and upward head motion; Δ-yaw to a head rotation to the left and right; and Δ-roll to a head tilt motion from one shoulder to another. Participants were instructed to adopt the degree of head inclination that felt more natural to them when inclining their heads to the right or the left. When participants moved their heads at different motion rates, they were asked to do it as naturally as possible. Therefore, each subject had their own pace for slow, medium, and fast head motion. Two

***Table 12.*** *Summary of head actions performed during the validation study.*

| *Head actions* | *Head inclinations* | *Head motion rates* |
|---|---|---|
| Immobility | Center, tilted to the right, tilted to the left | - |
| Δ-Pitch, Δ-Yaw, Δ-Roll | - | Slow, medium, fast |

trials of each head action type with its corresponding head inclination or motion rate were performed. These resulted in a total of 24 head motion recordings, each with a duration of 30 seconds captured per participant. A total of six labels, corresponding to the three head motions and three head positions of interest, were assigned to the collected data.

### 4.1.3. Results and Discussion

The analysis, model design, and fusion model were coded in MATLAB. The DT classifiers were trained and validated with the collected dataset consisting of IMU signal segments lasting 30 seconds, which were further segmented according to the buffer size to be analyzed.

#### *4.1.3.1. Best feature sets and classification results per model stage*

*Static Stage* - To evaluate the best set of features for the detection of head position and to train a classification model for the task, the collected IMU signal segments containing data of motions were labeled as "other". Therefore, stage 1 classifies data segments using 4 data labels (left, right, center, and other).

The feature importance analysis consistently revealed across different data buffer sizes that, to recognize head positions versus general motion, gyroscope's features tend to be the most important signal features, followed by accelerometer features, and lastly magnetometer features. Based on the results of the feature importance analysis, ten different feature subsets containing from 3 to 12 features were selected and used to train DT classifiers. Results using training data at different buffer sizes and feature sets varied from 60.50% to 99.56% classification accuracy. Across all trained classifiers using different buffer sizes, results consistently show that the most important features across all buffer sizes are accelerometer $SE_i$ and $Av_i$ in the x-axis and y-axis, accelerometer RMS in the y-axis, gyroscope $SE_i$ in all axes, and gyroscope $Mnorm_{xyz}$.

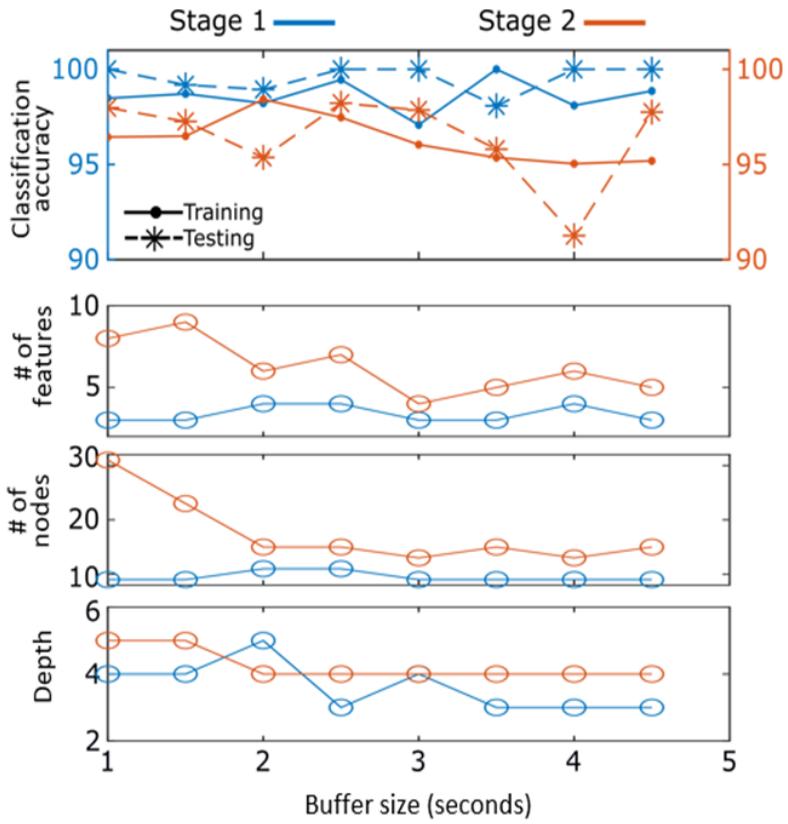


***Figure 20.*** *Results of classification accuracy, number of features used for by the DT model, number of nodes, and depth of the best performing models across data buffer sizes.*

*Dynamic stage* – Classification was performed using 3 data labels (Δ-pitch, Δ-yaw, Δ-roll). The feature importance analysis revealed that as the data buffer increases in size, frequency features become more important whereas in short window sizes time-domain and synchronicity features may be more important. Likewise, for the detection of head motion it was noted that as the buffer size increases, signal features from the magnetometer sensor become less relevant. Classification results using training data at different buffer sizes and feature sets varied from 91.35% to 98.22%.

**Figure 20** shows the classification accuracy results and the number of features, number of nodes, and depth of the best-performing classifier for a given buffer size, expressed in seconds. To

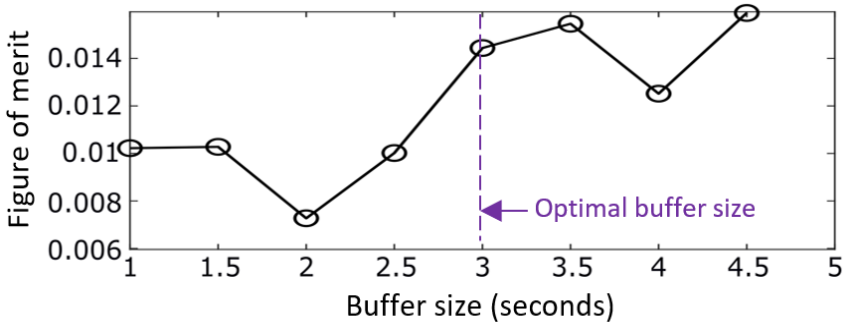


***Figure 21***. *Results of FoM, which contributed to evaluate classifier performance versus complexity. The higher the FoM value, the optimal the classifier.*

***Table 13***. *Summary model parameters for the HAD unit using a buffer size of 3 seconds.*

| *Stage* | *Testing accuracy* | *Training time* | *Features* | *# Nodes* | *Depth* |
|---|---|---|---|---|---|
| 1 | 100% | 0.57488s | 3 | 9 | 4 |
| 2 | 97.86% | 0.61673s | 4 | 15 | 4 |
| *Overall* | 97.91% | *Macro F1-score* | *Macro Recall* | *Macro Precision* | *Classification time* |
| | | 98.5% | 98.67% | 98.67% | 0.0034s |

determine an optimal buffer size based on DT classifier accuracy and complexity, a Figure of Merit (FoM) was established and defined as

$$FoM = \sum_{i=1}^{2} \frac{TeA_i}{TrA_i * N_{f_i} * N_{n_i} * N_{d_i}} \quad (3)$$

where $i$=1 is referring to the results of stage 1 and $i$=2 to the results of stage 2, *TeA* is testing accuracy, *TrA* is training accuracy, $N_f$ is the number of features used for classification, $N_n$ is the number of nodes of the DT, and $N_d$ is the depth of the DT. Training and testing accuracy were included in the FoM to account for cases where the best performing classifier was an overfitted one, as is the case for stage 1 and buffer size of 3.5s and stage 2 for buffer sizes of 2s and 4s. **Figure 21** shows the results for the FoM, where the higher the value, the more optimal is the classifier. However, because latency is an important factor in the real-time detection of head actions, a buffer size of 3s provides the best tradeoff between DT model performance and

complexity for both stages of the fusion model. Because the buffer has a 50% overlap with previous data, the HAD unit updates at a rate of 1.5s.

*4.1.3.2.Head action detection processing unit*

**Table 13** shows a summary of parameters for the final implemented DT classifiers at each of the stages in the fusion model architecture. Based on the results of the FoM, a buffer size of 3s was selected for the final implementation. The three features used for the classification of head position (stage 1) were $Av_x$ and $Av_y$ from the accelerometer and $Mnorm_{xyz}$ from the gyroscope. On the other hand, $RMS_z$ from the accelerometer and $SE_x$, $Corr\_coef_{xz}$, and $Speak\_2h_x$ from the gyroscope were used for the classification of head motion (stage 2). Therefore, magnetometer data were excluded from the final design. The HAD unit, using a fusion model architecture, has an overall testing accuracy of 97.91% and an F1-score of 98.5%.

The performance of the design HAD unit is on par with previous works. However, the architecture of the HAD unit allows for easy re-training to add recognition of additional head actions by having specialized head action classification models.

This represents the first effort in establishing methods for the design of head activity detection in real time using the sensor setup established for our behavior monitoring system. Because the data collected for the design of the HAD unit was in a controlled environment, no extensive data annotation procedure was required, accelerating the design procedure. As the accuracy of the trained model is high, it is important to highlight that this performance could decrease when employed using data from the wild, as other factors not accounted for in the lab will influence the results. However, by employing a fusion model approach, one of the stages or both could be easily retrained with data from the wild.

### 4.2.Real-Time User-Independent Speech Intonation Recognizer

Speech signals carry important social information that is expressed through verbal and nonverbal communication. Verbal communication includes the use and understanding of words, whereas nonverbal communication in speech refers to the way words are said, e.g., the tempo and the intonation used while communicating verbally. In many areas of research, such as natural speech processing and affective computing, the automatic identification of speech intonations has played an important role in creating effective human-machine interactions and in recognizing emotional states from the speakers. Other areas of research, such as those that focus on understanding and monitoring social interactions, have started to incorporate information related to voice tonality in the analysis of human behaviors. Speech intonation carries information about our social intentions and feelings. Moreover, the way people talk contributes to building rapport and establishing social likeability. Still, the automatic and real-time recognition of speech intonations and their emotional content is an active and challenging area of research, especially for natural environments [212].

Intonation, in general, refers to the rise and fall in voice inflection, which happens consistently throughout speech. However, experts in the area have not agreed on a universal definition for intonation. From reviewed literature, two types of intonation classes are studied and for which automatic recognition systems have been designed: (1) intonations described by pitch contour and (2) intonations described by perceived affective state or intention.

In general, works that focus on the study and recognition of pitch contour are intended for speech synthesizer systems. In the English language, there are four basic intonation classes that describe pitch contour: Glide-up, Glide-down, Dive, and Take-off. Glide-up refers to the rise of pitch values, which is associated with the production of question and encouragement statements.

Glide-down refers to the fall of the pitch contour values and is attributed to the production of a general statement. Dive refers to a combination of fall and rise pitch contour and is associated with warning and commanding statements. Lastly, Take-off refers to having a sustained pitch level and gradually increasing it, which is generally associated with negative affective states. While a variety of works have a focus on designing pitch contour recognition systems [213]–[215], the mapping between paralinguistic functions and pitch contours varies within a language and differs cross-linguistically [216].

On the other hand, works that focus on studying and recognizing intonations as an affective function tend to use categorical intonation classes. A variety of works have focused on studying positive and negative intonations [141], [144], [145], but a wide range of other affective states have been explored [140], [142], [189], [217]. For example, Wu and Liang [142] utilized speech-derived information for the recognition of four emotional states: neutral, happy, angry, and sad. The dataset was collected from 8 Chinese-speaker volunteers in a laboratory environment. The authors made use of acoustic-prosodic information and semantic labels as features of the utterances (sentences) that formed part of their dataset. Acoustic-prosodic features were extracted offline and classifiers such as GMM, SVM, and MLP were trained to recognize the four classes, achieving a range of accuracies going from 68.73% to 78.16%. SVM was the classifier with the highest accuracy. Because none of the classifiers were optimal for recognizing all emotional states [218], [219], a Meta Decision Tree (MDT) was used for classifier fusion, achieving a recognition accuracy of 80%. However, the designed algorithm is not suitable for real-time processing based on the use of semantic labels and the lack of real-time methods to automatically identify utterances.

Lanjewar et al. [189] used the Berlin Emotion Speech Database (BES), which is an acted emotional content database with around 500 utterances in German portraying emotions of

happiness, anger, disgust, fear, sadness, surprise, and neutral. The authors focused on all but disgust and used spectral features such as MFCC, pitch, and Wavelet coefficients to train a GMM and a K-NN classifier. Recognition accuracies were 66% and 52% for GMM and K-NN, respectively. Here, the authors showed how GMM dominates the recognition of angry and sad emotions, whereas K-NN dominates the recognition of happy and angry emotions. However, these models were not designed for real-time operation and do not account for natural expressions of emotion during social interactions.

A common approach in designing speech intonation recognizers is the use of supervised machine learning methods. When supervised methods are employed, two principal areas require attention: (1) data collection and annotation and (2) machine learning model design.

Traditionally, speech intonation recognizers (including speech emotion recognizers) have made use of acted datasets to train their recognition models. However, systems trained on acted data do not translate well to real-life situations [220], since "full-blown" emotions rarely appear in everyday interactions [221]. Only in the last 10 years, the speech emotion recognition research area has started to see a shift towards the use of natural datasets [212]. Even though naturally collected datasets exist, they tend to be from call centers, TV shows [217], or interactions with virtual agents, which do not capture the nature of group interaction environments.

In general, data collection and annotation are key to supervised machine learning design pipelines and well-developed annotation guidelines are critical for its success. However, there are no standard annotation guidelines for the design of speech intonation recognition models. In addition, there are a variety of factors that impact the results of a data annotation project, which include, for example, the annotation tools employed, the human annotators, and the specific application [116], [222].

Here, the collection of a natural dataset and the development of an annotation pipeline are presented. A natural dataset is collected from research group meetings, where ideas were being exchanged providing an opportunity to capture reactions to disagreements in a workplace environment. Because of the lack of annotation guidelines to identify speech intonation, two types of annotation modes were analyzed: annotation of audio in the order it occurred and in a randomized order. We hypothesized that as annotators get used to people's way of talking, the likeability will increase. This could affect how datasets are labeled and the overall results of an analysis of dyadic and group social interactions. For instance, annotating in sequential order may help annotators gain a sense of familiarization with the person or persons involved in the interaction, whereas annotating in random order could maintain a sense of distance from the individuals involved in the interaction since the context of the meeting is lost. Intonations of interest include a combination of affective states with interrogative expressions. Based on annotation analysis, a dataset was constructed to train a model for the real-time recognition of intonations. Because of the goal of implementing real-time algorithms for the real-time monitoring of human behaviors, the design of the model was performed using a resource-aware approach where the effect of different sampling rates and the reduction of feature dimensionality was evaluated using the resulting classification accuracy of the trained models.

### 4.2.1. Designing a Real-time Speech Intonation Recognition Algorithm

#### *4.2.1.1.Pre-processing*

To study the effect that different sampling rates have over the recognition of intonations carrying affective state information, collected signals were low-pass filtered and downsampled to 8 kHz, 4 kHz, 3.2 kHz, and 2 kHz. Real-time audio signal processing requires the data to be processed using small data frames. Typically, audio signals are processed in frames of ~30ms to

~80ms with overlaps between each consecutive frame. Here, we make use of a 40ms frame with 50% overlap. These frames are used to detect speech in the audio signal to later perform speech segmentation.

*4.2.1.2. Speech detection and segmentation*

Speech segmentation, also known as audio segmentation, refers to the task of dividing the audio signal into segments that will be used for feature extraction and classification. Speech segmentation can be performed in two ways, using an utterance-based approach or a windowing-based approach. Utterance-based approaches require the implementation of an automatic speech recognizer (ASR), which increases system complexity and may represent a threat to privacy to users because the goal is to recognize linguistic units such as vowels, phonemes, words, and phrases [140], [141]. On the other hand, windowing-based approaches make use of windows of data defined by time (milliseconds to seconds), windows of speech activity (defined by thresholds in pauses or silence periods), or windows of voiced/unvoiced signals. Windowing-based approaches tend to be fast and computationally efficient, however, efficiency is compromised when high amounts of memory are needed to extract features of interest effectively. To have results that compare to speech segmentation performed by ASR, windows of speech need to be long enough to contain voiced-unvoiced segments and breath periods, which are elements that comprise an utterance. To implement a speech segmentation method that can operate in real-time, first, an energy-based voice activity detector (VAD) method is designed. The design of an energy-based VAD involved the evaluation of an energy threshold. A pre-set threshold based on noise statistic studies [223] and a threshold calculated using histogram and maxima estimation were evaluated. Then, silence periods were measured, and their distribution was studied to determine a threshold that will be used for the speech segmentation. Lastly, to complete the segmentation process, a

distribution of the duration of identified speech periods was studied to determine the minimum length of a speech period to be considered an utterance of interest.

*4.2.1.3.Feature extraction and selection*

Inspired by previous research and the review of signal features for audio data presented in Section 3.2.2.1, a set of features that have proven to contribute the most to identifying changes in speech affective states were used in this work. This set of features includes a combination of prosodic features (e.g., energy and pitch), voice quality features (e.g., zero-crossing rate), frequency spectrum coefficients, and cepstral features.

For each data frame, the energy was calculated and used to identify voice activity as described in the previous section. If voice activity was detected, then zero-crossing (ZC) was calculated and if the value laid below a pre-defined threshold of 35, obtained from [223], then the data frame was classified as a voiced speech segment. Voiced segments are periods of speech generated using the vibration of vocal cords. If the ZC value was over the threshold, then the data frame was classified as an unvoiced speech segment, which are periods of speech generated using air passed through the vocal cords. For all identified voiced segments, the pitch was determined using autocorrelation. Before the calculation of pitch, the data frame was filtered using a band-pass filter with cutoff frequencies of 50 and 900Hz. Note that a wide range of pitch detection algorithms have been studied and that no available pitch detection scheme can be expected to give perfect pitch period estimates. To approximate the calculated pitch value to what is perceived by human hearing, the pitch value obtained through autocorrelation was transformed using the following Mel-scale [224]:

$$Mel\ pitch = 2595 * \log_{10}(1 + 0.0014 * pitch_{autocorrelation}) \quad (4)$$

and the Δ-Mel pitch values and ΔΔ-Mel pitch values were also calculated. The Δ-Mel pitch value represents the difference between two consecutive Mel pitch values (from two consecutive data

***Table 14**. List of features extracted per data frame where speech was detected.*

| *Feature type* | *Feature name* |
|---|---|
| Prosodic | Signal energy |
| | Average value |
| | Mel pitch (pitch value on Mel scale) |
| | Δ-Mel pitch |
| | ΔΔ-Mel pitch |
| | Voiced |
| | Unvoiced |
| Voice quality | Zero-crossing rate |
| Frequency spectrum coefficients | Mean of power spectral density |
| Cepstral coefficients | Mel-frequency cepstral coefficients (MFCC) |

frames) and the ΔΔ-Mel pitch value represents the difference between two consecutive Δ-Mel pitch values. For all data frames, the average amplitude of the speech signal is also calculated.

To calculate features in the frequency domain, a pre-emphasis filter and a hamming window were applied to emphasize high-frequency components in the speech signal that otherwise would be dominated by low-frequency ones and to reduce spectral leakage when calculating features in the frequency domain, respectively. Then, the magnitude of the Fourier transform of the signal in the data frame was calculated and used to determine the mean of the power spectral density. Lastly, for each data frame, 13 Mel-frequency cepstral coefficients (MFCC) were calculated. **Table 14** shows the list of features extracted per data frame.

As the previous set of features is calculated per data frame, each estimated utterance is represented by time series of the aforementioned features. To prepare this extracted data for classification, once an utterance is determined, statistics of its corresponding feature series are calculated. Conversation features such as speaking rate and pausing rate are also calculated for each utterance. The speaking rate is calculated by dividing the number of voiced frames by the number of unvoiced frames. On the other hand, the pausing rate is calculated by dividing the number of silence frames by the total number of frames in the estimated utterance. **Table 15** shows

the list of features extracted for each utterance based on the time series constructed from features presented in **Table 14**.

To eliminate redundancy in extracted features and reduce feature dimensionality to minimize computational complexity, a correlation analysis across features presented in **Table 15** was performed to eliminate highly correlated features. A high correlation was considered to be any value over 0.8 or under -0.8.

*4.2.1.4.Classification of intonations*

For comparison, classification models were trained using the complete set of extracted features presented in **Table 15** and the reduced one obtained through correlation analysis. To evaluate how well different models fit different classes, a variety of models were trained to classify 4 intonation classes, 3 intonation classes, and a combination of 2 classes. In addition, models were

***Table 15**. List of features extracted for each utterance based on the feature time series.*

| *Feature type* | *Feature name – time series* | *Final set of features* | *Qty* |
|---|---|---|---|
| Prosodic | Signal energy | Mean, std, max, min, median, range | 6 |
| | Mel pitch (pitch value on Mel scale) | Min, max, range, mean, median, std, number of peaks, mean peak value, std of peak values, median of peak values, mode of peak values | 11 |
| | Δ-Mel pitch | Number of peaks of the absolute value, mean peak value, std of peak values, and mean, std, max, min, median, range of last 5 non-zero value | 9 |
| | ΔΔ-Mel pitch | Number of peaks of the absolute value, mean peak value, std of peak values | 3 |
| Conversation | Speaking rate | Voiced/Unvoiced | 1 |
| | Pausing rate | Silence frames/total frames | 1 |
| Voice quality | Zero-crossing rate | Mean, std, max, min, median, range | 6 |
| Frequency spectrum coefficients | Mean of power spectral density | Mean, std, max, min, median, range | 6 |
| Cepstral coefficients | Mel-frequency cepstral coefficients (MFCC) | Mean, std, max, min, median, range of first 4 MFCC coefficients | 24 |

also trained with data at different sampling rates to evaluate how reduced data rates may influence the classification performance of speech intonation.

Models that were used for evaluation included Support Vector Machine (SVM), K-Nearest Neighbor (KNN), linear discriminant, Naïve Bayes, Random Forest, and Gaussian Mixture Models (GMMs). For classifiers such as SVM, KNN, and Naïve Bayes different kernels were also evaluated.

#### 4.2.2. Study Procedure for Audio Collection

Given the lack of datasets with conversations in natural environments and with multiple individuals, we recorded audio and video data from virtual research group meetings at Michigan State University (MSU). The study procedure was approved by the MSU Institutional Review Board. Individuals participating in the virtual meetings were instructed to carry on their normal conversations. Research group meetings were of interest because of the number of ideas being exchanged, providing an opportunity to capture subtle reactions of agreement and disagreements in a workplace environment. The recordings were performed through the Zoom video conferencing program, which also allowed the recording of a separate audio file for each participant in the meeting.

A total of five meetings were recorded, over one month, with an average duration of 57 minutes and with 4 to 5 individuals per meeting, as shown in **Table 16**. Subjects used their audio recording equipment and participated in the meeting from a variety of locations. This provides a dataset that contains a variety of background acoustics and microphones, among others, which results in a representation of the technologies in the wild. All the audio meeting recordings were obtained at a sampling frequency of 32 kHz.

***Table 16.*** *Summary of recorded meetings' information. Identification of speech segments/audio clips was performed by two annotators, A and B.*

| *Meeting* | *Duration* | *# of participants* | *Annotator* | *# of audio clips* |
|---|---|---|---|---|
| 1 | 00:39:33 | 5 | A | 377 |
| 2 | 01:00:14 | 5 | A | 508 |
| 3 | 01:14:18 | 4 | A | 566 |
| 4 | 00:58:56 | 4 | B | 395 |
| 5 | 00:51:46 | 4 | B | 676 |

4.2.3. Procedure for Annotating Speech Intonations

The annotation procedure was divided into (1) partitioning the audio recordings into small speech segments and (2) human labeling of the intonation of those speech segments. A total of six annotators were used during the annotation procedure, two for the partitioning of the audio recordings and four for the labeling of intonations. The annotation scheme covers four general and 10 specific intonations.

*4.2.3.1.Selection of labels for intonation*

An audio dataset with labeled intonations is important for the development of algorithms capable of inferring, for example, emotional state-related information from audio signals. Inferring speech intonations forms part of the identification of nonverbal communication and the design of human-machine interfaces. Because the interest of this work is in designing an intonation recognition model that can contribute to the understanding of human behaviors and the establishment of rapport during social interactions, we created an initial list of specific intonations that could impact the establishment and perceptions of rapport. The list consisted of the following intonations: neutral, surprise, excitement, disappointment, affirmative, laugh, commanding, encouraging, doubtful, mad, and question. This initial list was utilized as a guide to performing manual audio segmentation. The initial list of intonations inspired the creation of a second specific

list of intonations that substituted question and neutral for frustration and none. This also inspired the creation of a general list of intonations that included: neutral, positive, negative, and question.

*4.2.3.2.Manual audio segmentation*

Audacity, an open-source digital audio editor, was used to perform manual audio segmentation. Annotators were instructed to identify speech segments, using the "Add Label at Selection" feature from Audacity, based on perceived intonations based on the initial list of specific ones: neutral, surprise, excitement, disappointment, affirmative, laugh, commanding, encouraging, doubtful, mad, and question. Speech segments were identified for each of the separate audio files recorded by each participant during each of the meetings. The identified speech segments were then saved in a text file containing the initial time, the end time, and the perceived intonation of the segment. The annotation of the meetings was divided into two groups, wherein Annotator A identified speech segments in the first three meetings and Annotator B in the last two meetings. **Table 16** shows the total number of identified speech segments per meeting. A MATLAB script and the generated text file containing the times of the identified speech segments were used to generate corresponding individual audio clips. **Figure 22** shows a summary of this first part of the annotation procedure.

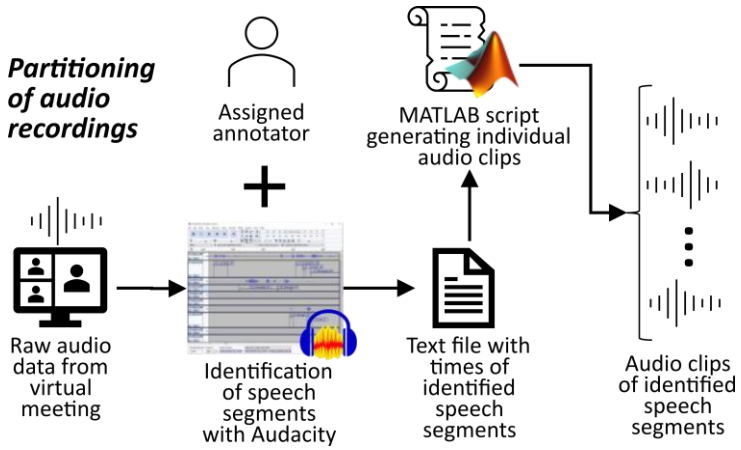


***Figure 22**. Diagram summarizing the first part of the annotation procedure, which constitutes partitioning the audio recordings from the virtual meetings into audio clips containing specific speech intonations.*

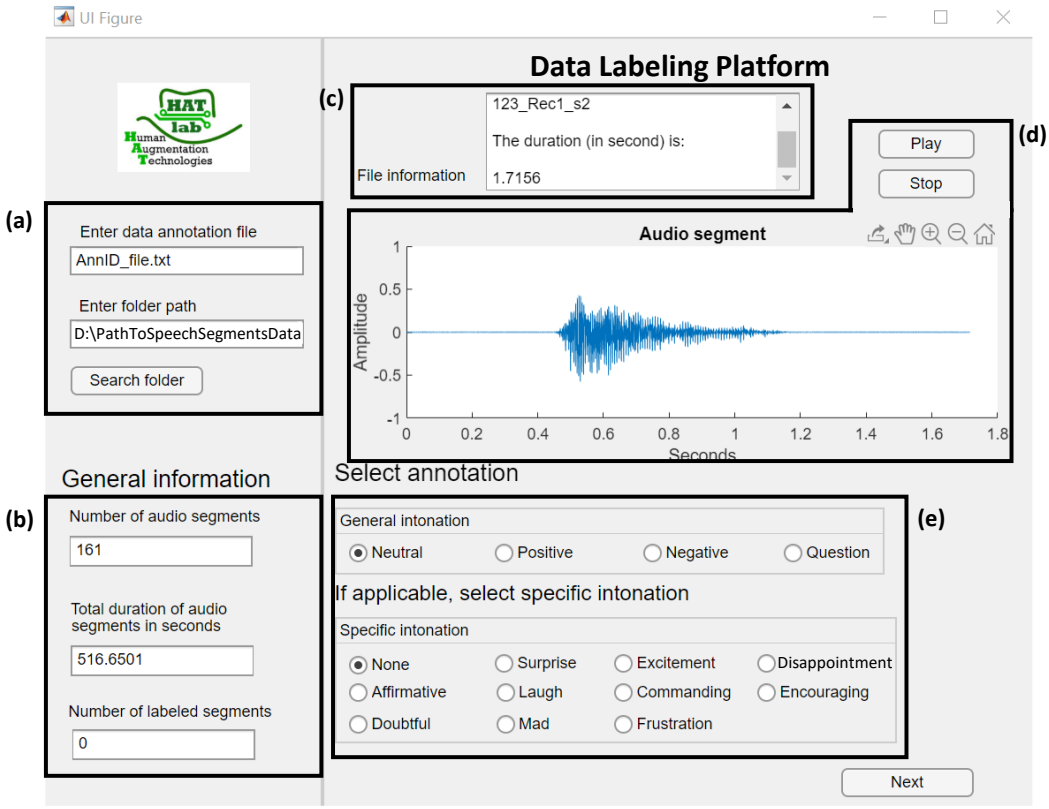


***Figure 23.*** *Designed interface for the labeling of intonation in audio clips. (a) The path to the folder containing the audio clips is given to the labeling program together with the name of the data annotation text file. The folder path is used to locate the audio clips that will be labeled by the annotator; (b) general information about the files contained in the given folder paths; (c) general information about the audio clip that is displayed; (d) plot of the audio clip and the buttons to play or stop the audio; and (e) labeling of general and specific perceived intonation.*

#### *4.2.3.3.Labeling of intonations*

The App Designer tool of MATLAB 2019b was used to develop a graphical user interface to facilitate the labeling process and ensure a consistent level of annotation. **Figure 23** shows the designed interface for the labeling of intonation in the audio clips. This data labeling program is a customized interface that takes the path to the folder containing the audio clips that will be labeled and display them. The labeling program also takes the name of the annotation text file created beforehand, which will be used to record the perceived intonation for each audio clip. Overall information about the audio clips in the folder path and information about the specific audio clip

being displayed is also displayed. The annotator can play or stop the audio clip/segment that will be labeled at any given moment. The labeling program supports two levels of annotation: a general intonation label and a specific intonation label. General intonations include neutral, positive, negative, and question. Specific intonations include surprise, excitement, disappointment, affirmative, laugh, commanding, encouraging, doubtful, mad, frustration, and none. Because this work was interested in understanding the impact of intonation perception when labeling data, two identical interfaces were designed, one with the ability to display the audio clips in their order of occurrence in the recorded meetings and another one with the ability to randomize the order in which the audio clips are presented.

All audio clips were labeled using both interfaces (i.e., the interface presenting the audio in sequential order and the other interface presenting the audio in random order). A total of four annotators (C, D, E, and F) participated in the labeling of intonations, wherein two of them labeled the audio segments in both sequential and random order, another one only labeled segments presented sequentially, and the last one only labeled segments presented randomly. This resulted in each audio clip being labeled three times, both when presented sequentially and randomly. The interface outputs the text file containing the assigned labels for general and specific intonations. These files were then used to perform an analysis of inter-annotator agreement (IAA) and label assignment based on the majority of the annotators. Because of human and labeling interface errors, not all audio clips were labeled by three annotators. Therefore, those audio clips lacking a third annotator were dropped from further analysis. **Figure 24** shows a summary of this second part of the annotation procedure. The code and executable files to run the data labeling program can be found at https://gitlab.msu.edu/davilasy/audio-data-labeling-tool.

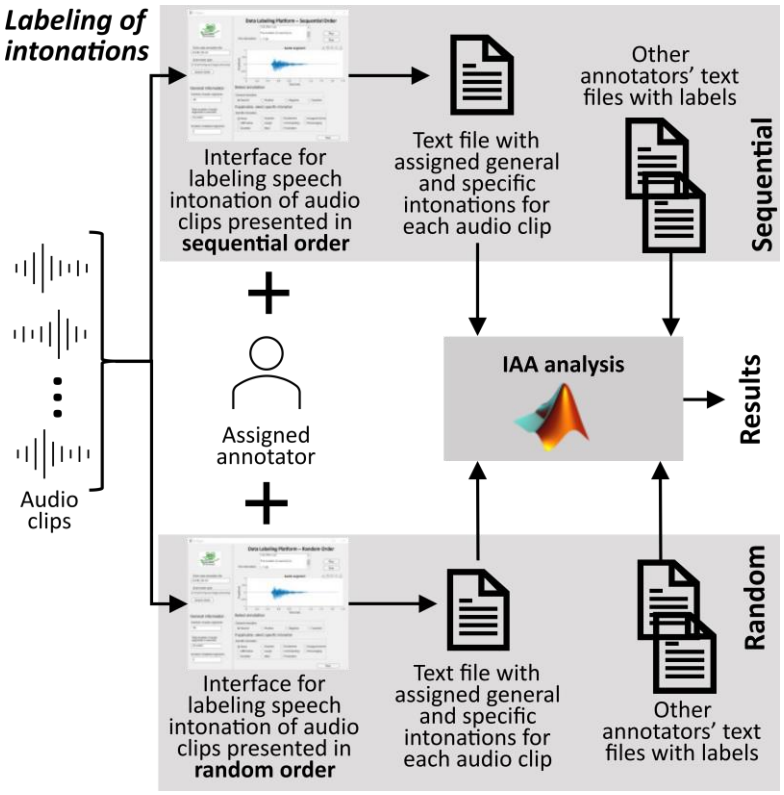


***Figure 24.*** *Diagram summarizing the second part of the annotation procedure, which involves labeling audio clips in the order in which they were produced in the meeting and in a random order. Files with label information from all annotators were combined for analysis.*

### 4.2.4. Analysis of Annotations

#### *4.2.4.1.Inter-annotator agreement (IAA)*

To determine if there is a significant effect when labeling audio clips presented in sequential and random order, two measurements of IAA were applied: pair-wise Cohen's kappa and Fleiss' kappa coefficients. IAA measures how well two or more annotators make the same annotation for a certain category. In this work, categories constitute four general intonations and 11 specific intonations.

Pair-wise Cohen's kappa coefficient ($k_C$) is a statistical measure of reliability between two annotators for categorical items [225]. The definition of $k_C$ is:

$$k_C = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

where $p_o$ is the observed proportionate agreement between raters and $p_e$ is the probability of random agreement. Cohen's kappa coefficient was calculated per pairs of annotators, meetings, annotation order (random and sequential), and annotation mode (general and specific).

On the other hand, Fleiss' kappa ($k_F$) is calculated over a group of multiple annotators assigning categorical ratings to a fixed number of items. The definition of $k_F$ is:

$$k_F = \frac{\bar{P} - \bar{P}_E}{1 - \bar{P}_E} \tag{5}$$

where $\bar{P}$ is the overall observed agreement chances per category divided by the number of categories and $\bar{P}_E$ is the average chance agreement over all categories. Fleiss' kappa was calculated per meeting, annotation order, and annotation mode.

A paired two-sample t-test was utilized to determine if the mean of the annotator agreement across all meetings for the different annotation orders and annotation modes were statistically significant. The difference between the means was determined to be statistically significant if the t-test resulted in a p-value of less than 0.05.

*4.2.4.2.Selection of labels for speech segments*

Three types of datasets were constructed, for both sequential and random annotation orders, based on the labels provided by the annotators and the IAA analysis. For simplicity, we enumerated the datasets using 1, 1.1, and 2. Dataset 1 contains all the audio segments labeled using general intonation and Dataset 1.1 contains all the audio segments where two or more annotators agreed on a general intonation label, meaning that all audio segments where annotators did not agree at all on an intonation were eliminated from this dataset. Audio segments where two annotators agreed on an intonation were then given the respective label and the third annotator was ignored. Dataset 2 contains all the audio segments labeled using a specific intonation. Datasets 1.1 of both

the order and random annotation order sets were used to construct the final dataset used to train the intonation recognition model.

*4.2.4.3.Rate of change in perceived general intonations*

To study if there is an increase or decrease in how positive, negative, and question intonations are perceived when labeling audio segments presented in a sequential order versus a random order, the rate of change in labeled general intonations was evaluated. This was studied using the constructed Datasets 1.1, where at least two annotators agreed on a label. This analysis was performed by counting the number of neutral, positive, negative, and question intonation labels that were assigned per meeting for both datasets. Then, to determine if the difference between the mean of these quantities was statistically significant, a t-test was performed.

4.2.5. Results and Discussion

*4.2.5.1. Audio data annotation*

A central question of this study was whether there is a significant effect on perception when labeling audio clips in the order they were generated versus labeling the audio clips in a

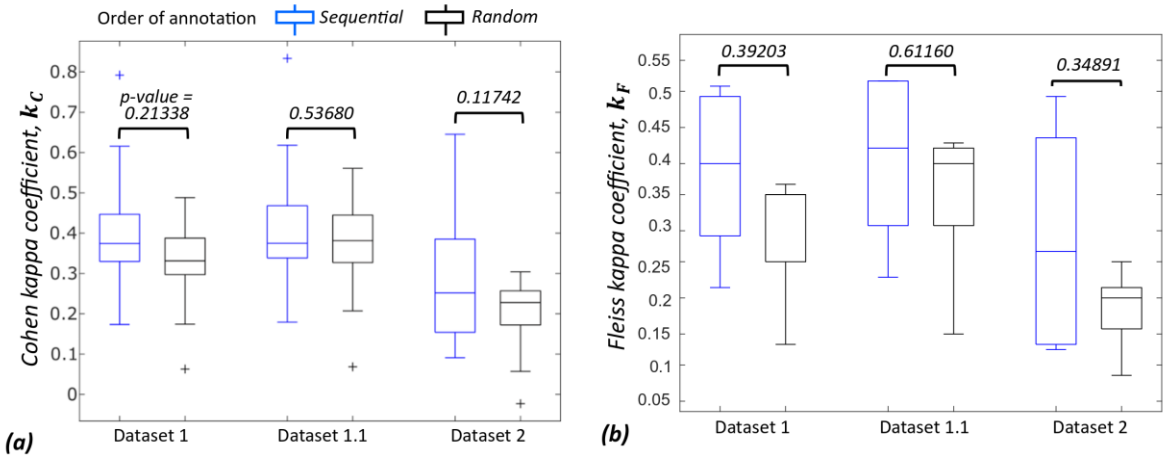


***Figure 25****. (a) Cohen kappa IAA results. Each box in the plot is summarizing the agreement between pair of annotators obtained across specific sets of labeled audio clips for different meetings. P-values are shown per pair of annotation mode. (b) Fleiss kappa IAA results. Each box in the plot summarizes the overall agreement of annotators per meeting and for the specific sets of labeled audio clips. P-values are shown per pair of annotation mode.*

randomized order. **Figure 25** shows the results of the IAA analysis. For both, the Cohen kappa and the Fleiss kappa, it can be observed that for all cases of datasets the IAA is slightly greater among audio segments labeled in a sequential order than in random order. A t-test demonstrated that the degree of agreement calculated by Cohen kappa and Fleiss kappa, across all cases of datasets, is not statistically significant. Therefore, our data suggests that annotators' agreement level does not change significantly by labeling audio in sequential or random order.

On the other hand, dropping audio segments from the initial dataset where none of the annotators agreed on a general intonation to create Dataset 1.1, resulted in a noticeable increase in IAA for the randomly labeled set. **Table 17** shows the number of audio clips that were dropped from both the sequential and random order labeled sets in Dataset 1 to create Dataset 1.1 because none of the annotators agreed on a label. In total, 42 in the sequential-order labeled set and 149 annotated items in the random-order labeled set were dropped. This constitutes a drop of 1.68% and 6.44% of the total number of audio segments in the sequential-order and random-order labeled sets, respectively. Although the difference in IAA between the two groups in Dataset 1.1 is not statistically significant (as shown in **Figure 25**), the difference in the number of audio segments that were dropped out resulted to be statistically significant. This shows that a higher level of IAA is achieved when speech intonation is labeled in sequential order.

***Table 17****. Summary of the total number of audio segments dropped from the final datasets because none of the annotators agreed on a label.*

| | *# of audio segments dropped* | |
|---|---|---|
| *Meeting* | *Sequential* | *Random* |
| 1 | 10 | 36 |
| 2 | 13 | 30 |
| 3 | 10 | 33 |
| 4 | 8 | 21 |
| 5 | 1 | 29 |
| *Mean* | 8.4 | 29.8 |
| *p-value* | 0.001582 | |

Note that, when looking across Dataset 1, Dataset 1.1, and Dataset 2, the Cohen kappa values fluctuate between -0.023 and 0.8338. Although a perfect agreement is not possible, typically, IAA is expected to be between 60% and 80% for the usefulness of the dataset in machine learning [222]. However, Uebersax [226] suggested that kappa values may be low even though there are high levels of agreement between annotators. In addition, most interpretations are performed considering 2-annotators and 2-categories were used to calculate kappa [227]. It was also noted in [228] that the number of categories and subjects will affect the magnitude of the kappa value. For example, the kappa is higher when there are fewer categories. However, as this is one of the best measurements of annotator agreement in the literature, it was employed in this study.

Although the level of agreement was higher for the datasets labeled in sequential order, the number of positive and negative general intonations was higher when audio clips were labeled in random order. This is illustrated using Dataset 1.1 in **Figure 26**, which shows the number of identified neutral, positive, negative, and question intonations for both sequential and random annotation order. A t-test revealed that the decrease in neutral-labeled audio clips and the increase in the number of positive, negative, and question-labeled audio clips are statistically significant. To gain a more accurate assessment of the increase in positive and negative intonations as a function of labeling in a sequential or random order, **Table 18** shows the percentage of positive, negative, and question assigned intonations within the non-neutral total number of labeled audio segments. Overall, 21.29% of the sequentially labeled set was assigned a non-neutral label. In contrast, 37.14% of the randomly labeled set was assigned a non-neutral label. As shown in **Table 18**, the random order labeled set contains 3% more positive labeled intonations and 8% more negative labeled intonations than the sequential order labeled set. These results suggest that an annotators' perception of speech intonation varies depending on the presence of the conversation's

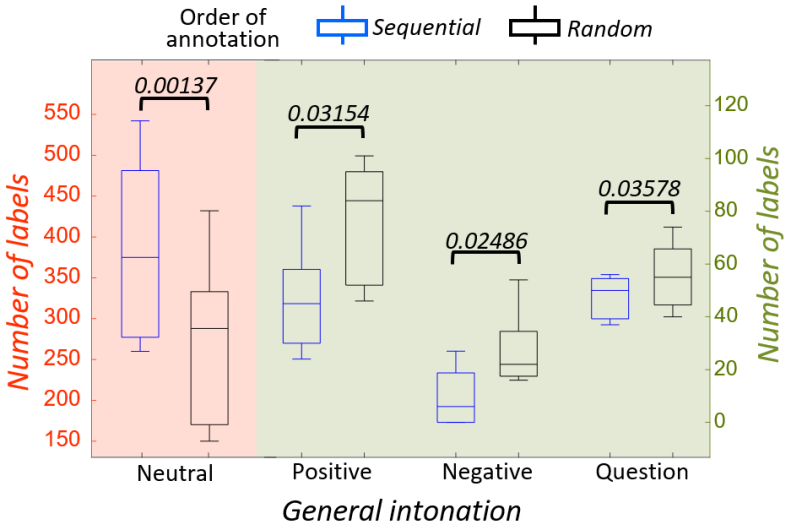


***Figure 26.*** *Number of identified neutral, positive, negative, and question intonations across annotators from Dataset 1.1 for both sequential and random order of annotation. P-values are shown per pair of labeled datasets for each of the different types of general intonations.*

***Table 18.*** *Percentage of positive, negative, and question labeled audio segments present in the non-neutral labeled portion of Dataset 1.1.*

| *Order of annotation* | *Total # of non-neutral labeled audio segments* | *Positive* | *Negative* | *Question* |
|---|---|---|---|---|
| Sequential | 519 | 44.86% | 9.44% | 45.66% |
| Random | 794 | 47.61% | 17.38% | 35.01% |

context, whereas in the absence of it, annotators may be more receptive to the nonverbal cues of the speech than to the meaning of the words. Consequently, annotators may identify more non-neutral speech intonations when audio segments are presented in random order.

Dataset 2 was used to study how identifying and assigning specific intonations is affected by annotating sequentially or randomly. In general, 75.42% of the sequential-order labeled set and 77.95% of the random-order labeled set were assigned a specific intonation by at least one annotator. **Figure 27** shows a summary of the percentage of specific intonations assigned by one, two, and three annotators. The plot shows how single annotators dominate the assignment of specific intonations. A comparison between annotations performed sequentially and randomly

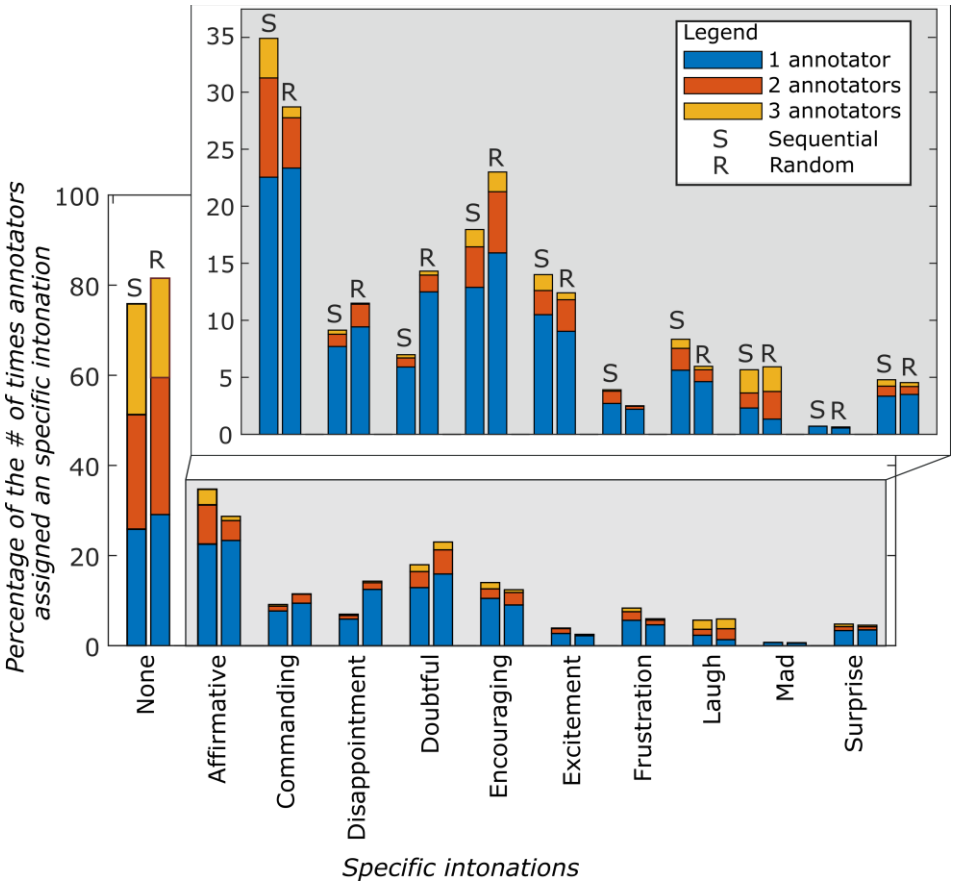


***Figure 27.*** *Summary of the percentage of specific intonations assigned by one, two, and three annotators. The plot shows how single annotators dominate the assignment of specific intonations, which may explain the low levels of IAA in Dataset 2.*

shows how intonations such as doubtful, commanding, and disappointment were more frequently identified when labeling in random order; affirmative, encouraging, excitement, and frustration were more frequently identified when labeling in sequential order. On the other hand, intonations of surprise, laugh, or madness seems to have been identified at a similar rate by both orders of annotation. However, only 31.96% and 27.06% of the total number of audio segments were assigned a specific intonation by two or more annotators for sequential and random labeling, respectively. When compared to the percentage of non-neutral general intonations assigned, the total percentage of labeled specific intonations is greater for sequential order of annotation. In

***Table 19.*** *Summary of the total number of audio clips that in both sequential and random order of annotation were assigned the same labeled (intersection) and the total size of the dataset if annotations from both sets were combined.*

| | *Order of annotation* | | *Intersection & Union* | |
|---|---|---|---|---|
| *General intonation* | *Sequential* | *Random* | ∩ | ∪ |
| Neutral | 1919 | 1344 | 1233 | 1538 |
| Positive | 233 | 378 | 143 | 463 |
| Negative | 49 | 138 | 30 | 152 |
| Question | 237 | 278 | 185 | 325 |
| **Total** | 2438 | 2138 | 1591 | 2478 |

contrast, the random order of annotation increases the assignment of non-neutral general intonations but decrease the assignment of specific intonation.

Dataset 1.1 was selected, over Datasets 1 and 2, to construct the final dataset to train the intonation recognition model because it contains the highest levels of IAA. To gain an understanding of how much overlap exists between the sequentially and randomly labeled sets in Dataset 1.1, **Table 19** shows the intersection and the union of both sets for each of the labels across all meetings. It can be noted that 92% of the randomly labeled set overlaps with the sequentially labeled set for neutral intonation, whereas for positive, negative, and question intonation the overlap is 38%, 22%, and 65%, respectively. To increase the number of positive, negative, and question-labeled items in the final dataset, the union of both sets was then taken as the final dataset for the intonation recognition model design. To assign labels to those audio segments outside of the interception set, we looked at the group-level agreement calculated using Fleiss kappa. For Dataset 1.1, the average Fleiss kappa across the five meetings for the sequentially labeled set is 0.403, whereas for the randomly labeled is 0.35. Therefore, for the audio segments outside the interception, the labels in the sequential-order set were given priority. However, if the label of an audio segment outside of the interception was neutral in the sequential-order set and the random-order set was non-neutral, the non-neutral label was assigned to that particular audio segment. This

resulted in a total of 1538 neutral, 463 positive, 152 negative, and 325 question-identified audio segments.

*4.2.5.2.Real-time intonation recognizer framework*

*4.2.5.2.1. Speech segmentation*

To identify speech segments in an automated and real-time fashion, a VAD was implemented with thresholding rules to determine what to consider speech and what to consider an estimated utterance. First, the recorded signals corresponding to the prepared dataset were downsampled from 32 kHz to 8 kHz. Because the audio obtained from Zoom has a high signal-to-noise ratio, a manual threshold of 0.01 was selected for the VAD. Then, thresholds for minimum speech time and maximum silence time to estimate utterances were determined by evaluating the distribution of the minimum speech periods and maximum silence periods present in the manually identified audio segments. **Figure 28** shows the distribution of the identified silent periods in the manually segmented audio. To determine a threshold of maximum silent duration before considering a new estimated utterance, the 90-percentile of the distribution was calculated, setting up the threshold to be 0.58s. Therefore, if a silent period passes the threshold of 0.58s, the next detected speech

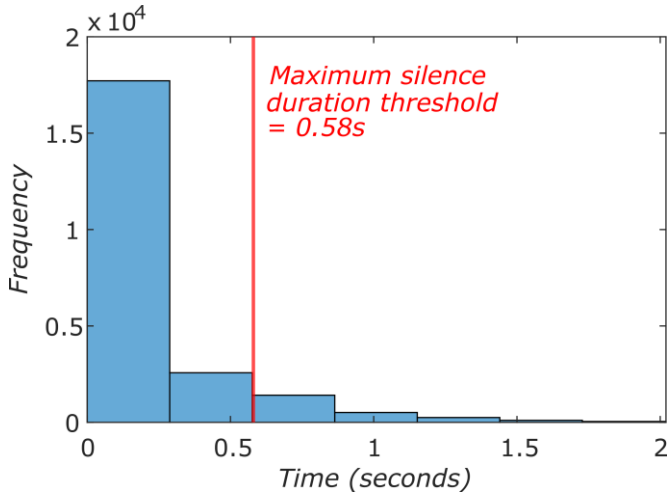


***Figure 28.*** *Distribution of the identified silent periods in the manually segmented audios and display of the 90-percentile of the distribution, which was set as the threshold for maximum silence duration before considering a new utterance.*

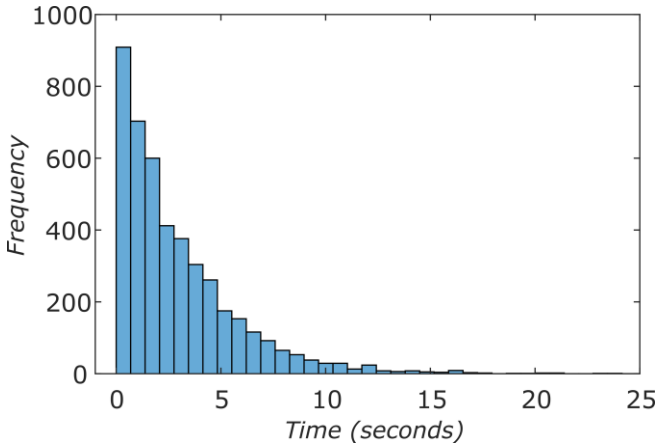


***Figure 29.*** *Distribution of audio segment lengths obtained by the speech segmentation block.*

period is considered a new utterance. Because there may have been buffers of data that are detected as speech but that are accurate noises, a minimum length of time with detected speech was determined. The dataset contains back-channel signals (i.e., laughs, "yes", "no", etc.), therefore the length of such back-channels was considered to set up the minimum speech threshold, which was set to be 4 windows of data or 0.12s. **Figure 29** shows the distribution of the audio segment lengths when the maximum silence duration and the minimum speech thresholds were applied to the dataset.

*4.2.5.2.2. Sampling rate, signal feature, and classification models evaluation*

Using the 8kHz signals, a total of 67 features calculated from time series features were extracted and evaluated using correlation analysis. The correlation analysis revealed that 25 out of the 67 evaluated features were highly correlated. To evaluate the effect that eliminating those 25 features may have on the classification performance, a variety of classifiers were trained to recognize 4 classes (negative, positive, question, neutral), 3 classes (negative, positive, question), and 2 classes (combinations of pairs of negative, positive, question, and neutral classes) using both sets of features (complete and reduced). **Figure 30** shows the results of classification accuracy for

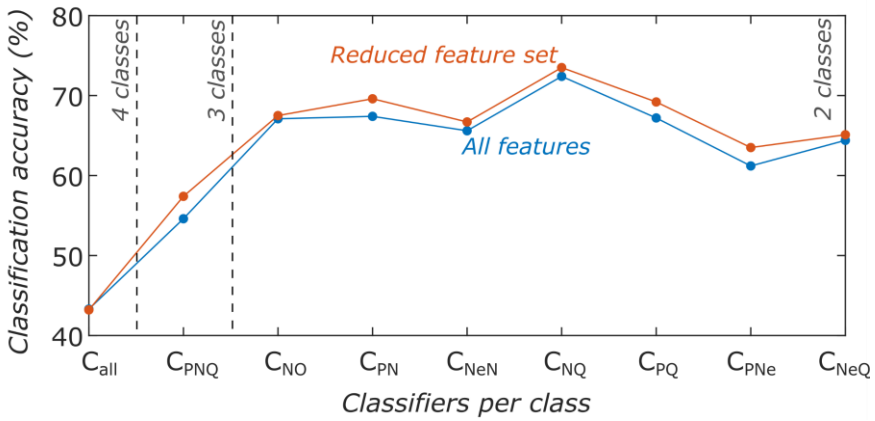


***Figure 30**. Evaluation of different classification model accuracies using two sets of features: (1) all features extracted and (2) a reduced feature set obtained by eliminating highly correlated features. Abbreviations: $C_{all}$ – classifier for four classes, $C_{PNQ}$ – classifier for positive, negative, and question classes, $C_{NO}$ – classifier for negative and all other classes combined, $C_{PN}$ – classifier for positive and negative, $C_{NeN}$ - classifier for neutral and negative, $C_{NQ}$ - classifier for negative and question, $C_{PQ}$ - classifier for positive and question, $C_{PNe}$ - classifier for positive and neutral, $C_{NeQ}$ - classifier for neutral and question.*

models trained using a different number of classes and trained using the initial 67 features and the reduced set of 42 features (shown in **Table 20**). It can be observed that for all cases of trained classifiers, the reduced set performs comparable to or better than the original set. Therefore, no loss in classification accuracy is obtained when reducing the feature set using correlation analysis. The type of classification model from which the displayed accuracies in **Figure 30** were obtained varies across all classifiers per class. However, the predominant model was SVM with a Gaussian kernel. All models were trained using 70% of the data with a 10-fold cross-validation approach to minimize overfitting results.

To evaluate the effect that reducing the sampling rate may have on classification accuracy, an SVM with a Gaussian kernel was selected based on the results from the correlation analysis. Furthermore, features that did not present a Gaussian distribution from the 42 features listed in **Table 20** were eliminated from this part of the evaluation. Features that did not follow a Gaussian

***Table 20**. Final list of features used for classification of intonations. This final list was obtained after eliminating 25 highly correlated feature sets out of 67.*

| *Feature type* | *Feature name – time series* | *Final set of features* | *Qty* |
|---|---|---|---|
| Prosodic | Signal energy | Mean, min | 2 |
| | Mel pitch (pitch value on Mel scale) | Min, max, mean, std, number of peaks, mean peak value, std of peak values, mode of peak values | 8 |
| | Δ-Mel pitch | Mean peak value, std of peak values, and mean, std, and median of last 5 non-zero value | 5 |
| | ΔΔ-Mel pitch | std of peak values | 1 |
| Conversation | Speaking rate | Voiced/Unvoiced | 1 |
| | Pausing rate | Silence frames/total frames | 1 |
| Voice quality | Zero-crossing rate | Mean, std, min, median | 4 |
| Frequency spectrum coefficients | Mean of power spectral density | Mean, min, median | 3 |
| Cepstral coefficients | Mel-frequency cepstral coefficients (MFCC) | Mean, std, max, min of first 4 MFCC coefficients and range of the $4^{th}$ MFCC coefficients | 17 |

distribution included the number of peaks in Mel pitch, frequency spectrum coefficients, signal energy, speaking rate, pausing rate, and std of the last 5 non-zero values of Δ-Mel pitch. In total, the models were trained with 33 features. The interest in exploring the effect of sampling rate over classification accuracy comes from the goal of designing a computationally and real-time resource-aware intonation recognition unit.

**Table 21** shows the results of the sampling rate analysis. Note that even when the overall classification accuracy of the models across the evaluated sampling rate does not change significantly, the precision accuracies do change for positive and negative intonations. The precision accuracy to recognize positive intonations decreases, although not consistently, as the sampling rate is decreased. On the other hand, the models seem to become more sensitive to negative intonations as the sampling rate is reduced. Because the accurate recognition of positive

***Table 21**. Comparison of the results of classification accuracy when sampling rate of the input signal is varied. The overall accuracy of the classification model does not seem to suffer a significant reduction; however, the precision accuracy of positive audio segments/estimated utterances do decrease by at least 1/3.*

| *Sampling rate* | *Model accuracy (validation training/testing)* | *Precision accuracy* | | | |
|---|---|---|---|---|---|
| | | *Question* | *Positive* | *Negative* | *Neutral* |
| 8kHz | 40.53%/41.76% | 46.81% | 45.74% | 51.06% | 23.40% |
| 4kHz | 40.18%/42.45% | 54.26% | 21.28% | 63.70% | 21.28% |
| 3.2kHz | 39.04%/39.23% | 46.81% | 24.47% | 54.41% | 24.47% |
| 2kHz | 38.13%/41.09% | 42.55% | 31.91% | 57.55% | 24.47% |

and negative intonations is important for evaluating the level of positivity contributing to the rapport between dyads and groups, 8 kHz was used in further analysis.

To investigate how well specific models adapt to the recognition of specific intonations, classification models were trained for the recognition of four, three, and two classes. **Table 22** presents a summary of the results. The displayed summary suggest that models trained to classify two classes achieve a more balance precision accuracy. In addition, models that focus on positive, negative, and question intonations also achieve a higher level of balance accuracies across its classes. A possible reason for the low classification accuracies of the neutral class is that based on

***Table 22.** Summary of best performing classification model with their respective model and precision accuracy.*

| *Model type* | *Model accuracy* | *Precision accuracy* | | | |
|---|---|---|---|---|---|
| | | *Negative* | *Positive* | *Question* | *Neutral* |
| Medium Gaussian SVM | 41.76% | 47% | 46% | 51% | 23% |
| Medium Gaussian SVM | 57.40% | 60% | 50% | 62% | |
| Medium Gaussian SVM | 65.10% | | | 64% | 66% |
| Medium Gaussian SVM | 69.60% | 71% | 68% | | |
| Linear SVM | 66.70% | 78% | | | 55% |
| Coarse Gaussian SVM | 63.50% | | 65% | | 62% |
| Medium Gaussian SVM | 73.50% | 71% | | 76% | |
| Medium Gaussian SVM | 69.20% | | 68% | 70% | |
| Linear SVM | 67.50% | 86% | 40% | | |

analysis of annotations, audio segments that were mark as neutral often carried a specific intonation that could be grouped with negative or positive types of intonation.

To better understand the advantages and disadvantages of this work, **Table 23** compares this approach with others in the literature. Works are compared based on the type of data used for training, the real-time capability, the sampling rate, the linguistic unit used for feature extraction, the number and type of extracted features, the type of classifier, the type of classes, and the percentage of accuracy. The works in [142], [144], [189], [229] do not perform real-time processing and most of those works made use of acted databases, not accounting for speakers with different cultural backgrounds, variations in the recording environment, and variation in microphone-distance. On the other hand, the works that perform real-time processing [230], [231] have focused on the recognition of affective state classes, instead of combination with other types of intonations. For example, the work by Alonso et al. [230] demonstrated the use of just 6 features for the classification of 5 affective classes, achieving from 41.06% to 52.43% classification accuracy when processing natural speech datasets. However, the used sampling rate and type of linguistic unit for feature extraction suggest that the classification of affective states is performed at a high rate compared to the other works presented in **Table 23**. The application of human/group behavior monitoring does not require such a high recognition rate for the quantification of positivity levels contributing to the rapport between people in an interaction. The work presented in this chapter work uniquely focuses on combining affective classes with question intonation. The interest in recognizing question intonations was to better understand the dynamics of a conversation and patterns of answering positively or negatively. In addition, the processing of a natural dataset at the low sampling frequency of 8kHz and estimation of a sentence-level utterance, whereas other works used voiced frames for classification or the acted speech segment from their

respective datasets, represent an advantage for real-time processing. In terms of classification performance, although difficult to compare because of the nature of the constructed dataset and

***Table 23.*** *Comparison of selected works in the research area of speech emotion recognition. Abbreviation for classifiers: Support Vector Machine (SVM), Meta Decision Tree (MDT), Gaussian Mixture Models (GMM), Auto-Associative Neural Networks (AANN), Sequential Minimal Optimization (SMO).*

| Reference | Real-time | Dataset | Sampling rate | Linguistic unit | # Features | Features type | Classifier | Classes | Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| [142] | No | Natural | 16kHz | - | 22 | Prosodic, voice quality | SVM | Positive negative | 52% |
| [144] | No | Acted | 16kHz | Sentence level | 253 | Prosodic, semantic | SVT+ MDT | Neutral, happy, angry, sad | 80% |
| [189] | No | Acted | - | Sentence level | - | Prosodic, cepstral, wavelet | GMM | Happiness, anger, fear, sadness, surprise, neutral | 66% |
| [229] | No | Acted | 8kHz | Sentence level | 400 | Voice quality, spectral | SVM +AANN | Anger, disgust, fear, happy, neutral, sadness | 84% |
| [230] | Yes | Natural | 16kHz | Voiced frame | 6 | Prosodic, spectral | SVM | Anger, boredom, happy, neutral, sadness | 41.06% - 52.43% |
| [231] | Yes | Acted | - | Sentence level | - | Prosodic | SMO | Happy, sad, surprise, fear, disgust, anger, neutral | 67% |
| This work | Yes | Natural | 8kHz | Estimation of sentence | 42 | Prosodic, voice quality, cepstral | SVM | Positive, negative | 70% |
| | | | | | | | | Negative, question | 74% |
| | | | | | | | | Positive, negative, question | 57% |
| | | | | | | | | Positive, negative, question, neutral | 42% |

the types of classes being classified, the results of this work are better or comparable to those previously in the literature. However, improvements should be made in the number of features and type of classifier used to improve computational efficiency.

### 4.3.Overall Discussion

In general, this work focused on two main technical points when designing sensor signal processing algorithms for the recognition of local transformed features. The first technical point relates to the collection of data to train models to recognize behavioral cues of interest and the second to the evaluation of signal processing and machine learning model parameters to increase computational efficiency.

In this chapter, the collection of data to train machine learning models can be classified/divided into two ways: acted/evoke data collection and natural data collection. Acted/evoke data was collected for the training of the head action detection model, while natural data was collected for the training of the speech intonation detection model. Acted/evoke datasets are good for fast prototyping because the onset of events of interest or "classes of interest" is known from the data collection processes. For example, the dataset collected to train the HAD unit was performed in a manner that evoked the actions of nodding (Δ-pitch), shaking (Δ-yaw), and rolling (Δ-roll) the head. Therefore, the onset of the action of interest was known and no data annotation process was needed to prepare the dataset for processing. However, training models with acted/evoke data may not perform as expected when running/implementing these models in the wild because it does not carry the level of noise or variation in events that may be encountered in a natural environment.

On the other hand, natural datasets are better at representing the reality of day-to-day interaction. However, the preparation of natural datasets is subjective to annotators, creating a high level of variability in assigned labels among annotators, and is time-consuming. For example, the

preparation of the dataset collected to train the speech intonation recognition model required the participation of a total of at least six annotators: two to perform manual segmentation and four to perform annotation of speech intonations. Therefore, well-established data annotation procedures can help decrease variability in assigned data labels and help establish a minimum number of required annotators to obtain an optimal dataset. This chapter shows how the level of inter-annotator agreement varies depending on the protocol used for data annotation. In the case presented here, segments of speech were labeled in the order in which they occurred, as well as in random order. This revealed that when speech segments are labeled in random order there are more non-neutral intonations identified than when the segments are labeled in order, possibly confirming that the lack of context in the speech segments influences the perception of intonations. Labeling intonations of speech segments presented in random order may be preferable when designing human behavior monitoring systems that are free of speech recognition units or any other methodology that provides information about the context of a conversation.

When evaluating different signal processing and machine learning parameters to decrease computational complexity while maintaining good accuracy, this work looked at data buffer sizes for real-time processing, the number of signal features used for classification, the type of features, and the complexity of selected models. Optimal machine learning models make use of a combination of the optimal aforementioned parameters. However, on occasions, optimized parameters can reduce the ability to generalize the classification models depending on the dataset used for training. For example, feature reduction techniques used to reduce the computational complexity of an overall machine learning pipeline can increase the classification accuracy of the training dataset, but they can also decrease the ability of the model to be transferable to cases outside the ones in which the model was trained. This can be particularly true when using

acted/evoked datasets. Therefore, it is recommended to use natural data to confirm the performance of optimized models designed with acted/evoked datasets. On the other hand, a combination or fusion of optimized classification models provides the opportunity to simplify re-training processes, if necessary or desired, and reduce computational time and power consumption. This work implemented this methodology in the design of the real-time HAD unit.

### 4.4.Summary

This chapter presents the design and implementation of real-time data processing blocks to recognize head activity and intonations using IMUs and audio signals, respectively. The HAD unit was trained with collected data from a laboratory environment. The HAD unit recognizes three static positions and three dynamic motions (i.e., Δ-pitch, Δ-yaw, Δ-roll) with an accuracy of 97.91%. On the other hand, a real-time speech intonation recognizer was trained using natural data collected during research team meetings and labeled using affective states and an interrogative expression. The natural dataset was constructed by analyzing two methods of labeling intonations: in sequential order or random order of occurrence. To the best of our knowledge, this is the first reported effort that studies the effects of labeling speech intonations using different orders of presentation, i.e., preserving the context of the interaction when labeling in sequential order or eliminating context when labeling in random order. Results revealed that labeling in sequential order leads to a higher level of inter-annotator agreement, wherein labeling in random order leads to a higher level of non-neutral intonations being recognized by two or more annotators. As the use of nonverbal behaviors to train machines for the recognition of human behaviors excludes contextual information, this may suggest that, in preparing natural datasets for training such systems, labeling in random order may be preferred. Furthermore, the trained speech intonation recognizer achieved a 70% classification accuracy when classifying positive and negative classes

and 57% when classifying between positive, negative, and question intonations. This also represents the first effort in combining affective classes with an interrogative intonation. The next chapter shows insights into the design and execution of a social interaction study to expand the available datasets for the complete design and implementation of the real-time machine learning framework.

# 5. SOCIAL INTERACTION STUDY: METHODS, DESCRIPTION OF DATA COLLECTION, AND ANALYSIS

The human studies and collected datasets presented in Chapter 4 served as the basis to start the design of models, able to identify individual and nonverbal behavioral cues of interest, that form part of the machine learning framework presented in Chapter 3. However, to explore and draw relationships between nonverbal cues from multiple individuals involved in an interaction and multiple channels of communication, a more comprehensive dataset needs to be utilized. Currently, no publicly available dataset exists that meets the needs of this work, that is, a dataset composed of audio, IMU, and physiological data from a head-mounted device, emotional state labels, and rapport labels. Therefore, in this chapter, the design and execution of a social interaction study are presented, together with the description of sensor data and survey data collected.

## 5.1. Study Methods

The goals of this human study were (1) to collect audio, visual, and physiological sensor data while a group of individuals was interacting for a given period of time, (2) to provide an environment where low and high levels of rapport could be evoked, and (3) collect self-reported data about the liking between dyads in a group and their perceived dyadic and group rapport level.

### 5.1.1. The Basis for Recruitment of Participants

For this study, dyads were considered the basic unit of interest to understand group consonance. Because our interest is in group interactions, groups were required to be composed of a minimum of 3 individuals, which contains 3 dyadic interactions. This work aimed at collecting data from at least 60 dyadic interactions which led to the aim of forming groups of 4 individuals, which each contains 6 dyadic interactions. However, due to human factors such as participants' availability or not being able to complete the study, some groups were composed of 3 individuals.

This resulted in the study collecting data from a total of 10 groups formed with 3 to 4 individuals, which required a sample size of ~40 individuals.

5.1.2. Study Overview

The study procedure was approved by the Michigan State University (MSU) Institutional Review Board (IRB) and conducted under strict physical distance and following privacy protocol guidelines. To form the 10 groups of 3 to 4 individuals, the study was divided into two parts: (1) consent to participate in the study and the administration of two questionnaires and (2) the interaction between participants, where multi-sensor data was collected, and additional administration of questionnaires. Participants' interaction consisted of two periods of 20 minutes, where in each period participants were discussing a topic statement given to them. **Figure 31** describes the parts involved in the study and their respective approximate duration.

The study was advertised through email around various departments across MSU and in flyers posted around university buildings. Therefore, participants were recruited from the MSU campus, however, there were no requirements for subjects to be students or MSU affiliated in any way. Interested participants were first asked to fill out a contact release form that briefly defined the goals of the study and participant criteria. The contact release form also allowed potential participants to submit their contact information and confirm that they met eligibility criteria. Individuals were eligible to participate in the study if they were 18 years of age or older and could be physically present on the MSU campus at the time of the second part of the study. Participants

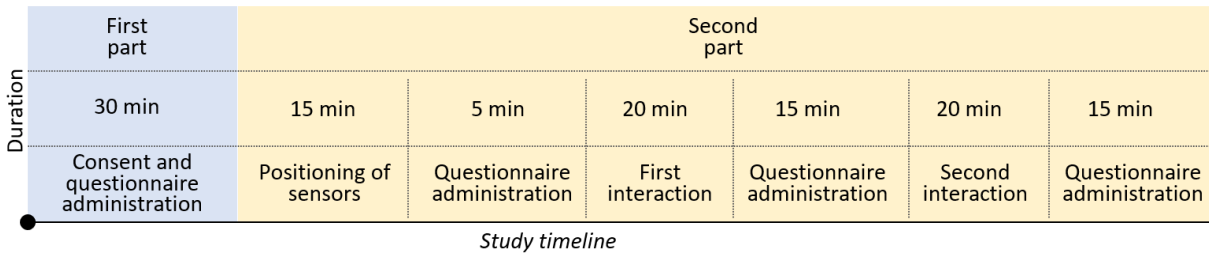


***Figure 31.*** *General description of the social interaction study timeline.*

were individually contacted to schedule a 30-minute Zoom meeting to perform the first part of the study.

### 5.1.3. First Part of the Study

During the first part of the study, the consent form was discussed and signed by the participant. Then, the participant was provided with two questionnaires. The first one was a Demographic questionnaire that collected information about their gender, age, ethnicity, educational background, and current employment status. The second one was a Topic questionnaire that asked participants to provide their opinion (how much they agree or disagree) using an 11-point Likert scale on a series of topic statements that included gun control, vegetarianism, animal testing, universal healthcare, death penalty, religious freedom, professional sports, vaccines, college athletes, environment, animal hunting, exercise, TV shows, travel, video games, food, outdoor activities, and social interactions (see APPENDIX A). After the questionnaires were completed, the participant was asked for their availability to perform the second part of the study. The responses to the Topic questionnaire together with the availability of the participants were used to form the 10 groups of 3 to 4 participants.

### 5.1.4. Topic Statement Selection and Group Formation

During the second part of the study, each group participated in two interactions, wherein two different topics were discussed. The first topic was intended to be one that not all individuals in the group agreed on and the second one was intended to be one where all participants had a similar opinion. Therefore, groups were formed by matching individuals' responses from the Topic questionnaire to invoke the desired level of interaction at each discussion section. The goal was to invoke conflict during the first discussion section that could affect the establishment of rapport but invoke an increase in rapport during the second interaction. **Table 24** shows a summary of the

***Table 24***. *Summary of topics selected for discussion and the average level of group agreement.*

| *Group* | *Num of individuals* | *Topic #1* | *Disagreement* | | *Topic #2* | *Agreement* | |
|---|---|---|---|---|---|---|---|
| | | | *Average* | *Range* | | *Average* | *Range* |
| 1 | 4 | Death penalty | 4.25 | 8 | Environment | 1 | 2 |
| 2 | 4 | Death penalty | 3.5 | 7 | Vaccines | 1 | 2 |
| 3 | 4 | Animal hunting | 4.5 | 9 | Environment | 1 | 2 |
| 4 | 3 | Gun control | 6 | 7 | College athletes | 5 | 0 |
| 5 | 4 | Animal hunting | 3.75 | 10 | Universal healthcare | 8.5 | 5 |
| 6 | 3 | Death penalty | 4 | 7 | Vaccines | 8 | 6 |
| 7 | 3 | Animal testing | 3 | 3 | Environment | 2 | 5 |
| 8 | 4 | Vaccines | 5.25 | 6 | Death penalty | 7.25 | 4 |
| 9 | 3 | Death penalty | 5.7 | 10 | Gun control | 1 | 2 |
| 10 | 4 | Animal testing | 5 | 5 | Environment | 1.75 | 5 |
| | | | ***Average*** | 7.2 | | ***Average*** | 3.3 |

topics selected for discussion for each group and their level of agreement in opinion. In general, the first topic was selected by looking at an average level of agreement of 5 (neutral opinion), but with a range value in opinions of 5 or more, which indicates the presence of diverse opinions. The second topic was selected by looking at an average level of agreement close to 1 or 10 with a range value of less than 5, or an average level of agreement of 5 with a range value of less than 1. However, there were cases where the availability of the participants limited the groups that could be formed and the variety of opinions available, which was the case of Group 6 and Group 7 for the second and first topics of discussion, respectively.

### 5.1.5. Second Part of the Study and Main Procedure

The second part of the study, which constituted the main part of the study, took place in a large laboratory space with four separate rooms. Each participant was assigned to a room that was equipped with a computer, the Zoom meeting software, a microphone, a webcam, a BrainBit headband, a Shimmer device, and the infrastructure to collect data through LSL. A study team member helped the participants to put on the wearable sensors (the BrainBit and Shimmer), as shown in **Figure 11**. Participants were then instructed to fill out an emotional state questionnaire (see APPENDIX B) containing a 9-point self-assessment manikin arousal, valence, and dominance (AVD) scale [232] and an 11-point rating tool based on the circumplex model of emotion [39], [233]. In the 9-point Likert arousal, valence, and dominance scale, participants were instructed to use arousal to describe how intense is their current emotion, using 1 as low and 9 as high, valence to describe how negative or positive is their current emotion, using 1 as negative and 9 as positive, and dominance to describe the degree to which their current emotion controls their thoughts and actions, using 1 as low and 9 as high. In the 11-point rating tool based on the circumplex model of emotion, participants were instructed to "rate how are you feeling at this moment using the following scale." Eight 11-point Likert scales were presented evaluating the following items: tense-calm, nervous-relaxed, stressed-serene, upset-contented, sad-happy, depressed-elated, lethargic-excited, and bored-alert, where the far left of the scale (score of 1) belong to the negative feeling and the far right (score of 11) to the positive one.

Participants were given the first topic statement for discussion and instructed to write at least three reasons to back up their opinion of the issue. Participants were also instructed to discuss the topic statement among themselves, to share their opinion during the interaction, and to persuade

those with a difference in opinion that their personal view was more reasonable. The instructions were given as follows, where the topic statement for "death penalty" is used as an example:

*"Consider the following statement: "The death penalty should be used to deter heinous crimes." for which you expressed on a scale from 0 (very strongly disagree) – 10 (very strongly agree) that your opinion is better described by a _X_ (inkling to agree/disagree/neutral).*

*Your first task is to make a note (below) of at least three reasons why you have this opinion. Then, during the virtual meeting, your task is to discuss these reasons and sway other attendees towards your point of view if differences in opinion are found. During the virtual meeting discussion, you should also try to learn the specific reasons other attendees express their opinions."*

where X represents the score given on how much they agree or disagree. The group was then left to discuss the topic statement for ~20 minutes. At the end of the discussion, participants were asked to fill out the emotional state questionnaire and a rapport questionnaire.

The criterion to measure rapport was performed using items derived from [120] as described in [121]. Some items were prefaced with the instruction to "rate yourself in the interaction on the following characteristics." The items were smooth, bored, cooperative, satisfied, comfortable, awkward, engrossed, involved, friendly, active, and positive. The remaining items were prefaced with the instruction to "rate the interaction between you and X on the following characteristics," where X represented one of the other two (for groups of 3) or three (for groups of 4) individuals in the interaction. The items included well-coordinated, boring, cooperative, harmonious, unsatisfying, uncomfortably paced, cold, awkward, engrossing, unfocused, involving, intense, unfriendly, active, positive, dull, worthwhile, and slow. Responses were recorded on five-point Likert scales.

The rapport questionnaire included the question “How much are you enjoying the discussion?,” which answer was recorded on an 11-point Likert scale. Also, a liking score was obtained from each individual in the interaction in relation to everybody else. This score was obtained by a five-point Likert scale when asked “Do you like your interaction with subject X?,” where X represented one of the other two (for groups of 3) or three (for groups of 4) individuals in the interaction (see APPENDIX C).

After the questionnaires were filled out, a second topic statement for discussion was given and participants were asked to follow previous discussion instructions. In the end, participants filled out the emotional state and rapport questionnaires.

### 5.2.Data Collection and Description

All collected data were managed by the study coordinator from a central computer and saved using an XDF data format [234]. The multi-sensor hardware and software infrastructure presented in Chapter 3 was the one used for the collection and management of sensor data. This process was transparent to the participants. Data collection through LSL was primarily performed during the two interaction periods, however, data was also collected after each interaction, while participants were filling out the questionnaires, for data quality assurance purposes. For each group, a total of four XDF files were generated: two corresponding to the interaction periods and two corresponding to the administration of the questionnaires. In addition, the entire study was recorded through Zoom, where the video of the meeting was obtained for annotation purposes, in addition to audio for each participant. However, because the Zoom meeting and the four periods of sensor data were recorded separately, the video from Zoom needed to be synchronized with the periods of sensor data. This synchronization was manually performed using the audio recorded from LSL and aligning them with the audio obtained from the Zoom recording. The

***Table 25***. *Summary of corrupted data and lost data from the first three groups due to a technical issue. "-" indicates no loss and "x" indicates corrupted data or lost data.*

| *Group* | *Participant* | *Interaction* | *Audio* | *PPG* | *Acc* | *Gyr* | *Mag* | *EEG* |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1st | - | x | x | x | x | - |
| | | 2nd | - | x | x | x | x | - |
| | 2 | 1st | - | x | - | - | - | x |
| | | 2nd | - | x | - | - | - | x |
| | 4 | 1st | - | - | x | - | - | - |
| | | 2nd | - | - | x | - | - | - |
| 2 | 2 | 1st | - | - | - | x | x | - |
| | | 2nd | - | - | - | x | x | - |
| 3 | 1 | 2nd | - | x | x | x | x | - |
| | 4 | 1st | - | - | - | x | x | - |
| | | 2nd | - | - | - | x | x | - |

synchronization was performed for the periods of data recorded during the interactions. Therefore, for each group, two videos were produced after synchronization, each corresponding to the two interactions that they carried on.

The overall dataset consists of 20 group discussions in English, 2 per group, each lasting on average 21 minutes. This results in an average total of 420 minutes of audio, visual, and physiological data. However, due to technical problems, part of the data from groups one to three (summarized in **Table 25**) was corrupted. Data corruption and, on occasions, loss of data problems seemed to be related to the order in which the sensors, especially the shimmer was prepared for connection to the multi-sensor system. It was determined that the PPG connector from the Shimmer device needed to be connected before turning it on and connecting it to its respective computer.

## 5.3.Summary and Analysis of Questionnaires' Data

### 5.3.1. Demographics

The second part of the study had a total participation of 35 individuals (21 males, 12 females, and 2 that identified as other). One of the female participants formed part of two groups.

Participants' age ranged from 18 to 44 years, where 15 individuals were in the range of 18-24 years old, 17 were in the range of 25-34 years old, and 3 were in the range of 35-44 years old. Participants' ethnicities were predominantly White and Asian with 15 and 10 participants, respectively. Other represented ethnicities included Black or African American, Native Hawaiian or Pacific Islander, American Indian or Alaska Native, and combinations of all of them. Participants' highest level of education ranged from having a high school diploma to have a doctorate or professional degree, where 10 participants indicated that they have some college credits, 10 had bachelor's degrees, and 10 had master's degrees. In terms of employment, 29 of the participants identified as students, 3 as having a part-time job, and 3 as having a full-time job.

#### 5.3.2. Emotional State

The emotional state questionnaire employed two scales a 9-point Likert AVD scale and an 11-point rating tool based on the circumplex model of emotion. To better display whether a participant was feeling more of a negative or positive feeling, the responses to the 9-point Likert AVD scale were transformed and centralized to 0, meaning that the scale was modified to go from -4 to 4, instead of 1 to 9. Likewise, the 11-point rating tool scale was modified to go from -5 to 5, instead of 1 to 11. **Table 26** and **Table 27** show a summary (average and standard deviation) of the responses provided by the participants, for each instance in which the emotional questionnaire was filled out. **Table 26** and **Table 27** also show the results of a two-sample t-Test for equal means that was applied to two sets of data to find which emotional states were significantly affected throughout the interactions. The two sets of data were (1) the responses to the emotional state questionnaire before and after the first interaction and (2) the responses obtained after the first interaction and after the second interaction. Results show that based on the responses to the arousal, valence, and dominance scale, there was a statistically significant change in arousal and valence

***Table 26.*** *Summary of the responses provided by the participants for the 9-point Likert AVD scale. This table shows the average and standard deviation of the provided responses to the items in the scale for each instance in which the emotional state questionnaire was filled out. Also, the P-values resulted from the t-Test applied between each of the instances in which the questionnaire was filled out are shown. These results demonstrate that there is a significant change in arousal and valence before and after the 1st interaction.*

| *Scale Items* | *Before 1st interaction* | *After 1st interaction* | *After 2nd interaction* | *P-value between before and after 1st interaction* | *P-value between 1st and 2nd interaction* |
|---|---|---|---|---|---|
| *Arousal* | -1.94±1.56 | -0.28±1.92 | -0.23±2.04 | ***0.00013*** | 0.8999 |
| *Valence* | 0.53±1.42 | 1.06±1.47 | 1.63±1.55 | ***0.0463*** | 0.4170 |
| *Dominance* | -0.79±1.86 | 0.17±1.54 | -0.22±1.76 | 0.2622 | 0.0874 |

***Table 27.*** *Summary of the responses provided by the participants for the 11-point rating tool based on circumplex model of emotion. This table shows the average and standard deviation of the provided responses to the items in the scale for each instance in which the emotional state questionnaire was filled out. Also, the P-values resulted from the t-Test applied between each of the instances in which the questionnaire was filled out are shown. These results demonstrate that there is a significant change in two of the items before and after the 1st interaction and three of the items before and after the 2nd interaction.*

| *Scale Items* | *Before 1st interaction* | *After 1st interaction* | *After 2nd interaction* | *P-value between before and after 1st interaction* | *P-value between 1st and 2nd interaction* |
|---|---|---|---|---|---|
| *Tense-Calm* | 0.33±2.1 | -0.75±1.92 | 0.47±2.37 | 0.3821 | ***0.0189*** |
| *Nervous-Relaxed* | -0.64±2.22 | -0.33±1.82 | 0.92±1.99 | 0.5251 | ***0.0070*** |
| *Stressed-Serene* | -0.61±2.03 | -0.5±1.7 | 0.54±1.96 | 0.5623 | ***0.0289*** |
| *Upset-Contented* | 0.14±2.09 | 0.19±1.51 | 0.92±2.01 | 0.8973 | 0.0886 |
| *Sad-Happy* | -0.39±1.52 | 0.14±1.68 | 0.53±1.95 | 0.1657 | 0.3671 |
| *Depressed-Elated* | -0.53±1.67 | -0.17±1.92 | 0.2±1.95 | 0.4239 | 0.5720 |
| *Lethargic-Excited* | -1.06±1.87 | -0.19±1.41 | -0.03±1.68 | ***0.0305*** | 0.6501 |
| *Bored-Alert* | -0.36±1.64 | 0.97±1.99 | 0.81±2.14 | ***0.0028*** | 0.7331 |

levels from before to after the first interaction. Related to the rating tool based on the circumplex model of emotion, results show a statistically significant change in the lethargic-excited and bored-

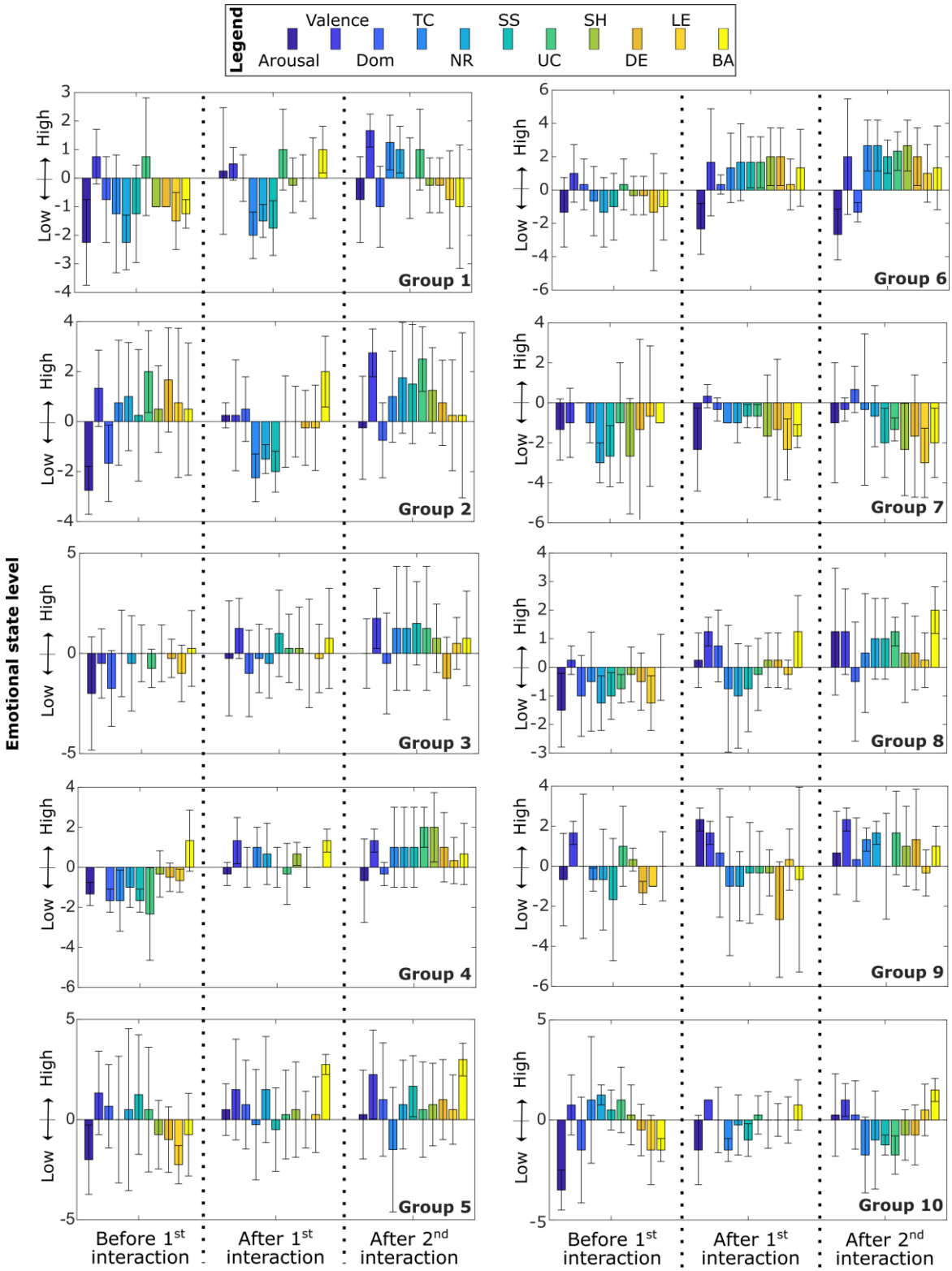


***Figure 32**. Summary of responses to the Emotional state questionnaire. The bars represent the average score given by all the participants of the group and the error bars represent the stand deviation of those responses.*

alert levels from before to after the first interaction and in the tense-calm, nervous-relaxed, and

stressed-serene levels from after the first interaction to after the second one.

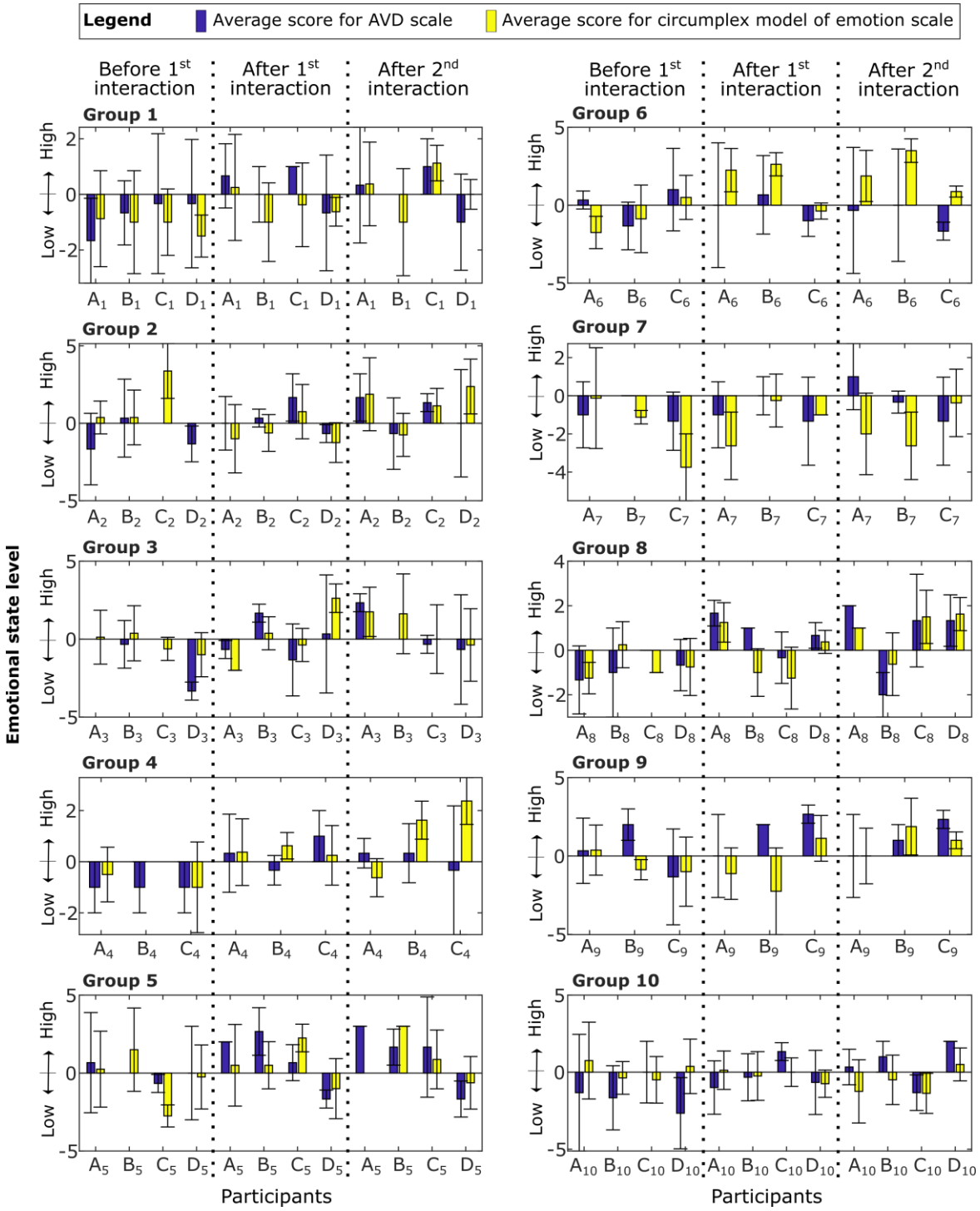


***Figure 33.*** *Overall scores with standard deviation bars for the AVD and circumplex model of emotion scales per individual per group for before and after the interaction sections. The overall scores were determined by calculating the average of the responses given by the participants for the items on each of the two scales.*

Because this work is interested in looking at the group-level behavioral factors that influence rapport, the averages of individuals' responses to the emotional state questionnaire are presented per group in **Figure 32**. The information presented in **Figure 32** provides insight into the level of positivity within the group. Generally, it can be observed that before and after the 1st interaction there is a variation of low and high affective states across individuals of a group and groups, while after the 2nd interaction there was a tendency to be at a high emotional state except for Groups 7 and 10.

To gain insight into the individual-level changes in emotional state across the study, **Figure 33** shows overall scores with standard deviation bars for the AVD and circumplex model of emotion scales per individual per group before and after the interaction sections. The overall scores were determined by calculating the average of the responses given by the participants for the items on each of the two scales.

5.3.3. Rapport

The responses to negative adjectives of the rapport questionnaire (i.e., boring, unsatisfying, uncomfortably paced, cold, awkward, unfocused, intense, unfriendly, dull, and slow) were, first, reverse scored. Then, the average of the responses to the items in the questionnaire was taken as the perceived score of rapport for the dyads under consideration. This yielded two rapport scores for each dyad inside a group.

Because of the existence of the social-desirability bias [235], which is the tendency of survey/questionnaire participants to answer questions in a way that will be favorably viewed by others, the distribution of the dyadic reported rapport values was studied, and the lower 25-percentile taken as the threshold to determine low rapport values. Social-desirability bias can be expressed by over-reporting a "good" behavior or under-reporting a "bad" behavior. In this case,

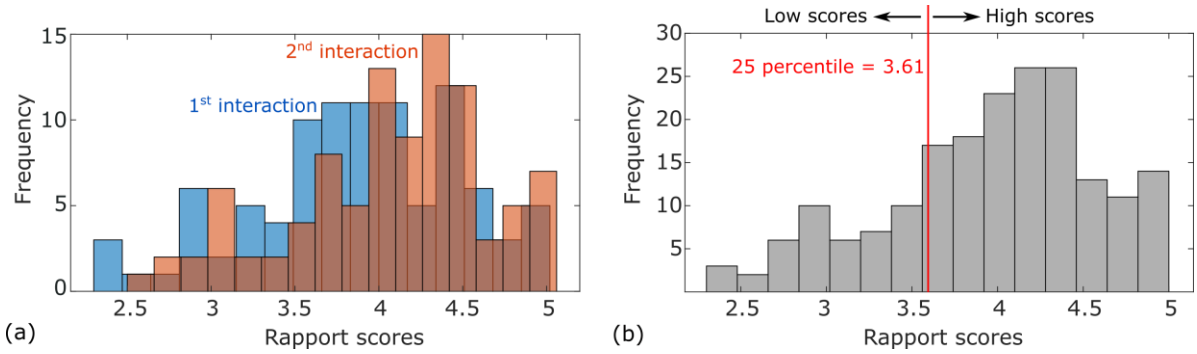


***Figure 34**. Distribution of calculated dyadic rapport scores. (a) Shows the distribution of the rapport scores corresponding to each of the two interaction periods; (b) shows the overall distribution of calculated dyadic rapport scores across the study (first and second interaction's rapport scores) and the value for the 25-percentile, which is used as a threshold to group rapport scores into low and high values.*

the overall average value reported for dyadic rapport in the study was 3.97±0.62 and a median value of 4.03 on a 5-point Likert scale. Therefore, most of the reported values were on the higher side of the scale and the reason to consider the lower 25-percentile as low rapport values. **Figure 34** shows the distribution of rapport scores for each of the interaction periods and all together as a whole, which was used to calculate the 25-percentile. The 25-percentile threshold value was determined to be 3.61. Based on this threshold, it was determined that low reported rapport scores amount to 32 in the first interaction period and 19 in the second interaction period from a total of 192 reported scores of dyadic interactions across the study, which includes two rapport values for each dyad in a group.

To characterize the level of rapport experienced by the groups, this work determined individual-experienced and dyadic-experienced rapport levels. The individual-experienced rapport levels were calculated per participant using the reported dyadic rapport values and were divided into active rapport values and passive rapport values. Active rapport values refer to the average reported experienced rapport by an individual towards the other people in the group, whereas passive rapport values refer to the average reported experienced rapport value from the people in

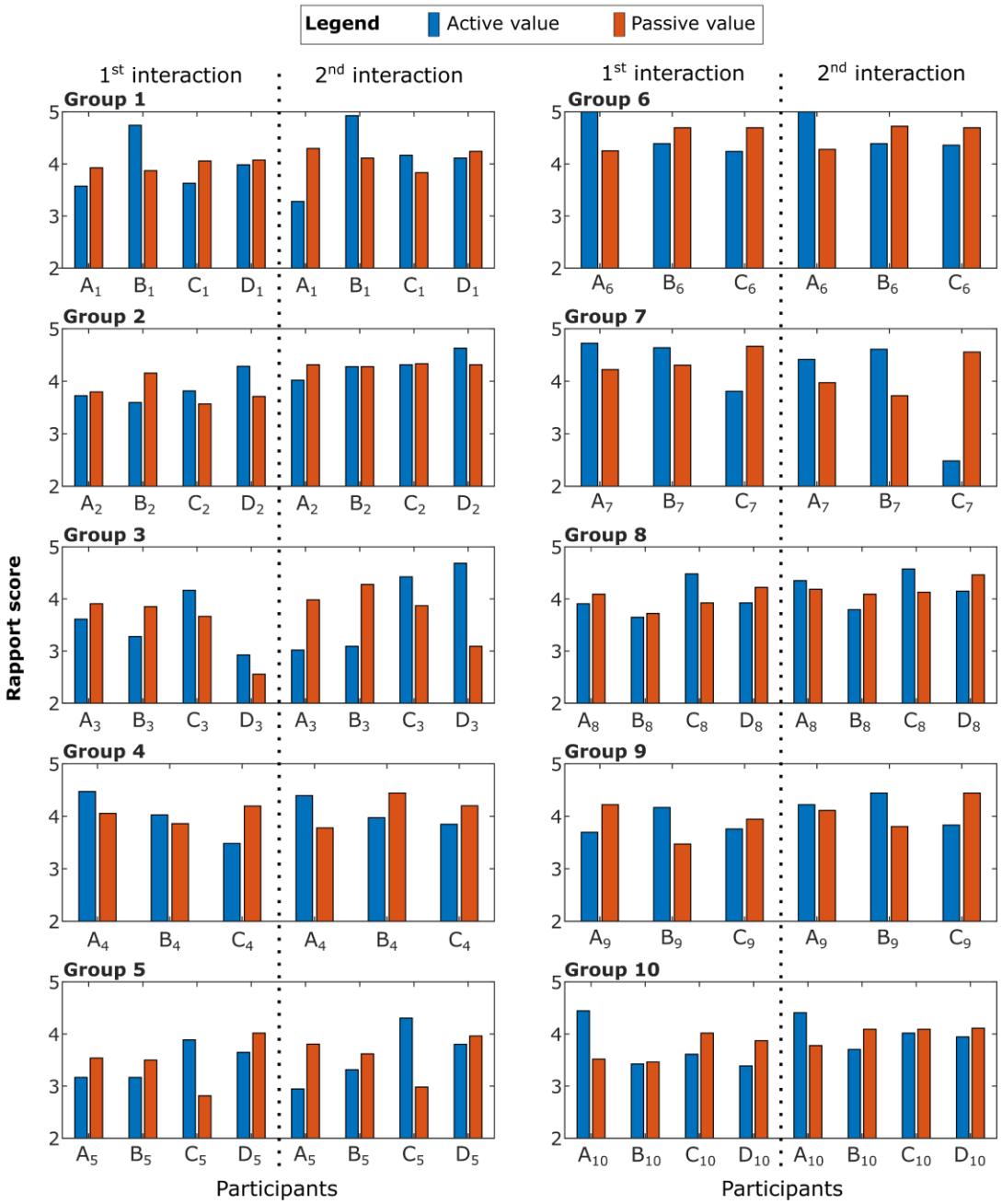


***Figure 35****. Summary of active and passive rapport scores corresponding to each individual in a group and for both of their interaction sections. Data used to construct these plots can be found in https://gitlab.msu.edu/davilasy/human-study-de-identified-data.*

the group towards the individual. Similarly, liking per individual was calculated as passive and active values. This yielded four values per individual per interaction section, two describing rapport levels and two describing liking values. **Figure 35** and **Figure 36** show the active and

passive rapport and liking scores corresponding to each individual, per group and interaction section, respectively. These active and passive values are used to compare how the rapport and liking connections felt by one person compared to what others felt towards that one person.

On the other hand, the dyadic-experienced rapport levels refer to the dyadic values directly obtained from the rapport questionnaire. These values help determine how close dyads felt the

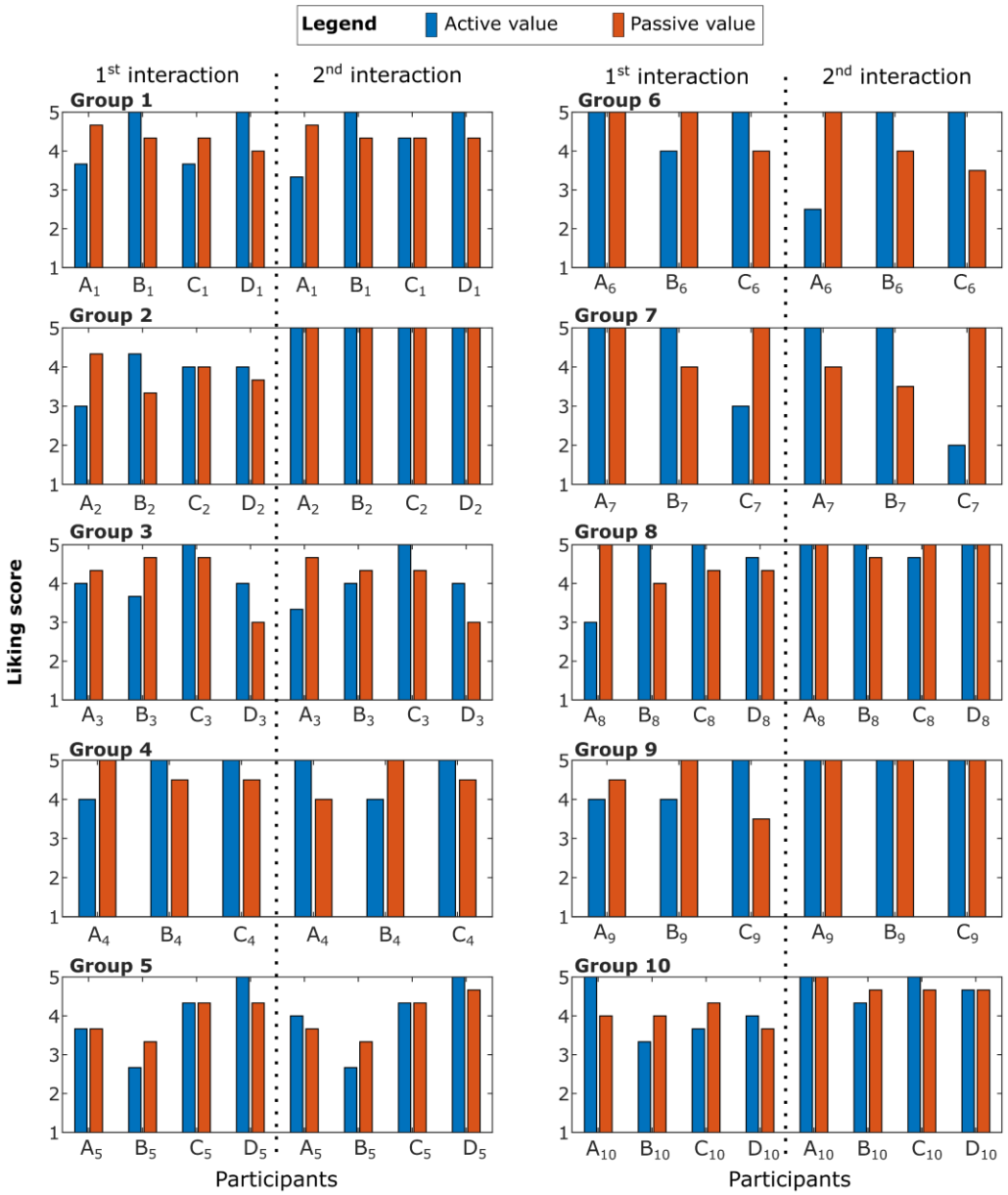


***Figure 36***. *Summary of active and passive liking scores corresponding to each individual in a group and for both of their interaction sections. Data used to construct these plots can be found in https://gitlab.msu.edu/davilasy/human-study-de-identified-data.*

***Table 28.*** *Dyadic strength of rapport based on reported self-assessments during the first interaction, which was intended to be one where there was disagreement among members of a group.*

| *Group* | *Interaction* | *Positive dyads* | *Negative dyads* | *Variant dyads* |
|---|---|---|---|---|
| **1** | **1** | $B_1C_1$ | $A_1C_1$, $A_1D_1$ | $A_1B_1$, $B_1D_1$, $C_1D_1$ |
| **2** | **1** | $A_2C_2$, $A_2D_2$, $B_2C_2$, $C_2D_2$ | $A_2B_2$ | $B_2D_2$ |
| **3** | **1** | $A_3B_3$, $A_3C_3$ | $A_3D_3$, $A_3C_3$, $C_3D_3$ | $B_3C_3$ |
| **4** | **1** | $A_4B_4$, $A_4C_4$ | - | $B_4C_4$ |
| **5** | **1** | $B_5D_5$ | $A_5B_5$, $A_5C_5$, $A_5D_5$ | $B_5C_5$, $C_5D_5$ |
| **6** | **1** | $A_6B_6$, $B_6C_6$ | - | $A_6C_6$ |
| **7** | **1** | $A_7B_7$, $B_7C_7$ | - | $A_7C_7$ |
| **8** | **1** | $A_8C_8$, $A_8D_8$, $B_8C_8$, $C_8D_8$ | $A_8B_8$, $B_8D_8$ | - |
| **9** | **1** | $A_9B_9$, $A_9C_9$, $B_9C_9$ | - | - |
| **10** | **1** | $A_{10}B_{10}$, $A_{10}D_{10}$, $C_{10}D_{10}$ | $B_{10}C_{10}$, $B_{10}D_{10}$ | $A_{10}C_{10}$ |
| | **Total** | 24 | 13 | 11 |

***Table 29.*** *Dyadic strength of rapport based on reported self-assessments during the second interaction, which was intended to be one where there was a high level of agreement among members of a group.*

| *Group* | *Interaction* | *Positive dyads* | *Negative dyads* | *Variant dyads* |
|---|---|---|---|---|
| **1** | **2** | $A_1D_1$, $B_1C_1$, $B_1D_1$, $C_1D_1$ | - | $A_1B_1$, $A_1C_1$ |
| **2** | **2** | $A_2B_2$, $A_2C_2$, $A_2D_2$, $B_2C_2$, $B_2D_2$, $C_2D_2$ | - | - |
| **3** | **2** | - | $A_3B_3$ | $A_3C_3$, $A_3D_3$, $B_3C_3$, $B_3D_3$, $C_3D_3$ |
| **4** | **2** | $A_4B_4$, $A_4C_4$, $B_4C_4$ | - | - |
| **5** | **2** | $B_5D_5$ | $A_5B_5$, $A_5D_5$, $B_5C_5$ | $A_5C_5$, $C_5D_5$ |
| **6** | **2** | $A_6B_6$, $B_6C_6$ | - | $A_6C_6$ |
| **7** | **2** | $A_7B_7$ | - | $A_7C_7$, $B_7C_7$ |
| **8** | **2** | $A_8B_8$, $A_8C_8$, $A_8D_8$, $B_8C_8$, $B_8D_8$, $C_8D_8$ | - | - |
| **9** | **2** | $A_9B_9$, $A_9C_9$, $B_9C_9$ | - | - |
| **10** | **2** | $A_{10}B_{10}$, $A_{10}C_{10}$, $A_{10}D_{10}$, $B_{10}C_{10}$, $B_{10}D_{10}$, $C_{10}D_{10}$ | - | - |
| | **Total** | 32 | 4 | 12 |

strength of rapport and help classify the dyadic interactions into positive dyads (both rated the interaction high), negative dyads (both dyads rated the interaction low), and variant dyads (one

dyad rated the interaction as high, another rated the interaction as low). To identify the aforementioned group of dyads, the average of the rapport values obtained from dyads (each participant rated their interaction with the other members of the group) was calculated. Likewise, the difference between the rapport values from dyads was also calculated. Then, the 25-percentile of the calculated average values was used as a threshold to identify positive and negative dyads and the 75-percentile of the calculated difference of rapport values was used to identify variant dyads. The 25-percentile of the dyadic average rapport values was 3.72 and the 75-percentile of the difference in rapport values was 0.833. **Table 28** and **Table 29** show a summary of the dyads that are classified as positive, negative, or variant dyads during the first and second group interaction, respectively. In total, 56 dyadic interactions were classified as positive dyads, 17 as negative dyads, and 23 as variant dyads. From the identified negative dyads, 13 manifested during the first group interaction and 4 during the second interaction. In general, it can be observed how the number of negative dyads is dominated by the first interaction, which proves that the second interaction was designed to increase rapport levels. From this analysis, it can also be observed that negative dyadic interactions were not developed in Group 4, Group 6, Group 7, and Group 9, although in three of those groups there is at least one variant dyadic interaction. The following repository contains raw data of individual dyadic scores, active and passive rapport scores, and the calculated dyadic strength: https://gitlab.msu.edu/davilasy/human-study-de-identified-data.

### 5.4.Data Labeling

Two major efforts were performed to annotate data of interest. The first annotation effort consisted of using external observers to annotate the perceived rapport level between dyads and the overall group interaction. The second annotation effort focused on annotating the head actions of individuals involved in the interactions by using the collected videos.

### 5.4.1. Labeling of Rapport Values using External Observers

Two external observers (one female and one male) were recruited for this task. External observers were instructed to watch the videos of the group interactions collected during the study and, using a similar version of the rapport questionnaire given to the study participants, score perceived rapport levels of the overall group interaction and between dyads. Therefore, each external observer watched a total of 20 videos, each lasting 21 minutes on average.

In the rapport questionnaire used by the external observers, the item intending to capture the overall group rapport level was prefaced with the instruction to "rate the overall interaction on the following characteristics." The items were smooth, bored, cooperative, satisfied, comfortable, awkward, engrossed, involved, friendly, active, and positive. The remaining items intending to capture perceived dyadic rapport level were prefaced with the instruction to "rate the interaction between subject X and subject Y on the following characteristics," where X and Y represented two of the individuals in the interaction. The items included well-coordinated, boring, cooperative, harmonious, unsatisfying, uncomfortably paced, cold, awkward, engrossing, unfocused, involving, intense, unfriendly, active, positive, dull, worthwhile, and slow. Responses were recorded on five-point Likert scales. Therefore, each external observer provided a total of one perceived overall value of rapport and three (for groups of 3) or six (for groups of 4) perceived dyadic rapport levels per video of the interaction section.

The scores from the two annotators were combined by calculating the mean between the rated items for each group and observed dyadic interaction. Then, similar to the method employed in Section 5.3.3, the overall value of rapport and perceived dyadic rapport levels were calculated by obtaining the mean of the values assigned to the items in the questionnaire. To find a threshold to group overall rapport values and perceived dyadic rapport levels into high and low, the 25-

percentile of each set was found. The 25-percentile of the overall rapport values was found to be 3.52, whereas the perceived dyadic rapport level was 3.36. **Table 30** and **Error! Reference source not found.** show a summary of the dyads that are classified as positive or negative dyads during the first and second interactions, respectively. In total, 72 dyadic interactions were classified as positive and 24 as negative based on the perceived rapport scores. In addition, 5 of the 20 group interactions were grouped as having low overall group rapport. These values are intended to serve as objective scores of rapport within the groups. When results from external annotators are compared to the self-reported rapport values (shown in **Table 28** and **Table 29**), both agree that during the first interaction 20 of the dyads are positive ones and 7 are negative ones, whereas during the second interaction both agree in that 28 of the dyads are positive ones and 4 are negative ones.

### 5.4.2. Labeling of Head Actions

As shown in Chapter 4, because head actions contribute to rapport establishment, the head actions of a subset of groups were labeled. Two annotators were employed for this task, and each was assigned a different set of groups for labeling. Annotators were instructed to watch the recorded videos of the interactions and use an annotation template made in an excel file to annotate the beginning and final time of a recognized movement or position. A recognized movement or position was annotated using a general label, a detailed label, and a direction label. Therefore, each identified head action was assigned three labels. Head actions were labeled for all the individuals involved in the video interactions.

General labels included no movement, one-time movement, repeating the motion, and other motions. Annotators were instructed to use *no movement* when the participant was in a steady position. *The one-time movement* was used when the participant went from left to right, up to

down, or vice versa, on a single movement. *The repeating motion* was assigned when a subject was head nodding, head shaking, or performing a cyclical motion. Finally, *other motion* was used for observed body position adjustments, chair motions, or any other inconsistent motion that could affect the position or motion of the head.

Detailed labels include steady, tilt shoulder, tilt yaw, bow, nod, shake, roll, body adjustment, inconsistent motion, and chair motion. Annotators were instructed to use *steady* when the participants were not showing movement. *Tilt shoulder* was used when the head was inclined to one of the shoulders (ear close to shoulder) and *tilt yaw* when there was a one-time head movement to the left or the right. *The bow* was used when there was a one-time head movement up or down. *Nod* was used for repeating movements in the pitch axis, *shake* for repeating movements in the

***Table 30.*** *Perceived rapport scores of the 1st interaction of each group obtained from two external annotators.*

| *Group* | *Interaction* | *Positive dyads* | *Negative dyads* | *Overall perceived group rapport-level* |
|---|---|---|---|---|
| **1** | **1** | $A_1C_1$, $A_1D_1$, $B_1C_1$, $B_1D_1$, $C_1D_1$ | $A_1B_1$ | **3.41** |
| **2** | **1** | $A_2C_2$, $A_2D_2$, $B_2C_2$, $B_2D_2$, $C_2D_2$ | $A_2B_2$ | 4 |
| **3** | **1** | $C_3D_3$ | $A_3B_3$, $A_3C_3$, $A_3D_3$, $B_3C_3$, $B_3D_3$ | **2.91** |
| **4** | **1** | $A_4B_4$, $A_4C_4$, $B_4C_4$ | - | 3.91 |
| **5** | **1** | $C_5D_5$ | $A_5B_5$, $A_5C_5$, $A_5D_5$, $B_5C_5$, $B_5D_5$ | **2.55** |
| **6** | **1** | $A_6B_6$, $A_6C_6$, $B_6C_6$ | - | 3.68 |
| **7** | **1** | $A_7B_7$, $A_7C_7$, $B_7C_7$ | - | 3.86 |
| **8** | **1** | $A_8C_8$, $A_8D_8$, $B_8C_8$, $B_8D_8$, $C_8D_8$ | $A_8B_8$ | 4.36 |
| **9** | **1** | $A_9B_9$, $A_9C_9$, $B_9C_9$ | - | 3.77 |
| **10** | **1** | $A_{10}C_{10}$, $A_{10}D_{10}$, $B_{10}C_{10}$, $B_{10}D_{10}$, $C_{10}D_{10}$ | $A_{10}B_{10}$ | 4.05 |
| | **Total** | 34 dyadic interactions | 14 dyadic interactions | 3 group interactions with perceived low rapport levels |

***Table 31**. Perceived rapport scores of the 2nd interaction of each group obtained from two external annotators.*

| *Group* | *Interaction* | *Positive dyads* | *Negative dyads* | *Overall perceived group rapport-level* |
|---|---|---|---|---|
| **1** | **2** | $A_1C_1$, $A_1D_1$, $B_1C_1$, $B_1D_1$, $C_1D_1$ | $A_1B_1$ | 4 |
| **2** | **2** | $A_2C_2$, $A_2D_2$, $B_2C_2$, $B_2D_2$, $C_2D_2$ | $A_2B_2$ | 4.36 |
| **3** | **2** | $A_3C_3$, $A_3D_3$, $B_3C_3$, $B_3D_3$, $C_3D_3$ | $A_3B_3$ | 3.68 |
| **4** | **2** | $A_4B_4$, $A_4C_4$, $B_4C_4$ | - | 4.05 |
| **5** | **2** | $C_5D_5$ | $A_5B_5$, $A_5C_5$, $A_5D_5$, $B_5C_5$, $B_5D_5$ | **2.32** |
| **6** | **2** | $A_6B_6$, $A_6C_6$, $B_6C_6$ | - | 3.64 |
| **7** | **2** | $A_7B_7$, $A_7C_7$, $B_7C_7$ | - | **2.82** |
| **8** | **2** | $A_8C_8$, $A_8D_8$, $B_8C_8$, $B_8D_8$, $C_8D_8$ | $A_8B_8$ | 4.41 |
| **9** | **2** | $A_9B_9$, $A_9C_9$, $B_9C_9$ | - | 4.09 |
| **10** | **2** | $A_{10}C_{10}$, $A_{10}D_{10}$, $B_{10}C_{10}$, $B_{10}D_{10}$, $C_{10}D_{10}$ | $A_{10}B_{10}$ | 4.68 |
| | **Total** | 38 dyadic interactions | 10 dyadic interactions | 2 group interactions with perceived low rapport levels |

yaw axis, and *roll* for repeating movements in the roll axis. On the other hand, *body adjustment* was assigned to any movement originating from an individual adjusting their body position, *chair motion* was assigned to any head/body movement originating from moving or rotating the chair in which the participants were sited, and the *inconsistent motion* was assigned to any set of motions that could not be clearly separated into a nod, shake, roll, etc. **Figure 37** shows an example of detailed labels aligned with raw sensor data from the second interaction of Group 5.

Direction labels include left, right, up, down, front, back, and changing. These were assigned in combination with the general and detailed labels to identify the direction of the motion, especially for the one-time movements. From the seven videos that were watched, there exist over 8700 identified head actions with respective labels.

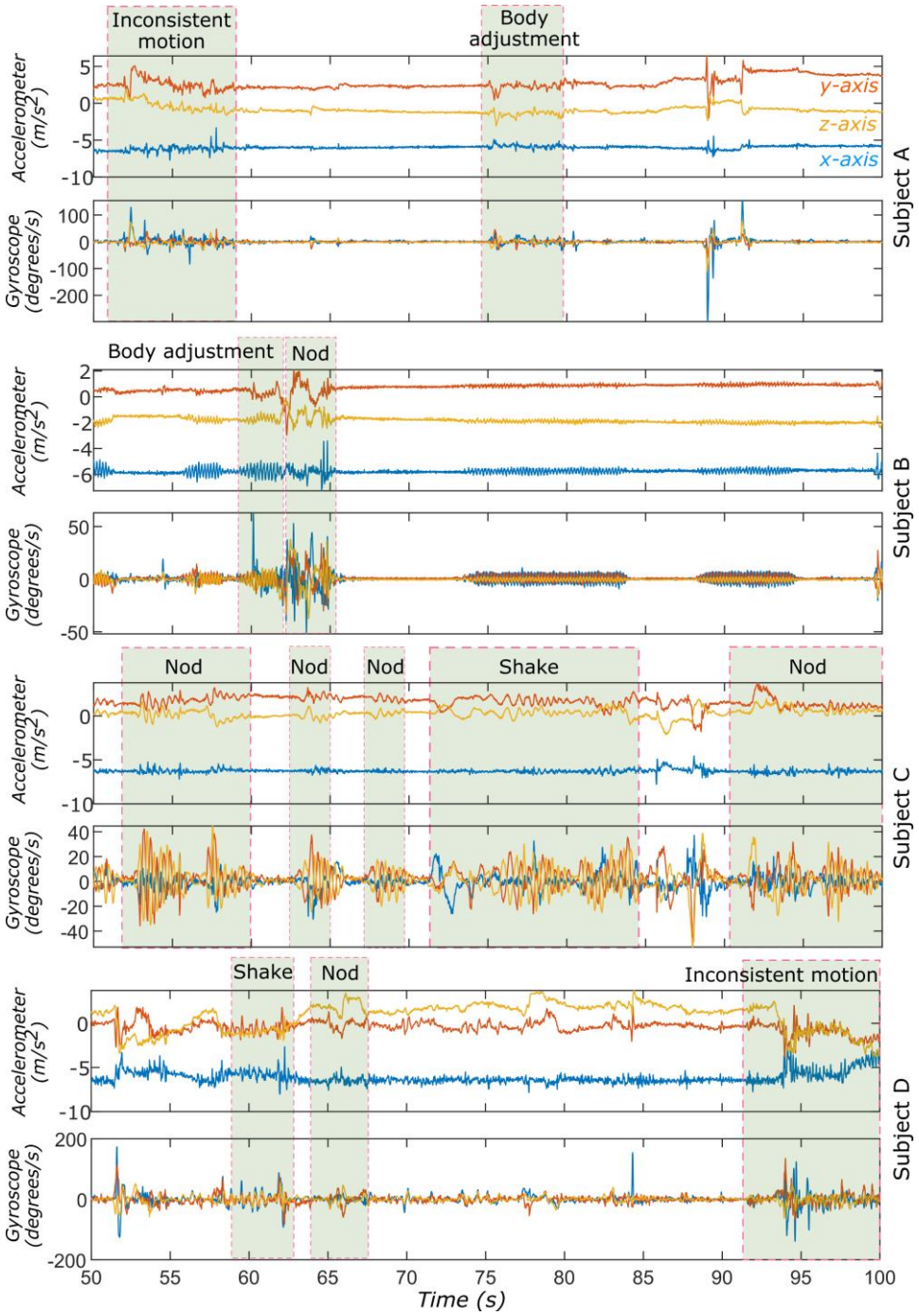


***Figure 37****. Example of raw IMU signals from the second interaction of participants in Group 5 and assigned 'detailed' labels.*

Note that assigned labels were not cross-validated due to a lack of human resources to contribute to this task. However, in the future, the cross-validation scheme presented in Chapter 4 for audio data can also be applied to this case.

## 5.5. Processing IMU Data using the Designed HAD Unit and Preliminary Establishment of Rapport Relationship

To advance the design of the machine learning framework for the group interaction monitoring system, the head-action detection (HAD) unit developed in Chapter 4 was evaluated using the natural data presented in this chapter. This model evaluation constitutes initial efforts in employing the collected dataset for the recognition of local transformed features. Data collected from Group 4 and Group 5 were selected for this analysis because they represent two different types of groups. While Group 4 appears to be a passive/collaborative group with high levels of shared rapport, Group 5 appears to be a confrontational one with variations in reported rapport levels.

### 5.5.1. Evaluation of HAD Unit

The HAD unit was evaluated using data from each individual in the selected groups; this includes data from 7 individuals interacting for ~20 minutes. The output of the HAD unit was compared to the labels assigned by an annotator, as explained in Section 5.4.2. The HAD unit was evaluated by its accuracy in recognizing (1) static or dynamic movement, (2) static with position versus motion, (3) static versus three motions, and (4) all six classes for which it was trained. Therefore, the performance of the HAD unit using this natural dataset was evaluated for the recognition of 2 classes, 4 classes, and 6 classes, for which it was ultimately trained.

Per design, data from the groups were processed in a real-time fashion using a data processing frame of 3 seconds with a 50% overlap. Only data from the accelerometer and gyroscope sensors were processed and only 7 features were extracted, in total, from each data frame.

### 5.5.2. Synchronicity of Dyadic Head Activity and Relationship to Rapport

Results from the best set of classes (2, 4, or 6 classes, as described in the previous section) were used to determine if there exists a mathematical relationship between the synchronicity of head activity of the dyads and the reported rapport values. The synchronicity of head activity between dyads is calculated by (1) measuring the dynamic time warping (DTW) between the signals, (2) using the obtained DTW results to correct for phase-shifts and signal length on each of the head activity time series, and (3) calculating the correlation coefficient between the phase-shifted signals. The DTW is an algorithm that measures the similarity between two time series and provides information about which data points from time series A match more closely with data points from time series B. The use of DTW has been employed in research related to the recognition of human activity [236].

The final correlation coefficient values are considered to represent a degree of coordination between dyads. These values are then matched with the dyadic rapport strength values obtained from self-reported data, as explained in Section 5.3.3.

### 5.5.3. Results and Discussion

#### *5.5.3.1.Validation of HAD unit*

Results of the validation of the HAD unit, in terms of classification model accuracy, precision accuracy, and recall accuracy, for each of the different classes are presented for both interactions of Group 4 (in **Table 32** and **Table 33**) and the second interaction of Group 5 (in **Table 34**). The first interaction of Group 5 was not evaluated because labels of head activity were incomplete.

In order to evaluate the accuracy of the HAD unit results, the time stamp of the annotations of head activity for Groups 4 and 5 needed to be aligned to the output of the HAD unit. The real-time classification of head activity provides an output every 1.5 seconds after the first 3 seconds

of processing. On average, the duration of a labeled action was of 4.1330 seconds. Therefore, the time stamp of the HAD unit was used to create a transformed set of annotations that aligns with the HAD unit output. In addition, because the labels assigned for head activity include other activities in addition to nod, shake, and roll motions, anything that was labeled otherwise was converted to a general motion label. The general motion labels were included in the assessment of the HAD unit's performance when evaluating the set of classes that also included a general motion

***Table 32***. *Accuracy, precision, and recall values obtained from evaluating the data from participants in Group 4, first interaction.*

| *Number of classes* | | *Host 1* | *Host 2* | *Host 3* | *Average* |
|---|---|---|---|---|---|
| 2 | Accuracy | 61.88% | 79.94% | 73.57% | 71.80% |
| | Precision | Steady: 37.7%<br>Motion: 71.9% | Steady: 23.8%<br>Motion: 90.8% | Steady: 19.1%<br>Motion: 78.7% | |
| | Recall | Steady: 35.6%<br>Motion: 73.7% | Steady: 33.3%<br>Motion: 86.0% | Steady: 8.2%<br>Motion: 91.1% | |
| 4 | Accuracy | 60.82% | 79.42% | 72.38% | 70.87% |
| | Precision | Tilt shoulder left: -<br>Steady neutral: 39.1%<br>Tilt shoulder right: -<br>Motion: 71.9% | Tilt shoulder left: 12.5%<br>Steady neutral: 22.9%<br>Tilt shoulder right: -<br>Motion: 90.8% | Tilt shoulder left: -<br>Steady neutral: 7.8%<br>Tilt shoulder right: -<br>Motion: 78.7% | |
| | Recall | Tilt shoulder left: -<br>Steady neutral: 33.3%<br>Tilt shoulder right: -<br>Motion: 73.7% | Tilt shoulder left: 33.3%<br>Steady neutral:28.6%<br>Tilt shoulder right: -<br>Motion: 86% | Tilt shoulder left: -<br>Steady neutral: 11.1%<br>Tilt shoulder right: -<br>Motion: 91.1% | |
| 4 | Accuracy | 37.34% | 43.15% | 46.57% | 42.35% |
| | Precision | Steady: 37.7%<br>nod: 43.3%<br>shake: 5.2%<br>roll: 52.2% | Steady: 23.8%<br>nod: 62.4%<br>shake: 24.4%<br>roll: 52.9% | Steady: 19.7%<br>nod: 56.2%<br>shake: 9.6%<br>roll: 20% | |
| | Recall | Steady: 44.6%<br>nod: 43.7%<br>shake: 30%<br>roll: 11.2% | Steady: 48.3%<br>nod: 60.9%<br>shake: 52.4%<br>roll: 7.4% | Steady: 11.1%<br>nod: 80.8%<br>shake: 12.8%<br>roll: 2.1% | |

***Table 32**. (cont'd).*

| | | | | | |
|---|---|---|---|---|---|
| 6 | Accuracy | 35.86% | 42.22% | 44.9% | 40.99% |
| | Precision | Tilt shoulder left: -<br>Steady neutral: 39.1%<br>Tilt shoulder right: -<br>nod: 43.3%<br>shake: 5.2%<br>roll: 52.2% | Tilt shoulder left: 12.5%<br>Steady neutral: 22.9%<br>Tilt shoulder right: -<br>nod: 62.4%<br>shake: 24.4%<br>roll: 52.9% | Tilt shoulder left: -<br>Steady neutral: 7.8%<br>Tilt shoulder right: -<br>nod: 56.2%<br>shake: 9.6%<br>roll: 20% | |
| | Recall | Tilt shoulder left: -<br>Steady neutral: 41.4%<br>Tilt shoulder right: -<br>nod: 43.7%<br>shake: 30%<br>roll: 11.2% | Tilt shoulder left: 50%<br>Steady neutral: 41.4%<br>Tilt shoulder right: -<br>nod: 60.9%<br>shake: 52.4%<br>roll: 7.4% | Tilt shoulder left: -<br>Steady neutral: 12.9%<br>Tilt shoulder right: -<br>nod: 80.8%<br>shake: 12.8%<br>roll: 2.1% | |

class, as was the case of the first two sets of classes evaluated: (1) static or dynamic movement, (2) static with position versus motion. Otherwise, for the other two sets of classes evaluated (i.e., (3) static versus three motions and (4) all six classes) that contained specific motion types, any instance with other motions outside of nod, shake, and roll was not used for model evaluation.

Results as shown in **Table 32** - **Table 34** revealed that, on average for, the detection of steady versus motion the HAD unit achieves a validation accuracy of 71.80%, 70.83%, and 56.64% for Group 4-1$^{st}$ interaction, Group 4-2$^{nd}$ interaction, and Group 5-2$^{nd}$ interaction, respectively. Group 5-2$^{nd}$ interaction average accuracy falls to 56.64% because the model appears to not be able to recognize head motions from one of the participants' data. For the static with position versus motion classes, the average accuracy for Group 4-1$^{st}$ interaction, Group 4-2$^{nd}$ interaction, and Group 5-2$^{nd}$ interaction was 70.87%, 68.28%, and 48.71%, respectively. The last set of 4 classes tested obtained average classification accuracies of 42.35%, 41.72%, and 54.45% for Group 4-1$^{st}$

interaction, Group 4-2nd interaction, and Group 5-2nd interaction, respectively. Lastly, the set of 6 classes obtained an average accuracy of 40.99%, 37.57%, and 38.35% for Group 4-1st interaction, Group 4-2nd interaction, and Group 5-2nd interaction, respectively.

It is expected that as more classes are added to the evaluation, the validation accuracy will drop, especially since the HAD unit was trained with data with clearly identifiable motions and the data used in this case is a naturally collected one. If recalled, the testing accuracy of the HAD

***Table 33**. Accuracy, precision, and recall values obtained from evaluating the data from participants in Group 4, second interaction.*

| *Number of classes* | | *Host 1* | *Host 2* | *Host 3* | *Average* |
|---|---|---|---|---|---|
| 2 | Accuracy | 65.39% | 78.93% | 68.17% | 70.83% |
| | Precision | Steady: 62.3%<br>Motion: 67% | Steady: 19.1%<br>Motion: 90.1% | Steady: 12.5%<br>Motion: 70.7% | |
| | Recall | Steady: 49.8%<br>Motion: 77.2% | Steady: 26.5%<br>Motion: 85.6% | Steady: 1.9%<br>Motion: 94.6% | |
| 4 | Accuracy | 59.1% | 77.84% | 67.90% | 68.28% |
| | Precision | Tilt shoulder left: -<br>Steady neutral: 48.5%<br>Tilt shoulder right: -<br>Motion: 67% | Tilt shoulder left:<br>Steady neutral: 14.7%<br>Tilt shoulder right: -<br>Motion:90.1% | Tilt shoulder left: -<br>Steady neutral: 4.5%<br>Tilt shoulder right: 33.3%<br>Motion: 70.7% | |
| | Recall | Tilt shoulder left: -<br>Steady neutral: 43%<br>Tilt shoulder right: -<br>Motion: 77.2% | Tilt shoulder left: -<br>Steady neutral:23.3%<br>Tilt shoulder right: -<br>Motion: 85.6% | Tilt shoulder left: -<br>Steady neutral: 7.7%<br>Tilt shoulder right: 0.5%<br>Motion: 94.6% | |
| 4 | Accuracy | 47.54% | 47.36% | 30.25% | 41.72% |
| | Precision | Steady: 62.3%<br>nod: 31.2%<br>shake: -<br>roll: 46.7% | Steady: 19.1%<br>nod: 67.6%<br>shake: 19.7%<br>roll: 47.8% | Steady: 12.5%<br>nod: 35.8%<br>shake: 1.8%<br>roll: 33.3% | |
| | Recall | Steady: 66.2%<br>nod: 40.8%<br>shake: -<br>roll:8.2% | Steady:37.3%<br>nod: 67.6%<br>shake: 35.1%<br>roll: 10.2% | Steady: 2.9%<br>nod: 79.8%<br>shake: 10%<br>roll: 3.2% | |

***Table 33***. *(cont'd).*

| | | | | | |
|---|---|---|---|---|---|
| 6 | Accuracy | 37.28% | 45.59% | 29.83% | 37.57% |
| | Precision | Tilt shoulder left: -<br>Steady neutral: 48.5%%<br>Tilt shoulder right: -<br>nod: 31.2%<br>shake: -<br>roll: 46.7% | Tilt shoulder left: -<br>Steady neutral: 14.7%<br>Tilt shoulder right: -<br>nod: 67.6%<br>shake: 19.7%<br>roll: 47.8% | Tilt shoulder left: -<br>Steady neutral: 4.5%<br>Tilt shoulder right: 33.3%<br>nod: 35.8%<br>shake: 1.8%<br>roll: 33.3% | |
| | Recall | Tilt shoulder left: -<br>Steady neutral: 59.4%<br>Tilt shoulder right: -<br>nod: 31.2%<br>shake: -<br>roll: 46.7% | Tilt shoulder left: -<br>Steady neutral: 30.4%<br>Tilt shoulder right: -<br>nod: 67.6%<br>shake: 35.1%<br>roll: 10.2% | Tilt shoulder left: -<br>Steady neutral: 10%<br>Tilt shoulder right: 0.8%<br>nod: 79.8%<br>shake: 10%<br>roll: 3.2% | |

unit during the design process was 97.91%. Natural collected head activity data may contain micro-motions embedded in specific motions of interest that will act as artifacts or noise in the signals. The designed HAD unit does not account for such artifacts, thus, the dropped in accuracy results when compared to the testing accuracies during the design process. Another possible explanation for the drops in accuracies is related to the feature set used for classification. During the design process, the model was optimized to decrease computational complexity. Thus, it is possible that the selected feature set cannot generalize enough to accommodate the characteristics of this natural dataset. Nevertheless, the results of the classification of 2 classes (steady versus motion) were used to investigate a preliminary relationship between head activity coordination and rapport strength between dyads.

*5.5.3.2.Synchronicity of motion and rapport*

**Table 35** presents the results of the synchronization measurement (explained in Section 5.5.2) between dyadic head activity as detected by the HAD unit for just two classes (steady versus motion). For the values presented in **Table 35**, a correlation analysis was applied between the reported synchronicity values and the corresponding rapport scores, resulting in a correlation coefficient of -0.2629. This provides an inconclusive relationship between synchronicity of

***Table 34**. Accuracy, precision, and recall values obtained from evaluating the data from participants in Group 5, second interaction.*

| *Number of classes* | | *Host 1* | *Host 2* | *Host 3* | *Host 4* | *Average* |
|---|---|---|---|---|---|---|
| 2 | Accuracy | 63% | 70.88% | 30.44% | 62.26% | 56.64% |
| | Precision | Steady: 76.4% Motion: 56.1% | Steady: 55.7% Motion: 80.3% | Steady: 81.9% Motion: 26.2% | Steady: 87.2% Motion: 37.2% | |
| | Recall | Steady: 47.3% Motion: 82.2% | Steady: 63.8% Motion: 74.4% | Steady: 8.4% Motion: 94.6% | Steady: 58.3% Motion: 74.2% | |
| 4 | Accuracy | 62.16% | 65.78% | 25.90% | 41.01% | 48.71% |
| | Precision | Tilt shoulder left: - Steady neutral: 76% Tilt shoulder right: - Motion: 56.1% | Tilt shoulder left:18.2% Steady neutral: 49.5% Tilt shoulder right: - Motion: 80.3% | Tilt shoulder left: 16.7% Steady neutral: 17.0% Tilt shoulder right: 46.2% Motion: 26.2% | Tilt shoulder left: 40% Steady neutral: 47.5% Tilt shoulder right: 9.7% Motion: 37.2% | |
| | Recall | Tilt shoulder left: - Steady neutral: 45.9% Tilt shoulder right: - Motion: 82.2% | Tilt shoulder left: 19% Steady neutral: 53.7% Tilt shoulder right: - Motion: 74.4% | Tilt shoulder left: 8.3% Steady neutral: 12% Tilt shoulder right: 1% Motion: 94.6% | Tilt shoulder left: 3.3% Steady neutral: 53.6%% Tilt shoulder right: 1.5% Motion: 74.2% | |

***Table 34***. *(cont'd).*

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | Accuracy | 62.63% | 50.55% | 29.55% | 75.09% | 54.45% |
| | Precision | Steady: 76.4%<br>nod: 52%<br>shake: -<br>roll: - | Steady: 55.7%<br>nod: 29.6%<br>shake: 34.8%<br>roll: - | Steady: 81.9%<br>nod: 5.7%<br>shake: 9.4%<br>roll: 50% | Steady: 87.2%<br>nod: 19.7%<br>shake: -<br>roll: - | |
| | Recall | Steady: 62.6%<br>nod: 68.7%<br>shake: -<br>roll: - | Steady: 92.6%<br>nod: 25.6%<br>shake: 24.2%<br>roll: - | Steady: 31.7%<br>nod: 26.3%<br>shake: 44.4%<br>roll: 4.2% | Steady: 87.5%<br>nod: 21.2%<br>shake: -<br>roll: - | |
| 6 | Accuracy | 61.43% | 40% | 12.15% | 39.82% | 38.35% |
| | Precision | Tilt shoulder left: -<br>Steady neutral: 76.0%<br>Tilt shoulder right: -<br>nod: 52%<br>shake: -<br>roll: - | Tilt shoulder left: 18.2%<br>Steady neutral: 49.5%<br>Tilt shoulder right: -<br>nod: 29.6%%<br>shake: 34.8%<br>roll: - | Tilt shoulder left: 16.7%<br>Steady neutral: 17.0%<br>Tilt shoulder right: 46.2%<br>nod: 5.7%<br>shake: 9.4%<br>roll: 50% | Tilt shoulder left: 4%<br>Steady neutral: 47.5%<br>Tilt shoulder right: 9.7%<br>nod: 19.7%<br>shake: -<br>roll: - | |
| | Recall | Tilt shoulder left: -<br>Steady neutral: 60.6%<br>Tilt shoulder right: -<br>nod: 68.7%<br>shake: -<br>roll: - | Tilt shoulder left: 25%<br>Steady neutral: 79.2%<br>Tilt shoulder right: -<br>nod: 25.6%<br>shake: 24.2%<br>roll: - | Tilt shoulder left: 25%<br>Steady neutral: 40.9%<br>Tilt shoulder right: 3.8%<br>nod: 26.3%<br>shake: 44.4%<br>roll: 4.2% | Tilt shoulder left: 3.9%<br>Steady neutral: 81.4%<br>Tilt shoulder right: 2.5%<br>nod: 21.2%<br>shake: -<br>roll: - | |

motions and reported rapport strength. Further analysis into the synchronicity of specific types of motion, such as head nodding, and its relationship to speech activity is recommended. Establishing a mathematical relationship between the measurable behavioral cues and the reported rapport values will allow for future real-time estimation of rapport values.

***Table 35**. Summary of obtained correlation coefficient of head activity (steady versus motion) detected from dyads and the corresponding dyadic strength of rapport based on reported self-assessment.*

| *Group/Interaction* | *Dyads* | *Synchronicity* | *Rapport* |
|---|---|---|---|
| 4/1st | BC | 0.7592 | 3.5833 |
| | AC | 0.9141 | 4.2222 |
| | AB | 0.8958 | 4.3055 |
| 4/2nd | BC | 0.8407 | 4.25 |
| | AC | 0.9537 | 4.0065 |
| | AB | 0.8865 | 4.1666 |
| 5/2nd | AC | 0.9743 | 3.7075 |
| | CD | 0.9828 | 3.8333 |
| | CB | 0.9768 | 3.3888 |
| | AD | 0.9799 | 3.6111 |
| | AB | 0.9802 | 2.8055 |
| | BD | 0.9921 | 4.2042 |

### 5.6.Overall Discussion

The design of the social interaction study protocol for the collection of natural data was guided by two factors: (1) the need to evoke changes in rapport levels among members of a group and (2) the need to collect self-reported data that could be used to validate changes in rapport levels among members of a group and/or be used as data labels for machine learning algorithms. The analysis of self-reported data reflects changes in rapport levels among individuals of a group between the first and the second interaction periods, as well as changes in individuals' emotional states. This serves as evidence of the effectiveness of the protocol to evoke changes in rapport levels and, more specifically, in demonstrating a trend where the first interaction carries lower rapport levels when compared to those reported during the second interaction. However, replicating this study with a higher number of groups is highly recommended to ensure the significance of the statistical analysis results. In addition, the design of a group interaction monitoring system will be benefited from adding data from a higher number of groups to increase the existing number of sensor data streams usable for the training of machine learning models. Increasing the number of groups

studied using this protocol could open the opportunity to study how demographics influence the established levels of rapport and the overall likeness of the group interaction.

On the other hand, increasing the number of groups used in this study will also increase the time needed to prepare the collected data for processing. As this work uses a traditional supervised machine learning pipeline, data labeling becomes an essential task. This chapter established data labeling protocols to accompany the design of algorithms for group behavior monitoring systems. Labeling was focused on obtaining rapport values from external observers and labels of head actions. Nevertheless, the data annotation protocol developed in Chapter 4 for the labeling of speech intonation can be applied to the dataset collected during the study. In addition, the increase in collected data opens opportunities to employ unsupervised machine learning methods such as neural networks and reduce the time that is required for annotating datasets.

Chapter 4 presented the design of the HAD unit, trained using acted/evoke data. In this chapter, the trained HAD unit was evaluated using the collected natural data from the social interaction study and employed to study the correlation between head activity and rapport scores. It was noted that (1) the performance of the HAD unit is lower than the one obtained during training and testing using the acted/evoked dataset and (2) the performance of the HAD unit varies across individuals in the interactions. It is recommended to employ a data normalization method across the collected dataset and re-train the HAD unit to increase the level of generality of such a model. In addition, as mentioned in Chapter 4, the level of computational optimization of the HAD unit could also limit its ability to generalize when presented with unseen and noisy data. Furthermore, the HAD unit was trained with a limited number of acted/evoked head actions. Therefore, this may serve as supportive evidence for the recommendation of using natural data to train classification models.

## 5.7.Summary

This chapter presents the design and execution of a social interaction study where sensor data was collected using the sensor framework presented in Chapter 3. Video, audio, movement, and physiological data were collected, together with self-reported scores of emotional states and rapport strength between dyads. A general analysis of changes in emotional states revealed that participants experienced significant changes in five emotions evaluated using the Circumplex emotional state scale. Self-reported rapport scores were analyzed, and it was found that five out of the ten groups contain low and variant dyadic interactions. Overall, out of the 96 dyadic interactions captured by this study, 56 were positive, 17 were negative, and 23 were variant. Moreover, most of the reported negative dyadic interactions happened during the first interaction, which was intended to evoke low rapport value. In addition, labels for perceived rapport values were assigned by external individuals, as well as labels for head actions. The trained HAD unit was evaluated with this natural dataset and detected patterns of motion were used to calculate synchronicity between dyads. Calculated synchronicity values were correlated with reported rapport scores between dyads; however, the results were inconclusive. This dataset serves to continue the design of a machine learning framework for the recognition of behavioral cues and to investigate relationships between recognized behavioral cues and components of rapport.

# 6. SUMMARY AND FUTURE WORK

## 6.1.Summary

This dissertation presents the design of a new human/group behavior monitoring platform to address existing challenges in the monitoring of group interactions for the improvement of social awareness and human health. The presented human/group behavior monitoring platform combines a multi-sensor system with a machine learning framework, covering all from sensor selection to algorithm design. First, rapport is established as a social construct of interest to understand the quality of social interactions. Fundamentals of human behavior and initial efforts on designing wearable real-time social monitoring systems are introduced. A comprehensive literature study was later conducted to define the state-of-the-art in sensors and algorithms and find existing design challenges. The transdisciplinary approach taken to study the social science theory behind group behaviors and the technology to monitor nonverbal behaviors informed the design of a multi-sensor system.

A new multi-sensor system for the study of group interactions was designed and implemented. The multi-sensor system combines six sensor modalities: microphone, accelerometer, gyroscope, magnetometer, photoplethysmography (PPG), and electroencephalography (EEG); it also synchronizes the sensor data through the use of Lab Streaming Layer (LSL) and allows for recording from multiple sensor nodes. Each sensor node receives 16 data streams: one from audio, four from EEG, nine in total from accelerometer, gyroscope, and magnetometer, and two corresponding to PPG (one filtered data stream and one pre-processed signal estimating heart rate). In addition, a machine learning framework for the training and design of real-time recognition of human and group behavior was also described. Of particular interest was the design of data

processing units to determine Type B features, which are high-level transformed features determined from features extracted from raw sensor data (Type A features).

Using the developed multi-sensor behavior monitoring system and support components, two human studies were conducted to establish the processes for which machines can be trained to recognize nonverbal behavior indicators and support the development of data processing units for the extraction of Type B features. The first set of human studies was conducted with 8 participants and consisted of recording (1) audio from virtual group meetings and (2) pre-defined head actions using inertial movement units (IMUs). A process to annotate speech intonations was established through the evaluation of labels assigned to the data collected from virtual group meetings. Using an inter-agreement annotator analysis, for the first time in the literature, two different modes of speech intonation annotation were evaluated. This analysis led to the construction of a dataset containing neutral, positive, negative, and question-labeled audio segments, which was used for the design of a real-time user-independent speech intonation recognizer unit. To the best of our knowledge, the designed speech intonation recognizer represents the first real-time model trained using just nonverbal information, an English-speaker dataset collected from group meeting interactions, not acted, containing speech from culturally diverse individuals, a combination of phrases with back-channel signals, and a combination of affective classes with an interrogative intonation. On the other hand, IMU data collected from pre-defined head actions were used to design a user-independent real-time head-action detection (HAD) unit based on a new fusion model architecture approach. Both units were designed taking a resource-aware approach for real-time processing where window sizes for data processing, types and number of features, and complexity of the classification models were taken into consideration.

The second human study consisted of collecting audio, visual, movement, and physiological sensor data while groups of individuals were interacting with each other in environments where low and high levels of rapport could be evoked. A total of 10 groups composed of 3 to 4 individuals participated in this study. This is the first study that collects data from IMU and physiological data from a head-mounted device in combination with audio from a personal computer and establishes the processes for which self-reported emotional state labels, self-report and externally assigned rapport labels, and head action labels are obtained. The HAD unit was used to explore the relationships between head actions and perceived rapport levels between dyads of a group. The contributions of this dissertation will advance the design of human behavior monitoring systems for group interactions and facilitate real-time feedback to increase self-awareness and promote successful social interactions. This work provides an infrastructure for the design of group behavior monitoring systems through which in-person and virtual group interactions could be studied and monitored.

### 6.2.Contributions

This dissertation bridges the gap between social science, communication science, and the engineering field by establishing a novel sensing and data collection platform for real-time monitoring of group interactions. **Figure 38** presents a summary of the areas in which this dissertation has made contributions to advance the design of human/group behavior monitoring systems. Innovations of this work include contributions in:

- ***Analyzed*** *the* ***complete*** *body of literature in the field of human behavior wearable monitoring technologies, which* ***provided new*** *visions and insights to* ***converge*** *research in the disciplines of social psychology, communication, and engineering*

To provide a clear understanding of state-of-the-art technologies for human behavior monitoring and promote convergence research into new technologies that can overcome current challenges, this dissertation provides an extensive review of the literature associated with monitoring human behavior. This dissertation uniquely presents a new comprehensive, transdisciplinary, perspective with a focus on identifying critical design considerations in real-time human behavior monitoring systems. Starting with an overview of social psychology theories that have established the framework to study human behaviors and their manifestations during social interactions, this dissertation then establishes a taxonomy of human behavior monitoring technologies based on these psychological theories. It also provides an insightful categorization of sensors and an informative analysis of signal characteristics, features, and computational models that have been reported in the field of human behavior monitoring. Analysis of recognition accuracies for existing computational models in the area of human behavior monitoring is also presented. Moreover, this dissertation focused on sensor hardware and real-time signal processing technologies that have proven most effective for embedded monitoring of human behaviors while

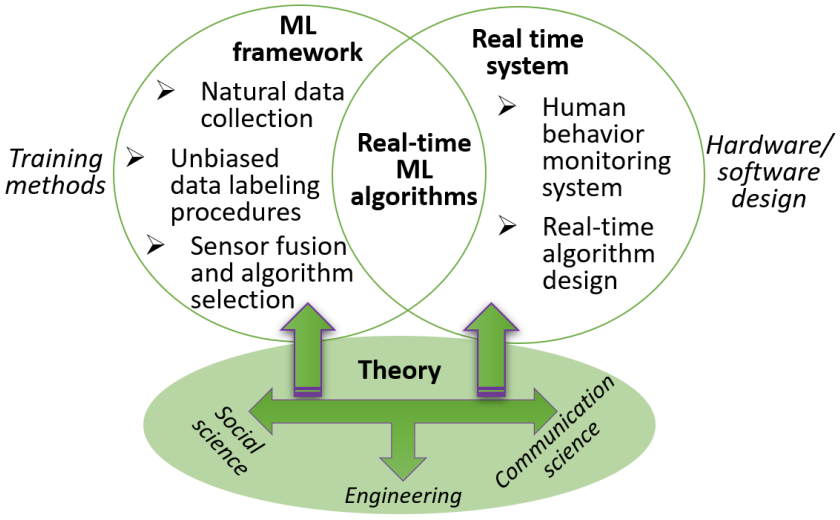


***Figure 38***. *Diagram summarizing the elements involved in the real-time human/group interaction monitoring platform developed in this work. Highlighted in* ***bold*** *are the four areas where this work made its research contributions.*

highlighting challenges and opportunities in near-future wearable applications. The performed analysis inspired the design of the real-time human/group interaction monitoring platform. This extensive review resulted in a publication with the citation: S. Dávila-Montero, J. A. Dana-Lê, G. Bente, A. T. Hall, and A. J. Mason, "Review and Challenges of Technologies for Real-Time Human Behavior Monitoring," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 1, pp. 2-28, Feb. 2021.

- ***Developed** the **first real-time enabled and accessible** multi-sensor system for group behavior analysis and **designed new real-time** algorithms to recognize behavioral cues associated with group consonance using sensor data*

Existing human behavior monitoring systems lack real-time capabilities and configurations to monitor complex human behaviors that could lead to identifying complex group dynamics. In addition, no existing systems are accessible for other researchers to reproduce with easy. Towards the goal of overcoming existing challenges for the real-time monitoring of complex social interactions, this dissertation introduces the first real-time enabled and accessible multi-sensor framework that allows the study and real-time analysis of both in-person and virtual interactive environments. The framework leverages the use of existing open-source commercially available wearable sensors, which were selected based on an analysis of the relation of sensing modalities to behavioral information and evaluation of commercially available wearable sensors. Sensing modalities include a microphone, accelerometer, gyroscope, magnetometer, PPG, and EEG. Moreover, this dissertation presents design details on the implementation of sensor integration and the use of networking protocols to manage 16 sensor data streams collected from each sensor node of this framework. The framework can manage at least 4 sensor nodes allowing the study of group interactions of 3 and 4 individuals. This multi-sensor framework system allows for easy

reproducibility because of the benefits of off-the-shelf sensors and networking resources. This resulted in a publication with the citation: S. Dávila-Montero, S. Parsnejad, E. Ashoori, D. Goderis, and A. J. Mason, "Design of a Multi-Sensor Framework for the Real-time Monitoring of Social Interactions," *IEEE International Symposium in Circuits and Systems (ISCAS)*, 2022.

Furthermore, this work introduces the first machine learning framework to monitor individual components of rapport and developed real-time computational blocks to identify two types of behavioral cues: head nods and speech intonations. Both achieve detection and recognition results that are comparable to existing ones in the literature. However, this work presents an analysis of variations in sampling rates, optimal window length for real-time processing, feature selection, and classification models to reduce computational complexity during real-time processing. Therefore, the design of the aforementioned algorithms was performed using a resource-aware approach considering the constraints of designing a human behavior monitoring system.

- ***Established*** *and, for the* ***first time, evaluated*** *data labeling methods for the establishment of a machine learning training framework for behavior monitoring, which included the* ***development*** *of a* ***new*** *human study protocol for the collection of group behavioral information of interest that evoke variations in experienced rapport levels*

Existing recognition systems of behavioral cues present limitations in real-time processing and use of natural data or data from the wild. Moreover, well-established protocols and machine learning training frameworks for the collection and preparation of natural data for the design of human behavior monitoring systems do not exist. This work developed the first machine learning training framework for the collection and labeling of natural human behavior data. This framework

was used for the collection of audio data and training of a speech intonation recognizer, where methods for effective labeling of speech intonations were evaluated.

In addition, well-described data protocols did not exist for the design of human studies focused on evoking low rapport during group naturalistic interactions. This work shows the methods for participant recruitment and study execution using the developed multi-sensor framework. This work established a new dataset containing audio, video, movement, and physiological data, self-reported emotional states and rapport scores, and externally assigned rapport scores and head action labels. Analysis of self-reported rapport scores was developed and showed that five out of 10 groups developed a negative and variant dyadic rapport.

### 6.3.Other Achievements

#### 6.3.1. Engineering and data science

- ***Developed*** *and* ***implemented*** *a* ***new user-friendly*** *labeling framework and applied it to label speech intonation in audio data*

To facilitate the annotator's access to data to be labeled and to maintain consistency in the way data was presented to the annotators, a graphical user interface (GUI) was developed. This GUI was utilized to label intonations in pre-identified audio segments. However, the GUI framework could be modified to label other types of 1-D signals, images, and video segments.

- ***Developed*** *and* ***implemented*** *stand-alone applications for the connection and management of sensor signal collection and processing*

The commercially available wearable sensors provided Application Programming Interfaces (APIs) that were used to establish sensor connections with MATLAB and LSL through designed stand-alone applications. The implemented stand-alone applications allow the real-time monitoring of collected and processed sensor signals.

### 6.3.2. Mentoring

This dissertation opened opportunities for undergraduate students interested in gaining experience in the areas of sensor integration, machine learning, and programming. **10 undergraduate students** were **mentored** and assisted to work on the following topics:

- Social interaction monitoring using audio signals
- Processing of EEG signals
- Wearable interpersonal monitor for enhanced teamwork
- Design of a multi-sensor head-mounted wearable device for the monitoring of human behaviors
- Design of a visual feedback interface to increase social behavior awareness
- Data labeling
- Cross-check and annotation agreement analysis

## 6.4.Applications and Social Implications

The contributions and the platform established in this dissertation could impact a variety of research areas and applications at the interception of the social sciences, communication, and engineering. The creation of the accessible multi-sensor platform allows for an increase in research collaboration, research reproducibility, and advancement in the areas of human-computer interaction, affective computing, and social signal processing. Such a platform, validated and combined with ethnographic methods, has the potential to serve as a tool for the study of group interactions in diverse scenarios. By extracting a myriad of informative individual, dyadic, and group behavioral cues, new methodological standards for psychology and communication research could be established. For example, the factors influencing team performance and subtle negative behaviors affecting social interactions could be further studied with the platform introduced in this

dissertation and results used to better understand how technology could help individuals increase their situational awareness. In addition, this platform could facilitate research into the establishment of a feedback mechanism for sharing information with individuals of a group about factors influencing their interaction. This will promote wellness and economic growth in the future work frontier where collaboration effectiveness of diverse knowledge-based teams will be critical to continued innovation and national financial security. Other areas of applications include training employees and other individuals to increase social skills, identify disruptive behaviors, and techniques to deal with conscious and unconscious biased behaviors. Furthermore, areas in the healthcare industry could also be benefited from technologies for the monitoring of human and group behaviors since a variety of health conditions influence the way individuals behave. Therefore, recognizing extreme changes in behaviors could help in the diagnosis of health conditions and identification of neurological and developmental disabilities or disorders (e.g., depression and autism, respectively).

In the case of monitoring behaviors to increase situational awareness, it is worthwhile to mention that the goal of this technology is not to control individuals' behaviors, but rather to inform/provide information about behavioral cues that without technology would be otherwise lost. Because the technology is designed to provide information about behavioral cues, individuals will be given the opportunity to change their behavior as they find it appropriate. In addition, during the design of the human/group interaction monitoring platform, individuals' privacy was considered essential. Because of that, just nonverbal messages were considered for the monitoring of human behaviors, and the use of speech recognition systems was avoided when designing the real-time speech intonation recognizer.

## 6.5.Future Work

This research has established the foundations for wearable real-time group behavior monitoring. However, a variety of labeling tasks, labeling analysis, algorithms development, and system testing have been left for future work and the start of new research projects.

To achieve the goal of deploying a human behavior monitoring system, the following suggestions are to continue this work:

- *Preparation of sensor data from the social interaction study*

Related to the sensor data collected from the study described in Chapter 5, labels related to head activity need to be assigned for all groups. In addition, labels related to perceived positivity and rapport levels in periods of 2 to 5 minutes should be assigned to study how the evolution of perceived behaviors influences final rapport levels.

Furthermore, the expansion of the performed human study is recommended by collecting data from at least 10 more groups. It is also recommended to modify the study protocol to include a confederate in each group interaction so a higher amount of low rapport instances can be obtained.

- *Standards for data labeling and its analysis*

This work will be benefited from the creation of more standards for data labeling using people's perception of the quality of social interactions and identification of informative nonverbal behaviors. For example, an analysis to determine how much variation in perceived rapport values exists when labeling using just audio versus audio plus video.

the following question could be asked for the labeling of perceived rapport levels:

- *Social interaction recognition: algorithms and model implementation*

Related to the pre-processing of sensor data, methods to extract attention levels and changes in emotional reactions from EEG should be implemented. In addition, for the processing of audio

signals, noise removal filters and other advanced algorithms should be implemented. On the other hand, mathematical relationships between Type A features and Type B features to aspects of rapport such as positivity and coordination should be investigated. A mathematical relationship between sensor data to specific aspects of rapport will allow the creation of a rapport equation that could be used in the future to provide feedback on how to improve rapport in dyadic and group interactions.

To expand on this work, the following suggestions are given:

- *Hardware and software optimization*

It is recommended to design a single wearable device that integrates the sensing modalities selected in this work. In addition, it is recommended the design databases that will store sensor information from the customized wearable device.

- *Complete close loop of the monitoring system*

To achieve the goal of bringing awareness to individuals during group interactions, a feedback mechanism should be put into place. Experiments to determine effective feedback modalities and information that could improve social and self-awareness should be investigated. In addition, experiments testing the designed feedback mechanism should be performed.

## BIBLIOGRAPHY


[1] S. G. Rogelberg, "Why your meetings stink-and what to do about IT Strategies for engagement," *Harv Bus Rev*, vol. 97, no. 1, pp. 140–143, 2019.

[2] V. Rousseau, C. Aubé, and A. Savoie, "Teamwork behaviors: A review and an integration of frameworks," *Small Group Res*, vol. 37, no. 5, pp. 540–570, 2006.

[3] S. N. Young, "The neurobiology of human social behaviour: An important but neglected topic," *Journal of Psychiatry and Neuroscience*, vol. 33, no. 5, pp. 391–392, 2008.

[4] D. Umberson and J. K. Montez, "Social relationships and health: A flashpoint for health policy," *J Health Soc Behav*, vol. 51(1_suppl, pp. S54–S66, 2010.

[5] L. M. Hernandez and D. G. Blaze, Eds., "The impact of social and cultural environment on health," in *Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate*, no. 2, National Academies Press (US), 2006.

[6] N. Dasgupta, "Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations," *Soc Justice Res*, vol. 17, no. 2, pp. 143–169, 2004.

[7] R. Wheeler, "We all do it: Unconscious behavior, bias, and diversity," *Law Libr J*, vol. 107, no. 2, pp. 15–36, 2015.

[8] N. Alduncin, L. C. Huffman, H. M. Feldman, and I. M. Loe, "Executive function is associated with social competence in preschool-aged children born preterm or full term," *Early Hum Dev*, vol. 90, no. 6, pp. 299–306, 2014.

[9] P. M. Merikle, D. Smilek, and J. D. Eastwood, "Perception without awareness: perspectives from cognitive psychology." [Online]. Available: www.elsevier.com/locate/cognit.

[10] T. Pyszczynski, J. Greenberg, and S. Solomon, "Proximal and distal defense: A new perspective on unconscious motivation," *Curr Dir Psychol Sci*, vol. 9, no. 5, pp. 156–160, 2000.

[11] T. D. Wilson and N. Brekke, "Mental contamination and mental correction: Unwanted influences on judgments and evaluations," *Psychol Bull*, vol. 116, no. 1, pp. 117–142, 1994.

[12] Y. Chuo *et al.*, "Mechanically flexible wireless multisensor platform for human physical activity and vitals monitoring," *IEEE Trans Biomed Circuits Syst*, vol. 4, no. 5, pp. 281–294, 2010.

[13] J. Y. Kim, C. H. Chu, and M. S. Kang, "IoT-based unobtrusive sensing for sleep quality monitoring and assessment," *IEEE Sens J*, vol. 21, no. 3, pp. 3799–3809, 2021.

[14] C. Setz, B. Arnrich, J. Schumm, R. la Marca, G. Troster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *IEEE Trans. Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410–417, 2010.

[15] N. A. Selamat and S. H. M. Ali, "Automatic food intake monitoring based on chewing activity: A survey," *IEEE Access*, vol. 8, pp. 48846–48869, 2020.

[16] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[17] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "Identifying human behaviors using synchronized audio-visual cues," *IEEE Trans Affect Comput*, vol. 8, no. 1, pp. 54–66, Jan. 2017.

[18] C. Hessler and M. Abouelenien, "Using thermal images and physiological features to model human behavior: A survey," *Proceedings - IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018*, pp. 278–281, 2018.

[19] I. Poggi and F. D. Errico, "Social signals: A psychological perspective," in *Computer Analysis of Human Behavior*, 2011, pp. 185–225.

[20] A. Best, S. F. Warta, K. A. Kapalo, and S. M. Fiore, "Of mental states and machine learning: How social cues and signals can help develop artificial social intelligence," *Proceedings of the Human Factors and Ergonomics Society*, pp. 1361–1365, 2016.

[21] T. Kim, D. O. Olguín, B. N. Waber, and A. Pentland, "Sensor-based feedback systems in organizational computing," in *2009 International Conference on Computational Science and Engineering*, 2009, pp. 966–969.

[22] J. Frey, M. Grabli, R. Slyper, and J. Cauchard, "Breeze: Sharing biofeedback through wearable technologies," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[23] Y. Hao, D. Wang, and J. G. Budd, "Design of intelligent emotion feedback to assist users regulate emotions: Framework and principles," in *International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 938–943.

[24] G. Chanel, S. Pelli, N. Ravaja, and K. Kuikkaniemi, "Social interaction using mobile devices and biofeedback: Effects on presence, attraction and emotions," in *BioSPlay Workshop, Fun and Games Conference*, 2010, pp. 5–9.

[25] J. Terken, J. Sturm, and I. Patras, "Multimodal support for social dynamics in co-located meetings," *Pers Ubiquitous Comput*, vol. 14, no. 8, pp. 703–714, 2010.

[26] J. Sturm, O. H. Herwijnen, A. Eyck, and J. Terken, "Influencing social dynamics in meetings through a peripheral display," in *Proceedings of the 9th international conference on Multimodal interfaces*, 2007, pp. 263–270.

[27] F. Cvrčková, V. Žárský, and A. Markoš, "Plant studies may lead us to rethink the concept of behavior," *Front Psychol*, vol. 7, pp. 10–13, 2016.

[28] A. Bandura, “Human agency in social cognitive theory,” *American Psychologist*, pp. 1175–1184, 1989.

[29] A. Bandura, “Social cognitive theory,” in *Annals of child development*, vol. 6, R. Vasta, Ed. Greenwich, CT: JAI Press, 1989, pp. 1–60.

[30] W. Mischel, “The interaction of person and situation,” in *Personality at the Crossroads: Current Issues in Interactional Psychology*, D. Magnusson and N. S. Endler, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1977, pp. 333–352.

[31] T. J. Bouchard and J. C. Loehlin, “Genes, evolution, and personality,” *Behav Genet*, vol. 31, no. 3, pp. 243–273, 2001.

[32] A. H. Eagly and S. Chaiken, *The Psychology of Attitudes*. Hardcourt Brace Jovanovich College Publishers, 1993.

[33] B. L. Fredrickson, “The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions,” *American Psychologist*, vol. 56, no. 3, pp. 218–226, 2001.

[34] W. B. Cannon, “The James-Lange theory of emotions: A critical examination and an alternative theory,” *Am J Psychol*, vol. 39, no. 1, pp. 106–124, 1927.

[35] S. W. Porges, “Orienting in a defensive world: Mammalian modifications of our evolutionary heritage. A Polyvagal Theory.,” *Psychophysiology*, vol. 32, no. 4, pp. 301–318, 1995.

[36] S. W. Porges, “Polyvagal theory,” *Biol Psychol*, vol. 74, no. 2, pp. 116–143, 2007.

[37] G. Chanel and C. Mühl, “Connecting brains and bodies: Applying physiological computing to support social interaction,” *Interact Comput*, vol. 27, no. 5, pp. 534–550, 2015.

[38] J. A. Russell, “A circumplex model of affect,” *J Pers Soc Psychol*, vol. 39, no. 6, pp. 1161–1178, 1980.

[39] A. Mehrabian and J. A. Russell, “The three emotional dimension,” in *An approach to environmental psychology*, Cambridge, MA: MIT Press, 1974.

[40] R. Plutchik, “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” *Am Sci*, vol. 89, no. 4, pp. 344–350, 2001.

[41] G. A. van Kleef, “How emotions regulate social life: The Emotions as Social Information (EASI) model,” *Curr Dir Psychol Sci*, vol. 18, no. 3, pp. 184–188, 2009.

[42] N. H. Frijda, *The Emotions*. Cambridge University Press, 1986.

[43] P. L. Perrewé and P. E. Spector, “Personality research in the organizational sciences,” in *Research in Personnel and Human Resources Management*, vol. 21, G. R. Ferris and J. J. Martocchio, Eds. Elsevier Science/JAI Press, 2002, pp. 1–63.

[44] T. A. Judge, J. E. Bono, R. Ilies, and M. W. Gerhardt, “Personality and leadership: A qualitative and quantitative review,” *Journal of Applied Psychology*, vol. 87, no. 4, pp. 765–780, 2002.

[45] M. C. Ashton and K. Lee, “Empirical, theoretical, and practical advantages of the HEXACO model of personality structure,” *Personality and Social Psychology Review*, vol. 11, no. 2, pp. 150–166, May 2007.

[46] D. L. Paulhus and K. M. Williams, “The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy,” *J Res Pers*, vol. 36, no. 6, pp. 556–563, Dec. 2002.

[47] J. B. Rotter, “Generalized expectancies for internal versus external control of reinforcement,” *Psychological Monographs: General and Applied*, vol. 80, no. 1, pp. 1–28, 1966.

[48] J. K. Alberts, T. K. Nakayama, and J. N. Martin, *Human Communication in Society*, 3rd ed. Pearson, 2012.

[49] P. Watzlawick, J. B. Bavelas, and D. D. Jackson, *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies and Paradoxes*. W. W. Norton & Company, 1967.

[50] A. Pentland, *Honest Signals: How They Shape our World*. MIT Press, 2010.

[51] E. Goffman, *The Presentation of Self in Everyday Life*. New York: Anchor Books, 1959.

[52] C. Crivelli and A. J. Fridlund, “Facial displays are tools for social influence,” *Trends Cogn Sci*, vol. 22, no. 5, pp. 388–399, 2018.

[53] G. Mohammadi and A. Vinciarelli, “Automatic personality perception: Prediction of trait attribution based on prosodic features,” *IEEE Trans Affect Comput*, vol. 3, no. 3, pp. 273–284, 2012.

[54] A. Vinciarelli, H. Salamin, and M. Pantic, “Social Signal Processing: Understanding social interactions through nonverbal behavior analysis,” *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 231287, no. 231287, pp. 42–49, 2010.

[55] R. A. Calvo and S. D’Mello, “Affect detection: An interdisciplinary review of models, methods, and their applications,” *IEEE Trans Affect Comput*, vol. 1, no. 1, pp. 18–37, 2010.

[56] A. Pentland, “Social Signal Processing,” *IEEE Signal Process Mag*, vol. 24, no. 4, pp. 108–111, 2007.

[57] G. Bente, “New tools – new insights: Using emergent technologies in nonverbal communication research,” in *Reflections on Interpersonal Communication*, S. W. Wilson and S. W. Smith, Eds. San Diego: Cognella, 2019, pp. 161–188.

[58] K. Yun, K. Watanabe, and S. Shimojo, “Interpersonal body and neural synchronization as a marker of implicit social interaction,” *Sci Rep*, vol. 2, no. 959, pp. 1–8, 2012.

[59] S. M. Thurman and H. Lu, “Perception of social interactions for spatially scrambled biological motion,” *PLoS One*, vol. 9, no. 11, pp. 1–12, 2014.

[60] A. Innocenti, E. de Stefani, N. F. Bernardi, G. C. Campione, and M. Gentilucci, “Gaze direction and request gesture in social interactions,” *PLoS One*, vol. 7, no. 5, pp. 1–8, 2012.

[61] E. de Stefani and D. de Marco, “Language, gesture, and emotional communication: An embodied view of social interaction,” *Front Psychol*, vol. 10, no. 2063, pp. 1–8, 2019.

[62] C. Frith, “Role of facial expressions in social interactions,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3453–3458, 2009.

[63] R. E. Jack and P. G. Schyns, “The human face as a dynamic tool for social communication,” *Current Biology*, vol. 25, no. 14, pp. R621–R634, 2015.

[64] P. Filippi, “Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language,” *Front Psychol*, vol. 7, no. 1393, pp. 1–19, 2016.

[65] N. Henriksen, “Style, prosodic variation, and the social meaning of intonation,” *J Int Phon Assoc*, vol. 43, no. 2, pp. 153–193, 2013.

[66] G. Bente, N. C. Kramer, and F. Eschenburg, “Is there anybody out there,” *Mediated interpersonal communication*, pp. 131–157, 2008.

[67] P. Ekman and W. v. Friesen, “The repertoire of nonverbal behavior: Categories, origins, usage, and coding,” *Semiotica*, vol. 1, pp. 49–98, 1969.

[68] S. Duncan, “Some signals and rules for taking speaking turns in conversations,” *J Pers Soc Psychol*, vol. 23, no. 2, pp. 283–292, 1972.

[69] G. Bente, H. Leuschner, A. al Issa, and J. J. Blascovich, “The others: Universals and cultural specificities in the perception of status and dominance from nonverbal behavior,” *Conscious Cogn*, vol. 19, no. 3, pp. 762–777, 2010.

[70] B. M. Depaulo and H. S. Friedman, “Nonverbal communication,” in *The Handbook of Social Psychology*, D. T. Gilbert, S. T. Fiske, and G. Lindzey, Eds. McGraw-Hill, 1998, pp. 3–40.

[71] A. Vinciarelli and A. S. Pentland, “New social signals in a new interaction world: The next frontier for Social Signal Processing,” *IEEE Syst Man Cybern Mag*, vol. 1, no. 2, pp. 10–17, 2015.

[72] L. Siemens, “The balance between on-line and in-person interactions: Methods for the development of digital humanities collaboration,” *Digital Studies/Le champ numérique*, vol. 2, no. 1, 2011.

[73] R. L. Daft, R. H. Lengel, and L. K. Trevino, “Message equivocality, media selection, and manager performance: Implications for information systems,” *MIS Quarterly*, vol. 11, no. 3, pp. 354–366, 1987.

[74] J. Teevan *et al.*, “The new future of work: Research from Microsoft into the pandemic’s impact on work practices,” 2021. [Online]. Available: https://www.microsoft.com/en-us/research/publication/the-new-future-of-work-research-from-microsoft-into-the-pandemics-impact-on-work-practices/.

[75] B. O’Connail, S. Whittaker, and S. Wilbur, “Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication,” *Hum Comput Interact*, vol. 8, no. 4, pp. 389–428, 1993.

[76] E. S. Rintel, “Conversational management of network trouble perturbations in personal videoconferencing,” in *ACM International Conference Proceeding Series*, 2010, pp. 304–311.

[77] P. Murali, J. Hernandez, D. McDuff, K. Rowan, J. Suh, and M. Czerwinski, “AffectiveSpotlight: Facilitating the communication of affective responses from audience members during online presentations,” in *CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[78] M. C. Lashley, “Observational research, advantages and disadvantages,” in *The SAGE Encyclopedia of Communication Research Methods*, M. Allen, Ed. SAGE Publications, Inc, 2018, pp. 1113–1115.

[79] D. Albudaiwi, “Surveys, advantages and disadvantages of,” in *The SAGE Encyclopedia of Communication Research Methods*, M. Allen, Ed. SAGE Publications, Inc, 2018, pp. 1735–1736.

[80] N. Kehtarnavaz and M. Gamadia, “Real-time image and video processing: From research to reality,” in *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 5, Morgan & Claypool, 2005, pp. 1–108.

[81] L. E. Holmquist, J. Falk, and J. Wigström, “Supporting group collaboration with interpersonal awareness devices,” *Pers Ubiquitous Comput*, vol. 3, no. 1–2, pp. 13–21, 1999.

[82] R. Want, A. Hopper, V. Falcão, and J. Gibbons, “The active badge location system,” *ACM Transactions on Information Systems (TOIS)*, vol. 10, no. 1, pp. 91–102, 1992.

[83] H. Jang, S. P. Choe, S. N. B. Gunkel, S. Kang, and J. Song, "A system to analyze group socializing behaviors in social parties," *IEEE Trans Hum Mach Syst*, vol. 47, no. 6, pp. 801–813, 2017.

[84] R. W. Devaul, S. J. Schwartz, and A. S. Pentland, "MIThril: Context-aware computing for daily life," 2001. https://www.media.mit.edu/wearables/mithril/MIThril.pdf.

[85] R. DeVaul, M. Sung, J. Gips, and A. Pentland, "MIThril 2003: Applications and architecture," in *Seventh IEEE International Symposium on Wearable Computers*, 2003, p. 4.

[86] N. Eagle and A. Pentland, "Wearables in the workplace: sensing interactions at the office," in *IEEE International Symposium on Wearable Computers*, 2003, pp. 256–257.

[87] M. Laibowitz and J. A. Paradiso, "The UbER-Badge, a versatile platform at the juncture between wearable and social computing," in *International Conference on Pervasive Computing*, 2004, pp. 1–6.

[88] J. Gips and A. Pentland, "Mapping human networks," in *Fourth Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2006*, 2006, pp. 159–168.

[89] M. Laibowitz, J. Gips, R. Aylward, and A. Pentland, "A sensor network for social dynamics," in *2006 5th International Conference on Information Processing in Sensor Networks*, 2006, pp. 483–491.

[90] T. Choudhury and A. Pentland, "Characterizing social networks using the sociometer," in *North American Association of Computational Social and Organizational Science (NAACSOS)*, 2004, pp. 1–4.

[91] W. Dong, B. Lepri, T. Kim, F. Pianesi, and A. S. Pentland, "Modeling conversational dynamics and performance in a Social Dilemma task," in *5th International Symposium on Communications Control and Signal Processing, ISCCSP 2012*, 2012, pp. 1–4.

[92] Y. Zhang *et al.*, "TeamSense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2018, vol. 2, no. 3, pp. 1–22.

[93] T. Kim, A. Chang, L. Holland, and A. S. Pentland, "Meeting Mediator: Enhancing group collaboration using sociometric feedback," in *Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work*, 2008, pp. 457–466.

[94] O. Lederman, D. Calacci, A. MacMullen, D. C. Fehder, F. Murray, and A. S. Pentland, "Open Badges: A low-cost toolkit for measuring team communication and dynamics," *arXiv preprint*. 2017.

[95] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland, “Rhythm: A unified measurement platform for human organizations,” *IEEE Multimedia*, vol. 25, no. 1, pp. 26–38, 2018.

[96] A. Madan and A. (Sandy) Pentland, “VibeFones: Socially aware mobile phones,” in *2006 10th IEEE International Symposium on Wearable Computers*, 2006, pp. 109–112.

[97] J. Müller, S. Fàbregues, E. A. Guenther, and M. J. Romano, “Using sensors in organizational research-clarifying rationales and validation challenges for mixed methods,” *Front Psychol*, vol. 10, no. 1188, pp. 1–14, 2019.

[98] J. Gu *et al.*, “Wearable Social Sensing and its application in anxiety assesment,” in *2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 2017, pp. 305–308.

[99] J. Gu *et al.*, “Wearable Social Sensing: Content-based processing methodology and implementation,” *IEEE Sens J*, vol. 17, no. 21, pp. 7167–7176, 2017.

[100] L. Jiang *et al.*, “Wearable long-term social sensing for mental wellbeing,” *IEEE Sens J*, vol. 19, no. 19, pp. 8532–8542, Oct. 2019.

[101] L. Fraiwan, T. Basmaji, and O. Hassanin, “A mobile mental health monitoring system: A smart glove,” in *Proceedings - 14th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2018*, Jul. 2018, pp. 235–240.

[102] D. Girardi, F. Lanubile, and N. Novielli, “Emotion detection using noninvasive low cost sensors,” in *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, 2017, pp. 125–130.

[103] R. S. McGinnis *et al.*, “Wearable sensors and machine learning diagnose anxiety and depression in young children,” in *IEEE EMBS International Conference on Biomedical and Health Informatics*, 2018, pp. 410–413.

[104] S. S. Panicker and P. Gayathri, “A survey of machine learning techniques in physiology based mental stress detection systems,” *Biocybern Biomed Eng*, vol. 39, no. 2, pp. 444–469, 2019.

[105] A. A. Torres-García, O. Mendoza-Montoya, M. Molinas, J. M. Antelis, L. A. Moctezuma, and T. Hernández-Del-Toro, *Pre-processing and feature extraction*. Elsevier Inc., 2022.

[106] H. Malik, N. Fatema, and A. Iqbal, *Advances in machine learning and data analysis*. Elsevier Inc., 2021.

[107] “The ultimate guide to data labeling for machine learning.” https://www.cloudfactory.com/data-labeling-guide (accessed Oct. 19, 2021).

[108] “Proven AI & data best practices training & certification for project managers and team leaders.” https://www.cognilytica.com/ (accessed Oct. 19, 2021).

[109] L. Zhang *et al.*, "'BioVid Emo DB': A multimodal database for emotion analyses validated by subjective ratings," in *2016 IEEE Symposium Series on Computational Intelligence*, 2017, pp. 1–6.

[110] F. Ringeval *et al.*, "The AV + EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 3–8.

[111] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[112] S. Koelstra *et al.*, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans Affect Comput*, vol. 3, no. 1, pp. 18–31, 2012.

[113] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans Affect Comput*, vol. 3, no. 1, pp. 42–55, 2012.

[114] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int J Speech Technol*, vol. 21, no. 1, pp. 93–120, 2018.

[115] J. Kossaifi *et al.*, "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 3, pp. 1022–1040, 2021.

[116] Y. Noh *et al.*, "Enhancing quality of corpus annotation: Construction of the multi-layer corpus annotation and simplified validation of the corpus annotation," in *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, 2020, pp. 216–224.

[117] E. Douglas-Cowie *et al.*, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," *Affective Computing and Intelligent Interaction*, pp. 488–500, 2007.

[118] S. Davila-Montero, J. A. Dana-Le, G. Bente, A. T. Hall, and A. J. Mason, "Review and challenges of technologies for real-time human behavior monitoring," *IEEE Trans Biomed Circuits Syst*, vol. 15, no. 1, pp. 2–28, 2021.

[119] J. E. Grahe and F. J. Bernieri, "The importance of nonverbal cues in judging rapport," *J Nonverbal Behav*, vol. 23, no. 4, pp. 253--269, 1999.

[120] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychol Inq*, vol. 1, no. 4, pp. 285–293, 1990.

[121] F. J. Bernieri, J. M. Davis, J. S. Gillis, and J. E. Grahe, "Dyad rapport and the accuracy of its judgment across situations: A lens model analysis," *J Pers Soc Psychol*, vol. 71, no. 1, pp. 110–129, 1996.

[122] P. Müller, M. X. Huang, and A. Bulling, “Detecting low rapport during natural interactions in small groups from non-verbal behaviour,” in *23rd International Conference on Intelligent User Interfaces (IUI ’18)*, 2018, pp. 1–12.

[123] H. Hung and D. Gatica-Perez, “Estimating cohesion in small groups using audio-visual nonverbal behavior,” *IEEE Trans Multimedia*, vol. 12, no. 6, pp. 563–575, 2010.

[124] D. B. Jayagopi, B. Raducanu, and D. Gatica-Perez, “Characterizing conversational group dynamics using nonverbal behaviour,” in *2009 IEEE International Conference on Multimedia and Expo, ICME 2009*, 2009, pp. 370–373.

[125] 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 International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 2011, pp. 613–616.

[126] A. Cerekovic, O. Aran, and D. Gatica-Perez, “Rapport with virtual agents: What do human social cues and personality explain?,” *IEEE Trans Affect Comput*, vol. 8, no. 3, pp. 382–395, Jul. 2017.

[127] N. Wang and J. Gratch, “Rapport and facial expression,” 2009.

[128] R. Zhao, T. Sinha, A. W. Black, and J. Cassell, “Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior,” 2016.

[129] R. Zhao, A. Papangelis, and J. Cassell, “Towards a dyadic computational model of rapport management for human-virtual agent interaction,” in *Intelligent Virtual Agents*, T. Bickmore, S. Marsella, and C. Sidner, Eds. Springer, Cham, 2014, pp. 514–527.

[130] M. Cristani, R. Raghavendra, A. del Bue, and V. Murino, “Human behavior analysis in video surveillance: A Social Signal Processing perspective,” *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[131] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, “Automatic analysis of facial actions: A survey,” *IEEE Trans Affect Comput*, vol. 10, no. 3, pp. 325–347, 2019.

[132] S. Yang *et al.*, “IoT structured long-term wearable social sensing for mental wellbeing,” *IEEE Internet Things J*, vol. 6, no. 2, pp. 3652–3662, Apr. 2019.

[133] S. Ha *et al.*, “Integrated circuits and electrode interfaces for noninvasive physiological monitoring,” *IEEE Trans Biomed Eng*, vol. 61, no. 5, pp. 1522–1537, 2014.

[134] R. Sharma, V. I. Pavlovic, and T. S. Huang, “Toward multimodal human-computer interface,” *Proceedings of the IEEE*, vol. 86, no. 5, pp. 853–869, 1998.

[135] A. Jaimes and N. Sebe, “Multimodal human-computer interaction: A survey,” *Computer Vision and Image Understanding*, vol. 108, no. 1–2, pp. 116–134, 2007.

[136] D. Gatica-Perez, “Analyzing group interactions in conversations: A review,” in *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006, pp. 41–46.

[137] D. Gatica-Perez, *Modelling Interest in Face-to-Face Conversations from Multimodal Nonverbal Behaviour*, 1st ed. Elsevier, 2010.

[138] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huanag, “Human-centred intelligent Human Computer Interaction ($HCI^2$): How far are we from attaining it?,” *Int J Auton Adapt Commun Syst*, vol. 1, no. 2, p. 168, 2008.

[139] T. Theodorou, I. Mporas, and N. Fakotakis, “An overview of automatic audio segmentation,” *International Journal of Information Technology and Computer Science*, vol. 6, no. 11, pp. 1–9, Oct. 2014.

[140] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. 941–944.

[141] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[142] C. H. Wu and W. bin Liang, “Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels,” *IEEE Trans Affect Comput*, vol. 2, no. 1, pp. 10–21, Jan. 2011.

[143] H. Balti and A. S. Elmaghraby, “Speech emotion detection using time dependent self organizing maps,” in *IEEE International Symposium on Signal Processing and Information Technology, IEEE ISSPIT 2013*, 2013, pp. 470–478.

[144] M. Tahon, G. Degottex, and L. Devillers, “Usual voice quality features and glottal features for emotional valence detection,” in *Speech Prosody*, 2012, pp. 1–4.

[145] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog,” in *International Conference on Spoken Language Processing*, 2002, pp. 2037–2040.

[146] S. Sahoo and A. Routray, “Detecting aggression in voice using inverse filtered speech features,” *IEEE Trans Affect Comput*, vol. 9, no. 2, pp. 217–226, Apr. 2018.

[147] D. B. Jayagopi and D. Gatica-Perez, “Mining group nonverbal conversational patterns using probabilistic topic models,” *IEEE Trans Multimedia*, vol. 12, no. 8, pp. 790–802, Dec. 2010.

[148] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, “Using linguistic cues for the automatic recognition of personality in conversation and text,” *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[149] D. Furnham, “Language and Personality,” in *Handbook of Language and Social Psychology*, H. Giles and W. Robinson, Eds. Winley, 1990.

[150] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, “From nonverbal cues to perception: Personality and social attractiveness,” in *Cognitive Behavioural Systems*, 2012, pp. 60–72.

[151] K. Tusing, “The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence,” *Hum Commun Res*, vol. 26, no. 1, pp. 148–171, 2000.

[152] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, “Estimating dominance in multi-party meetings using speaker diarization,” *IEEE Trans Audio Speech Lang Process*, vol. 19, no. 4, pp. 847–860, 2011.

[153] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, “Modeling dominance in group conversations using nonverbal activity cues,” *IEEE Trans Audio Speech Lang Process*, vol. 17, no. 3, pp. 501–513, Mar. 2009.

[154] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, “A nonverbal behavior approach to identify emergent leaders in small groups,” *IEEE Trans Multimedia*, vol. 14, no. 3, pp. 816–832, 2012.

[155] D. Hillard, M. Ostendorf, and E. Shriberg, “Detection of agreement vs. disagreement in meetings: Training with unlabeled data,” in *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, 2003, pp. 34–36.

[156] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, “Detecting group interest-level in meetings,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, 2005, pp. I489–I492.

[157] L. S. Kennedy and D. P. W. Ellis, “Pitch-based emphasis detection for characterization of meeting recordings,” in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2003*, 2003, pp. 243–248.

[158] A. Vinciarelli, “Capturing order in social interactions,” *IEEE Signal Process Mag*, vol. 26, no. 5, 2009.

[159] H. D. Critchley, “Electrodermal responses: What happens in the brain,” *Neuroscientist*, vol. 8, no. 2, pp. 132–142, 2002.

[160] I. R. Kleckner *et al.*, “Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data,” *IEEE Trans Biomed Eng*, vol. 65, no. 7, pp. 1460–1467, 2018.

[161] S. Dávila-Montero, S. Parsnejad, and A. J. Mason, “Exploring the Relationship between Speech and Skin Conductance for Real-Time Arousal Monitoring,” in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[162] P. Slovák, P. Tennent, S. Reeves, and G. Fitzpatrick, "Exploring skin conductance synchronisation in everyday interactions," in *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, 2014, pp. 511–520.

[163] H. J. Pijeira-Díaz, H. Drachsler, S. Järvelä, and P. A. Kirschner, "Investigating collaborative learning success with physiological coupling indices based on electrodermal activity," in *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, 2016, pp. 64–73.

[164] E. Haataja, J. Malmberg, and S. Järvelä, "Monitoring in collaborative learning: Co-occurrence of observed behavior and physiological synchrony explored," *Comput Human Behav*, vol. 87, pp. 337–347, 2018.

[165] H. J. Pijeira-Díaz, H. Drachsler, S. Järvelä, and P. A. Kirschner, "Sympathetic arousal commonalities and arousal contagion during collaborative learning: How attuned are triad members?," *Comput Human Behav*, vol. 92, pp. 188–197, 2019.

[166] "10/20 System Positioning Manual," 2012. https://www.trans-cranial.com/docs/10_20_pos_man_v1_0_pdf.pdf (accessed Jun. 29, 2020).

[167] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *Neuroimage*, vol. 34, no. 4, pp. 1600–1611, Feb. 2007.

[168] M. Teplan, "Fundamentals of EEG measurement," *Measurement Science Review*, vol. 2, no. 2, pp. 1–11, 2002.

[169] S. Valenzi, T. Islam, P. Jurica, and A. Cichocki, "Individual classification of emotions using EEG," *J Biomed Sci Eng*, vol. 07, pp. 604–620, 2014.

[170] L. Zou, X. Chen, G. Dang, Y. Guo, and Z. J. Wang, "Removing muscle artifacts from EEG data via underdetermined joint blind source separation: A simulation study," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 1, pp. 187–191, 2020.

[171] Y. Yang and J. Zhou, "Recognition and analyses of EEG&ERP signals related to emotion: From the perspective of psychology," in *First International Conference on Neural Interface and Control*, 2005, pp. 96–99.

[172] R. N. Duan, J. Y. Zhu, and B. L. Lu, "Differential entropy feature for EEG-based emotion classification," in *International IEEE/EMBS Conference on Neural Engineering, NER*, 2013, pp. 81–84.

[173] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans Affect Comput*, vol. 5, no. 3, pp. 327–339, 2014.

[174] W. L. Zheng and B. L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with Deep Neural Networks," *IEEE Trans Auton Ment Dev*, vol. 7, no. 3, pp. 162–175, 2015.

[175] B. Pholpoke, T. Songthawornpong, and W. Wattanapanitch, “A micropower motion artifact estimator for input dynamic range reduction in wearable ECG acquisition systems,” *IEEE Trans Biomed Circuits Syst*, vol. 13, no. 5, pp. 1021–1035, 2019.

[176] U. Satija, B. Ramkumar, and M. Sabarimalai Manikandan, “A review of signal processing techniques for Electrocardiogram signal quality assessment,” *IEEE Rev Biomed Eng*, vol. 11, pp. 36–52, 2018.

[177] D. Nikolova, P. Petkova, A. Manolova, and P. Georgieva, “ECG-based emotion recognition: Overview of methods and applications,” in *ANNA ’18; Advances in Neural Networks and Applications 2018*, 2018, pp. 1–5.

[178] C. Xiefeng, Y. Wang, S. Dai, P. Zhao, and Q. Liu, “Heart sound signals can be used for emotion recognition,” *Sci Rep*, vol. 9, no. 1, Dec. 2019.

[179] J. Cai, G. Liu, and M. Hao, “The research on emotion recognition from ECG signal,” in *International Conference on Information Technology and Computer Science, ITCS 2009*, 2009, vol. 1, pp. 497–500.

[180] D. S. Quintana, A. J. Guastella, T. Outhred, I. B. Hickie, and A. H. Kemp, “Heart rate variability is associated with emotion recognition: Direct evidence for a relationship between the autonomic nervous system and social cognition,” *International Journal of Psychophysiology*, vol. 86, no. 2, pp. 168–172, 2012.

[181] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, “Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges,” *Information Fusion*, vol. 35, pp. 1339–1351, 2017.

[182] J. Kim, “Bimodal emotion recognition using speech and physiological changes,” in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. 2007, pp. 265–280.

[183] G. Chanel, M. Bétrancourt, T. Pun, D. Cereghetti, and G. Molinari, “Assessment of computer-supported collaborative processes using interpersonal physiological and eye-movement coupling,” in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 116–122.

[184] G. Chanel, S. Avry, G. Molinari, M. Betrancourt, and T. Pun, “Multiple users’ emotion recognition: Improving performance by joint modeling of affective reactions,” in *2017 7th International Conference on Affective Computing and Intelligent Interaction*, 2017, pp. 92–97.

[185] T. Vogt, E. André, and N. Bee, “EmoVoice - A framework for online recognition of emotions from voice,” in *Perception in Multimodal Dialogue Systems. PIT 2008. Lecture Notes in Computer Science*, vol. 5078, E. André, L. Dybkjær, W. Minker, H. Neumann, R. Pieraccini, and M. Weber, Eds. Berlin, Heidelberg: Springer, 2008.

[186] M. Abdelwahab and C. Busso, “Supervised domain adaptation for emotion recognition from speech,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2015, pp. 5058–5062.

[187] L. Cen, F. Wu, Z. L. Yu, and F. Hu, *A Real-Time Speech Emotion Recognition System and its Application in Online Learning*. Elsevier Inc., 2016.

[188] R. Rajak and R. Mall, “Emotion recognition from audio, dimensional and discrete categorization using CNNs,” in *2019 IEEE Region 10 Conference (TENCON 2019)*, 2019, pp. 301–305.

[189] R. B. Lanjewar, S. Mathurkar, and N. Patel, “Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques,” *Procedia Comput Sci*, vol. 49, pp. 50–57, 2015.

[190] S. Jing, X. Mao, and L. Chen, “Prominence features: Effective emotional features for speech emotion recognition,” *Digital Signal Processing: A Review Journal*, vol. 72, pp. 216–231, 2018.

[191] I. McCowan *et al.*, “Automatic analysis of multimodal group actions in meetings,” *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 3, pp. 305–317, 2005.

[192] K. Katevas, K. Hänsel, R. Clegg, I. Leontiadis, H. Haddadi, and L. Tokarchuk, “Finding Dory in the crowd: Detecting social interactions using multi-modal mobile sensing,” in *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems*, 2019, pp. 37–42.

[193] M. Wöllmer *et al.*, “Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies,” in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2008, pp. 597–600.

[194] K. Brady *et al.*, “Multi-modal audio, video and physiological sensor learning for continuous emotion prediction,” in *Proceedings of the 6th International Workshop Audio/Visual Emotion Challenge*, 2016, pp. 97–104.

[195] W. L. Zheng and B. L. Lu, “A multimodal approach to estimating vigilance using EEG and forehead EOG,” *J Neural Eng*, vol. 14, no. 2, pp. 1–14, 2017.

[196] A. Huk, K. Bonnen, and B. J. He, “Beyond trial-based paradigms: Continuous behavior, ongoing neural activity, and natural stimuli,” *Journal of Neuroscience*, vol. 38, no. 35, pp. 7551–7558, Aug. 2018.

[197] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, “A review on wearable photoplethysmography sensors and their potential future applications in health care,” *Int J Biosens Bioelectron*, vol. 4, no. 4, pp. 195–202, 2018.

[198] M. van Dooren, J. J. G. (Gert J. de Vries, and J. H. Janssen, “Emotional sweating across the body: Comparing 16 different skin conductance measurement locations,” *Physiol Behav*, vol. 106, no. 2, pp. 298–304, 2012.

[199] “Emotiv.” https://www.emotiv.com/.

[200] “Enobio 8.” https://www.neuroelectrics.com/solutions/enobio/8/.

[201] “E4 wristband.” https://www.empatica.com/research/e4/.

[202] C. Kothe, D. Medine, C. Boulay, M. Grivich, and T. Stenner, “LabStreamingLayer’s Documentation,” 2019. https://labstreaminglayer.readthedocs.io/index.html.

[203] “Shimmer Engineering Team (2021). Shimmer MATLAB Instrument Driver,” *MATLAB Central File Exchange*, 2021. https://www.mathworks.com/matlabcentral/fileexchange/43712-shimmer-matlab-instrument-driver (accessed Jun. 03, 2021).

[204] S. Dávila-Montero, S. Parsnejad, E. Ashoori, D. Goderis, and A. J. Mason, “Design of a multi-sensor framework for the real-time monitoring of social interactions,” in *International symposium on circuits and systems*, 2022, pp. 615–619.

[205] I. Severin, “Head gesture-based on IMU sensors: A performance comparison between the unimodal and multimodal approach,” in *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, 2021, pp. 3–6.

[206] K. Sancheti, K. S. Krishnan, A. Suhaas, and P. Suresh, “Hands-free cursor control using intuitive head movements and cheek muscle twitches,” *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2018-Octob, no. October, pp. 356–361, 2019.

[207] C. L. Fall *et al.*, “A multimodal adaptive wireless control interface for people with upper-body disabilities,” *IEEE Trans Biomed Circuits Syst*, vol. 12, no. 3, pp. 564–575, 2018.

[208] T. Yokozuka, E. Ono, Y. Inoue, K. I. Ogawa, and Y. Miyake, “The relationship between head motion synchronization and empathy in unidirectional face-to-face communication,” *Front Psychol*, vol. 9, no. SEP, pp. 1–10, 2018.

[209] A. Borowska-Terka and P. Strumillo, “Person independent recognition of head gestures from parametrised and raw signals recorded from inertial measurement unit,” *Applied Sciences (Switzerland)*, vol. 10, no. 12, 2020.

[210] J. R. Terven, B. Raducanu, M. E. Meza, and J. Salas, “Evaluating real-time mirroring of head gestures using smart glasses,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 452–460.

[211] S. Ionut-Cristian and D. Dan-Marius, “Using inertial sensors to determine head motion—a review,” *J Imaging*, vol. 7, no. 12, 2021.

[212] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digital Signal Processing: A Review Journal*, vol. 110, pp. 1–28, 2021.

[213] C. Yarra and P. K. Ghosh, "Automatic intonation classification using temporal patterns in utterance-level pitch contour and perceptually motivated pitch transformation," *J Acoust Soc Am*, vol. 144, no. 5, pp. EL471–EL476, 2018.

[214] G. Szaszák, D. Sztahó, and K. Vicsi, "Automatic intonation classification for speech training systems," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 1899–1902.

[215] E. Rodero, "Intonation and emotion: Influence of pitch levels and contour type on creating emotions," *Journal of Voice*, vol. 25, no. 1, pp. e25–e34, Jan. 2011.

[216] L. Liu, A. Götz, P. Lorette, and M. D. Tyler, "How tone, intonation and emotion shape the development of infants' fundamental frequency perception," *Front Psychol*, vol. 13, no. June, pp. 1–14, 2022.

[217] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int J Speech Technol*, vol. 15, no. 2, pp. 99–117, 2012.

[218] R. López-Cózar, Z. Callejas, M. Kroul, J. Nouza, and J. Silovský, "Two-level fusion to improve emotion classification in spoken dialogue systems," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5246 LNAI, no. 1, pp. 617–624, 2008.

[219] I. Luengo, E. Navas, and I. Hernáez, "Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 332–335, 2009.

[220] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Commun*, vol. 40, no. 1–2, pp. 117–143, 2003.

[221] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," 2000.

[222] Z. Zhang, S. Chapman, and F. Ciravegna, "A methodology towards effective and efficient manual document annotation: Addressing annotator discrepancy and annotation quality," in *International Conference on Knowledge Engineering and Knowledge Management*, 2010, pp. 301–315.

[223] L. R. Rabiner and R. W. Schafer, "Speech-background/Silence discrimination," in *Theory and Applications of Digital Speech Processing*, Pearson Higher Education, Inc., 2011, pp. 586–595.

[224] D. J. Hermes and J. C. van Gestel, “The frequency scale of speech intonation,” *J Acoust Soc Am*, vol. 90, pp. 97–102, 1991.

[225] M. McHugh, “Interrater reliability: the kappa statistic,” *Biochem Med (Zagreb)*, vol. 22, no. 3, pp. 276–282, 2012.

[226] J. S. Uebersax, “Validity inferences from interobserver agreement,” *Psychol Bull*, vol. 104, no. 3, pp. 405–416, 1988.

[227] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[228] K. L. Gwet, “Benchmarking inter-rater reliability coefficients,” in *The definitive guide to measuring the extent of agreement among raters*, Fourth edi., Gaithersburg, MD: Advance Analytics, LLC, 2014, pp. 163–182.

[229] S. R. Krothapalli and S. G. Koolagudi, “Characterization and recognition of emotions from speech using excitation source information,” *Int J Speech Technol*, vol. 16, no. 2, pp. 181–201, 2013.

[230] J. B. Alonso, J. Cabrera, C. M. Travieso, K. López-de-Ipiña, and A. Sánchez-Medina, “Continuous tracking of the emotion temperature,” *Neurocomputing*, vol. 255, pp. 17–25, 2017.

[231] K. Bahreini, R. Nadolski, and W. Westera, “Towards real-time speech emotion recognition for affective e-learning,” *Educ Inf Technol (Dordr)*, vol. 21, no. 5, pp. 1367–1386, 2016.

[232] M. Bradley and P. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” *J Behav Ther Exp Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[233] S. Pawar, T. Jacques, K. Deshpande, R. Pusapati, and M. J. Meguerdichian, “Evaluation of cognitive load and emotional states during multidisciplinary critical care simulation sessions,” *BMJ Simul Technol Enhanc Learn*, vol. 4, no. 2, pp. 87–91, 2018.

[234] C. Kothe and C. Brunner, “XDF (Extensible Data Format),” 2014. https://code.google.com/archive/p/xdf/.

[235] I. Krumpal, “Determinants of social desirability bias in sensitive surveys: A literature review,” *Qual Quant*, vol. 47, no. 4, pp. 2025–2047, 2013.

[236] S. Li, Y. Ma, H. Huang, and S. Li, “An Improved DTW Method for Human Behavior Recognition,” in *2019 2nd International Conference on Intelligence Systems Research and Mechatronics Engineering*, 2019, pp. 187–192.

[237] F. Ringeval *et al.*, “Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data,” *Pattern Recognit Lett*, vol. 66, pp. 22–30, 2015.

[238] F. Moukayed, H. Yun, T. Bisson, and A. Fortenbacher, “Detecting academic emotions from learners’ skin conductance and heart rate: A Data-driven approach using Fuzzy Logic,” in *Proceedings of DeLFI Workshops*, 2018, pp. 1–10.

[239] B. Zhong *et al.*, “Emotion recognition with facial expressions and physiological signals,” in *2017 IEEE Symposium Series on Computational Intelligence*, 2017, pp. 1–8.

[240] H. Yun, A. Fortenbacher, N. Pinkwart, T. Bisson, and F. Moukayed, “A pilot study of emotion detection using sensors in a learning context: Towards an affective learning companion,” in *DeLFI/GMW Workshops*, 2017, pp. 1–11.

[241] W. Mou, H. Gunes, and I. Patras, “Alone versus In-a-group : A comparative analysis of facial affect recognition,” in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 521–525.

[242] W. Wei and Q. Jia, “Weighted Feature Gaussian Kernel SVM for Emotion Recognition,” *Comput Intell Neurosci*, vol. 2016, pp. 1–8, 2016.

[243] S. Chen, Y. L. Tian, Q. Liu, and D. N. Metaxas, “Recognizing expressions from face and body gesture by temporal normalized motion and appearance features,” *Image Vis Comput*, vol. 31, no. 2, pp. 175–185, 2013.

[244] M. Shimura, F. Monma, S. Mitsuyoshi, M. Shuzo, T. Yamamoto, and I. Yamada, “Descriptive analysis of emotion and feeling in voice,” in *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2010*, 2010, pp. 1–4.

[245] J.-C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, and S. Abrilian, “Manual annotation and automatic image processing of multimodal emotional behaviors: Validating the annotation of TV interviews,” *Personal Ubiquitous Comput.*, vol. 13, no. 1, pp. 69–76, 2009.

[246] M. M. Khan, R. D. Ward, and M. Ingleby, “Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature,” *ACM Trans Appl Percept*, vol. 6, no. 1, pp. 1–22, 2009.

[247] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, “Technique for automatic emotion recognition by body gesture analysis,” in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008, pp. 1–6.

[248] Y. S. Shin, “Facial expression recognition based on emotion dimensions on manifold learning,” in *International Conference on Computational Science*, 2007, pp. 81–88.

[249] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, “Web-based database for facial expression analysis,” in *IEEE International Conference Multimedia and Expo*, 2005, pp. 5–8.

[250] L. Chaby, M. Chetouani, M. Plaza, and D. Cohen, “Exploring multimodal social-emotional behaviors in autism spectrum disorders: An interface between social signal processing and psychopathology,” in *2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, 2012, pp. 950–954.

[251] A. Mahdhaoui, F. Ringeval, and M. Chetouani, “Emotional speech characterization based on multi-features fusion for face-to-face interaction,” in *2009 International Conference on Signals, Circuits and Systems*, 2009, pp. 1–6.

[252] M. Dahmane, P.-L. St-Charles, M. Lalonde, K. Heffner, and S. Foucher, “Arousal and valence estimation for visual non-intrusive stress monitoring,” in *2019 9th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2019, pp. 1–6.

[253] F. Ahmed, A. S. M. H. Bari, and M. L. Gavrilova, “Emotion recognition from body movement,” *IEEE Access*, vol. 8, pp. 11761–11781, 2020.

[254] B. D. Yetton, J. Revord, S. Margolis, S. Lyubomirsky, and A. R. Seitz, “Cognitive and physiological measures in well-being science: Limitations and lessons,” *Front Psychol*, vol. 10, no. 1630, pp. 1–18, 2019.

[255] T. Keshari and S. Palaniswamy, “Emotion recognition using feature-level fusion of facial expressions and body gestures,” in *Proceedings of the Fourth International Conference on Communication and Electronics Systems (ICCES 2019)*, 2019, pp. 1184–1189.

[256] M. Raja and S. Sigg, “RFexpress! - RF emotion recognition in the wild,” in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017*, 2017, pp. 38–41.

[257] A. Pradhan, A. Singh, and S. Saraswat, “Emotion recognition through wireless signal,” in *2017 4th International Conference on Signal Processing and Integrated Networks, SPIN 2017*, 2017, pp. 91–95.

[258] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, “Prediction of the leadership style of an emergent leader using audio and visual nonverbal features,” *IEEE Trans Multimedia*, vol. 20, no. 2, pp. 441–456, 2018.

[259] L. Batrinca, N. Mana, B. Lepri, N. Sebe, and F. Pianesi, “Multimodal personality recognition in collaborative goal-oriented tasks,” *IEEE Trans Multimedia*, vol. 18, no. 4, pp. 659–673, 2016.

[260] O. Aran and D. Gatica-Perez, “Fusing audio-visual nonverbal cues to detect dominant people in small group conversations,” in *2010 International Conference on Pattern Recognition*, 2010, pp. 3687–3690.

[261] F. Pianesi, B. Lepri, A. Cappelletti, M. Zancanaro, and N. Mana, “Multimodal recognition of personality traits in social interactions,” in *Proceedings of the 10th International Conference on Multimodal Interfaces*, 2008, pp. 53–60.

[262] H. Hung *et al.*, “Using audio and video features to classify the most dominant person in a group meeting,” in *Proceedings of the 15th ACM International Conference on Multimedia*, 2007, pp. 835–838.

[263] D. Zhang, D. Gatica-perez, S. Bengio, and D. Roy, “Learning influence among interacting Markov Chains,” in *Advances in Neural Information Processing Systems*, 2006, pp. 1577–1584.

[264] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, “Towards measuring human interactions in conversational settings,” in *IEEE International Workshop on Cues in Communication*, 2001, pp. 1577–1584.

[265] Z. Shen, A. Elibol, and N. Y. Chong, “Inferring human personality traits in human-robot social interaction,” in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2019, pp. 578–579.

[266] G. Leone, S. Migliorisi, and I. Sessa, “Detecting social signals of honesty and fear of appearing deceitful: A methodological proposal,” in *7th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2016 - Proceedings*, 2016, pp. 289–294.

[267] L. Ahonen, B. U. Cowley, A. Hellas, and K. Puolamäki, “Biosignals reflect pair-dynamics in collaborative work: EDA and ECG study of pair-programming in a classroom environment,” *Sci Rep*, vol. 8, no. 1, pp. 1–16, 2018.

[268] J. Malmberg, S. Järvelä, J. Holappa, E. Haataja, X. Huang, and A. Siipo, “Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning?,” *Comput Human Behav*, vol. 96, pp. 235–245, 2019.

[269] M. T. Knierim, D. Jung, V. Dorner, and C. Weinhardt, “Designing live biofeedback for groups to support emotion management in digital collaboration,” in *International Conference on Design Science Research in Information System and Technology*, 2017, pp. 479–484.

[270] A. (Sandy) Pentland, “Social dynamics: Signals and behavior,” in *International Conference on Developmental Learning*, 2004, pp. 1–5.

[271] A. Marcos-Ramiro, D. Pizarro, M. Marron-Romera, and D. Gatica-Perez, “Let your body speak: Communicative cue extraction on natural interaction using RGBD data,” *IEEE Trans Multimedia*, vol. 17, no. 10, pp. 1721–1732, 2015.

[272] E. Shmueli, V. K. Singh, B. Lepri, and A. Pentland, “Sensing, understanding, and shaping social behavior,” *IEEE Trans Comput Soc Syst*, vol. 1, no. 1, pp. 22–34, 2014.

[273] U. Avci and O. Aran, “Effect of nonverbal behavioral patterns on the performance of small groups,” in *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, 2014, pp. 9–14.

[274] C. Busso, P. G. Georgiou, and S. S. Narayanan, “Real-time monitoring of participants’ interaction in a meeting using audio-visual sensors,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, vol. 2, pp. 685–688.

[275] Y. Shi, S. Das, S. Douglas, and S. Biswas, “An experimental wearable IoT for data-driven management of autism,” in *2017 9th International Conference on Communication Systems and Networks, COMSNETS 2017*, Jun. 2017, pp. 468–471.

[276] G. Schiavo, “Socially-aware interfaces for supporting collocated interaction,” in *IUI Companion ’14: Proceedings of the companion publication of the 19th international conference on Intelligent User Interfaces*, 2014, pp. 65–67.

[277] S. O. Ba and J. M. Odobez, “Multiperson visual focus of attention from head pose and meeting contextual cues,” *IEEE Trans Pattern Anal Mach Intell*, vol. 33, no. 1, pp. 101–116, 2011.

[278] M. T. Curran, J. R. Gordon, L. Lin, P. K. Sridhar, and J. Chuang, “Understanding digitally-mediated empathy: An exploration of visual, narrative, and biosensory informational cues,” in *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, 2019, pp. 1–13.

[279] A. Vinciarelli, M. Pantic, and H. Bourlard, “Social signal processing: Survey of an emerging domain,” *Image Vis Comput*, vol. 27, no. 12, pp. 1743–1759, 2009.

[280] S. W. Byun, S. P. Lee, and H. S. Han, “Feature selection and comparison for the emotion recognition according to music listening,” in *2017 International Conference on Robotics and Automation Sciences*, 2017, pp. 172–176.

## APPENDIX A: TOPIC QUESTIONNAIRE

This appendix section contains the instructions and the list of statements used in the Topic questionnaire:

**Instructions:**

*You are given a series of topic statements, using the provided scale, indicate how much you agree or disagree with the statement. If you find that a topic statement might be uncomfortable or offensive to discuss with someone having a different opinion, you have the opportunity to not respond to how much you agree or disagree with the statement. If you do not respond, you are opting out of discussing that topic statement.*

*Please note that all personal information will be kept completely confidential and none of the responses you provide will be connected to your name or email address.*

*Disclaimer: These topic statements do not represent the official policy or position of Michigan State University, the College of Engineering, the Electrical and Computer Engineering department, or the Study Team members.*

**(Topic) Statements**:

- *(Gun control) The government should regulate firearms through stricter gun control laws, including more extensive background checks and regulations on assault weapons.*
- *(Gun control) The "right of the people to keep and bear arms" means that the government cannot regulate firearms in any way.*
- *(Vegetarianism) Meet is a normal part of my diet and important for a healthy life.*
- *(Vegetarianism)Vegetarianism is more sustainable for food production and reduces cruelty to animals.*

- *(Animal testing) Animal testing is unethical.*
- *(Animal testing) Although animals may feel pain or die as a result of it, animal testing is necessary in order to save human lives.*
- *(Universal healthcare) Access to affordable, quality healthcare should be a fundamental service provided by the government.*
- *(Universal healthcare) Tax money should not be used to provide healthcare for everyone; people should be responsible for themselves.*
- *(Death penalty) No matter the crime, the death penalty should never be applied because killing is wrong.*
- *(Death penalty) The death penalty should be used to deter heinous crimes.*
- *(Religious freedom) My personal religion is the one true religion.*
- *(Religious freedom) All people should feel free to practice any faith or to have no faith without fear of peer or government coercion.*
- *(Vaccines) Some vaccines save lives and should be mandatory to protect the population.*
- *(Vaccines) Individuals should have the right to choose whether or not to be vaccinated.*
- *(Animal hunting) Animal hunting is a fun sport and part of the American culture.*
- *(Animal hunting) Animal hunting constitutes animal abuse and should be prohibited.*
- *(Professional sports*) *Professional sports are a great source of entertainment.*
- *(Professional sports*) *People should not be paid to play sports.*

- *(College athletes) College athletes should not be allowed to receive payment from sponsors because it ruins the purity of the game.*
- *(College athletes) College athletes work hard and generate income for the university and should be compensated for their efforts.*
- *(Exercise) I consider exercising part of my daily routine.*
- *(Exercise) I rarely even think about exercising.*
- *(TV shows) I love a good TV show or movie when I have time.*
- *(TV shows) I consider watching fictional/reality TV a waste of time.*
- *(Travel) I love to travel, experience new cultures, and meet new people.*
- *(Travel) Traveling is overrated. I prefer to stay near home.*
- *(Video games) I enjoy playing video games.*
- *(Video games) I consider video games a waste of time.*
- *(Food) I like trying new foods and going to different restaurants.*
- *(Food) I prefer to eat food that I know I like.*
- *(Outdoor activities) I enjoy outdoor adventurers and activities.*
- *(Outdoor activities) Outdoor adventures and activities are not worth the effort; my house is all the nature I need.*
- *(Social interactions) Virtual interactions are enough for me to fulfill my social needs.*
- *(Social interactions) I need in-person interactions to fulfill my social needs.*
- *(Environment) It is the duty of all humans to protect the environment and minimize our carbon footprint.*
- *(Environment) Human consumption does not affect the environment negatively.*

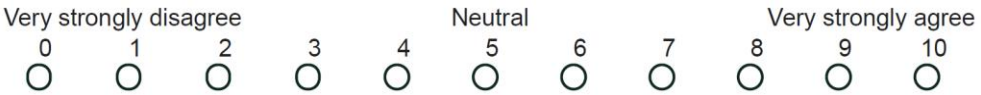


***Figure 39**. 11-point Likert scale used to collect the opinions of the participants about the given topics.*

Participants provided their opinion using an 11-point Likert scale, shown in **Figure 39**.

## APPENDIX B: EMOTIONAL STATE QUESTIONNAIRE

This appendix section contains the instructions and the items used in the Emotional State questionnaire:

**Instructions:**

*Rate how are you feeling at this moment using the following scales. Please note that all personal information will be kept completely confidential and none of the responses you provide will be connected to your name or email address.*

**Items:**

*Rate how are you feeling in terms of arousal, valence, and dominance:*

Participants provided their responses using the scales in **Figure 40**.

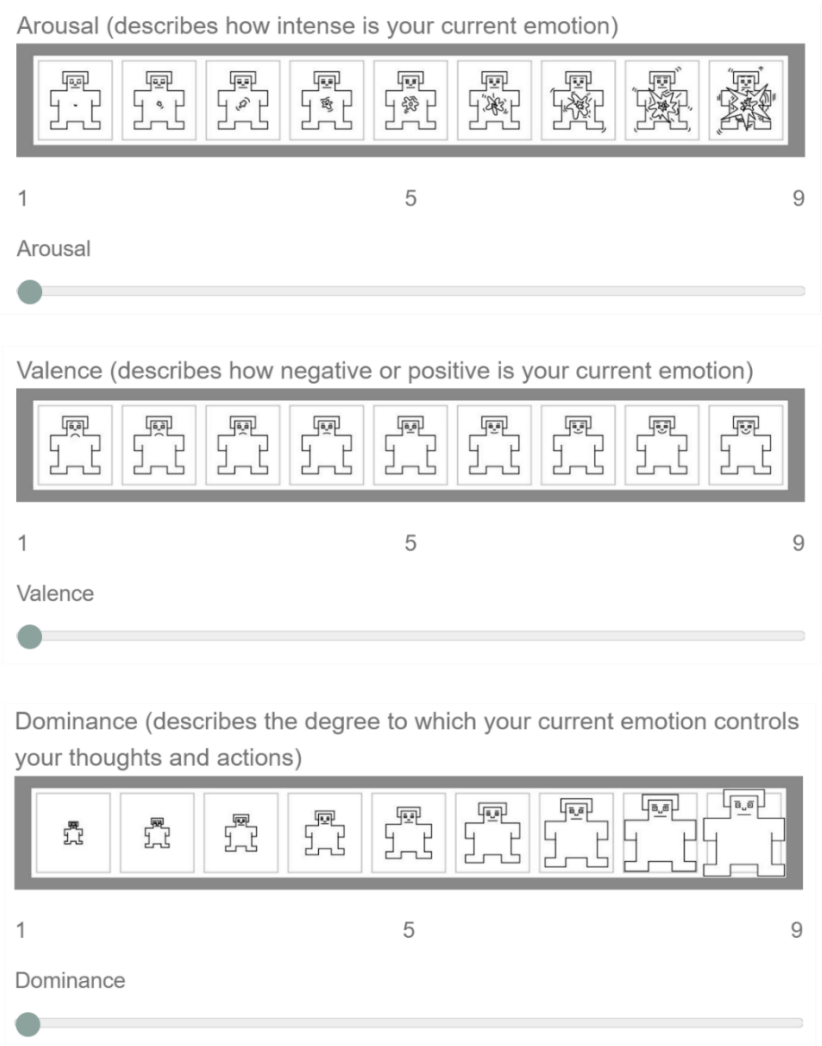


***Figure 40***. *9-point Self-Assessment Manikin scale for arousal, valence, and dominance.*

*Now, please rate how are you feeling at this moment using the following scale:*

Scale is shown in **Figure 41**.

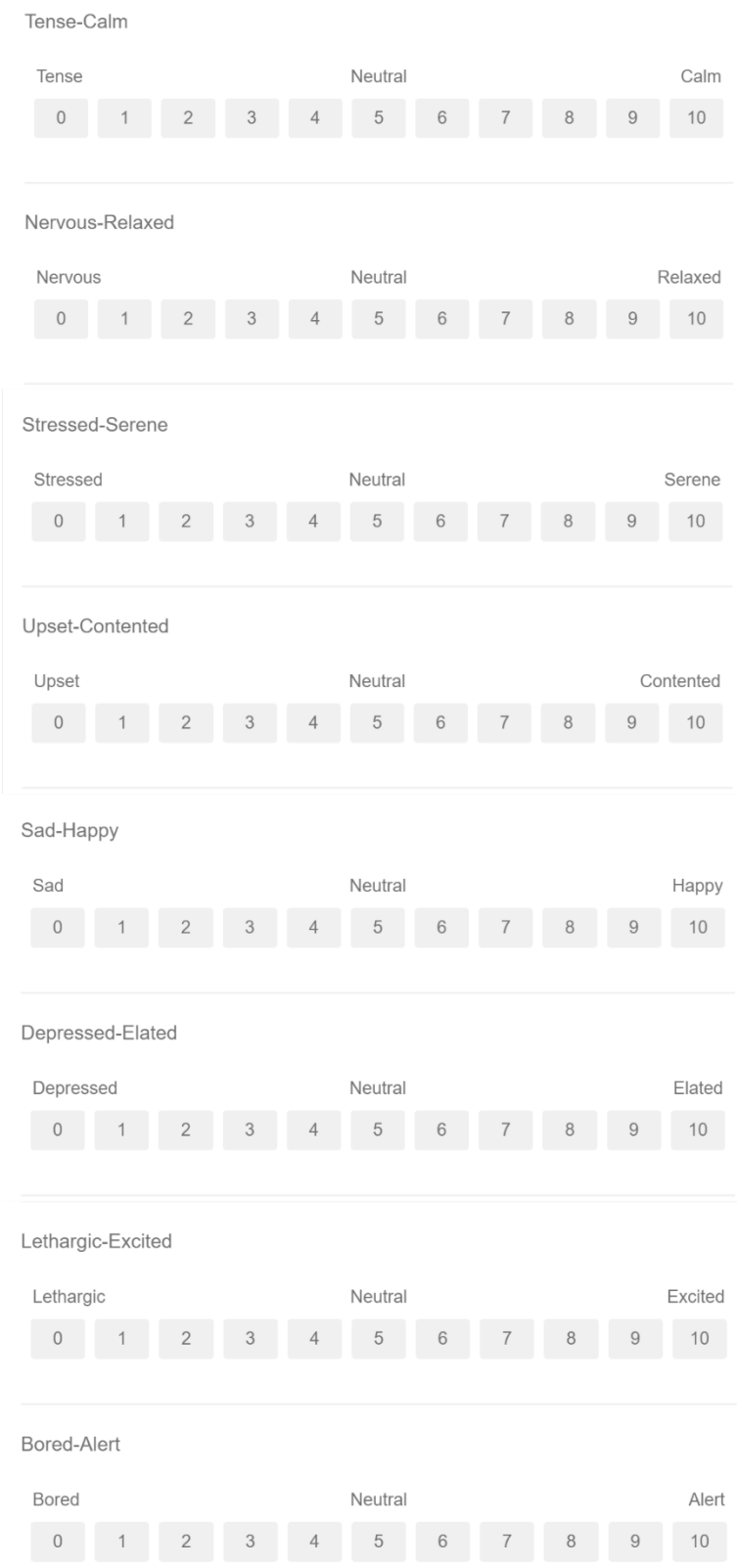


***Figure 41.*** *11-point rating tool based on the circumplex model of emotion.*

## APPENDIX C: RAPPORT QUESTIONNAIRE

This appendix section contains the instructions and the items used in the rapport questionnaire:

**Instructions:**

*Please note that all personal information will be kept completely confidential and none of the responses you provide will be connected to your name or email address.*

**Items:**

First, participants selected a reference letter that identified them, as shown in **Figure 42**.

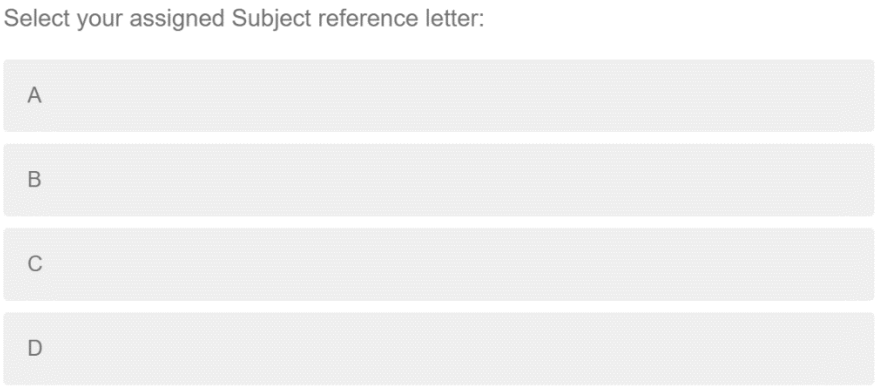


***Figure 42.*** *Selection of reference letter by the participant.*

Then, participants rated their own perceived level of engagement during the interaction using the items and scale shown in **Figure 43**, rated how much they enjoyed the interaction using the scale in **Figure 44**, and the level of linking of other people using the scale in **Figure 45**.

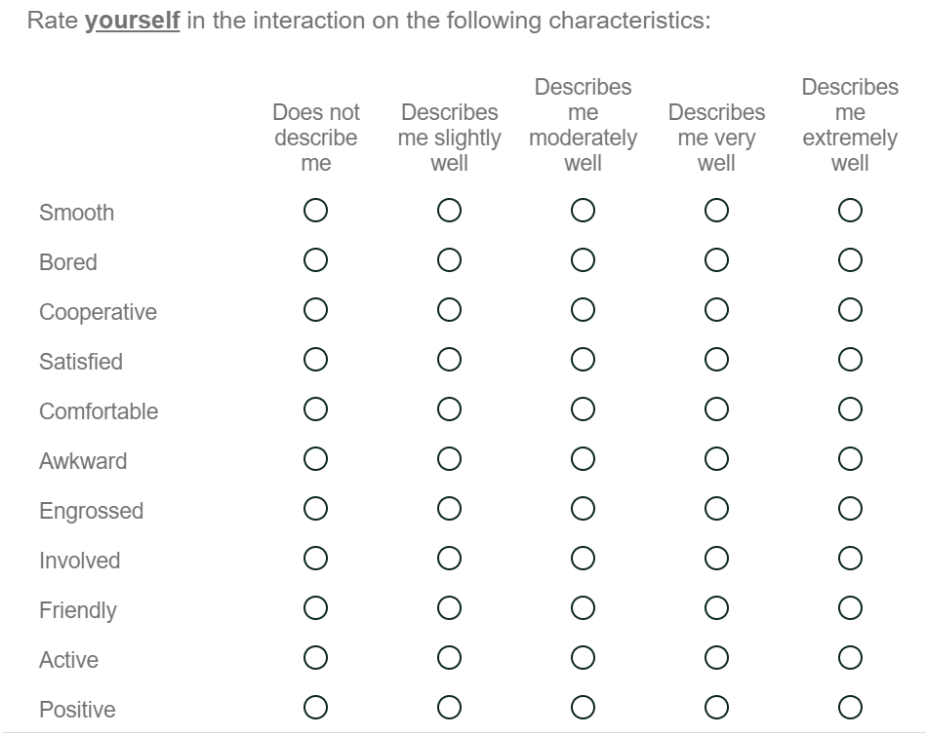


***Figure 43.*** *Items and scale used to rate oneself interaction performance during the discussion section.*

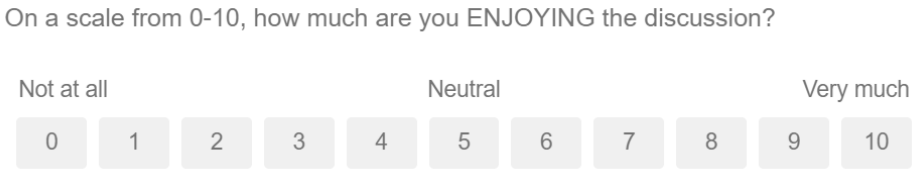


***Figure 44.*** *Scale to measure the overall feeling of enjoyment during the interaction.*

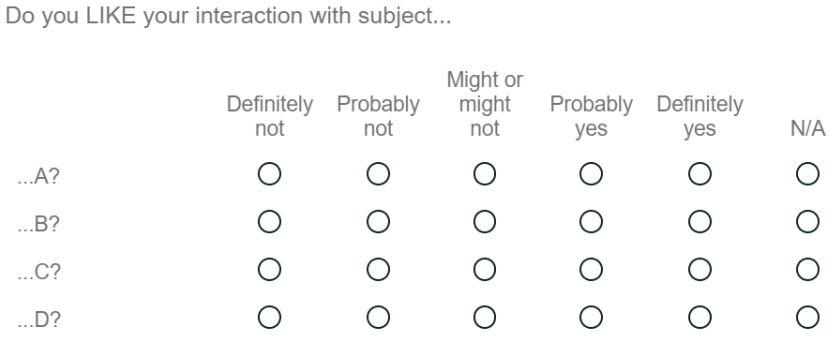


***Figure 45.*** *Liking scale.*

Lastly, participants rated their interaction with the other members of the group using the scale in **Figure 46**, where X represents the reference letter of one of the other two (for a group of 3) or three (for groups of 4) participants of the interaction. For example, if a participant's reference letter was A, then the X in the scale in **Figure 46** was a B, C, or D. The same scale appeared three times for groups of four and two times for groups of three.


***Figure 46****. Items and scale used to determine a value of rapport between dyads.*

## APPENDIX D: AVAILABLE RESOURCES GENERATED BY THIS WORK

**Repository #1**:

URL: https://gitlab.msu.edu/davilasy/sensor-connection-atlas

Description: Repository #1 contains code to connect sensors (Shimmer and BrainBit) to computers and synchronize their signals using LSL.

**Repository #2**:

URL: https://gitlab.msu.edu/davilasy/audio-data-labeling-tool

Description: Repository #2 contains the code used to create the audio data annotation tool to label speech intonations. Here, you will also find the executables to run the tool as a stand-alone application. The code and executable were developed in MATLAB 2019b using their Design application.

**Repository #3**:

URL: https://gitlab.msu.edu/davilasy/human-study-de-identified-data

Description: Repository #3 contains de-identified questionnaire data used to generate **Figure 35** and **Figure 36**.