COMPUTER ENABLED INTERVENTIONS TO COMMUNICATION AND BEHAVIORAL PROBLEMS IN COLLABORATIVE WORK ENVIRONMENTS

A Dissertation in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

by

Ashutosh Shivakumar

B.E., Visveswaraya Technological University, India, 2014

M.S.C.E., Wright State University, 2017

2022

Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

April 12, 2022

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Ashutosh Shivakumar ENTITLED Computer Enabled Interventions to Communication and Behavioral Problems in Collaborative Work Environments BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEFREE OF Doctor of Philosophy.

Yong Pei, Ph.D.
Dissertation Director

Yong Pei, Ph.D.
Director, Computer Science and Engineering Ph.D. Program

Barry Milligan, Ph.D.
Vice Provost for Academic Affairs
Dean of the Graduate School

Committee on
Final Examination

Yong Pei, Ph.D. (Advisor)

Nia S. Peters, Ph.D. (Co-Advisor)

Mateen M. Rizki, Ph.D.

Krishnaprasad Thirunarayan, Ph.D.

Paul J. Hershberger, Ph.D.

Shengrong Cai, Ph.D.

ABSTRACT

Shivakumar, Ashutosh, Ph.D., Department of Computer Science and Engineering, Wright State University, 2022. Computer Enabled Interventions to Communication and Behavioral Problems in Collaborative Work Environments.

Task success in co-located and distributed collaborative work settings is characterized by clear and efficient communication between participating members. Communication issues like 1) Unwanted interruptions and 2) Delayed feedback in collaborative work based distributed scenarios have the potential to impede task coordination and significantly decrease the probability of accomplishing task objective. Research shows that 1) Interrupting tasks at random moments can cause users to take up to 30% longer to resume tasks, commit up to twice the errors, and experience up to twice the negative effect than when interrupted at boundaries 2) Skill retention in collaborative learning tasks improves with immediate feedback dissemination.

To address the negative impact of these communication issues, this dissertation presents two multi-user, multi-tasking collaborative work scenarios and illustrates respective real-time fully functional computer supported cooperative work (CSCW) based prototypes. ACE-IMS leverages lexical affirmation cues which are indicative of task boundaries to intelligently identify “the right time to interrupt” and ReadMI assesses Motivational Interviewing (MI) based clinician-client dialogue in collaborative learning

environment to identify speaker intents like open-ended questions, close-ended questions, reflective statements and scale enquiring statements and provide quantitative feedback to assist the facilitator in comprehensive practitioner skill assessment. To implement these functionalities both systems leverage task-oriented dialogues as datasets and utilize natural language processing with latest developments in ubiquitous technologies like mobile-cloud computing, computational linguistics, and deep learning. This research goes a step further in demonstrating the usability of CSCW based system designs by reporting qualitative and quantitative user feedback data by deploying ReadMI in an actual collaborative learning environment. The participants agree that ReadMI based metrics provide a tangible way to measure practitioner progress and offsets facilitator workload, showing a strong potential to enhance collaborative work experience.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

"Coming together is a beginning. Keeping together is progress. Working together is success". – Henry Ford

My sincere gratitude to the 711th Human Performance Wing at the US Air Force Research Laboratory under contract number FA8650-14-D-6501, Department of Computer Science and Engineering, Wright State University, Dayton, Ohio, Boonshoft School of Medicine, Wright State University, Dayton, Ohio, whose financial support, and generosity was crucial to this research endeavor. Special thanks to the multi-disciplinary Graduate Dissertation Committee consisting of Dr. Mateen M. Rizki, Dr. Krishnaprasad Thirunarayan, Dr. Shengrong Cai, Dr. Paul J. Hershberger, Dr. Nia S. Peters, and Dr. Yong Pei for their invaluable feedback and guidance during the conceptualization and development of this dissertation. Many thanks to the team at the Boonshoft School of Medicine consisting of Angie Castle, Dr. Timothy N. Crawford, Dr. Josephine F. Wilson, Dr. Dean Bricker for introducing me to Motivational Interviewing and facilitating mechanisms for data acquisition and evaluation.

For their assistance in data collection, analysis, and system design, I would like to acknowledge and thank my fellow researchers at SmartLab, Department of Computer Science and Engineering, WSU, Dr. Miteshkumar Vasoya, Raveendra Medaramitta and Aishwarya Bositty.

This research experience has been nothing but enriching and productive thanks to the inputs from Dr. Paul J. Hershberger and Dr. Nia S. Peters. Through his infectious enthusiasm, knowledge, eye-opening insights, and patience Dr. Hershberger has been instrumental in guiding me though the nuances of Motivational Interviewing, for which I interviewed him, a lot! and Dr. Peters through her energetic, insightful and passionate discussions on interruption management systems, where we interrupted one another a lot, played an important role in introducing me to the world of interruption science and advising me on the intricacies in writing scientific publications.

Such valuable collaborations with a diverse, intelligent, and multi-faceted team of experts and researchers would have been impossible, if not for the support, encouragement, and guidance from my advisor Dr. Yong Pei, who has played a pivotal role in my growth as a researcher, student and as a person over the past 5 years at Wright State University. As a guide, debater, manager, and teacher Dr. Pei has helped me navigate through innumerable professional and personal tribulations and has celebrated my successes. Thank you.

Last but not the least, I would like to dedicate this research work to my parents Mr. Shivakumar B. H. and Mrs. Parimala G. N. for their unconditional love, support, and sacrifice.

1 INTRODUCTION

Computer Supported Cooperative Work (CSCW) is a design-based multi-disciplinary field of research that endeavors to understand the characteristics of cooperative work and utilize these insights to design computing systems that support groupwork among participants. Typically, participants engaged in cooperative work belong to diverse professional backgrounds, are situated in different work settings, and have different propensities and perspectives. However, they share a common focus in the mutual dependency towards the task outcome. Incorporating system design features (hardware and software) into computing systems to account for mutual dependency is one of the core aspects of CSCW [1].

Formalized in 1984, CSCW is an interdisciplinary field of study whose problem space demands a confluence of expertise from behavioral and social psychology, sociology, computer science and engineering. A perfect storm of technological advancements and business practices have contributed to the popularity of CSCW based focus in system design. Invention of the internet and advancements in system on chip (SoC) design synergized the implementation of telecommunication and computing into mobile technologies. The ever-expanding nature of modern business enterprises both local and global that require integration of various human expertise have mandated global businesses to build and incorporate efficient collaborative tools that transcend geographical barriers. Furthermore, the emergence of global pandemics like COVID-19 and the consequent

social-distancing measures have created a need to maintain face-to-face communication for social interactions both for business and casual settings, creating a market teleconferencing tools.

1.1 Nature of Cooperative Work Environments

According to Schmidt and Bannon [1], participants engaged in cooperative work are mutually dependent on its outcome and are encouraged to cooperate for its success. Cooperative work environments can be distributed physically in time and space and cooperative tasks require participants to combine expertise, perspectives, and skillsets. Hence cooperative work environments are a distributed set of semi-autonomous decision-making agents with respective unique localized goals, contingencies and capabilities. Consequently, distributed nature of work should be well articulated or described so that the combination of efforts is directed towards the fulfillment of the mutual dependency characterizing a cooperative task. The set of activities that manage and coordinate the distributed nature of cooperative work is referred to as "Articulation of Work" and participants employ appropriate means of communication to articulate their localized activities. For example, a muti-national business organization may employ technologies like email, file-sharing, or video-conferencing technologies to coordinate and articulate the activities of teams distributed over large geographical distances.

1.2 Real-time Communication in Cooperative Work Environments

In high risk and time-sensitive cooperative work settings characterized by distractions and barriers for communication, clear and unambiguous voice-based communication with focused messaging play a crucial role in information transfer. These work environments range from entertainment focused team-based sporting activities like soccer, American football including motorsports, to time sensitive and life supporting work environments like healthcare and search and rescue operations. In motorsports, the driver in the racing car is in constant two-way radio communication with a controller situated in the garage called race-engineer. During the race, the driver and the race-engineer continuously exchange information concerning vehicular parameters like engine temperature, fuel level and extra-vehicular parameters like race-track conditions like foreign debris, distance to the nearest competitor and pit-stop strategy to change degraded tires at the appropriate lap [2]. In healthcare settings, a team of surgeons and assisting medical personnel use clear and focused messaging to confirm surgical checklists and instrument names in perioperative environment. In search and rescue operations and special operations tactical communication, personnel engage in simultaneous data and voice transfer to communicate with a central coordinator and/or amongst themselves for enhanced situation awareness in a rapidly changing environments that may interfere with mission success[3].

1.3 Communication Breakdown in Cooperative Work Environments

Breakdown of communication while performing such highly critical tasks impede mechanisms of coordination thereby minimizing the chances of accomplishing task objectives [1]. The reasons for the breakdown may vary from environmental anomalies, malfunctioning communication devices or miscommunications pertaining to uncoordinated communication, namely, delayed or incorrect feedback from communication partners and disruptive interruptions [4][5]. Negative consequences of task breakdown in these cooperative work environments characterized by mutual dependency and requiring real-time communication are seldom localized and have the potential to spread to connected subtasks jeopardizing favored task outcomes. A delayed reply from the race-engineer to a driver-reported "engine issue" could contribute to vehicular mal-function eluding a race-win or contribute to a fatal driver injury. Interruptions in perioperative environments can have deleterious consequences on patient care and malfunctioning radio-transmitter/receivers could deprive search and rescue teams of crucial information pertaining to obstacles in time sensitive task environment.

Although aforementioned examples underrepresent commonly observed communication tasks, they do highlight unfavorable effects of miscommunication and the importance of unambiguous coordinating mechanisms. Interrupting tasks at random moments can cause users to take up to 30% longer to resume tasks, commit up to twice the errors, and experience up to twice the negative effect than when interrupted at boundaries

[6][7][8]. In classroom discourse it is shown that time-bound and appropriate feedback from teachers improve pupils' ability to construct knowledge and improve classroom interaction rather than unidirectional questioning [9].

Participating interlocutors in these communication tasks use language functionally to co-construct meaning and situational awareness in the form of dialogues which are multi-dimensional in nature composed of literal and intended meaning [10]. Discourse markers in these conversations can be exploited to identify evidence of dialogic coordination and intent. In terms of usability, individuals can be distributed or collocated, geographically, presenting a need to improve the corresponding social experience of face-to-face communication.

The aim of this dissertation is to present two computer supported cooperative environments where fully functional computing systems are designed to facilitate cooperative work by addressing the communication problems of disruptive interruptions and delayed incomplete feedback. "Affirmation Cue based Interruption Management System" (ACE-IMS) monitors task-oriented dialogue in two multi-user multi-tasking cooperative work environments and disseminates intelligent real-time interruptions intended to ensure minimum cognitive overload to the dialogue participants. "Real-time Assessment of Dialogue in Motivational Interviewing" or ReadMI is a computing system that facilitates Motivational Interviewing (MI) based collaborative learning environment involving a dialogic interaction between MI instructor and trainee practitioners. By

mitigating the delay and incompleteness of MI based feedback, which is typically computed and compiled manually by hand, ReadMI ensures a productive and interactive learning experience.

Research efforts in this direction have predominantly focused on developing individual functionalities with limited efforts invested towards functional prototyping. The solutions presented in this dissertation leverage the latest advancements in ubiquitous technologies such as mobile-cloud computing and artificial intelligence that confluence into functional protypes. This research work goes a step further in answering "whether the functional prototypes support cooperative work?" and "How do the users perceive the function of the developed prototypes?"

1.4 Dissertation Organization

This research document is organized in accordance with the fundamental design methodology adopted to conceptualize typical CSCW based solutions. In Chapter 1, the reader is introduced to CSCW systems and the importance of communication as a coordination requirement. Subsequent subsections describe the nature of cooperative work interactions characterizing the respective cooperative work environments for ACE-IMS and ReadMI. Chapter 2 is designed to inform the reader with relevant background on interruptions and interruption management systems, Motivational Interviewing based training as a cooperative learning environment, and parallel solution space addressing these issues. In Chapters 3 and 4, the reader is presented with a comprehensive description of the

design, development, and evaluation results of ACE-IMS and ReadMI respectively. Chapter 5 summarizes and discusses the implications of adopting CSCW based design framework to communication. Chapter 6 presents conclusions and Chapter 7 discusses the current limitations and future work.

1.5 Background

1.5.1 Disruptiveness of Interruptions in Multi-user Collaborative Environments

Interruption science explores the disruptiveness of interruptions on human performance. This research area is motivated by the reality that as users increasingly multitask with proactive systems, their tasks are being interrupted more often. An interruption within these interactions can be defined as an unanticipated request for task switching from a person, an object, or an event while multitasking [11]. The disruptiveness of interruptions has been widely studied, such as, the implications of interruptions on productivity [12][13][14] and affective state [6][15]. For instance, previous studies have illustrated that interrupting users engaged in tasks has a considerable negative impact on task completion time [16][14][13][17][18][19]. Other studies illustrate the implications of ill-timed interruptions particularly in medical settings from an inter – clinician communications perspective[20] and from work design and systems or processes' perspective in hospitals [21] or the cost of interruptions [14][15] [22][23] which some have suggested are attributed to differences in workload at the point of interruption [7].

A typical example of a multi-user multi-tasking environment where the disruptiveness of interruptions can have a deleterious effect is *emergency management* involving communication operators and first responders. Here the operator at the command center communicates with the emergency personnel on the ground and aligns his/her knowledge of the location of the hazard. The command center operator has two different tasks to perform simultaneously: 1.) Primary task: Location alignment with first responder concerning the emergency; 2.) Secondary task: Monitor the system for other emergencies or system maintenance alerts. Other multi-user, multitasking distributed interactions involve communication between air traffic controllers and pilots, unmanned aerial systems (UAV operators) and military ground troops, and technical support agents and customers. Frequently interrupting the users in these scenarios with orthogonal tasks [5] or interruption task can lead to cognitive overload with potentially devastating consequences where the participant may be distracted and overwhelmed to complete their primary task effectively and efficiently [24].

1.5.2 Delayed Feedback in Collaborative Learning Environments

Motivational Interviewing (MI) is a collaborative, goal-oriented, dialogue-based communication style, involving an MI practitioner and client, designed to strengthen the client's personal motivation and commitment to a specific goal or behavior change. An MI practitioner can be a caregiver like a doctor, nurse practitioner, a sports coach, teacher, or any individual who seeks to guide their client (a patient, child, athlete) towards behavior

change conversationally in non-confrontational ways through shared decision making. Since its invention in 1983 by Stephen Rollnick and William R. Miller as brief intervention strategy for treating alcohol addiction, the application of MI has diversified into the fields of education, health, social sciences, criminal justice, and more recently, sports [25] [26].

The collaborative and cooperative nature of MI style of communication stems from the fact that practitioners engage clients as active partners in conversations. Instead of directing or imposing knowledge on a passive client, the practitioner evokes the client's innate positive intentions towards behavior change. The practitioner is guided by the fundamental understanding that it is the client who has to invest in the recovery plan and honors client autonomy without resorting to coercion [27] [26]. Miller and Rollnick describe this collaborative, evocative and autonomy honoring aspect of MI as the Relational component/skillset or "Spirit" of MI [28]. The spirit of MI accounts for its qualitative aspects, essentially, "How?" an MI session must be conducted. By creating an environment of trust and empathy the practitioner then proceeds to guide the client towards their goal through skilled use of interaction techniques, hence the "goal-orientation" of MI.

This is referred to as the "Technical" component/skillset or "what?" must be said" part of the MI communication style [29] characterized by the usage of interaction techniques like asking evocative open-ended questions (example: "What does change look like for you?"), minimizing close-ended questions (example: "Did you take your medication?"), performing simple reflections (example: "It sounds like you want to make time exercise....") and enquiring for "readiness for change" with utterances like "On a scale

of 1 to 10 with 1 being least likely to change. Where do you stand?". The frequency of usage of a specific interaction technique or combination of techniques provides tangible MI performance metrics for post session analysis. Therefore, the relational component of MI accounts for the qualitative aspect of MI, the technical component provides a quantitative dimension [29].

The success of a typical motivational interviewing session is measured by the practitioner's ability to elicit "change talk" from the client. The definition of change talk as per Miller and Rollnick is "any self-expressed language that is an argument for change" [27], which, in this case is the client's self-expressed desire for change articulated through MI dialogue. To elicit change – talk, the practitioner must listen well, encourage longer client speaking time and use a combination of relational and technical components to build a "working alliance with the client". Unfortunately, practitioners, for example, physicians in healthcare settings are accustomed to recommending or prescribing lifestyle change strategies without taking into consideration the patient's motivation to change. This direct approach contributes to poor patient outcomes, thus, requiring a change in communication style that is non-intuitive to the practitioner. During MI training sessions, practitioners learn to change their natural inclination to lead the discussion and, instead, allow the conversation to be client centered. This means the practitioner must talk less, listen more, and ask open-ended questions – critical skills in the MI approach.

1.5.3 MI and Collaborative Learning Setting

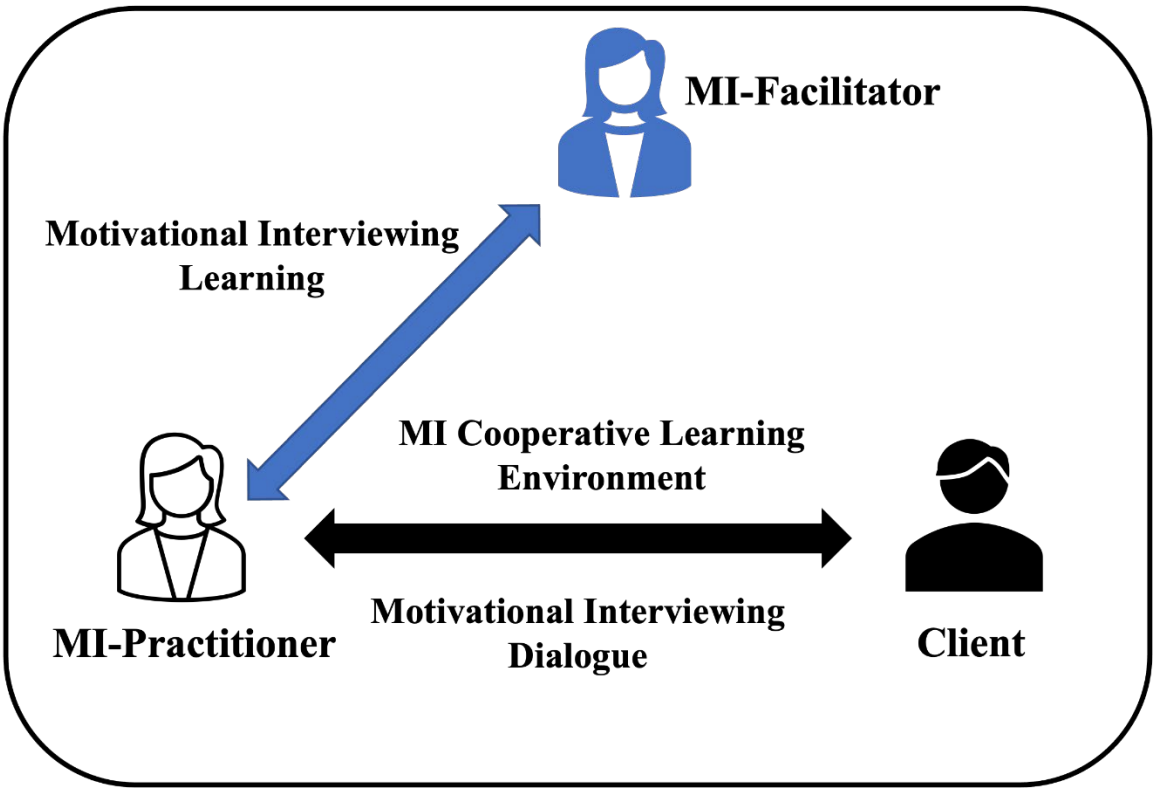


Figure 1.1: Motivational Interviewing Collaborative Learning Environment

For skillful delivery of MI sessions, practitioners are trained by an MI based communication style expert, called a facilitator. A typical MI training session consists of an MI trainee practitioner, MI facilitator or trainer and a client – a fellow MI trainee who plays the role of a client. An audio/video recording device is used for recording purposes. In terms of workflow, the practitioner engages in MI based conversation with the client via role plays while the facilitator observes and provides feedback based on the practitioner's performance. The facilitator must modify his/her instruction in a way that accurately fits

the learning demands of the practitioner. Figure 1.1 represents a typical motivational learning environment. This learning environment differs acutely from an instruction-led classroom where the emphasis is on the instructor and one way, one-size-fits-for-all instruction [30]. In an MI training environment, skill development is gained through hands-on experience via roleplays with fellow trainees and the MI trainer facilitates the learning experience with instruction and feedback on an "as needed" basis. Therefore, a typical MI training sessions can be framed as a collaborative learning environment [31].

1.6 Challenges in Timely Quantitative Feedback Dissemination

Presently, the facilitator disseminated MI based feedback is often limited to qualitative elements of the session, which is "How?" the practitioner spoke in the session – the relational component. Immediate feedback on technical elements remains a major challenge for the facilitator due to its quantitative nature translating into computational complexity and attention requirements for calculation. The session recordings are manually transcribed after role-play sessions [32][33], utterance-by-utterance by Motivational Interviewing Treatment Integrity (MITI)[47] -trained raters who also assign labels like "OPEN-ENDED QUESTION", "CLOSE-ENDED QUESTION", "REFLECTION", "SCALE" and "NONE" to each utterance and then compile frequencies (for example: Number of OPEN-ENDED QUESTIONS) and ratios (for example: Ratio between REFLECTIONS to CLOSE-ENDED questions). To put it into perspective, by observation, a typical MI session spans approximately 10 minutes with an average of 90

utterances. A classroom lecture session teaching MI lasts for at least 60 minutes. Manually transcribing utterances and generating quantitative metrics based feedback for 6 MI sessions in a classroom is a time-consuming endeavor, thereby delaying the delivery of instructional feedback to the practitioner [34]. Research studies have shown that skill retention and development improves when feedback is immediate [35] and change talk improves in clients with a synergized implementation of both relational and technical elements [29]. Hence feedback characterized by delay and devoid of technical elements eludes the practitioner and the facilitator a means for tangible assessment of MI delivery and maintain an awareness of task quality, contributing to the depletion and breakdown of the learning task within MI education.

Two possible solutions to improve feedback delivery are 1) Increase labor, i.e., hire more transcribers and raters. 2) Leverage rapid developments in technology, particularly ubiquitous technologies like the internet, mobile technology, artificial intelligence to automate the feedback process. The former solution is financially cost-intensive - the hourly cost of a human transcriptionist is 15 USD to 30 USD [36] and an automatic speech recognition algorithm-based service costs 2.16 USD [37]. The latter solution shows potential due to the convenience in terms of computational processing power, capacity to operate at scale, rate of adoption, and financial cost due to manufacturing in scale. This technology assisted delivery of MI is more formally abbreviated in the literature as TAMI. Since MI training environment is a collaborative learning environment, technology assisted

delivery of MI [38] in this research work, is referred to as “Computer Supported Collaborative Learning (CSCL) in MI” [39].

1.7 Proposed Solution

The first of the two systems presented are “Affirmation Cue-based Interruption Management System” or (ACE-IMS). ACE-IMS is a real-time interruption management system that monitors task-oriented dialogue in multi-user multi-tasking cooperative work environment and disseminates intelligent real-time interruptions optimized to ensure minimum cognitive overload to the dialogue participants. Across domains it is shown that affirmation cues like “got it”, “yeah”, “gotcha” signal task transition to another topic or dialogue-turn in task-oriented dialogues [40] [5]. ACE-IMS is designed to identify such points of interruptions in task-oriented dialogues to create opportunities for least disruptive interruptions. Performance evaluation of ACE-IMS is accomplished by comparing it with the baseline real-time prosody-only system of C-CIMS.

Secondly, to address the problem of delayed feedback, “Real-time Assessment of Dialogue in Motivational Interviewing” or ReadMI is presented. ReadMI is designed to address the prevailing limitations in MI training workflow that lacks a mechanism to deliver timely and accurate feedback based on quantitative MI technical components, within reasonable operating complexity and time, and sustainable acquisition and operation cost. By combining the calculated quantitative metrics with manually determined qualitative metrics, the MI expert instructor provides a comprehensive and time-bound

feedback, thereby enriching the MI collaborative learning experience. The absence of such tools results in MI being under-utilized to train workforce especially in work environments where behavior modification is necessary to facilitate lifestyle changes, for example, behavior modification of patients suffering from chronic conditions [41].

To assess task-oriented dialogues that emanate from communication based cooperative work environments, both ACE-IMS and ReadMI advance the state of art developments in "Behavioral Signal Processing" based technologies like Automatic Speech Recognition. By leveraging the highly accurate (WER = 5.6%) real-time ASR to transcribe participant's spoken utterances of an MI role-play session, the systems automatically translate biological signals like voice to text string data in real-time. We emphasize "real-time" to indicate the fact that it takes less than 300 ms for speech to text conversion, which appears as "instantaneous" operation to human perception.

The nature of the work environments in both ACE-IMS and ReadMI can be geographically remote and/or co-located. But they share a common characteristic, in that they are "synchronous", where interactions between participants is in real-time. Within the CSCW literature, the CSCW matrix [42] formally represents the classification of such computing supporting groupwork into 4 quadrants in accordance with interaction time frame (synchronous versus asynchronous) and geography (co-located versus distributed). Figure 1.2 shows the positioning of the two examples of the computing systems ACE-IMS and ReadMI in as per CSCW matrix. Further commonalities between the two

examples extend to the number of participants – multi-user and medium of

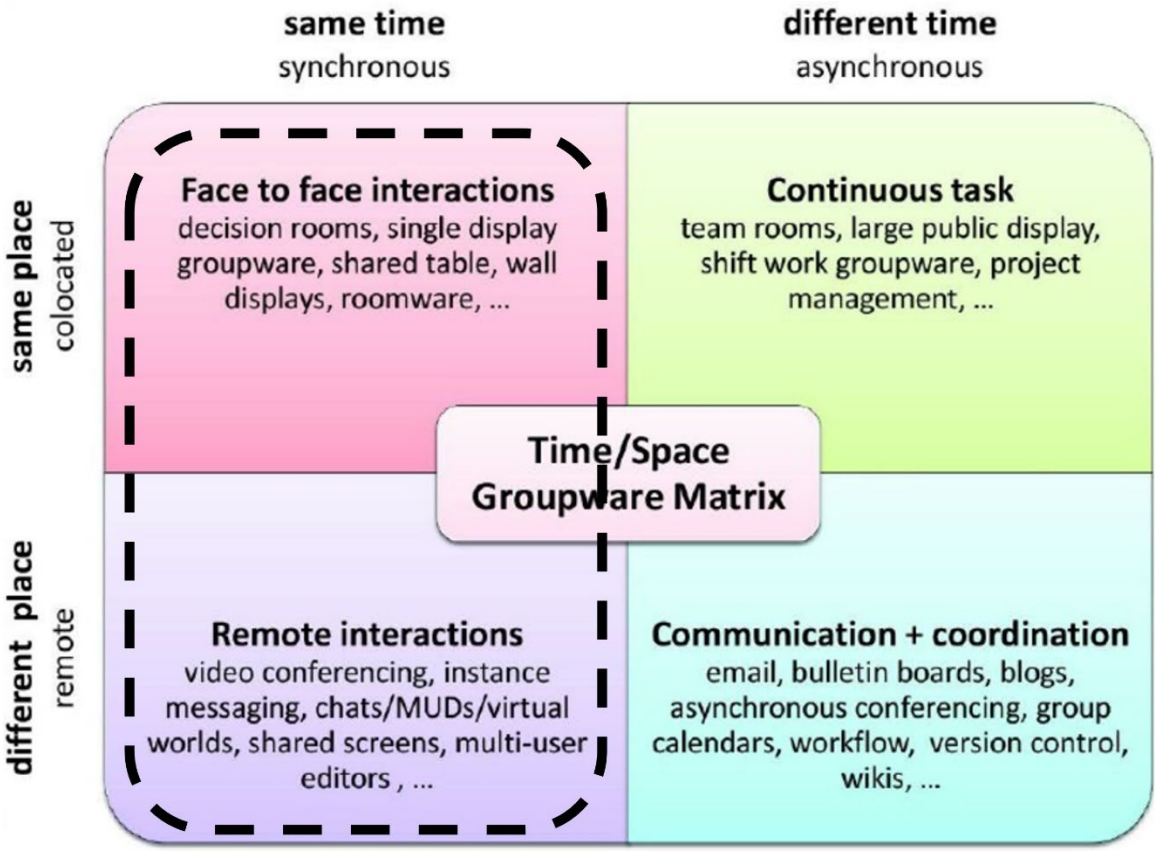


Figure 1.2: CSCW Matrix – ReadMI Usage Context in MI Collaborative Learning Environment [42]

1.8 Contributions

The intended contributions of this research work are:

1. Explicit exploration of communication problems of interruptions and delayed feedback dissemination within the framework of computer system design for collaborative work environments.
2. Presentation of the design and implementation details of fully functional prototypes of ACE-IMS and ReadMI as solutions to coordination issues in collaborative work due to verbal miscommunication.
3. Demonstration of the effectiveness of discourse markers as key features that characterize the nature of utterances in task – oriented dialogues.
4. Presentation of qualitative and quantitative results of user feedback for ReadMI in real-life cooperative work environment.

2 LITERATURE REVIEW

2.1 Interruption Management and Task Structure

To alleviate the consequences of disruptions, manipulating the timing of interruptions [12][13][14][7] using system-mediated interruptions [43] within multi-task environments [44] has been proposed and studied for different timing strategies. Interruption times explored include immediate delivery [14][45][46], random timing [12][14] [17][7][46][47], and delivery at task boundaries [14][6][7][48] as examples. The benefit of appropriately timed interruptions, particularly the task-boundary based approach, is evident in works such as [49].

One area of research that aims to alleviate the negative effects of these interruptions via system-mediated interruptions is the Interruption Management Systems (IMS) literature. The focus of this area is to leverage the available modalities of an interaction (i.e., visual, meta-data and speech) within domains of varying participants, tasks, and objectives to disseminate information at the least disruptive times. Methods have been proposed to determine the appropriate interruption timings via task structure inference, and a subset of this literature recommends point of interruptibility at boundaries within task execution. A task boundary is a time instance between two moments of task execution. Within single-user, multitasking interactions, task boundary modeling has been used to indicate appropriate points of interruptibility via system-state [14][6][7][48] and physiological data [50].

2.2 Collaborative Communication Interruption Management System

Until recently the exploration of task boundary modeling to infer interruption decisions has been limited to single-user, multitasking interactions. The Collaborative Communication Interruption Management or C-CIMS proposed by [5] extends the Interruption Management and Task Structure literature and aims to use task boundary modeling for interruption inference within distributed multi-user, multitasking interactions, as illustrated in Figure 2.1. It has laid out the foundation to extend the use of task boundary modeling for interruption inference within distributed (users can reside in different geographical locations at the same time), multi-user, multitasking domain.

C-CIMS [5] leverages speech information within the distributed multi-user, multitasking interactions and aims to infer a task boundary as candidate points of interruption. C-CIMS explored this problem using offline (non-real-time) and online (real-time) machine learning techniques which train and test their proposed model on the entire available data collection. The offline implementation of C-CIMS explored lexical features ("what is said?") and appears to offer increased performance in detecting task boundaries to infer interruption timings when compared to a prosodic-only implementation that leverages only prosodic information ("How it is said?"), such as: energy and pitch information. The real-time implementation of C-CIMS established a baseline performance for real-time IMS system in this domain. The limitation of real-time C-CIMS to prosody-

only model was reported due to prohibitive latency issues in processing lexical information in real-time [5].

Therefore, there is a need to explore a lexical-based interruption management system that can support real-time interactions. Additionally, within the offline models, the author in [5] inferred that affirmation cues are salient lexical predictors of a task boundary, which merits comprehensive investigation.

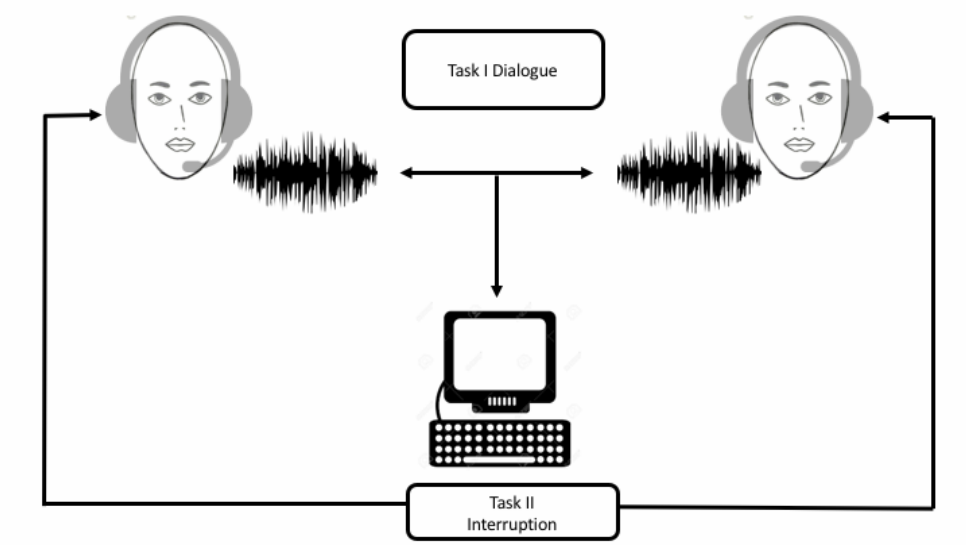


Figure 2.1 : Distributed Multi-user, Multi-tasking Interaction [5]

2.3 Collaborative Learning Environments and Motivational Interviewing

In this section, to understand the characteristics of "feedback" in Motivational Interviewing training as a collaborative learning environment and subsequently derive system design features from this collaborative introduction 1.) A brief introduction to the nature of collaborative learning environment is provided 2.) Behavioral coding systems used to generate quantitative metrics-based MI feedback is summarized 3.) Current efforts towards automation of Behavioral Coding systems are provided.

2.3.1 Collaborative Learning Environments

According to [51], the field of CSCW addresses how computer systems can support collaborative activities and facilitate their coordination. Computer systems aim to reduce task complexity and improve efficiency by offering better communication facilities with improved monitoring and awareness. Irrespective of formal or informal learning setting, educational institutions act as work environments where technology, social or cultural interactions moderate work [52]. Here the interaction between CSCW and the work of education holds promise for improving workflows in learning, a central activity in education. Learning can take the form of a collaborative activity where the collaborating individuals are groups of students and facilitators where participants elicit information from one another, monitor one another's activities and form collective knowledge by sharing experiences [53]. Due to this interaction-based knowledge creation among learners, collaborative learning is a learner centered approach where knowledge is considered a

social concept that is created, facilitated through peer interaction, cooperation, and evaluation [54]. Here the role of the instructor or teacher or tutor is as a mediator facilitating these interactions and fosters collaboration and exchange among members through techniques or strategies that include discussions, role play, peer review and jigsaw [55].

Research in CSCW systems that support group work in education settings has given rise to CSCL (Computer Supported Collaborative Learning). Both CSCW and CSCL support collaborative nature of work supported by computing systems, except CSCL research is grounded more specifically towards technological support of learning, pedagogy that takes place via social interactions [39].

2.3.2 Observational Coding and Motivational Interviewing Metrics

With Motivational Interviewing training environment conforming to the characteristics of a CSCL environment, it is imperative to understand the features of MI that can be leveraged to automate the feedback generation process. One such feature set is a byproduct of the "Observational coding system". The primary purpose of an observational coding system is to evaluate MI integrity/fidelity in training sessions. Pioneered by Carl Rogers [56] , observational coding task, involves an MI rater listening to audio tapes of MI sessions and manually - assigning categorical labels for clinician and client utterances. For example: "OPEN" for open ended questions like "How are you doing?" or "CLOSE" for close ended questions or "Yes/No" reply eliciting questions like "Are you addicted to cocaine?". Some examples of MI integrity evaluation observational

coding tools are Motivational Interviewing Skills Code (MISC) [57], Yale Adherence and Competence Scale(YACS) [58], or Motivational Interviewing Treatment Integrity (MITI)[47]. An utterance in the context of this research work is defined as a complete thought expressed by a clinician or client. For example, “How are you?” is one utterance.

As ReadMI was developed within medical settings, for the time-constrained medical students and aimed at minimizing information overload [59] of the facilitators, we have adopted a streamlined version of the observational coding system along with prosodic metrics that capture the essence of motivational interviewing training, which is, measuring the capability of practitioners to encourage “change talk” in clients. The proposed modified observational coding system includes 5 categories: 1) Simple Reflective statements, 2) Open questions, 3) Close questions, 4) Scale Sentences (i.e., the use of change ruler), and 5) NONE (i.e., statement).

Table 2.1: Examples of MI related Sentences

ReadMI Behavioral Codes	Utterance Examples
OPEN QUESTION	“How has drinking affected your work performance?”
CLOSE QUESTION	“Is drinking affecting your work performance?”
REFLECTIVE QUESTION	“It sounds like drinking is affecting you at work.”

SCALE SENTENCE	"From a scale of 1 to 10, with 1 being least likely and 10, most likely, how ready are you to give up drinking?"

Additionally, primitive prosodic MI metrics like: 1) Doctor Speaking time, and 2) Patient speaking time are included. The numerical counts of the mentioned observational codes, and prosodic elements of practitioner-client speaking time constitute MI metrics based real-time feedback, in addition to the full list of utterances for each category.

2.3.3 ReadMI Observational Coding System

Table 2.1 highlights the observational codes used in ReadMI. According to [27], linguistic devices like reflections, open questions and utterances that assess readiness of change like scale-ruler sentences ("on a scale of one to ten, . . . "), help practitioners communicate empathy and assume a nonjudgmental stance in practitioner-client dialogues.

- Simple reflections are statements where the practitioner repeats both the positive and/or negative implications of the client's addiction with an empathic tone. Simple reflections rephrase what the patient said and they add little to what was said, while complex reflections are used to inject some meaning or emphasis on what the patient has said. In this work, we will focus on simple reflections and complex reflection will be considered in future work. Although simple reflections essentially repeat information provided by the client and do not add new information, they

encourage the client to self-analyze and resolve internal conflicts or ambivalence to behavioral change. Simple reflections are most likely but not always indicated by phrases like “sounds like” and “looks like”. For example, sentences like “It sounds like your late-night drinking is affecting your sleep”.

- “Open questions” encourage clients to provide a variety of answers. These questions allow practitioners to understand client’s perspective on present behavior, encourage self-exploration or seek additional details to aid diagnosis and/or behavioral change.
- Close questions are “Yes/No” answers seeking questions. According to MI experts, they are the least preferred utterances as clients may not convey additional information by just answering yes/no, which is contrary to the client-centric approach of MI that encourages client to speak. Hence clinical experts discourage MI trainees from such usage.
- Scale sentences are used to measure client’s readiness for change. These sentences present a numerical scale for the client to analyze and state their readiness. When the client chooses a score – typically between 1 and 10 - it provides a numerical context for the clinician to gauge client’s interest to change or sustain the status quo.
- Finally, a default NONE class label has been added to label any remaining utterances that do not belong to any of the above stated labels, which mainly are the educational or directive statements.

2.3.4 ReadMI Prosodic Metrics

The premise of MI is to encourage the patient to speak more and allow them to do the "work" of behavioral change. Practitioners tend to falsely assume that they allow the patient to speak more, it [60]is they who have the higher percentage of speaking time. Hence, a practitioner and client speaking time measurement in percentage is included as a metric. This is an important component of the feedback as MI is primarily a client centered approach and speaking time is a direct indicator.

2.3.5 Behavioral Signal Processing

According to [57], usage of computational methods for signal analysis and decision-making falls within the purview of "Behavioral Signal Processing (BSP). A BSP task involves detection of overt (voice, facial expressions, body posture) and covert (heart rate, electrodermal response, brain activity) signals, manifestations of human behavior. Speech signals are multimodal, complex and context specific, composed of Lexical and prosodic features. Lexical features indicate "What is said?", referring actual words used, while prosodic features refer to "How it is said?", referring pitch and inflection. Speech signal processing and information extraction provides a window into human behavioral expression. In motivational interviewing, reflective sentences and open-ended questions (both implicit and explicit) indicate empathetic tone of the practitioner contributing to increased client change-talk [26]. This research work focuses on BSP tasks corresponding to speech to text conversion of practitioner-client speech. As a BSP agent, ReadMI utilizes

Automatic Speech Recognition algorithms that convert human speech signals to text, paving the way for additional analysis of lexical content in the transcriptions. Its workflow involves capturing and transcribing practitioner-client dialogue and then coding the practitioner utterance as shown in Table 2.1. Hence, ReadMI aims to automate the process of behavioral coding and dissemination of feedback, thereby assisting training by dialogue flow tracking.

2.3.6 Automatic Behavioral Coding

The common technological enablers of BSP are Signal processing and machine learning. Previous literature [34], [61] has shown that, a combination of natural language processing machine learning models can be used to leverage the syntactic features (sentence structure, word or phrase detection and counting, topic modeling) and semantic features, dialogue acts [62], [63]in clinician-client utterances to facilitate automated observational coding with commendable accuracy and inter-rater agreement. These studies show that Natural Language Processing (NLP)-based models can emulate human rater expertise and automate the behavioral coding of MI dialogues. However, the major drawback in these studies is the usage of manually generated transcriptions of clinician-client dialogues as datasets for NLP models, thus they stop short of providing real-time transcriptions during MI training sessions. As one may recall, the major bottleneck in MI training pertains to the prohibitive costs of manual transcriptions of practitioner-client dialogues, thereby constraining the process in terms of latency, scale, time, and cost. To

alleviate this bottleneck, the sub-processes of both speech to text transcription and behavioral coding should be automated and completed in real-time.

Efforts have been underway in this direction. Researchers in [64], demonstrate the possibility of automating the process of behavioral coding. They have developed an ASR with automatic speaker diarization [65] capability which automatically transcribes psychotherapy session recordings and these transcriptions are used to identify ratings of "higher" or "lower" empathy for each session. They then proceed to compare the empathy detection accuracy between ASR-generated transcriptions (82.0%), ASR-generated transcriptions with human labelled speaker labels (80.5%) and human transcriptions (85.0%). The accuracy values appear to be closer, even though the utterances are evaluated for only one behavioral code "empathy". Detecting observational codes at a higher resolution, such as for MI training, however, will involve utterance-based detection, which requires ASR to perform at a higher accuracy to capture the lexical features. On that aspect the researchers in [64] are severely limited by in-house developed ASR accuracy (WER = 43.9 %), where WER refers to Word Error Rate [66]. The researchers in [64] demonstrate speech to text capabilities of the ASR on previously recorded psychotherapy sessions, but not live MI training sessions. Nevertheless, the closeness between ASR generated transcriptions and human transcriptions in [64] presents an encouraging sign for the viability of automating speech to text transcription in Motivational Interviewing sessions. In summary, by combining 1.) The knowledge of unique discourse markers in the form of

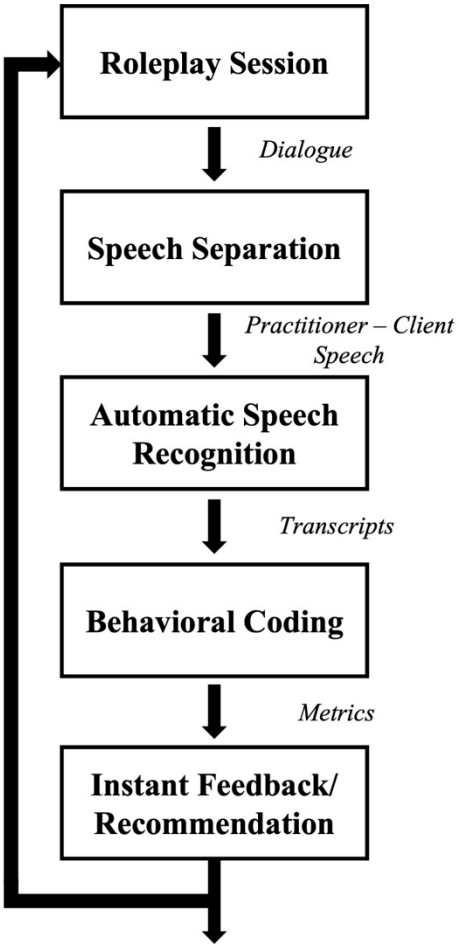


Figure 2.2: ReadMI Workflow

lexical features that characterize intents in utterances like "Task-boundary/Non-Task-boundary" or behavioral codes like "OPEN/CLOSE", "REFLECTIONS", "SCALE", "NONE" and 2.) Background literature of interruption management systems, behavioral coding and signal processing informing the characteristics of the cooperative communication tasks and computer systems-based solutions to support these tasks in the form of ACE-IMS and ReadMI as presented in the following chapters.

3 AFFIRMATION CUE BASED INTERRUPTION MANAGEMENT SYSTEM

The primary focus of this section is to explore the usage of affirmation cues to identify task boundaries in real-time for intelligent interruption dissemination in multi-user multi-tasking interactions. Task–oriented dialogues that simulate multi–user, multi–tasking dialogues, where participants communicate with one another verbally to accomplish a task at hand, are used as dialogue datasets. To accomplish this strategy, the following steps are used:

1. Assess and understand datasets: description of task boundary annotated task–oriented dialogue datasets are provided.
2. Analyze and gain insights of affirmation cues preceding task boundaries.
3. Provide a system design and prototype of an Affirmation Cues based Interruption Management System (ACE-IMS) to demonstrate real-time identification of task boundary.
4. Develop a progressive rule–set design of affirmation cues which forms the heart of the ACE-IMS.

Below subsections describe each of these steps in detail.

3.1 Dataset

The proposed ACE-IMS is trained and tested using two human–human task datasets from the research work in [5]: UMT and Tangram. The two datasets represent the domain of interest: distributed multi-user, multitasking task-oriented dialogues.

In these tasks, two distributed human participants communicate using push-to-talk to accomplish a common task and the machine disseminates information related to an orthogonal task or interruption task. A brief description of the datasets is as follows:

1. **Uncertainty Map Task (UMT):** UMT is a distributed multi-user collaborative communication task where the two participants align their knowledge to identify a target house while looking at the house from a different perspective: birds eye/aerial or forward facing / street view. In the task, two participants are presented with one of these 4 target house views: a.) aerial target vs. street view identification, b.) street view target–aerial identification, c.) street view target–street view identification, and d.) aerial/street view target and identification. There is a total of 67 dual-channel audio files (one audio channel per speaker) for the UMT task that are used in our project. Each audio file consists of task conversations corresponding to 10 target identification tasks as illustrated in Figure 3.1.
2. **Tangram:** Tangram task is a distributed multi-user collaborative communication task where participants use a push-to-talk to communicate on a task where they arrange the abstract shapes called Tangrams in corresponding order that is aligned

with each other as illustrated in Figure 3.2. 40 dual-channel audio files from Tangram are used in our project, each consisting of task operator–teammate conversations (one channel per speaker).

Figure 3.1 and Figure 3.2 are interface diagrams of the tasks respectively and more information about the data collection is available [5].

3.1.1 Metadata

The accompanying log file for the audio files in both datasets consists of task begin and end time. Task begin time is defined as the time at which both users were presented with a new set of targets (UMT task) or abstract shapes (Tangram task), and task end time is defined as the time instance when both participants mutually acknowledge the completion of a task with a mouse click on "Done" button [67].

The timing of delivering information related to an orthogonal task or interruption task while participants are performing the primary task (UMT or Tangram) is the focus of this research work. Adding another task, i.e., orthogonal task as specified in [5] would make them multi-user, *multitasking* interactions.

Since the overall objective of this research work is to present a real-time IMS system that leverages affirmation cues as lexical features to infer a task boundary, training and testing datasets are created and explored to identify the influence of affirmation cues on task boundaries.

Figure 3.2 : (A) Aerial Target View from UMT Task (B) Street View Identification from UMT Task [5]

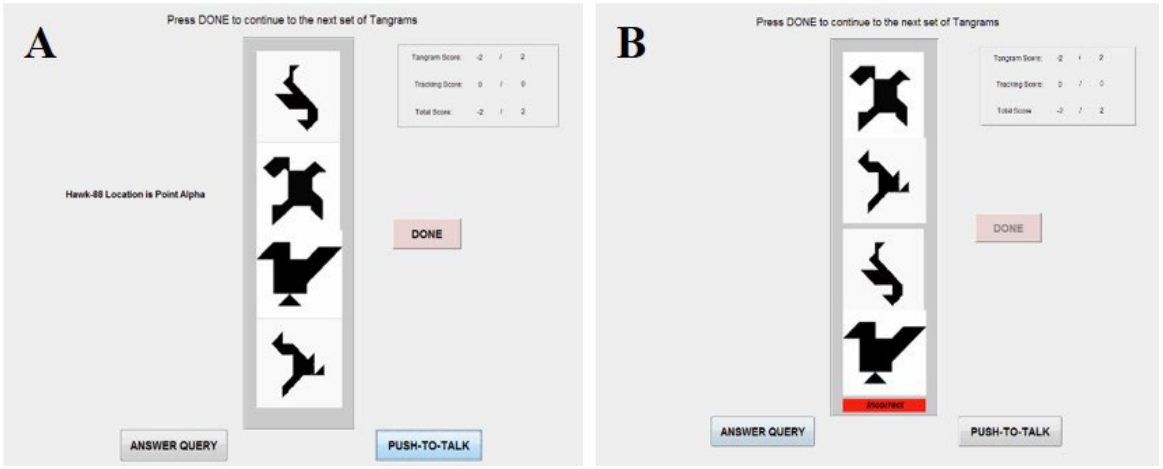


Figure 3.1 : (A)Tangram Task Teammate 1 Interface (B) Tangram Task Teammate 2 Interface [5]

3.1.2 Training and Test datasets

1. **Training Dataset:** A random portion of the UMT dataset 30 audio files (3066 utterances) out of 67, was designated as the training dataset and used to identify the affirmation cues and generate the rules of the classifier.
2. **Testing Datasets:** The remaining 37 UMT audio files (2904 utterances) were added to the original 30 to create a 67 audio files (5970 utterances) test dataset. Additionally, 40 audio files (4554 utterances) of the Tangram dataset were used as an additional test dataset to evaluate the generalizability of the identified affirmation cues as lexical features.

Table 3.1: Training and Test Datasets

Dataset	**Audio Files**	**Utterances**
Training Dataset	30 randomly selected audio files from the UMT dataset (approximately 6 hours)	3066
UMT Test Dataset	67 (approximately 13 hours)	5970
Tangram Test Dataset	40 (approximately 8 hours)	4554

Since multiple audio recordings may come from the same participant, the audio files within each dataset were randomly selected to incorporate more speaking styles for

variety in affirmation cues for both training and testing. Each channel of the dual–channel audio files from the training and test data was passed through automatic speech recognition for speech to text conversion. The resulting text transcripts of the two separate

channels were then interleaved together with the aid of timestamps to create the dialogue text transcripts. For clarification, the timestamps were sorted in chronological order and the corresponding utterances were added to create dialogue text transcripts. Each audio file had a corresponding dialogue transcript file. The utterance that preceded the task boundary timestamp, as provided in the corresponding task log files, was labelled as the task boundary utterance. For a single dataset, like training dataset, all dialogue transcript files corresponding to the audio files in the dataset are interleaved to form a 3066-utterance dataset. Similar operations were performed on UMT test and Tangram test datasets. The task start and end time in the dataset provide structure to a task-oriented dialogue. This gives us an excellent opportunity to delve into lexical affirmation cues preceding task end time.

3.2 Affirmation Cues Preceding Task Boundaries

As indicated by [40][67] humans tend to use affirmation cues such as *like, got it or, yeah* to signal transition to another topic or task and to signal turn-taking. Since the objective of the proposed interruption management system is to predict a task boundary or

a task transition as a candidate interruption point, we expect that detection of an affirmation cue can predict such moments.

To explore the use of affirmation cues and their relationship to task boundary utterances, we examine the existence of lexical features reflecting affirmation cues in the training dataset. The definition of a *task boundary* as presented in [67] is *a timestamp associated with both players clicking a button to indicate they are done with one task and ready to transition another task.* We then define *a task boundary utterance* as the *utterance immediately preceding this task boundary timestamp*.

The nine most frequent affirmation cue phrases present in task boundary utterances within the training dataset were manually identified and recorded. The list is shown in Table 3.2.

Then, the occurrence of the identified affirmation cues in the labelled task boundary utterances for the training dataset and the two test datasets are reported in Table 3.3. Here we use a term called *Coverage* to measure the extent of affirmation cue occurrence within the task boundary utterances, as defined by Equation (1).

$$Coverage = k/N \tag{1}$$

where k is the total number of task boundary utterances with the identified affirmation cues in the corresponding dataset, and N is the total number of task boundary utterances for the corresponding dataset. For the training dataset here, $N = 329$. Figure 3.3 shows the Coverage of the identified affirmation cues (as listed in Table 3.2) among task boundaries in the training dataset and the UMT and Tangram test datasets, respectively. Figure 3.3.A indicates that affirmation cues present in

69.9% of the total task boundary utterances in the training dataset (with total number of task boundary boundary utterances N = 329), the remaining 30.1% can be mapped to other unexplored

Table 3.2 : Affirmation Cues Count for N=329 Task Boundary Utterances Within Training Dataset

Affirmation Cues	**Frequency in Task Boundaries**
got it	180
got you	13
yep	14
gotcha	2
awesome	3
sounds good	4
done	9
great	3

Table 3.3: Total Number of Task Boundaries Corresponding to Training and Test datasets

Dataset	Total Number of Task Boundary Utterances With The Identified Affirmation Cues (k)	Total Number of Task Boundary Utterances(N)
UMT Training	230	329
UMT Test	508	808
Tangram Test	1020	1158

features. Figure 3.3.B shows that when the same identified affirmation cues are applied to the UMT test dataset task boundaries (N = 808), the coverage decreases by 7% to 62.9%. Figure 3.3.C indicates that, for Tangram test dataset, the same identified affirmation cues account for 88.1% of the total task boundaries (N = 1158), which is higher compared to both the UMT-based training dataset (a random selection from UMT as defined in Section 3.1.2) and the UMT test dataset. These results provide us with the following insights:

- The 9 affirmation cues present in Table 3.2 are strong feature candidates for identifying task boundary utterances
- The higher Coverage in the Tangram test dataset (88.1%) when using the affirmative cues obtained from the UMT-based training dataset implies that:

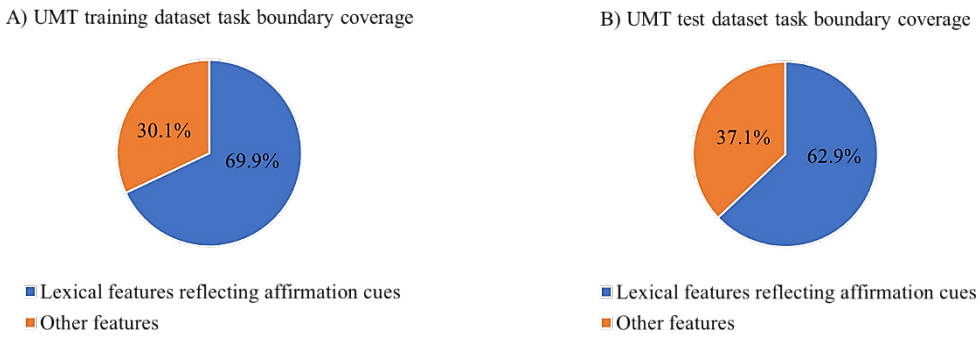
a) the identified lexical features from UMT-based training dataset generalize well and perform robustly across both datasets (UMT and Tangram).

b) the affirmation cues present at a higher rate in the task-boundary utterances of Tangram tasks, may imply that the dialogues of Tangram tasks could be more structured. Future investigation is warranted to identify the causes of such variations among task-oriented dialogues.

- The remaining task boundaries, those without affirmation cue phrases presented (e.g., 30.1% of task boundaries in the training dataset, 37.1% in the UMT test dataset and 11.9% in the Tangram test dataset), could be the focus of future work.

3.2.1 Interference from Backchannel Utterances

Further examination also indicate that the same affirmation cues present in task boundary utterances may also be used as backchannels in a task-oriented dialogue. In the context of



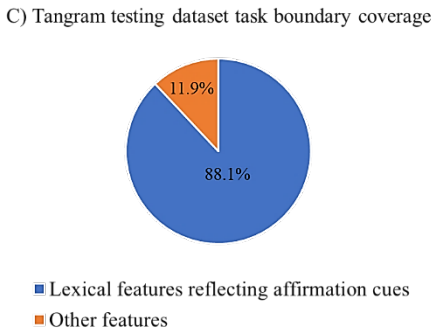


Figure 3.3 : Coverage of Identified Affirmation Cues in Task boundaries – A) Training Dataset, B) UMT test dataset, C) Tangram test dataset

this work, *backchannels* are defined as verbal cues that represent continuity in a task-oriented dialogue [40]. For example, the affirmation cue *yep* could indicate continuity in a conversation by the interlocutor while also functioning as an affirmation cue indicating a task boundary.

This could potentially result in false identification of a task boundary (i.e., false positive) when using affirmation cues; and eventually, lead to disruptive interruptions. Thus, while adding more affirmation cues into the feature set may improve the Coverage of the ACE-IMS, i.e., reduce the missed interrupting opportunities, it has the risk of increasing undesirable disruptive interruptions. Clearly, reducing false and missed interruptions are two conflicting objectives. Moreover, different distributed collaborative applications may prioritize them differently, for example, some professions may be tolerant to interruptions from frequent alarms even if they are false rather than miss an alarm altogether and leading to potentially disastrous situations. Therefore, there is a need for a

balanced and flexible design approach that supports application-specific operation requirements through convenient system adaptations.

In short, the coverage data shown in Figure 3.3 inform us that affirmation cues account for most of the task-boundary phrases. Thus, a system implementation that utilizes these features to disseminate real-time interruptions is described in detail in Section 3.3 and 3.4.

3.3 Real-time Affirmation Cues based Interruption Management System (ACE-IMS)

The proposed ACE-IMS solution addresses the issue of processing lexical information in real-time for the purpose of making intelligent interruption decisions. The developed prototype serves to validate its operation within real-time interactions. The prototype specifically emphasizes key phrases associated with task boundaries that reflect affirmation cues. For this reason, a rule-based classification approach is proposed which, in future work, can be expounded upon to consider other machine learning and deep learning modeling approaches to create a hybrid architecture. The system design is shown in Figure 3.4.

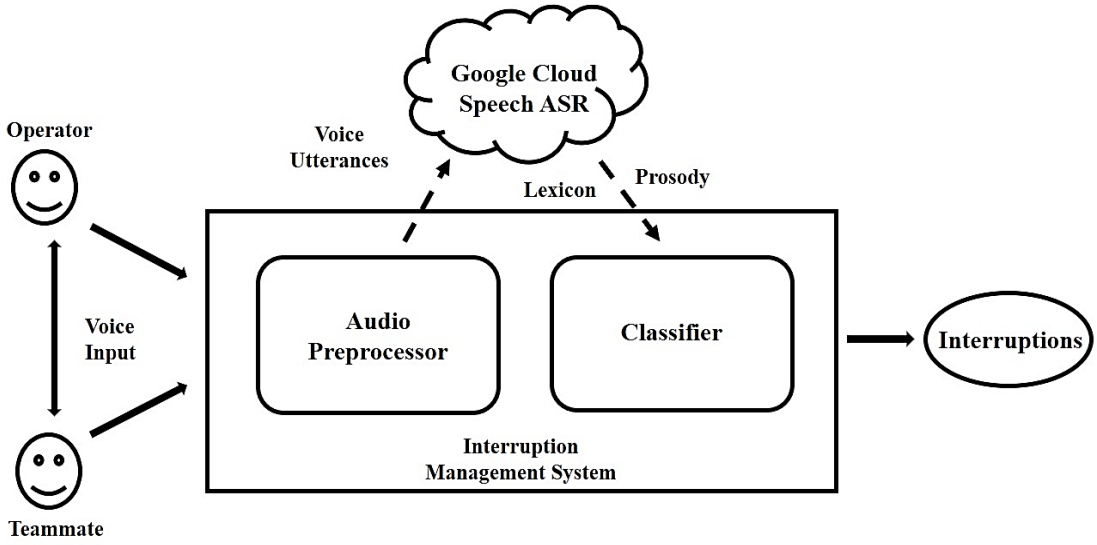


Figure 3.4 : System Design of Proposed Interruption Management System

The system consists of a multi-channel audio input, i.e., one audio input per user, that records voice data and relays its digital manifestation to an acoustic preprocessor. The acoustic preprocessor reduces the noise and fine-tunes the gain of the audio using Audacity API (Application Programming Interface) [68]. The preprocessed audio is then sent to cloud-based Automatic Speech Recognition (ASR) engine, e.g., the Google Cloud Speech service in our implementation, which uses a server-client implementation for speech to text transcription [37]. The real-time ASR is one of the key components that enables real–time operation of the proposed ACE-IMS in addition to the lexical analysis system. The adoption of the widely available cloud-based ASR services, such as Google Cloud speech, helps mitigate the potentially prohibitive computation burden of running an ASR on the local machine and, ultimately, the delay associated with high-accuracy speech recognition (*WER=4.9%*, where WER stands for "Word Error Rate")[69]. Experimental studies

conducted to observe the real-time ASR latency show that the delay is under *350 ms*. The resulting text utterances are then fed into a progressive rule-based classifier that controls and disseminates the interruptions in real-time.

To visualize and evaluate the feasibility of real-time task boundary detection capability, an Android-based prototype of the ACE-IMS is implemented and illustrated in Figures 3.5 and 3.6. The implementation consists of two Android tablets. Figure 3.5 and 3.6 illustrate the interfaces of the ACE-IMS for supporting the distributed operations of a UMT task. The dialogue visualizer on the left side of the interface displays the real-time speech-to-text output from ASR for the two-person dialogue within a distributed multiuser multitasking interaction. This visualizer can be toggled on and off. These capabilities allow researchers to view the dialogue and interruption decisions made by the IMS in real-time. The dialogue highlighted in red indicate the task boundaries identified by the ACE-IMS running in the background. On the right side of the interface in Figure 3.5 and Figure 3.6 is a customizable primary task interface (for example, Aerial view in Figure 3.5 or street view in Figure 3.6 for supporting the UMT tasks). This portion of the interface can be customized to any visual interface that is conducive to simulating a task within the domain of interest (Tangram/UMT). The prototype also provides features like data collection for both voice and text, which can be further used to expand the existing dataset, train the classifier and improve the accuracy of task boundary identification and other interruption inference models.

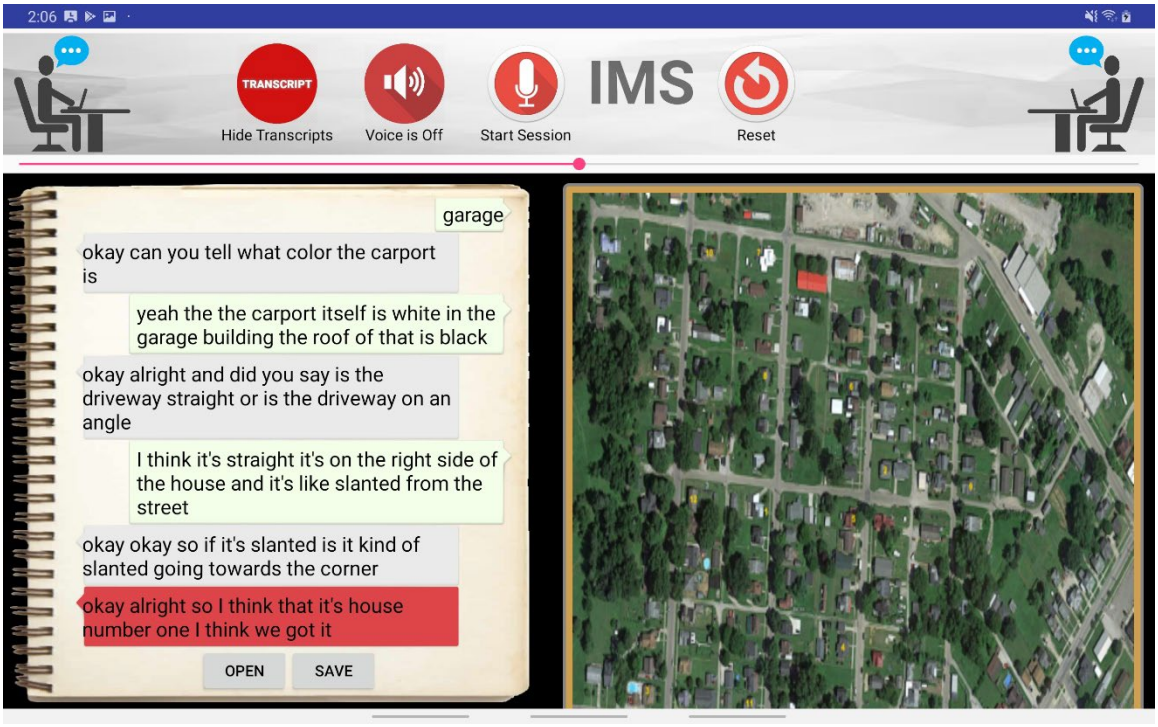


Figure 3.5 : ACE-IMS Prototype Supporting UMT Task (Aerial View)

3.4 Progressive Ruleset Design

Since a primary contribution of this work is to focus on affirmation cues as an indicator of a task boundary for intelligent interruption dissemination, the affirmation cues need to be identified in speaker utterances. Hence, a rule-based classifier is the first pass at implementing the proposed real-time ACE-IMS. A rule–based classifier provides the following advantages in the context of task boundary classification:

1. Deterministic decision-making based on domain specific features that indicate task completion,

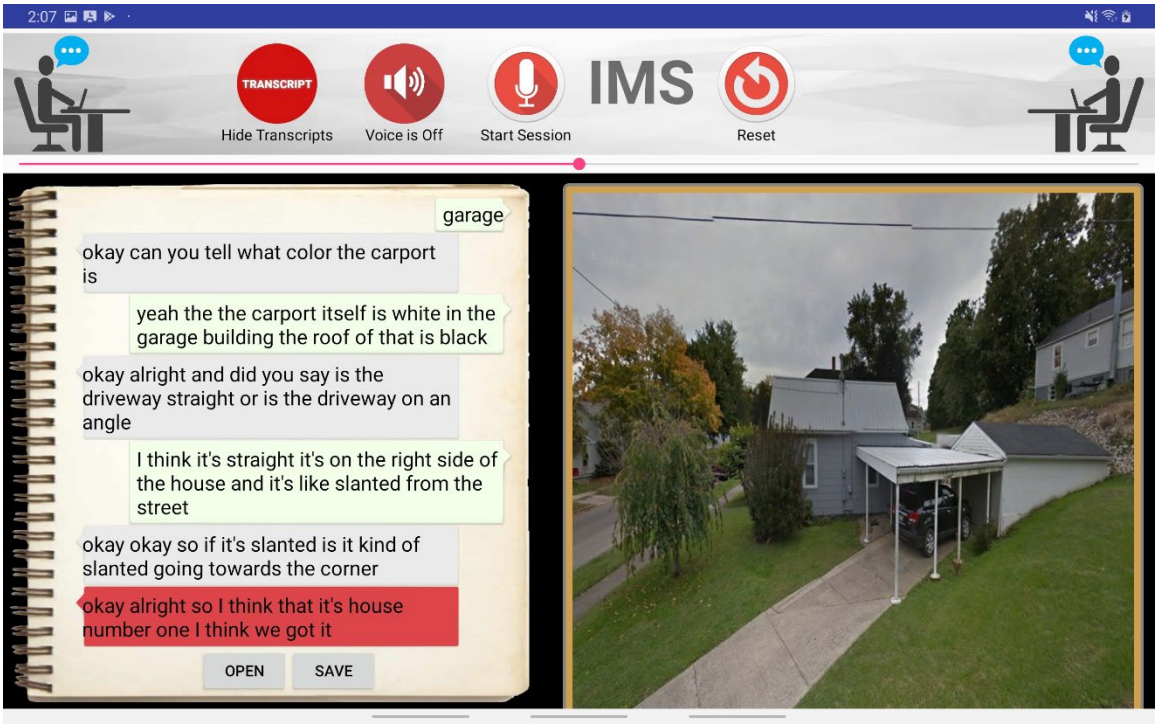


Figure 3.6 : ACE-IMS Prototype Supporting UMT Task (Street View)

2. Flexible operating points that optimize conflicting variables like missed interrupting opportunities and disruptive interruptions based on application requirements,
3. Flexible feature adaptation to accommodate multiple modalities of data and revise them based on the availability of new features.

These advantages make the rule-based classifier a viable candidate for classifying task boundaries in IMS. Mapping affirmation cues to utterances may seem to be a straight-forward task where one must parse utterances for affirmation cues to determine task-

boundaries. However, due to the interference from backchannels as discussed earlier in Section 3.2.1, it is not so trivial when optimality, scalability, adaptability and real-time needs of interruption dissemination are brought into focus. Consider the problem of designing the optimal collection of affirmation cue features. In the context of this research, based on the initial assessment of Coverage within the training dataset, 9 affirmation cues are shortlisted, as shown in Table 3.2. There is a chance that these affirmation cues may also appear as backchannels in non-task boundary utterances. Reducing false and missed interruptions are two conflicting objectives, eliciting the question: *what should be the objective function that is used to optimize the rule–set with conflicting goals of minimizing false interruptions and minimizing missed interruptions?*

Furthermore, if the application domain of the IMS dictates that interruptions must be disseminated frequently, but is tolerant to the number of false interruptions or vice versa, *How can the user modify the behavior of the classifier to facilitate application-specific interruption dissemination?* And most importantly *How can these rules be selected and arranged to enable real–time operation?* In the following subsections, these concerns are addressed by presenting a detailed description of progressive rule-set design and its effectiveness in addressing the challenges.

3.4.1 Objective Function and Optimization

To take a more balanced consideration for performance assessment, the F1 score, a combined measure of both false interruptions and missed interruptions, is used. To

determine the right sequence of affirmation cues-based rules that support progressive operation, the *steepest ascent method for multivariate optimization* [70] with the multivariate objective function as *F1 score* is adopted. For each iteration of the algorithm, the F1 score and the delta F1 score are calculated. The formula for F1 score is as shown in Equation (2):

$$F1\ score = (2 * Precision * Recall) / (Precision + Recall) \quad (2)$$

Where

$$Precision = True\ Positives / (True\ Positives + False\ Positives)$$

and

$$Recall = True\ Positives /(True\ Positives + False\ Negatives)$$

Here *True Positives* are utterances that are correctly identified as task boundaries, *False Positives* or false interruptions are non-task-boundary utterances which are wrongly identified as task boundaries, and $False\ Negatives$ or missed interruptions are task-boundary utterances which are wrongly identified as non- task-boundary utterances. Within each iteration, the ΔF1 score is given in Equation (3).

$$\Delta F1\ score = F1\ score(current\ set\ of\ lexical\ affirmation\ cues + new\ lexical\ affirmation\ cue) - F1\ score(current\ set\ of\ lexical\ affirmation\ cue) \quad (3)$$

Then, the *current set of lexical affirmation cues* is updated by adding the lexical feature that produces the maximal $\Delta F1\ score$ improvement at each iteration.

3.4.2 Iteration–Wise Description of Affirmation Cue based Ruleset

The iteration-wise development of lexical affirmation cue-based classifier is shown in Table 3.4 with the optimal ruleset highlighted in bold. Additionally, a graphical representation of affirmation cue development of Table 3.4 is rendered in Figure 3.7 where the X-axis consists of *Iteration Number or Operating Point* and the Y axis represents the performance measures in F1 score, Precision or Recall. Iteration Number corresponds to the iteration of steepest search algorithm. Operating point is the F1 score that characterizes the performance of the progressive ruleset-based classifier. In addition to F1 score, Precision and Recall are also presented to understand the contribution of each affirmation cue to false interruptions and missed interruptions in greater detail. The characteristics for the lexical affirmation cue classifier are displayed in dashed lines.

A detailed description of the operations performed in the first 2 out of 9 iterations of the steepest ascent search algorithm is described with the aid of an illustration in Figure 3.8.

Although ***Iteration 0*** is mentioned in the description as the initial iteration with an empty ruleset, it is done for theoretical purposes to serve the conceptual explanation of the steepest ascent search algorithm while for all purposes of implementation the operations begin from Iteration 1.

Iteration 1: The F1 scores of the individual affirmation cues are calculated by checking for the presence of each affirmation cue in each utterance of the 3066 UMT training dataset. At this iteration, the affirmation cue producing the highest value of

$\Delta F1\ score$= 65.4% is ***got it***. Therefore, *got it* is chosen as our base affirmation cue in the progressive rule-set and represented as a point on the F1 score line at Iteration 1 or operating point 1 in Figure 3.7.

Iteration 2: the affirmation cue "got it" from the first iteration is individually combined with the remaining affirmation cues, two at a time, to calculate the combined F1 score. At this iteration, the affirmation cue producing the highest value of $\Delta F1\ score$=1.6 is *{got it,* ***yep****}*, represented as the F1 score = 67.0% on Iteration 2 or operating point 2 in Figure 3.7.

This process of iteration-wise addition of affirmation cues to the existing set of affirmation cues contributes to our definition of a "Progressive rule-set" design. This progressive development of affirmation cues continues for 9 iterations (for 9 affirmation cues), as shown by a monotonically increasing F1 score line graph in Figure 3.7. However, on close observation we find that the trend decreases slightly after the F1 score of 70.2% corresponding to iteration number 7. Hence 70.2% is the maximum F1 score. The corresponding sequences are the optimum sequences and represented in bold in Table 3.4.

Table 3.4: Iteration-wise Progression of Affirmation Cues

Iteration Number or Operating Point	**Rule-set for lexical-only classifier**

0	NULL
1	got it
2	got it, yep
3	got it, yep, sounds good
4	got it, yep, sounds good, done
5	got it, yep, sounds good, done, got you
6	got it, yep, sounds good, done, got you, awesome
7	**got it, yep, sounds good, done, got you, awesome, gotcha**
8	got it, yep, sounds good, done, got you, awesome, gotcha, sweet
9	got it, yep, sounds good, done, got you, awesome, gotcha, sweet, great

3.4.3 Utterance Duration As an Extra Feature

To further mitigate the interference from back-channels, the duration of an utterance is introduced as an extra rule to help reduce false positives. The distribution of duration of the task boundary utterance is shown in Figure 3.9.A. Here 97.5% of task

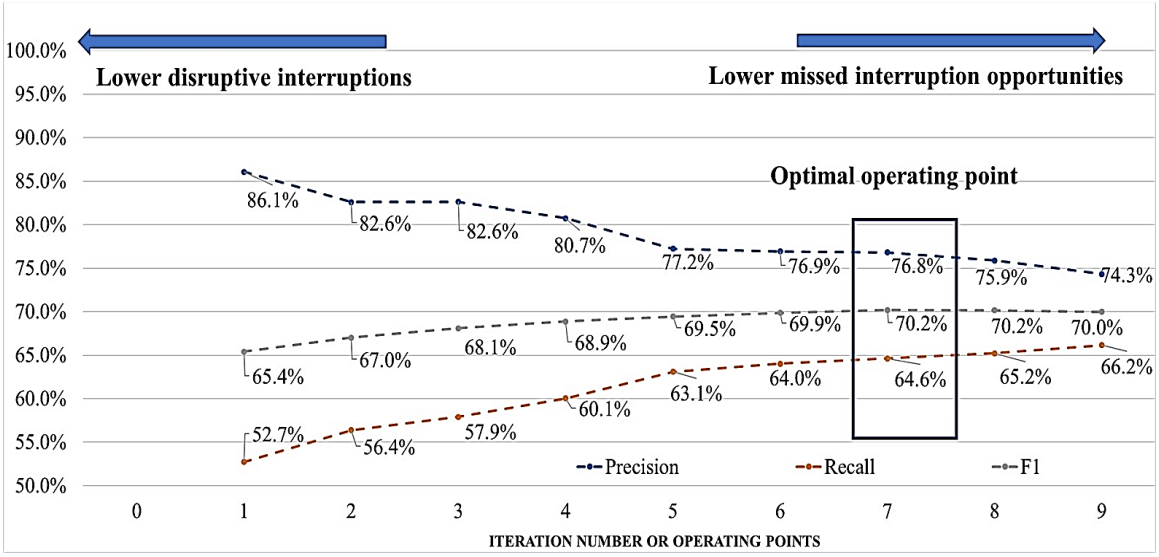


Figure 3.7 : Iteration-wise Graphical Representation of Metrics of Affirmation Cues in Steepest Ascent Search Algorithm

boundary utterances have a duration less than 10 seconds, which leads to a 2.5% extra missed interruptions detection if a threshold of 10s is used for task boundary utterances. However, the distribution of the duration for non–task boundary utterances that are false positives is examined, as shown in Figure 3.9.B, which clarifies that 17.7% of false positives are above the 10 second limit and will be filtered out of the identified task boundary if a threshold of 10 seconds is applied for task boundary utterances. Thus, it is expected that a portion of these false positive can be removed at a low cost of missed interruption opportunities. Therefore, we also look at the effect of adding one extra rule "DURATION is less than or equal to 10 seconds" into the ruleset.

Iteration 1		
Confirmatory Cues	**F1 Scores**	**$\Delta F1$ score**
got it	65.4	65.4
got you	6.8	6.8
yep	7.9	7.9
gotcha	1.2	1.2
awesome	1.8	1.8
sounds good	3.0	3.0
done	4.7	4.7
sweet	1.8	1.8
great	1.8	1.8

Iteration 2		
Confirmatory Cues	**F1 Scores**	**$\Delta F1$ score**
{got it + awesome}	65.9	0.5
{got it + yep}	67.0	1.6
{got it + done}	66.3	0.9
{got it + got you}	66.0	0.6
{got it + sounds good}	66.5	1.1
{got it + gotcha}	65.8	0.4
{got it + sweet}	65.4	0
{got it + great}	65.3	-0.1

Figure 3.8: First 2 of the 9 Iterations of Steepest Ascent Search Algorithm

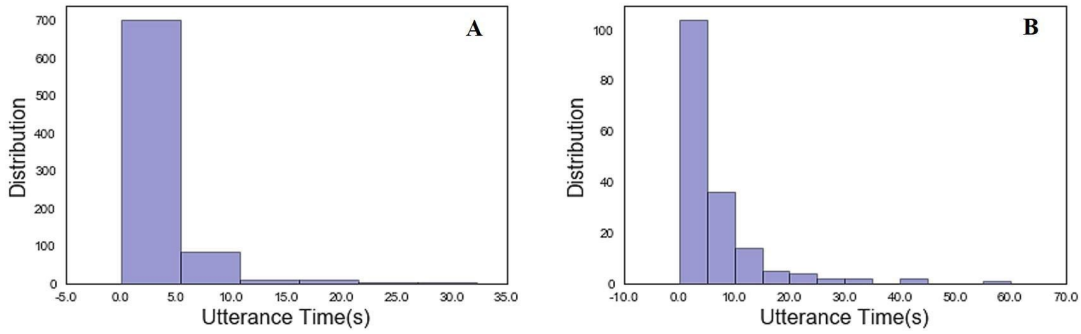


Figure 3.9 : A) UMT Task Boundary Frequency Distribution B) UMT False Positive Frequency Distribution

3.4.4 Implications of Progressive Rule–Set Design

It is evident that systematic selection of lexical affirmation cues is necessary to construct an accurate, scalable, and adaptable rule–based classifier for interruption dissemination. The steepest ascent search algorithm with F1 score as objective function facilitates this systematic selection of optimal progressive rule set. In this subsection, let us examine how the progressive rule-set based classifier addresses the challenges of optimality, complexity, adaptability, and real-time operation.

1. *Solution to optimal ruleset:* By utilizing F1 score as the objective function, the steepest ascent search algorithm produces a steepest ascending curve that peaks at the maximum F1 score for the ruleset at each iteration before it starts descending as shown Figure 3.7. As a result, the ruleset of affirmation cues sequenced until the maximum F1 score can be considered as a series of incremental optimal rulesets.

2. *Solution to* complexity [71]: The search process of the steepest search algorithm for $n = 9$ affirmation cues performs n affirmation cue F1 score calculations in the first iteration, $(n–1)$ calculations in the second iteration, $(n–2)$ in the third and so on. Hence, it can conclude that the search process has a quadratic complexity of $O(n^2)$, which has less complexity when compared to the exponential complexity of $O(2^n)$ when using a brute–force approach, to find all rule-set combination of all size. This difference in time complexity allows us the rule-set to operate relatively faster with more affirmation cues.

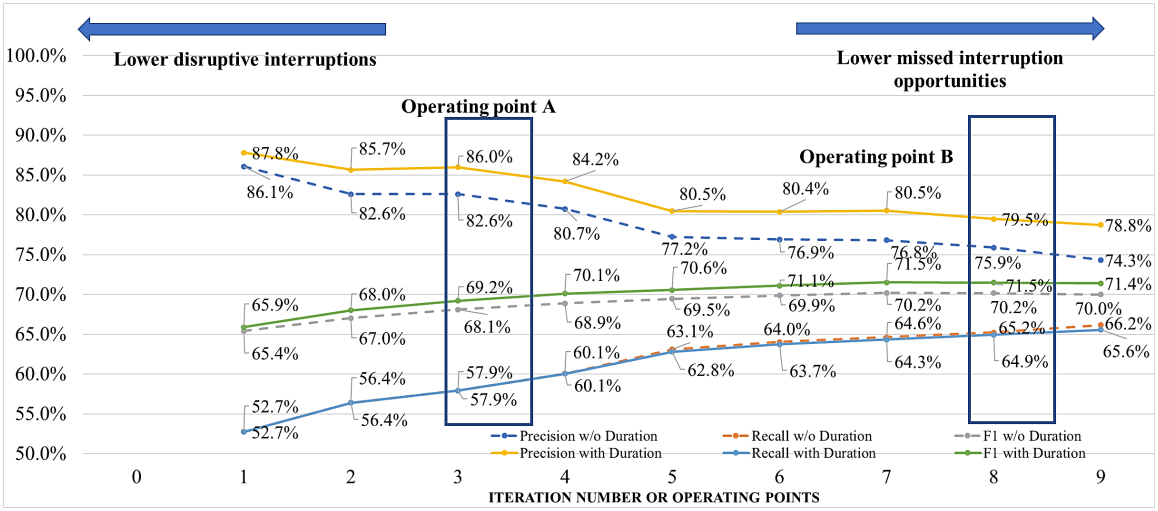


Figure 3.10 : Operating Point-based Adaptability of Progressive Ruleset

3. *Solution to adaptability:* In Figure 3.10 the X-axis as labelled as an operating point. This is done to emphasize operating point dependent behavior of the classifier. By choosing operating point "A" a lower disruptive interruption is favored over missed interruption opportunities and vice-versa by choosing operating point B. Hence, a classifier can be tuned to prioritize the needs of interruption dissemination for the application.

Thus, we have demonstrated that the steepest search algorithm can be utilized to generate a progressive rule set that facilitates real-time operation. It is optimal, scalable and adaptable to application specific requirements (Missed interruptions vs False interruptions). This ruleset should enable the classifier to distinguish between task–boundaries and non–task boundaries according to the selected operating point. Section 4 is utilized to present the performance evaluation of the designed ruleset against task–oriented dialogues and consequently its experimentation design and results.

3.5 Results

In this section, the experimental results of the proposed ACE-IMS are presented. Firstly, a brief description summarizing the experiments is made, then the generalizability of the progressive ruleset-based classifier is evaluated by comparing the results between the training and testing datasets. This followed by a performance comparison with the real-time C-CIMS, current baseline ACE-IMS.

As described in Section 3.1, two test datasets are used for the performance evaluation: UMT test dataset and Tangram test dataset. The UMT test dataset allows us to test if the classifier performance generalizes to other utterances from the same task, while the Tangram dataset allows us to assess classifier performance on task-oriented dialogue of a different task. To compute the classifier metrics of Precision, Recall and F1 score, the ACE-IMS assigned labels of *Task boundary versus Non – task boundary* is compared with the manual annotations of task boundary information, *Task boundary versus Non – task boundary* based on corresponding log files of the two datasets, UMT and Tangram, described in Section 3.1.

Table 4.1 summarizes the classification results for the ACE-IMS for the UMT training dataset, UMT test dataset and Tangram test dataset. The results for both lexical-only classifier and lexical–Duration classifier are presented, with the latter exploring additional potential to suppress false interruptions due to back–channels as discussed in Section 3.4.3.

Table 3.5 : Classification Results of Training Versus Test Datasets

Tasks	Features	Precision (%)	Recall (%)	F1 Score (%)	Data Split (Non-task boundary, task boundary)
UMT (Training)	Lexical	76.8	64.6	70.2	(2737, 329)
	Lexical + Duration	80.5	64.3	71.5	(2737,329)
UMT(Test)	Lexical	76.8	62.4	68.8	(5162,808)
	Lexical + Duration	80.2	60.2	68.9	(5162,808)
Tangram (Test)	Lexical	94.6	87.9	91.1	(3396,1158)
	Lexical + Duration	95.5	87.2	91.1	(3396,1158)

3.6 Inter Dataset Performance Evaluation

Firstly, let us evaluate the performance of the ACE-IMS by considering the lexical-only classifier across UMT test and Tangram test datasets. Results in Table 4.1 show that ACE-IMS perform robustly across both UMT test dataset and Tangram dataset. It achieves 68.8% F1 score for UMT test dataset, which is close to its performance of 70.2% for the UMT training dataset. It proves that ACE-IMS generalizes well for the same type of task. Furthermore, ACE-IMS achieves 91.1% F1 score for the Tangram test dataset, which validates the earlier Coverage of affirmation cues in the task-boundary utterances of

Tangram tasks. It shows that ACE-IMS generalizes well for a different type of task. Moreover, it demonstrates that ACE-IMS can take full advantage of the higher rate of affirmation cue usages in the Tangram tasks.

3.7 Inter Classifier Performance Evaluation

Next, the inter classifier performance is evaluated between lexical-only and lexical–Duration classifiers. Both classifiers achieve comparable optimal F1 scores on all three datasets, refer Table 4.1., with only marginal loss of Recall (which means it misses interruptions but to a lesser degree) which validate approach, motivated by the observations in Figure 3.9 in Section 3.4.3. Furthermore, on closer examination it can be discerned that this lexical-Duration classifier improves the precision score across all three datasets.

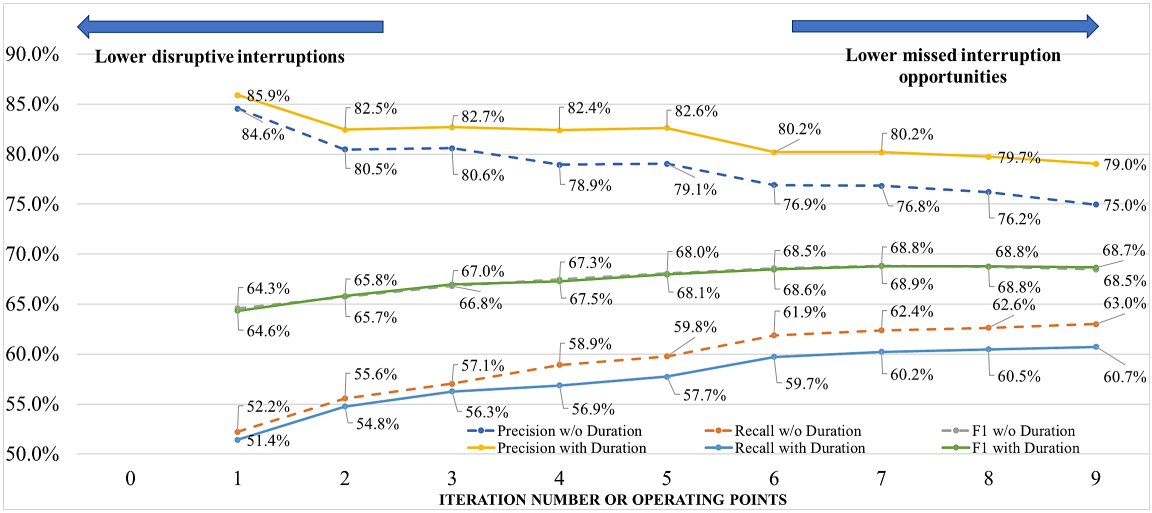


Figure 3.11 : Performance and Operation Adaptability of ACE-IMS (UMT Test Dataset)

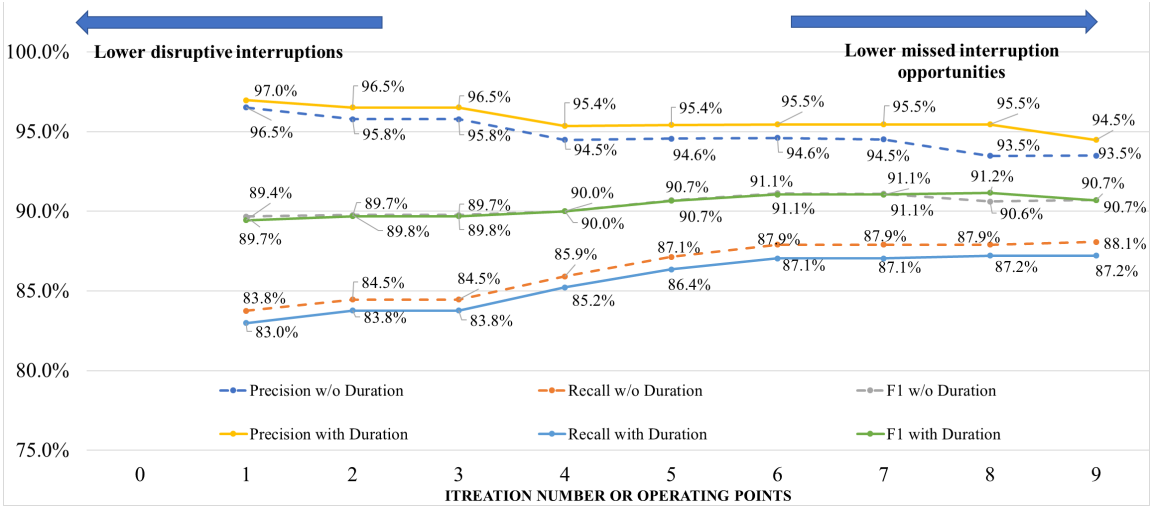


Figure 3.12 : Performance and Operation Adaptability of ACE-IMS (Tangram Test Dataset)

Although not to a greater extent, the Duration feature helps reduce the number of false interruptions. But, more importantly, by adding Duration feature, it achieves much more gracefully descending trend of the Precision lines as shown in Figures 3.10, 4.1 and 4.2. This is of importance when reducing false positives (which lead to disruptive interruptions) is of high priority.

3.8 Real–time ACE-IMS vs Real-time C-CIMS [5]

Since one of the primary foci of this research work is to study the role of lexical affirmation cues in identifying task boundaries and compare its performance against existing literature, our focus, in this sub-section is limited to compare the performance of the real-time lexical – only classifier against the real-time C-CIMS implementation. Table 4.2 summarizes the performance results.

Table 3.6 : Real-time IMS Results Comparison with Real-time C-CIMS (Peters 2017b)

Tasks	Features	Precision (%)	Recall (%)	F1 Score (%)	Data Split (Non-task boundary, task boundary)
Tangram	Real-time ACE-IMS (proposed solution)	94.6	87.9	**91.1**	(3396,1158)
	Real-time prosodic (C-CIMS)	79.1	70.8	74.7	(1205,811)
UMT	Real-time ACE-IMS (proposed solution)	76.8	62.4	**68.8**	(5162,808)
	Real-time prosodic (C-CIMS)	44.7	75.6	56.2	(3517,961)

The performance results in Table 4.2 clearly demonstrate that the proposed Lexical based ACE-IMS classifier outperforms the real-time C-CIMS implementation. The ACE-IMS shows improvements in F1 score against the C-CIMS for both Tangram test dataset and UMT test dataset, 16.4% and 12.6%, respectively. For the Tangram dataset, the proposed ACE-IMS shows an improvement of 15.5% for Precision - which means it disseminates less disruptive interruptions, while, at the same time, achieves 17.1% improvement for

Recall, i.e., missing less opportunities to interrupt. For the UMT test dataset, the proposed ACE-IMS shows an improvement in precision by 32.1% when compared to C-CIMS. Although C-CIMS shows a better Recall by 13.2% for UMT test dataset, it is largely due to its imbalanced treatment between Precision and Recall, which results in a loss of 12.6% in F1 score when compared to ACE-IMS.

The numerical results suggest that the proposed IMS generalizes well across the UMT and Tangram datasets and outperforms the existing real-time implementation of C-CIMS in identifying task-boundaries.

Table 3.7 : Tabular Coverage Results of Figure 3.3

Dataset	**Total Number of Task Boundary Utterances with The Identified Affirmation Cues(k)**	**Total Number of Task Boundary Utterances(N)**	**Coverage (k/N) (%)**
UMT Training	230	329	69.9
UMT Test	508	808	62.9
Tangram Test	1020	1158	88.1

4 DISSEMINATION OF INTANTANEOUS FEEDBACK THROUGH REALTIME ASSESSMENT OF DIALOGUE IN MOTIVATIONAL INTERVIEWING

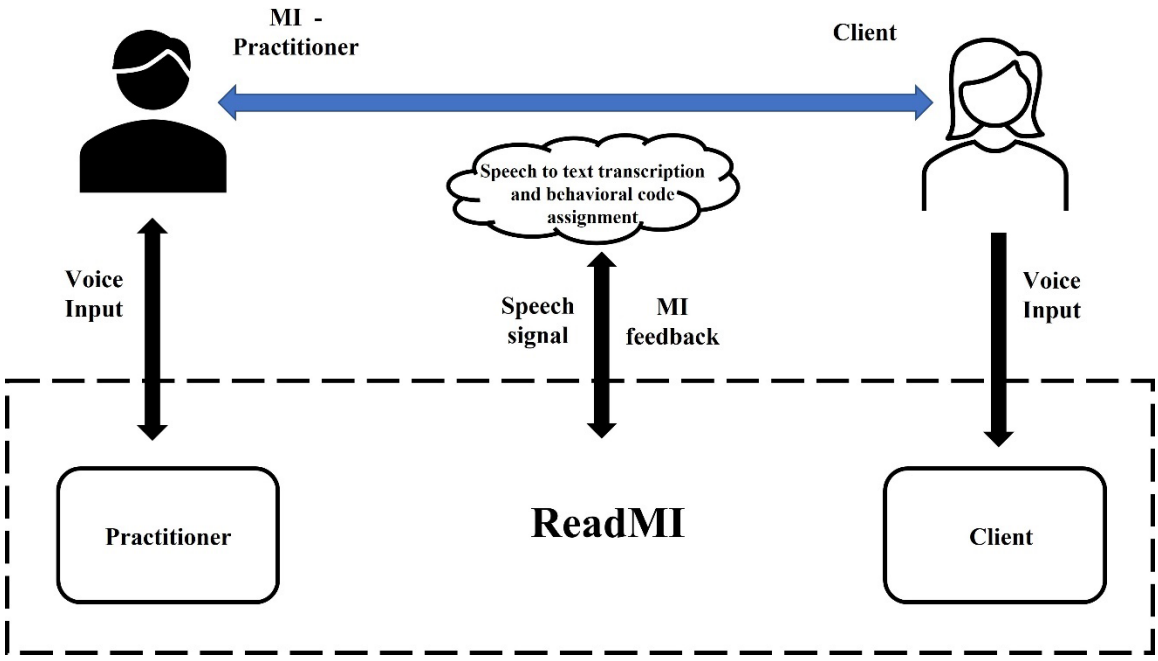


Figure 4.1: ReadMI System Architecture

4.1 System Architecture

By leveraging the high performance distributed computational capabilities of cloud computing infrastructure in synergy with ergonomically supportive mobile devices, a mobile cloud computing architecture is leveraged to implement ReadMI as shown in Figure. 4.1. Irrespective of whether ReadMI is used in a co-located (in-person) and

synchronous (same-time) setup where the practitioner, facilitator and client are in the same room and interacting at the same time or are in different geographical location, this system architecture holds true supporting both cooperative workflows. The flexibility in workflow is accommodated by the different interfaces. Mobile application is developed and used on tablets to support in-person training and a web-application is used for virtual training sessions to maximize usage so that the trainer and trainee have the flexibility of using their own familiar computing devices to avoid any significant need for technical support, while trying to create the same social interactions characterizing a physical in-person training scenario. This flexible architecture proved to be of great convenience during a global pandemic like COVID-19 which necessitated physical distancing measures.

In-built microphones in mobile devices act as voice inputs to both practitioner and the client. Microphones are single channel voice inputs. Voice signals in the form of voice packets are then sent to cloud computing resources implemented as servers via high-speed internet where the cloud hosted ASR algorithm converts speech signals to text data, sent back to the mobile devices, in real time (delay < 300ms). As it is the motivation of this research work to create a cost-effective workflow built on ubiquitous technologies, efforts have been directed towards choosing existing technologies and melding them into a seamless technology. Consequently, deep learning based automatic speech recognition algorithms, made available as cloud services from Google [37], IBM[72] , Microsoft [73], Amazon [74] are leveraged.

Utterances obtained from such services are sent simultaneously in real-time to the mobile devices as captions to enhance user experience and the ReadMI Behavioral coding classifier for feedback generation. Due to significant improvements in processing power, the classifier can be implemented locally in mobile devices or on other cloud computing services like Amazon EC2 instances to facilitate remote collaboration. Incoming ReadMI clinician utterances are coded as “Open-ended Question”, “Close-ended Question”, “Reflective statement”, “Scale statement” and “None”. These classified utterances are then displayed on the mobile device as ReadMI metrics in the form of number of instances of use and the corresponding utterances spoken by the clinician in the categories of OPEN QUESTION, CLOSE QUESTION, REFLECTIONS, SCALE. This organization and presentation of practitioner utterances is labelled as “MI feedback” in Figure 4.1. An in-depth look into the classifier design is covered in Section 4.3. In addition to the above stated ReadMI metrics, the practitioner speaking time and client speaking times are recorded, allowing MI trainees and the MI experts to evaluate each individual session.

To facilitate remote collaboration in MI training, video conferencing Application Programming Interfaces (APIs) [75] can be leveraged, thereby removing distance limitations on ReadMI. This would allow MI trainee clinicians, MI experts and clients to collaborate remotely utilizing audio-visual video signals to practice motivational interviewing skills without the need for in-person sessions. In the current, challenging times of global pandemic, the remote video conferencing adaptation of ReadMI has proved to be an invaluable tool for MI training.

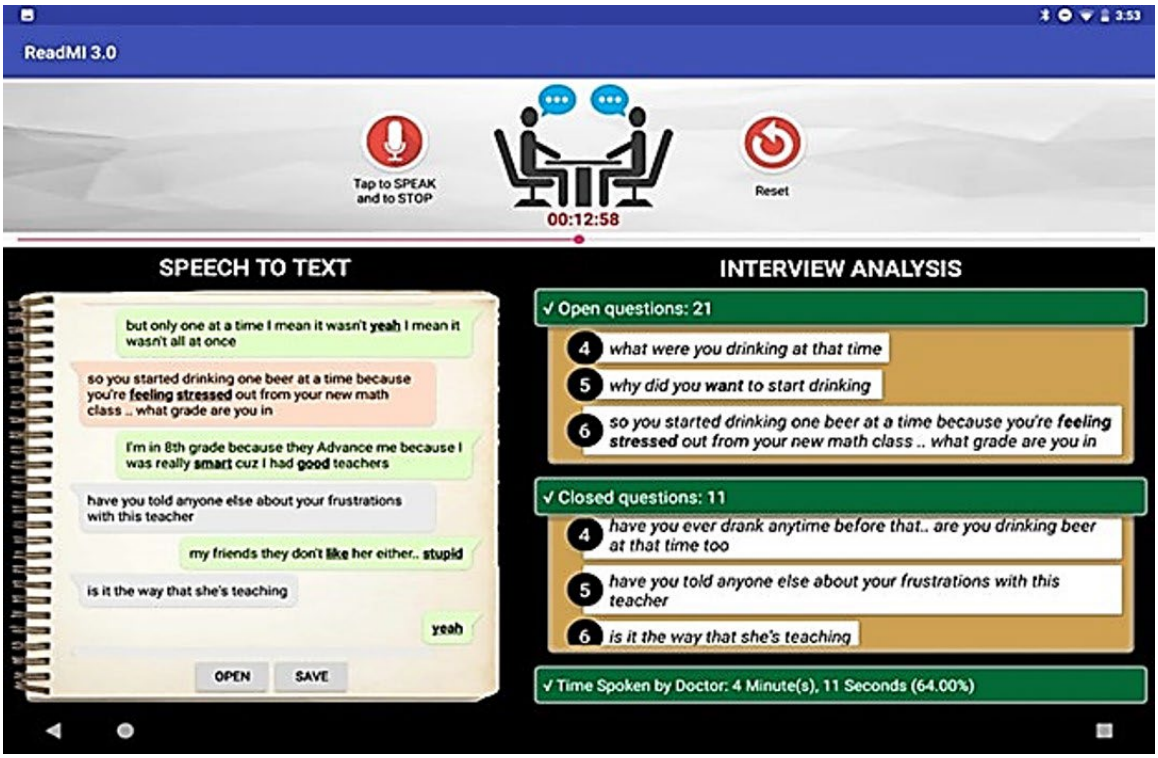


Figure 4.2: Android based ReadMI Prototype

4.2 Prototype

The primary focus of this research work is to present a robust, fully operational solution to support an MI training session in real-time and facilitate both in-person and remote learning sessions. To bring this to fruition the system architecture described in section 4.1 is leveraged to develop mobile-cloud based solutions for both in-person and

remote training solutions. To contextualize and demonstrate a computer supported

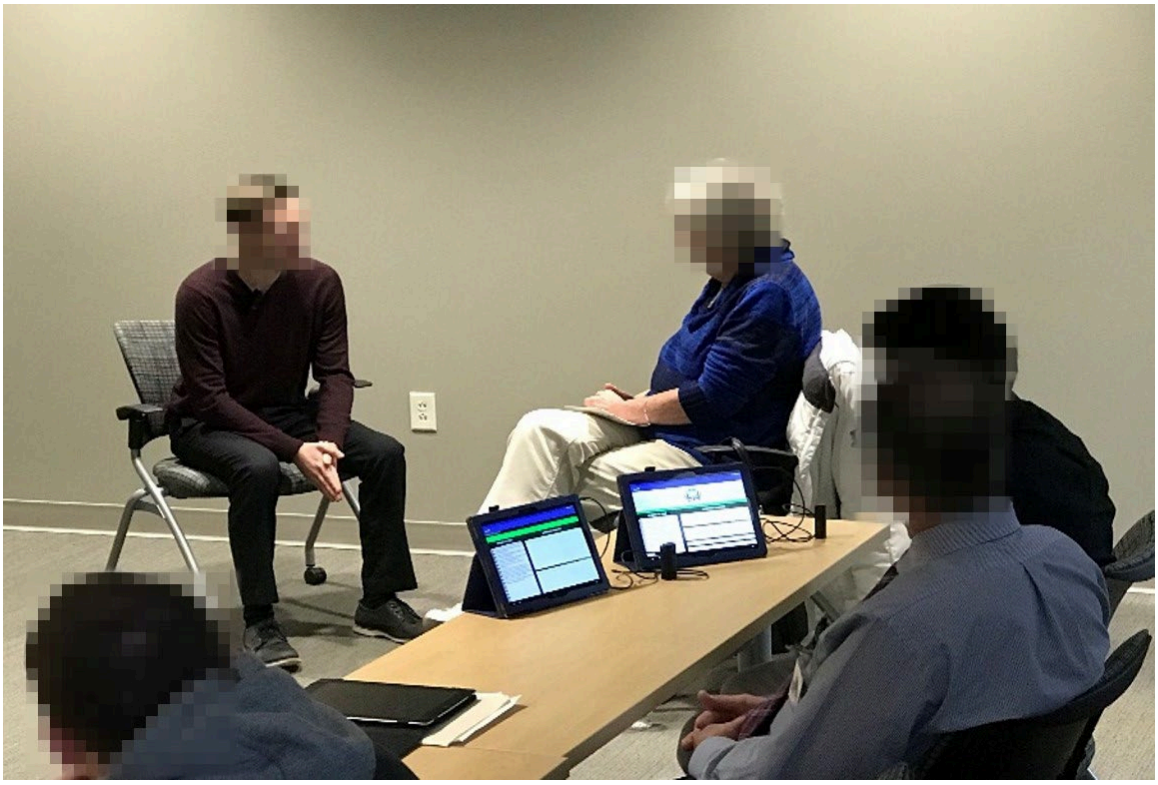


Figure 4.3: ReadMI In-person Training Session

collaborative learning workflow, the ReadMI prototype is developed, operated and validated in a healthcare setting where the MI practitioner can be a clinician like a doctor, nurse-practitioner or a therapist. The facilitator is an experienced MI expert instructor, and the client is most likely a patient. However, as stated in section 1.5 ReadMI's applications grow beyond traditional health-care setting.

4.2.1 In-person Training

An Android platform [76] based ReadMI application shown in Figure 4.4 facilitates in-person MI training sessions. This choice was made as Android is the largest mobile operating system in use by market share [76], which means it is supported by regular system and security software updates. This allows for greater adoption of ReadMI services. The user-interface as shown in Figure 4.4 is displayed on a tablet computer with a touch interface. It consists of a “Speech to Text” window and an “Interview Analysis” window. The “Speech to Text” window shows real-time clinician-client utterances as speech bubbles depicting a real-time sequential flow of conversation. The "speech bubble" design was inspired by graphic user interface (GUI) elements in text messaging applications to enable similar intuitive user experience**.** The clinician speech bubble GUI element is displayed to the left and the client to the right. Speech bubbles containing reflective utterances are highlighted in “Green” color for easier recognition. The "Speech to Text" window allows user to scroll through speech bubbles via touch interface. This scrolling motion-based navigation allows the MI expert facilitator and clinician to reference a previously spoken utterances for analysis. For example, if the clinician had spoken a closed question, the expert facilitator can refer to the close question and suggest an open-ended way of framing that question and suggest how the clinician could direct the subsequent conversation in an MI consistent way. Real-time transcriptions are obtained by utilizing the Google Cloud Speech ASR via the Google Cloud Speech API [37]. has the best accuracy of all competing ASR services with Word Error Rate(WER) = 5.6%, [66]. In the

interview analysis section of the application the number of "Open-ended" questions, "Close-ended" questions and the actual corresponding questions asked in the Clinician-client sessions are shown. A “Stop” sign glows to warn the clinician of successive delivery of unfavorable close-ended questions.

Figure 4.4: ReadMI Remote Training Session

Additionally, the “Open” Graphical User Interface (GUI) button allows the clinician MI trainee or the expert to load and review previous role-plays and the corresponding metrics for analysis after a session and the “Save” GUI button can be used for saving a roleplay session and its metrics to create new data points for data analysis and

facilitate further refinement of the ReadMI application in general, and the behavioral coding classifier. A typical MI training session with ReadMI is shown in Figure 4.3.

4.2.2 Virtual Remote Training

The urgent need for a remote collaboration implementation of ReadMI to support virtual remote MI training arose due to the COVID-19 global pandemic and the consequent social distancing restrictions that limit face-to-face interactions. Non-face-to-face environments limit social learning due to the inability of the participants to see one another [77]. Hence, there is a need to improve the ability of learners to see and express themselves socially and emotionally as "real people" through available communication means i.e. improve social presence [78]. [79] has reported positive outcome in student satisfaction and increased social presence on using WVC (Web-based Video Conferencing) for learning. Hence, motivated by the need to leverage the social learning benefits of face-to-face learning over large geographical distances the remote collaboration version of ReadMI was implemented as a web application using JITSI [15] video conferencing API, see Figure 4.4. Google Cloud Speech to Text API was used for live speech to text capabilities, an Amazon AWS EC2 [80] instance was used to host a JITSI server which allowed users to log in via designated Uniform Resource Locator (URL) [81]. In the JITSI based implementation of ReadMI, MI metrics along with the corresponding utterances spoken by the trainee are shown on the left in addition to the clinician trainee and client/patient speaking time as shown in percentage (%) of the session time. The trainee and the client

are connected and shown in the video conferencing window together with the MI expert supervising the session. ReadMI will join as a virtual participant to observe, process and provide feedback. Other participants may also join to observe the ongoing session. Their corresponding video streams are usually disabled due to bandwidth constraints and the convenience of participants of ongoing sessions.

4.3 Behavioral Coding Classifier Design

One of the primary goals of ReadMI is to automate the behavioral coding process of practitioner utterances in MI sessions to generate technical component-based metrics of MI, improving the comprehensiveness of feedback. As the focus of ReadMI application is to measure practitioner performance, the behavioral coding mechanism is restricted to practitioner utterances only. Client utterances too provide a way to measure practitioner performance, example, change-talk, but this is out of the purview of this dissertation. To assign behavioral codes to practitioner utterances, a behavioral coding classifier that uses a rule-based classifier to assign behavioral codes is implemented. The factors contributing to our decision are 1) MI dialogue is inherently task-oriented, where the practitioner is trained by the facilitator to use certain specific lexical features in spoken utterances. Hence the motivation was to capture these lexical feature patterns in the simplest way possible and map them directly to the speech utterances as a possible first step to establish a baseline. 2) The MITI behavioral coding and the expert modified MI metrics adapted for the version of MI shown in Table 2.[82]1 have characterized what lexical patterns.

Table 4.1:Excerpt of Lexical Feature Dictionary

OPEN QUESTION	CLOSE QUESTION	SIMPLE REFLECTIONS	SCALE
how	did you	sounds like	on a scale of 1 to 10
what	shall you	feels like	
why	does	it seems	
when	do	looks like	

constitute an OPEN-ENDED QUESTION, CLOSE-ENDED QUESTION, SIMPLE REFLECTION, SCALE SENTENCE. 3) The number of labelled ASR-based utterances (N=2181) proved a limitation in using more data-intensive Deep Learning algorithms [82]. Hence, as a possible first step, by leveraging the lexical patterns articulated in the MI literature and experts and accounting for the limitation we implement a Rule-based classifier. This as a possible first step in an iterative system development cycle, where progressively intelligent and data intensive algorithms will be used for implementing increasingly intelligent decision making, as we gather more data.

4.3.1 Challenge

Spoken utterances are characterized by irregularity in structure when compared to formal written discourse which are structured with respect to rules of grammar. The Google Cloud Speech ASR that produces real-time speech to text utterance, although rated at

(WER = 5.6%) [83] produces spelling errors and missed utterances which percolate into spoken utterances. Additionally, environment based technical difficulties like internet connectivity issues or errors in voice signal acquisition from input devices like microphones (low battery, improper positioning) contribute to the errors.

Therefore, the primary challenge in designing ReadMI behavioral coding classifier condenses to mapping unstructured ASR generated spoken utterances, with rules of structured English grammar to identify OPEN-ENDED, CLOSE-ENDED, REFLECTIVE, SCALE and NONE utterances.

4.3.2 Solution

As a first step, dictionaries containing explicit lexical features characterizing different class of utterances were created and added as context parameters [84] to Google Cloud Speech ASR. The context parameters ensure that the ASR assigns a higher weight to words with similar pronunciation and transcribes them with greater accuracy. The context parameters were used to address the problem of missing or misspelling the words by Google ASR. The lexical features as context parameters were obtained as inputs from multiple sources, 1) The MI experts facilitating our study; and 2) Motivational Interviewing literature [85] and formal English Grammar [86]. Examples for lexical features characterizing each class of MI utterances except for the "NONE" class are shown in Table 4.1. An utterance is defaulted to “NONE” class if it does not belong to any of the other classes. The utterance characterizing the scale class, utilized to assess readiness for change,

was characterized by the phrase “on a scale of 1 to 10” was observed in almost 100% of the scale indicating utterances.

Next, certain observations in the ASR generated practitioner utterances show a pattern characterizing the utterances as shown in Table 4.2.

Table 4.2: Patterns in Practitioner Utterances

Observations	Example
If an utterance begins with a closed question keyword then it is most likely a closed question	"Is there anyway I can help you?"
If an utterance contains a closed question keyword and the immediately following word is a pronoun (ex. He, she , you etc.) or a possessive pronoun (ex. Himself, herself, itself) then it’s a closed question	"I am doing great today by the way did you sleep well yesterday?"
If an utterance begins with an open question keyword and is followed by a pronoun or possessive pronoun then it is a statement	"How often you run is more important than how many miles you run"
If an utterance contains an open keyword and the succeeding word is a pronoun and or possessive pronoun then it is most likely a statement	“I am interested in knowing what your thoughts on alcohol addiction are”

If an utterance begins with an open question keyword then it is most likely an open question	"How are you doing today?"

The patterns from Tables 4.1 and 4.2 inform us with clues that can be incorporated into our rules. Typical parts of speech like "Nouns", "Pronouns", "Possessive Pronouns" in the "Observations" column in Table 4.2 provide insights on the effectiveness of parts of speech (POS) in identifying syntax-based patterns. To leverage POS-based patterns from spoken utterances, in line with English grammar an Apache OpenNLP POS tagger in Java Programming Language [87] is used. Furthermore, to identify the obtained POS tags from the Apache OpenNLP POS tagger the Penn Treebank [88] was used. A typical NLP based POS tagging method is shown in Table 4.3 with a legend describing individual tag description according to the Penn Treebank in Table 4.4.

Table 4.3: Computer Generated NLP based Part of Speech (POS) Tagging

Sentence	It	looks	like	you	are	well
POS Tags	[PRP]	[VB]	[IN]	[PRP]	[RB]	[JJ]

Table 4.4: Tag description according to Penn Treebank

Tag	**Description**

PRP	Personal Pronoun
VB	Verb (base form)
IN	Preposition
RB	Adverb
JJ	Adjective

4.3.3 Categorization of Utterances into Simple, Complex, and Compound Utterances

The ASR generated spoken utterances are broadly categorized into 1) Simple 2) Complex 3) Compound utterances, not to be confused with typical sentence types in English grammar. Utterance within the context of this research shares similar definitions to a “turn”, a continuous piece of speech beginning and ending with a pause. This categorization scheme is defined based on unique characteristics observed in training dataset.

1. Simple utterances: They are characterized by the absence of punctuation, example, period (.) and conjunction words like "and", "because”. Additional examples of conjunction words can be found here [86]. Simple utterances are mostly used to convey a complete thought. A scale-based utterance may lexically appear to have two complete thoughts (for example: Thought 1: "On a scale of 1 to 10 with 1 being not ready for change and 10 showing maximum readiness for change." which would

belong to class "NONE" and " Thought 2: Where do you stand?" which is an OPEN question) but in the present adaptation of MI for ReadMI scale sentences are assumed to convey one thought or function and that is to assess readiness for change. Some examples of simple utterance-based classification results are shown

Table 4.5: Simple Utterances

Utterances	Label
"how long have you been smoking.."	OPEN
"**do** you feel to start making these changes"	CLOSE
"**it sounds like** they are not big fans of cigarette smoking"	REFLECTIVE
"so **scale of 1 to 10**, how likely are you to give up drinking"	SCALE
"okay that makes sense."	NONE

in Table 4.5. The lexical features emblematic of each simple utterance is highlighted for convenience. The decision-making process for simple sentences is quite straightforward. The rule-based classifier identifies the characterizing lexical feature in an utterance and assigns a label accordingly.

2. Complex utterances: Within the context of this research, complex utterances are characterized by absence of conjunction words or punctuation like a period. It mainly consists of utterances that contain lexical features that signal utterances to belong to one class but grammatically point to a different class.

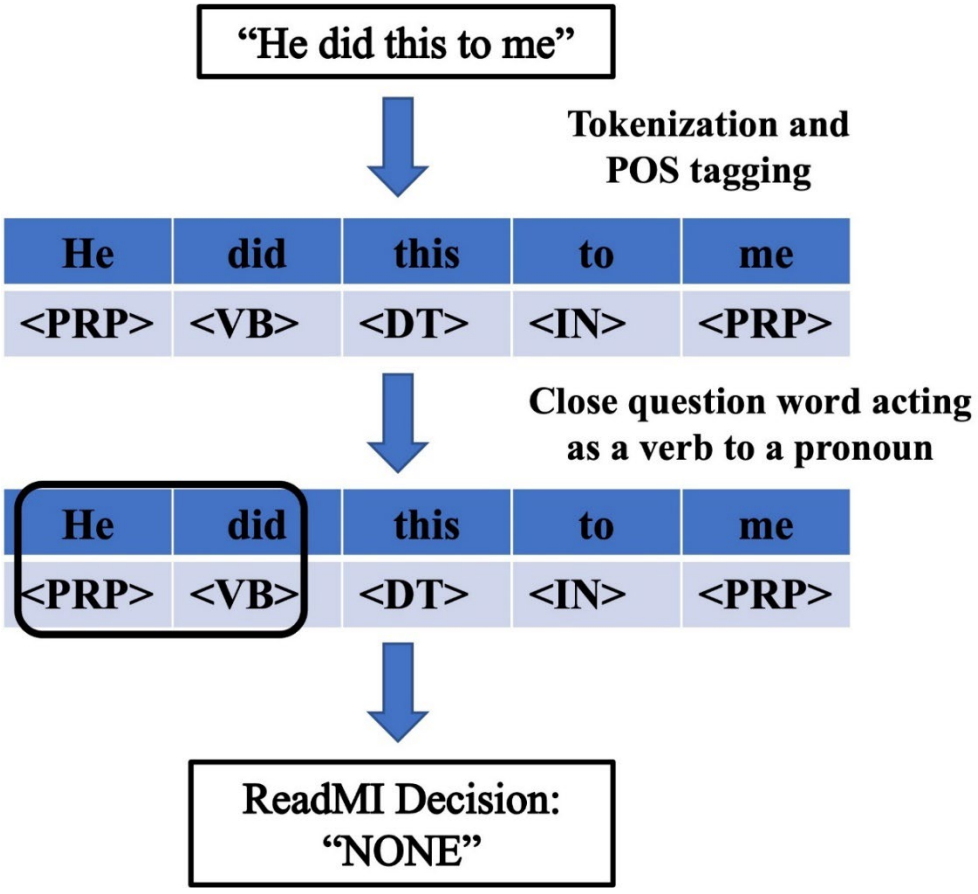


Figure 4.5: Complex Utterance Decision Flow with Part of Speech Tagging

For example, consider an utterance like "He did this to me", due to the presence of the lexical feature "did", a blind pattern matching algorithm may assign a CLOSE ENDED question tag to this utterance. However, by leveraging the POS of utterance we understand that "did" is not a question word but a verb supporting a pronoun "He", thereby disambiguating the utterance to a statement. Hence a "None"

behavior code is assigned. For clarity a sample classification decision flow diagram is shown in Figure 4.5.

3. Compound utterances: As shown in Table 4.6 compound utterances are characterized by combinations sub-utterances like simple utterances or simple utterances with complex utterances, for example, combination of Reflective utterance with an open question, closed question followed by open question or vice versa, statement followed by a question etc. These combinations of sub utterances are joined together by the presence of Conjunction words like “and”, “or”, “yet”, “ok” etc. and the presence of the punctuation: period (.) or comma (,) separating two utterances. Each part of the compound sentence is analyzed recursively to produce individual intermediate results which are class labels assigned to the constituent sub utterances before the final label is assigned according to the hierarchical utterance class order shown in Table 4.7. Based on the granularity of the feedback the facilitator wants to provide the trainee practitioner, a full sequence of codes may be presented, alternatively, a single combined code can be assigned to the entire utterance.

In training sessions, MI trainees are encouraged to use certain utterance types more to maximize empathy in the MI process. These utterances are prioritized over the others. In Table 4.7, Reflection is ranked at the top due to its primary role in encouraging the client to self-analyze and resolve internal conflicts or ambivalence to behavioral change. Next,

Scale utterances are encouraged to assess the patient readiness to change by quantification through 1 to 10. Finally, close question utterances elicit a “yes/no” answer from the client, thereby diminishing additional information gathering potential for the practitioner. Furthermore, they are commonly spoken and are intuitive when compared to Open-ended questions. Hence in a training environment to increase practitioner awareness towards Close-ended questions over open question utterance, classifier priority was set to penalize MI trainees on close questions. Figure 4.6 presents a detailed depiction of compound sentence decision making process for the utterance “It looks like you are drinking when nervous. How do you feel about that?”. In summary, the behavioral coding classifier leverages the lexical patterns determined within MI literature and MI experts to classify real-time ASR generated spoken utterances into OPEN, CLOSE, REFLECTIVE, SCALE and NONE classes. The challenges peculiar to real-time ASR generated spoken utterances are minimized by using lexical patterns as context phrases and leveraging POS metadata to disambiguate complex sentences that differ grammatically from the lexical features constituting them.

Within the broader context of this research work the behavioral codes generated from the classifier help facilitators in automatic coding of Motivational Interviewing sessions, a crucial aspect of the "instantaneous feedback" feature of ReadMI. Further, incorporating MI’s technical elements-based feedback consisting of behavior codes improves the comprehensiveness of the MI feedback. The adaptive system design, prototype and behavioral coding classifier ensure behavioral code assignment in a flexible

collaborative learning environment.

Table 4.6: Compound Utterances

Utterances	Label
"It looks like your are drinking when nervous. How do you feel about that?"	REFLECTIVE followed by OPEN question
"Okay, so it sounds like you use smoking as an outlet for yourself for pleasure or to relaxation, right?"	REFLECTIVE followed by CLOSE question
"Hi, I am Dr. Hudson, how are you doing today?"	STATEMENT followed by OPEN question

Table 4.7: Utterance Class Hierarchy for Compound Sentence Decision-Making

Priority	Utterance Class
1	Reflective
2	Scale
3	Close
4	Open
5	Statement

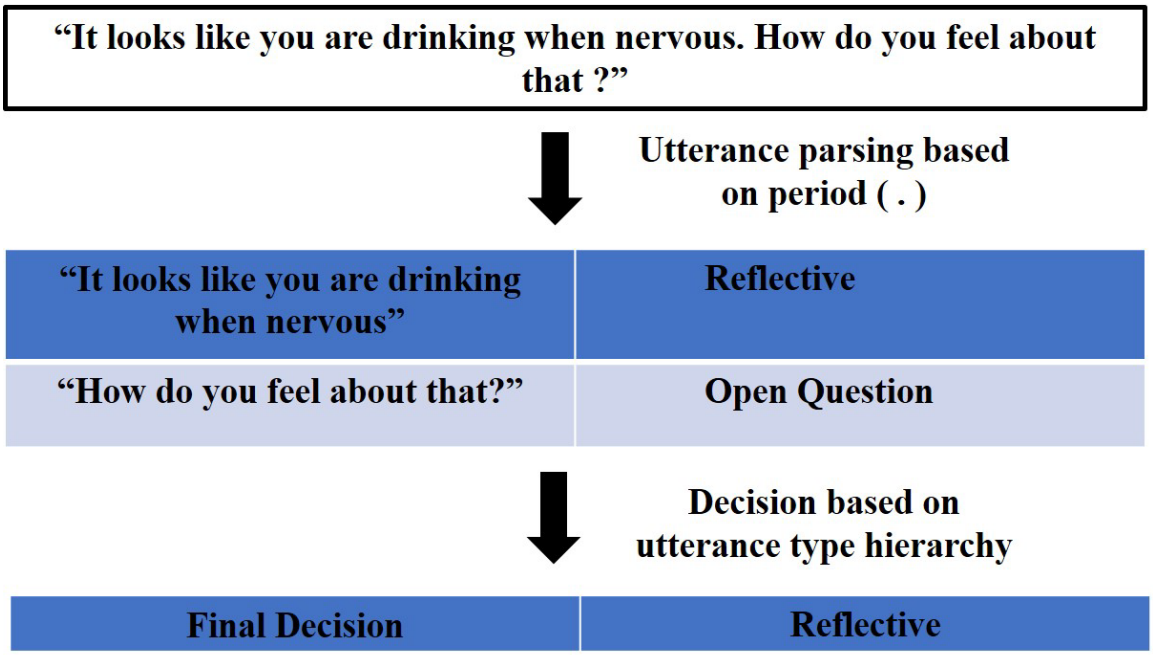


Figure 4.6: Compound Sentence Decision-Making Process

4.4 Experimental Study

An experimental study was designed and conducted to answer the question:

"Does quantitative metrics-based feedback from ReadMI, when combined with facilitator's subjective feedback significantly improve practitioner MI skills?"

Students are introduced to Motivational Interviewing in the first year of medical school and their training is reviewed during second year and the "clerkship bootcamp" just prior the beginning of their clerkship year. As part of their third-year curriculum students must take part in six-week rotations for each specialty. Each rotation consists of 15 students. During the family medicine rotation, third year students at Wright State Boonshoft School

of Medicine, Dayton, Ohio must undergo a 90-minute MI training session with an MI facilitator. Two students are scheduled for each session, where each student enacts the role of a doctor and a patient twice, resulting in four roleplays. The students are also provided with a preparatory material before each session consisting of an MI review sheet and a video. At the beginning of the academic year, 8 family medicine cohorts each consisting of 15 students are randomly assigned to a control group (about 48%) and an intervention group (about 52%). The demographics among the experimental study participants are shown in table 4.10.

After each roleplay, students in the intervention group were given ReadMI based feedback to test for improvements in MI performance from first to second MI session. ReadMI based feedback was given to the students of the control group after all four sessions to ensure that these metrics would not influence their roleplays. The MI facilitator played the role of a timekeeper and provided MI based qualitative feedback to both groups. Each roleplay consists of a simulated case, representing one of several prepared scenarios to each student, for example, 1) 35- year- old smoker who has been smoking for 20 years who uses smoking to relieve stress 2) 40-year-old with good job and decent marriage. Doesn't drink during the week but drinks heavily on the weekends, to the extent that most weekends are just a blur. During their role plays, each student's interview was captured and analyzed by ReadMI, initially planned, and developed for face-to-face use. The COVID- 19 [12] pandemic was the catalyst for adapting ReadMI for use on a virtual platform through JITSI [15] ensuring continuity with MI training sessions. MI metrics

produced by the ReadMI application were presented to the medical students as feedback, which consisted of number of open- and

Table 4.8: Demographics Among Experimental Study participants (N = 125)*

	N (%)
Age – mean (std)	26.7 (3.0)
Group Status	
Intervention	63 (51.6)
Control	59 (48.4)
Race	
Asian	24 (19.7)
Black	18 (14.8)
Hispanic	5 (4.1)
White	70 (57.4)
Other	5 (4.1)
Gender	
Male	55 (45.1)
Female	67 (54.9)
Native Language	
English	110 (90.9)

Non-English	11 (9.1)
*3 participants did not provide data on demographics	

number of reflective statements, and use of the change ruler. To validate the performance of ReadMI, two MI training facilitators read the transcripts created by the ReadMI application and rated physician responses as reflective, open-ended, closed-ended, scale (i.e., change-ruler), or none as shown in Table 2.1. Interrater reliability statistics were used to determine the accuracy of the ReadMI application's analysis of clinician responses. In terms of protocol, this research study involved measuring the effectiveness of comparison of instructional techniques. Hence the Wright State University Institutional Review Board deemed the study to be exempt from human subjects.

Table 4.9: Summary of Data from ReadMI Experimental Study

ReadMI Behavioral Codes	Utterance Examples
Total number of training sessions	88
Total number of roleplays	352
Total number of medical students	125
Total number of "Doctor" role utterances	2181

Table 4.10: Class Distribution of MI Utterance Labels from ReadMI Experimental Study

Utterance Class	Class Distribution
REFLECTIVE	440
OPEN	711
CLOSE	289
SCALE	75
NONE	666

4.5 Data

88 MI training sessions yielding 352 roleplays from 125 medical students produced a total of 2181 practitioner/clinician role utterances. As the primary application of ReadMI is to measure practitioner skill acquisition analysis is restricted to practitioner utterances in the resulting dialogue dataset. Each of these utterances were also rated by MI expert raters independently based on our coding scheme shown in Table 2.1. The data is summarized in Table 4.8 and the class distribution of MI utterance labels is described in Table 4.9. The ASR generated data collected in the experimental design was a significant first step in automating the process of observational coding. It demonstrates that present day ASR

algorithms have sufficient accuracy (WER = 5.6%) [83] to generate human- readable conversational utterances which sometimes can go upto 50 to 60 words in length. It further alludes to the fact that real-time speech to text transcription is possible alleviating the drawback of unscalable and expensive manual transcriptions.

ASR generated data collected in the Clerkship bootcamp as reported in the experimental study section is a significant first step in automating the process of observational. It demonstrates that present day ASR algorithms have sufficient accuracy (WER = 5.6%) to generate human- readable conversational utterances which sometimes can go upto 50 to 60 words in length and real-time speech-to-text transcription is possible alleviating the drawback of unscalable and expensive manual transcriptions.

4.6 Results

All N=2181 clinician-utterances labelled by two MI expert facilitators independently, were used to test the accuracy of the Behavioral Coding classifier. Classifier accuracy was determined through typical metrics Precision, Recall and F1 score, see Table 4.11. Furthermore, to compare the classifier accuracy with senior MI expert facilitators Inter-rater agreement metric, Cohen's kappa [89] was used to establish a benchmark. Cohen's kappa measurements are presented for both overall classifier performance and individual utterance class level performance and further extended to measure the agreement between two expert raters: a senior MI expert (>30 years of MI training experience) and junior MI expert (about 5 years of MI training experience).

Table 4.11: ReadMI Performance Metrics

	Precision (%)	Recall (%)	F1 Score (%)
Open-ended question	0.918	0.709	0.800
Close-ended question	0.730	0.626	0.674
Reflective statement	0.514	0.730	0.603
Scale statement	0.911	0.680	0.779
None (Physician speaking)	0.761	0.805	0.782

Table 4.12: Decision Matrix for ReadMI (read from Top) v. Senior MI Expert (read from Left)

	Open	Close	Reflection	Scale	None	Total
Open	504	29	105	4	69	711
Close	8	181	77	1	22	289
Reflection	27	16	321	0	76	440
Scale	3	3	17	51	1	75
None	7	19	104	0	536	666
Total	549	248	624	56	704	2181

4.6.1 ReadMI Performance Metrics

From Table 4.11 we infer that, the Behavioral coding classifier identifies Open-ended questions with highest accuracy (F1 Score = 80.0%) and precision = 91.8% and Recall of 70.9% implying a relatively balanced false positives and missed utterances, followed by scale statement with F1 Score = 77.9% and precision 91.1% and Recall at 68.0%. It identifies the reflective statements, characterized by a high degree of prosodic information with least accuracy at F1 score = 60.3%, with highest false positives at 51.4% at Recall = 73.0% second highest after "NONE" class at Recall 80.5%.

4.6.2 Interrater Agreement Between ReadMI v. Senior MI Expert

The overall Cohen's kappa across all classes scores a healthy 0.638 with Substantial agreement which tells us that ReadMI is comparable with human raters in classifying the ASR based utterance into one of the 5 MI metrics. From Table 4.13 it can be inferred that better performance is observed for open-ended (kappa = 0.721), closed-ended (kappa = 0.629), Scale (kappa = 0.772) and None (kappa = 0.683) utterances, which is expected and can be attributed to more structured syntax for these utterances, while kappa = .480 moderate agreement is reported for reflective statement.

4.6.3 Interrater Agreement Comparison between ReadMI v. Senior Expert and Senior Expert v. Junior Expert

The agreement between two expert raters was analyzed to assess the rule-based classifier's reliability with rater's agreement scores as the benchmark. As summarized in

Table 4.13: ReadMI-Senior MI-Expert Cohen's Kappa calculati

	ReadMI-Senior MI-Expert		
	% of Agreement	**Cohen's Kappa Statistic**	**Interpret the Cohen's Kappa**
Open-ended question	0.885	0.721	Substantial agreement
Close-ended Question	0.920	0.629	Substantial agreement
Reflective Statement	0.807	0.480	Moderate agreement
Scale statement	0.987	0.772	Substantial agreement
None (Physician speaking)	0.863	0.683	Substantial agreement

Table 4.14, similar to ReadMI, the expert rater demonstrate better agreement in open-ended (kappa = 0.833), closed-ended (kappa = 0.798), Scale (kappa = 0.906) and None (kappa = 0.854) with comparatively lower agreement on reflective statements (kappa= 0.625). By comparing the interrater agreement between ReadMI v. Senior Expert and Senior v. Junior Expert, it is observed that ReadMI closely tracks the performance of human expert, achieving either the same level of agreement (substantial agreement in the class of Close-ended question) or one notch lower (in all other classes).

In both the control and intervention groups, the average percentage of time the doctor spoke (48.2% at session 1 versus 41.8% at session 2) decreases and increases in the

Table 4.14: Comparison of Cohen's kappa scores

	Cohen's kappa (Senior v. Junior expert)	Interpret the Cohen's kappa (Senior v. Junior expert)	Cohen's Kappa (ReadMI v. Junior expert)	Cohen's kappa Interpretation (ReadMI v. senior expert)
Open-ended question	0.833	Almost perfect agreement	0.721	Substantial agreement
Close-ended Question	0.798	Substantial agreement	0.629	Substantial agreement
Reflective statement	0.625	Substantial agreement	0.480	Moderate agreement
Scale statement	0.906	Almost perfect agreement	0.772	Substantial agreement
None (Physician speaking)	0.854	Almost perfect agreement	0.683	Substantial agreement

percent of questions that were open questions (62.0% at session 1 and 69.0% at session 2). At the first session, there were some differences in the ReadMI metrics between the control and intervention groups, with the control group speaking longer (50.4 versus 46.1; $p = .04$), having more closed questions (5.2 versus 3.4; $p = .0002$), and having a lower percentage of questions being open (55.0% versus 68.0%; $p = .0003$) compared to the intervention group. The significant results of the experimental study with ReadMI is observed as follows: After training, the intervention group had significantly fewer close-ended questions (2.8 versus 5.0; $p<.0001$), and a significantly higher percentage of questions that were open-ended (80.0% versus 64.0%; $p = .0005$). There were no significant differences

observed between groups on the number of reflective statements or the use of change scales.

4.7 Analysis

4.7.1 Open-ended Question, Close-ended Question, Scale Utterance, None

Open-ended questions were mostly characterized by the presence of patterned explicit lexical features like “How. . . ”, “Why. . . .”, “What. . . ”. The corresponding high precision score of 91.8% indicating least false positives attests to the presence of chosen features. Further, ReadMI’s substantial agreement with the expert for open-ended questions throws light on its ability to use Parts of Speech to identify the complex utterances which are more nuanced than simple sentences but easily perceived by human raters. Similar trend extends to scale based statements usually characterized by the phrase "On a scale of 1 to 10..". The substantial agreement with the MI expert and the high precision score attests to the observation. Closed questions are characterized by utterances like “would you say. . . ?”, “Did you. . . .?”, “have you. . . .?”. Although ReadMI shows substantial agreement with the MI expert, it is comparatively lesser to open questions and scale statements. From the decision matrix in Table 4.12 it can be inferred that ReadMI identifies 77 close questions as reflections. Utterances like "Okay, you tried anything in the past?" which are absent of closed question keywords are characterized as "Reflective" but are rated as "Closed" by the experts. Although this first iteration of system design focusses on exploiting explicit lexical patterns, it is well established that human speech is characterized by additional

features like prosody, which highlights "How?" an utterance is articulated with regards to inflection, intonation, and pitch. Statements that are grammati- tically structured as simple sentences in addition of prosodic elements exhibit different functionality. Further utterances like “do you have you had a chance to exercise?”, which are termed as self-realization utterances where candidates correct themselves mid-utterance from forming closed questions to MI compatible open question. These utterances are composed of multiple sub-utterances constituting signature lexical features. Here since "....have.." occurs after a pronoun "you" and the utterance begins with a close question keyword "do" ReadMI identifies the utterance as close question as close questions have higher preference than open, from Table 4.7, however, the experts identify the question as "Open" when intonation and facial expression are considered. Therefore, abstract spoken utterances whose functions must be characterized by prosodic and visual cues pose a challenge to the Behavioral coding classifier of ReadMI. Further, utterances like "(ASR missed "on a scale of 1 to")....10. Where would you say you are right now in ready already. Would you be to make a change with your drinking?" - here the ASR misses out the initial part of the utterance thereby contributing to an erroneous classification. Since ReadMI shows substantial agreement with Open-ended questions, closed-ended question, Scale based sentence and the default utterances in NONE one may conclude that these utterances are feature rich in explicit lexical features. A sophisticated classification algorithm would likely improve accuracy, for example, by identifying the prosody.

Table 4.15: ReadMI Metrics among Medical Student Participants (N = 125)

	Total (n = 125)		Intervention (n = 65)		Control (n = 60)	
	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2
	Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)
Doctor Speak Time	48.2 (11.3)	41.8 (11.0)	46.1 (11.2)	40.1 (10.0)	50.4 (11.1)	43.5 (11.8)
Open Questions	6.8 (3.2)	8.3 (3.6)	7.0 (3.1)	8.5 (4.1)	6.6 (3.3)	8.1 (3.0)
Closed Questions	4.3 (2.8)	3.9 (2.9)	3.4 (2.4)	2.8 (2.3)	5.2 (2.9)	5.0 (3.1)
Total Questions	10.9 (4.4)	12.2 (4.3)	10.2 (4.1)	11.3 (4.4)	11.6 (4.6)	13.1 (4.0)
Reflections	6.1 (3.5)	6.5 (3.5)	4.6 (3.5)	5.0 (3.4)	7.8 (2.6)	8.1 (2.7)
Scale	0.7 (0.8)	0.7 (0.7)	0.8 (0.9)	0.8 (0.6)	0.7 (0.7)	0.6 (0.7)
Ratio of Open to Close Questions	2.6 (2.7)	3.7 (3.6)	3.2 (3.1)	4.8 (4.1)	1.9 (2.0)	2.5 (2.5)
Ratio of Reflection Questions	0.6 (0.4)	0.6 (0.4)	0.5 (0.4)	0.5 (0.4)	0.8 (0.5)	0.7 (0.3)
Percent of Open Questions	62.0 (20.0)	69.0 (20.0)	68.0 (20.0)	80.0 (20.0)	55.0 (19.0)	63.0 (17.0)

4.7.2 Reflective Utterances

They are contextual in function, characterized by a combination of prosodic and lexical features. As ReadMI is designed to identify prominent lexical patterns in utterances it is unable to capture this context as evidenced by moderate agreement between ReadMIand the expert rater, (kappa = 0.480) refer Table 4.13 and the increased false positives due to Precision score of 51.4%, Table 4.11. The comparatively lesser kappa scores between the senior and the junior expert (0.625) shows slightly higher disagreements unsurprisingly alluding to the complexity of reflective sentences. Complex reflection which are a variation of reflections are characterized by a higher degree of contextual and prosodic information. Hence such utterances are identified by ReadMI as NONE while identified as Reflective statements by human raters. The nature of prosodic information noticed in complex reflections can be attributed to acoustic prominence [90]. Acoustic prominence is achieved by varying the duration, pitch, pitch movement, and amplitude [90] of selective words like “Those are all very good and yummy, but not so great for us, right?”. This utterance appears syntactically close-ended question as it elicits a “yes/no” answer from the client. However, it was labelled as “Reflective” statements by the Human expert who rate utterances after every session by viewing the session audio-visual recordings. This multi-signal information stream characterizing reflective statements, relates to their contextual nature. Although lexical features like “. . . sounds like. . . .”, “. . . .seems like. . . ..”, “And you. . . .” identify reflective statements, it is unambiguously clear that prosodic and visual components are critical for identification.

In summary, although the rule-based classifier whose rules are designed based on explicit lexical features identified from MI literature and formal English grammar, it is to be interpreted as a first step to demonstrate classifier design in a ReadMI-like prototype. Preliminary data analysis highlights the need for incorporating prosody and identifying more sophisticated patterns which characterize the unstructured and contextual nature of human speech.

5 DISCUSSION

CSCW is first and foremost a design-based field whose primary goal is to understand the social and behavioral nature of social interactions and incorporate these characteristics into design of computing systems so that intelligent and seamless human-machine teaming can be achieved. A key component of these interactions is verbal communication and disruptions in verbal communication through interruptions and delayed feedback will impede collaborative activities in cooperative work settings. The crux of this research work lies in its demonstration of how social and behavioral characteristics that define collaborative interactions in the form of task-oriented dialogues amongst human collaborators can be translated into design of computing systems that focusses explicitly on solving said communication problems and enhance collaborative experiences. More importantly, it demonstrates that combinations of mature and ubiquitous technologies that exist today can be combined to create fully functional intelligent prototypes that can be tested in actual collaborative interactions. The infusion of social and behavioral characteristics of interactions means the GUI design, system design, utterance analysis and user experience are influenced by these design considerations. This section briefly summarizes the implications of ReadMI and ACE-IMS design from the perspectives of 1) User Interface 2) Social Presence 3) Task-oriented dialog features 4) User feedback 5) User Feedback and Behavior Modification.

5.1 User Interface

The primary requirement of CSCW based applications is to capture the interactional component of collaborative work and contribute to its enhancement in terms of scale and computational power. Both ACE-IMS and ReadMI demonstrate this requirement by capturing the dialogic form of interaction between participating individuals through GUI based design elements which show synchronous and chronological representation of dialogue flow enabled by real-time ASR and picturized by "speech balloons" based GUI elements. In case of intelligent interruptions dissemination, ACE-IMS, embedded in the synchronous chat application (which are inherently collaborative by nature) where the dialogue is predominantly task-oriented, can dispense interruptions at the points of interruptibility. The sequentially arranged dialogic flow of conversations assist participating interlocutors when they context-switch back to the existing task by providing visual evidence of previous utterances. Revaluation of points of interruptibility chronologically allows users to reconfigure ACE-IMS to suitably interrupt when conversations of high priority are carried out despite the presence of affirmation cues. For example, disseminating advertisements triggered by word contexts that enhance dialogic experience. In ReadMI's case, visual dialogic flow of conversation, coupled with the color changes in speech bubbles highlighting reflective and scale sentences and bifurcation of open and close ended questions, provides pedagogical evidence to the facilitator. As a post MI session training aid, practitioners can point to the exact point in conversation history and suggest possible changes to the phrasing of the subsequent utterances to improve MI

delivery, for example if a trainee practitioner during a key moment of interaction with the model patient phrases a question as closed in the session, the instructor can refer to that time-instance and suggest an open-ended way to phrase the question.

5.2 Social Presence

The gap between our knowledge of everyday social interactions and translation of this knowledge over to system design to enhance cooperation and collaboration is well articulated in [91]. For successful design of collaborative systems incorporating social presence is a priority [78]. ACE-IMS and ReadMI demonstrate system design elements that aims to minimize socio-technical gap. In addition to the dialogic flow-based representation of conversations, a system design which is primarily mobile based, ergonomically convenient and representative of majority of communication exchanges is brought to fruition through the mobile tablet-based implementation on the Android platform. As it is shown that face-to-face communication increase social presence, JITSI based distributed remote implementation of ReadMI is augmented with a video conferencing implementation which allows for both facilitators and practitioners to observe facial expressions along with voice to deliver and assess empathetic MI delivery. data obtained from such applications, fundamentally designed to incorporate social interactions, have to potential to generate data that can be used to fine tune ReadMI and ACE-IMS.

5.3 Language Features Used for Algorithm Development

Language is a functional instrument of social interactions. Both ACE-IMS and ReadMI incorporate language features that highlight functional aspects in the form of discourse markers that parameterize task-oriented dialogues. These discourse markers were used to identify points of interruptibility by ACE-IMS. As affirmation cues signal task transitions, the identified lexical affirmation cues accounting for 69.9%, 62.9% and 88.1% of the set of manually annotated task boundaries for UMT training, UMT test and Tangram test dataset, suggest that identified affirmation cues account for majority of the task boundary utterances and can contribute to intelligent interruption dissemination, provided that the task structure of task–oriented dialogues are similar to UMT and Tangram datasets. In case of ReadMI, discourse markers were used in combination with part of speech tagging to classify utterances based on behavioral codes used in MI for ReadMI. The usage of traditional question identifying discourse markers, particularly open question identifying ones like "what…", "who…", reflective sentence identifying utterances like "it seems like…" and scale-based sentences like "…on a scale of 1 to 10…." signal the identification of different utterances in MI based task-oriented dialogue.

5.4 User Feedback and Behavior Modification

To understand the user-based feedback of the computing systems used to facilitate social interactions ReadMI was chosen as an example. Research shows that immediate feedback is necessary for skill establishment and development [58]. By deploying ReadMI

in an actual collaborative learning environment - medical education settings as explained in Section 4.4, we present the findings on the effects of its instantaneous feedback, its perception by both the facilitators and the practitioners in the workflow, which are critical in understanding the human computer interface aspect of cooperative workflow.

The subjective qualitative feedback from the practitioners and facilitators in tables 5.1 and 5.2 informs that ReadMI provides a quantified perspective to the seemingly analogous practice of motivational interviewing. The quantitative feedback from the experimental study reported in Section 4.4 shows significantly fewer close-ended questions (2.8 versus 5.0; $p<.0001$) and significantly higher percentage of questions that were open-ended (80.0% versus 64.0%; $p = .0005$). These qualitative and significance results demonstrate the potential of ReadMI in altering the way Practitioner's approach MI. Behavior modification is defined as "the process of changing human behavior over the long term through motivational techniques, positive(rewards) and negative(consequences) reinforcement"[92]. By providing numerical feedback in MI sessions and if used consistently over sufficient period ReadMI shows a strong potential to positively influence practitioner behavior in the direction of MI consistent techniques.

Table 5.1: Medical Student Reviews from Experimental Study

Student Reviews
"The ReadMI experience opened my eyes to many aspects of MI I had not realized. **For one, it made me realize how frequently I use close-ended questions although I personally felt that I hardly ever do.**

It also showed me the importance of making the patient do the work, and that I am simply there to guide the conversation in the right direction. The patient time speaking vs doctor time speaking metric illustrated this clearly. I also learned many new tricks to illicit behavior discrepancies in my patients, and gained a better understanding of how big that gap should be for optimal results. I left the experience knowing what I did well (affirming statements), and what I needed to work on (re-phrasing my questions, guiding the patient to do the work)."
"The ReadMI experience continued to reinforce my positive thoughts regarding motivational interviewing and I will continue to utilize this technique on a regular basis. The qualitative data was comparable to my subjective experience and highlighted areas of improvement. **The qualitative data specifically revealed a need to utilize more reflective statements in my patient interviews.** The ReadMI experience exposed me to the benefits of SBIRT and I will begin to integrate this protocol into my patient interviews. This session will serve as motivation to continue utilizing motivational interviewing techniques when working with patients in need of behavior change."
"The ReadMI experience changed my thoughts on motivational interviewing by implementing the feedback I received from the facilitator and the tool into my second try. Trying it on a real patient would help be certain the changes implemented the second time made the conversation more fruitful, as well as see if it actually changed the patient's motivation. **Receiving the qualitative data compared to my subjective experience disconfirmed my feelings. I thought I would do most of the talking or did not provide enough reflections, however, the tool indicated otherwise.** The new insight I have regarding motivation interviewing is to be attentive to what the patient says, as well as always ask what he/she likes about the certain substance – this is something I had never considered doing in the past and it was valuable to the conversation."

5.4.1 Reviews from Medical Students and Facilitators.

Table 5.1 shows a sample non-exhaustive list of reviews by students who were responding to **"How did the ReadMI experience change your thoughts on Motivational Interviewing? How did seeing quantitative data compared to your subjective experience confirm your feelings? Disconfirm your feelings? What new insights do you have regarding Motivational Interviewing given the experience?"**, and the facilitators in Table 5.2 were responding to **"How did ReadMI affect your workflow?".** From the students' feedback one of main themes that emerges from Table 5.1 is that ReadMI's feedback brings MI interview into perspective, in terms of numbers and the instantaneous dissemination of knowledge appears to elicit the realization or "the lightbulb" moment as mentioned by one facilitator from Table 5.2. Lastly, from the perspective of the facilitators ReadMI seems to offset their workload by providing feedback on the technical components of MI, thereby allowing them to focus on the qualitative aspects if the Practitioner interview process.

Table 5.2: Facilitator reviews from Experimental Study

Facilitator Reviews
"Readmi aides in my facilitation of MI teaching by providing quantitative data back up on the subjective feedback I'm presenting students.It helps make the feedback more real and give the students tangible skills to focus on**. You can truly see the students "lightbulb" go off when they see the metrics back after their first interview.** It's fulfilling as the facilitator to see students learn to be exploratory with patients in a medically directed way. "

"As a training facilitator, the use of **ReadMI allows me to focus on qualitative aspectsof a student's use of the motivational interviewing approach because I am not also trying to track quantitative grammatical metrics. This process reduces the cognitive burden on facilitator.** The benefits to the student are that they receive both ReadMI metrics as well as substantive feedback from me regarding the spirit of motivational interviewing."

6 CONCLUSION

Communication issues in multi-user multi-tasking interactions manifesting through ill-timed interruptions and delayed feedback disrupt the ongoing task and degrade human productivity and affective state and has the potential to pose a serious threat to mutual dependency in cooperative work environments. To address these challenges, this dissertation work proposed ACE-IMS, a real-time interruption management system that utilizes typical affirmation cues present in human-human communication to construct a highly effective and efficient solution for disseminating interruptions at task-boundary. and ReadMI, a real-time dialogue assessment tool that enhances knowledge retention in MI practitioners through introspective real-time quantitative feedback in MI based training. This work, motivated by the observation that task-oriented dialogues are characterized by the presence of discourse markers like affirmation cues, reflective and scale phrases and can be used in combination with expert ruleset to build intelligent cooperative work supporting systems like Interruption Management Systems and Dialogue assessment tools. The proposed solutions have the potential to improve the effort in growing field of computer-supported distributed collaborations that strive to augment human cognitive capability with computational and sensory power of machines, while suppressing the risk of cognitive overload. Thus, the resulting cooperative teaming between humans and machines paves way for a new wave of human-machine teaming applications in stressful distributed multi-tasking environments, such as: emergency and disaster management, law enforcement, telemedicine, and other field operations.

7 FUTURE WORK

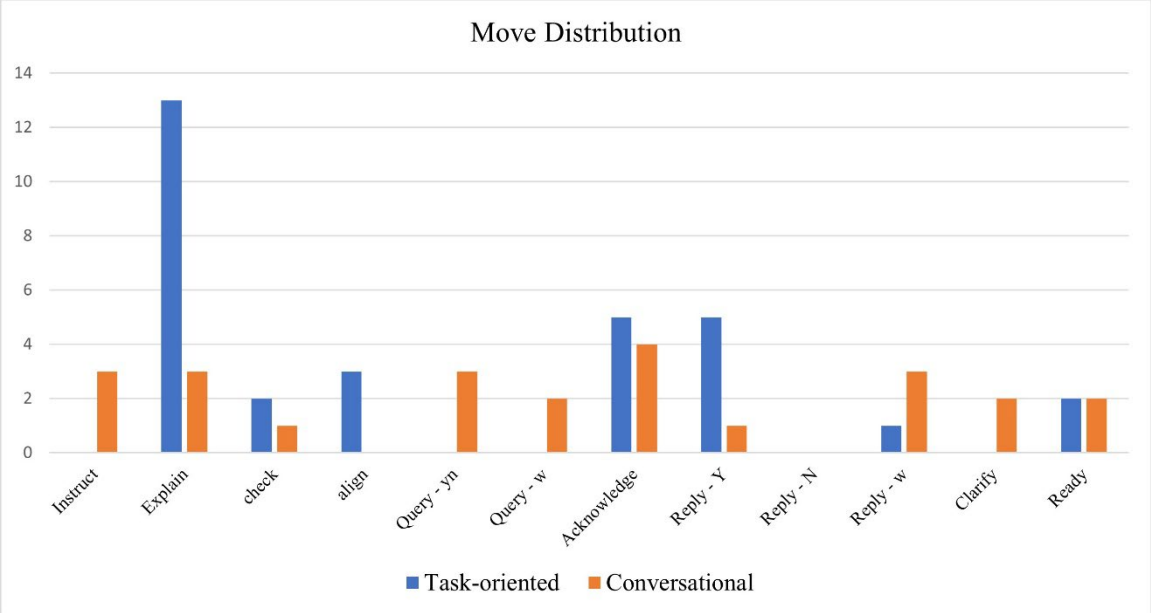


Figure 7.1: Move Structure Comparison between Task-oriented versus Conversational Datasets

Currently, the ACE-IMS and ReadMI prototypes are designed as frameworks, to demonstrate how existing technologies can be melded together to produce a system-based solution address the problem of unwanted interruptions and delayed feedback in multi-user cooperative work. This framework is presented as the first step for incorporating the latest advancements in technologies to build progressively intelligent interruption management systems and dialogue assessment tools. In this subsection, current limitations of ACE-IMS and ReadMI are discussed and potential improvements are suggested.

7.1 Dataset and Acquisition

The nature of the datasets used in ReadMI, and ACE-IMS are currently restricted to task-oriented dialogues constrained by fixed discourse markers. These constraints can be relaxed by incorporating dialogue datasets that resemble everyday conversations. Figure 7.1 shows the difference between conversational versus task-oriented dialogue in terms of frequency of moves or discourse structure. It is clear that intents or moves in task-oriented dialogue is structured while the conversational dialogue demonstrates a higher variance.

7.2 Data Analysis

The rule-based classifier in both ACE-IMS and ReadMI are implemented as proof-of-concepts and are considered as baselines towards future improvements. They can be made progressively intelligent with more data. Data intensive algorithms like Deep learning. and other machine learning algorithms can be implemented in place of Rule-based classifier. Deep learning algorithms are characterized by their capability to unearth complex patterns in datasets, predict data based on sequential patterns, incorporate multi-dimensional datasets like audio and video. ACE-IMS identifies affirmation cues indicative of task boundaries. From Figure 3.3 it was evident that most of the task boundaries are covered by the lexical affirmation cues as listed in Table 3.3. Hence, it would be of interest to analyze the remaining task–boundary utterance, i.e., those without the identified affirmation cues. One direction is to add prosodic features as described in [5] with the identified lexical affirmation cue features into a combined solution. The challenge could be “How to combine prosodic and lexical features to enable real-time operation?”.

Additionally, in this work, the task–oriented dialogues were analyzed on an utterance basis, which could be expanded to include the entire task conversation to explore the turn–taking characteristics and discourse structure. This will lead to a comprehensive understanding of the grounding mechanism. [93], and in turn help identify task-boundaries without the explicit presence of affirmation cues in the utterance. In case of ReadMI, client datasets can be analyzed to identify features for change-talk. As higher open questions, reflections and lower closed question directly predict change-talk, it could be an interesting analysis to assess the manifestation of facilitator-practitioner strength of cooperative work based on client change talk. Further, ReadMI detects comparatively lesser number of behavioral codes than in [34], [61], [94]. The decision to do so was made to minimize information overload to the users [59] as ReadMI was built as an actual usable prototype. However, based on user feedback the number of displayed metrics could be improved with appropriate graphical user interface elements to improve intuitiveness in gleaning information. Further experimental studies in the form of Randomized Control Trials must be conducted to understand the user interface implications of using ReadMI.

7.3 System Design and Application

From a system design and application perspective, Interruption management systems like ACE-IMS can be made available as a functional component of AI assistants and messaging applications, where these applications can recommend or suggest information by interrupting at the right time and help make conversations more informative

and constructive. ReadMI can be implemented as "Software as A Service"(SaaS) [95]. By leveraging the computational power of cloud computing ReadMI as a service can be licensed by educational institutions, workplaces, or private individuals on a "subscription" based service. The SaaS model built on cloud computing infrastructure has the potential to increase service penetration for broader workforce training.

REFERENCES

[1] K. Schmidt and L. Bannon, "Taking CSCW seriously," *Computer Supported Cooperative Work (CSCW)*, vol. 1, no. 1, pp. 7–40, Mar. 1992, doi: 10.1007/BF00752449.

[2] B. Spurgeon, "Behind Formula One Winners, Hard Work and Ties That Bind," *The New York Times*, Mar. 26, 2015. [Online]. Available: https://www.nytimes.com/2015/03/27/sports/autoracing/behind-formula-one-winners-hard-work-and-ties-that-bind.html

[3] "Snapshot: ATAK increases situational awareness, communication and alters understanding of actions across agencies," *Science and Technology*, Nov. 17, 2017. https://www.dhs.gov/science-and-technology/news/2017/11/17/snapshot-atak-increases-situational-awareness-communication

[4] C. Page, "'Roger, Roger. What's our vector, Victor?' — How pilots use the radios." https://thepointsguy.com/news/how-pilots-use-radios/ (accessed Mar. 05, 2022).

[5] N. S. Peters, "Collaborative Communication Interruption Management System (C-CIMS): Modeling Interruption Timings via Prosodic and Topic Modelling for Human-Machine Teams," Dissertations, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2017.

[6] P. D. Adamczyk and B. P. Bailey, "If Not Now, When? The Effects of Interruption at Different Moments Within Task Execution," in *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems, Vienna, Austria, 2004, pp. 271–278.

[7] B. P. Bailey and J. A. Konstan, "On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state," *Computers in Human Behavior*, vol. 22, no. 4, pp. 685–708, Jul. 2006.

[8] S. T. Iqbal and B. P. Bailey, "Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, Portland, OR, USA, 2005, pp. 1489–1492.

[9] H. Smith and S. Higgins, "Opening classroom interaction: the importance of feedback," *null*, vol. 36, no. 4, pp. 485–502, Dec. 2006, doi: 10.1080/03057640601048357.

[10] J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*, New Ed. Cambridge, UK: Cambridge University Press, 1970.

[11] E. Arroyo and T. Selker, "Attention and Intention Goals Can Mediate Disruption in Human-Computer Interaction," in *INTERACT 2011: Human-Computer Interaction – INTERACT 2011*, Berlin, 2011, vol. 6947, pp. 454–470.

[12] C. A. Monk, D. A. Boehm-Davis, and J. G. Trafton, "The Attentional Costs of Interrupting Task Performance at Various Stages," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2002.

[13] M. Czerwinski and E. Horvitz, "Instant messaging: Effects of relevance and timing.," in *People and computers XIV: Proceedings of HCI*, 2000, vol. 2, pp. 71–76.

[14] M. Czerwinski and E. Horvitz, “Instant Messaging and Interruption: Influence of Task Type on Performance,” in *OZHI 2000 conference proceedings*, 2000, vol. 356, pp. 361–367.

[15] F. R. Zijlstra, R. A. Roe, A. B. Leonora, and I. Krediet, “Temporal factors in mental work: Effects of interrupted activities.,” *Journal of Occupational and Organizational Psychology 72*, vol. 72, no. 2, pp. 163–185, 2010.

[16] E. Cutrell, E. Horvitz, and M. Czerwinski, “Notification, disruption, and memory: Effects of messaging interruptions on memory and performance.,” vol. 1, p. 263.

[17] J. G. Kreifeldt and M. E. Mccarthy, “Interruption as a test of the user-computer interface,” in *Proceedings of the Seventeenth Annual Conference on Manual Control*, 1981, pp. 655–667.

[18] D. C. McFarlane, “Interruption of people in human-computer interaction: A general unifying definition of human interruption and taxonomy.,” Cornell University, Office of Naval Research Arlington, VA 22217-5660, 1997.

[19] B. P. Bailey and S. T. Iqbal, “Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management,” *ACM Transactions on Computer-Human Interaction*, pp. 1–28, 2008.

[20] A. Edwards *et al.*, “Synchronous communication facilitates interruptive workflow for attending physicians and nurses in clinical settings,” vol. 78, no. 9, pp. 629–637, Sep. 2009.

[21] L. Hall *et al.*, “Interruptions and Pediatric Patient Safety,” *Journal of Pediatric Nursing*, vol. 25, no. 3, pp. 167–175, Jun. 2010.

[22] E. M. Altmann and J. G. Trafton, “Task interruption: Resumption lag and the role of cues.,” in *Annual Meeting of the Cognitive Science Society*, 2004, vol. 26.

[23] D. C. McFarlane, “Coordinating the Interruption of People in HumanComputer Interaction,” 1999, vol. 99, pp. 295–303.

[24] J. Laarni, H. Karvonen, S. Pakarinen, and J. Torniainen, “Multitasking and Interruption Management in Control Room Operator Work During Simulated Accidents,” in *Engineering Psychology and Cognitive Ergonomics - 13th International Conference, EPCE 2016 and Held as Part of HCI International 2016*, Toronto, Canada, Jan. 2016, pp. 301–310.

[25] S. Rollnick, J. Fader, J. Breckon, and T. B. Moyers, *Coaching Athletes to Be Their Best - Motivational Interviewing in Sports*. The Guildford Press, 2020.

[26] S. Rollnick and W. R. Miller, “What is Motivational Interviewing?,” *Behavioural and Cognitive Psychotherapy*, vol. 23, no. 4, pp. 325–334, 1995, doi: 10.1017/S135246580001643X.

[27] W. R. Miller and S. Rollnick, *Motivational Interviewing: Preparing people for change*, 2nd ed. New York: Guilford Press, 2002.

[28] Wi. R. Miller and G. S. Rose, “Toward a Theory of Motivational Interviewing,” *The American psychologist*, vol. 64(6), no. 1, Mar. 2009, doi: 10.1037/a0016830.

[29] M. C. Villarosa-Hurlocker, A. J. O'Sickey, J. M. Houck, and year = Theresa B. Moyers, "Examining the Influence of Active Ingredients of Motivational Interviewing on Client Change Talk," *Journal of Substance Abuse Treatment*, vol. 96, pp. 39–45, doi: 10.1016/j.jsat.2018.10.001.

[30] L. Harasim, "Online Education: Perspectives on a New Environment," 1990.

[31] Q. Wang, "Design and evaluation of a collaborative learning environment," *Computers and Education*, vol. 53, 2009, [Online]. Available: https://doi.org/10.1016/j.compedu.2009.05.023

[32] C. R. Rogers, *Counseling and Psychotherapy*. Rogers Press, 2007.

[33] H. Kirschenbaum, "Carl Rogers's Life and Work: An Assessment on the 100th Anniversary of His Birth," *Journal of Counseling & Development,* pp. 116–124, 2004.

[34] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, p. 49, Apr. 2014, doi: 10.1186/1748-5908-9-49.

[35] D. Kahneman and G. Klein, "Conditions for intuitive expertise: A failure to disagree.," *American Psychologist*, vol. 64, no. 6, pp. 515–526, Jun. 2009, doi: https://doi.org/10.1037/a0016755.

[36] "How much are transcriptions rates and what does it cost?" https://www.thumbtack.com/p/transcription-

rates#:~:text=Generally%20you%20can%20expect%20to,rate%20of%20%2415%2D%2430.

[37] "Cloud Speech-to-Text - Speech Recognition | Cloud Speech-to-Text," *Google Cloud*. https://cloud.google.com/speech-to-text/ (accessed Oct. 07, 2019).

[38] R. M. Shingleton and T. P. Palfai, "Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research," *Patient Education and Counseling*, vol. 99, no. 1, pp. 17–35, 2016, doi: https://doi.org/10.1016/j.pec.2015.08.005.

[39] L. Lipponen, "Exploring Foundations for Computer-Supported Collaborative Learning," in *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*, Boulder, Colorado, 2002, pp. 72–81.

[40] A. Gravano, J. Hirschberg, and S. Benus, "Affirmative Cue Words in Task-Oriented Dialogue," *Computational Linguistics*, vol. 38, no. 1, pp. 1–39, Mar. 2012.

[41] D. Tuccero, K. Railey, M. Briggs, and S. K. Hull, "Behavioral Health in Prevention and Chronic Illness Management: Motivational Interviewing," *Primary Care: Clinics in Office Practice*, vol. 43, no. 2, pp. 191–202, 2016, doi: https://doi.org/10.1016/j.pop.2016.01.006.

[42] R. M. BAECKER, J. GRUDIN, W. A. S. BUXTON, and S. GREENBERG, Eds., "Chapter 11 - Groupware and Computer-Supported Cooperative Work," in *Readings in Human–Computer Interaction*, Morgan Kaufmann, 1995, pp. 741–753. doi: https://doi.org/10.1016/B978-0-08-051574-8.50077-7.

[43] D. S. McCrickard, C. M. Chewar, J. P. Somervell, and A. Ndiwalana, “A model for notification systems evaluation—assessing user goals for multitasking activity.,” *ACM Transactions on Computer-Human Interaction (TOCHI) 10*, vol. 10, no. 4, pp. 312–338, 2003.

[44] D. C. McFarlane and K. A. Latorella, “The scope and importance of human interruption in human-computer interaction design,” *Human-Computer Interaction*, vol. 17, no. 1, 2002.

[45] L. Dabbish and R. E. Kraut, “Controlling interruptions: awareness displays and social motivation for coordination.,” in *In Proceedings of the 2004 ACM conference on Computer supported cooperative work*, 2004, pp. 182–191.

[46] K. A. Latorella, “Investigating Interruptions: An Example from the Flightdeck,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 40, no. 4, 1996.

[47] J. S. Rubinstein, D. E. Meyer, and J. E. Evans, “Executive control of cognitive processes in task switching.,” *Journal of experimental psychology: human perception and performance*, no. 4, p. 763, 2001.

[48] S. T. Iqbal and B. P. Bailey, “Leveraging characteristics of task structure to predict the cost of interruption,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montreal, Quebec, Canada, 2006, pp. 741–750.

[49] N. Peters, G. Romigh, G. Bradley, and B. Raj, “When to Interrupt: A Comparative Analysis of Interruption Timings Within Collaborative Communication Tasks,” 2017, vol. 497, pp. 177–187.

[50] P. D. Adamczyk, S. T. Iqbal, and B. P. Bailey, “A method, system, and tools for intelligent interruption management,” in *TAMODIA ’05 Proceedings of the 4th international workshop on Task models and diagrams*, Gdansk, Poland, 2005, pp. 123–126.

[51] P. H. Carstensen and K. Schmidt, “Computer Supported Cooperative Work: New challenges to systems design,” 1999.

[52] P. S. Wardrip, R. B. Shapiro, A. Forte, S. Maroulis, K. Brennan, and R. Roque, “CSCW and Education: Viewing Education as a Site of Work Practice,” in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion*, New York, NY, USA, 2013, pp. 333–336. doi: 10.1145/2441955.2442035.

[53] P. Dillenbourg, *Collaborative Learning: Cognitive and Computational Approaches. Advances in Learning and Instruction Series*. Emerald Group Publishing Limited, 1999.

[54] S. R. Hiltz, “Collaborative learning in asynchronous learning networks: Building learning communities,” in *Proceedings of WebNet 98 World Conference of the WWW, Internet, and Intranet*, 1998, pp. 7–12.

[55] D. Persico, F. Pozzi, and L. Sarti, “Monitoring collaborative activities in computer supported collaborative learning,” *Distance Education*, vol. 31, pp. 22–5, 2010.

[56] C. R. Rogers, *On Becoming a Person*. Mariner Books, 1961.

[57] J. M. de Jonge, G. M. Schippers, and C. P. D. R. Schaap, “The motivational interviewing skill code: Reliability and a critical appraisal,” *Behavioural and Cognitive Psychotherapy*, pp. 285–298, 2005.

[58] K. M. Carroll *et al.*, “A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions.,” *Drug and Alcohol Dependence*, vol. 57, no. 3, pp. 225–238, Jan. 2000, doi: 10.1016/s0376-8716(99)00049-6.

[59] “Information Overload.” 2021. [Online]. Available: https://www.interaction-design.org/literature/topics/information-overload

[60] S. Narayanan and P. G. Georgiou, “Behavioral Signal Processing: Deriving Human Behavioral Informatics From Speech and Language: Computational techniques are presented to analyze and model expressed and perceived human behavior-variedly characterized as typical, atypical, distressed, and disordered-from speech and language cues and their applications in health, commerce, education, and beyond,” *Proc IEEE Inst Electr Electron Eng*, vol. 101, no. 5, pp. 1203–1233, Feb. 2013, doi: 10.1109/JPROC.2012.2236291.

[61] D. C. Atkins, T. N. Rubin, M. Steyvers, M. A. Doeden, B. R. Baucom, and A. Christensen, “Topic models: A novel method for modeling couple and family text

data," *Journal of Family Psychology*, vol. 26, no. 5, 2012, [Online]. Available: https://doi.org/10.1037/a0029607

[62] D. C. and˘David C. Atkins and S. S. Narayanan, "A Dialog Act Tagging Approach to Behavioral Coding: A Case Study of Addiction Counseling Conversations," in *INTERSPEECH, 13th Annual Conference of the International Speech Communication Association*, Dresden, Germany, 2012, pp. 2251–2254.

[63] J. Cao, M. Tanana, Z. E. Imel, E. Poitras, D. C. Atkins, and V. Srikumar, "Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes." 2019.

[64] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "Rate My Therapist: Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing," *PLoS ONE*, vol. 10, no. 12, Dec. 2005, [Online]. Available: https://doi.org/10.1371/journal.pone.0143055

[65] "New advances in speaker diarization." 2020. Accessed: Oct. 28, 2020. [Online]. Available: https://www.ibm.com/blogs/research/2020/10/new-advances-in-speaker-diarization/

[66] B. Worthy, "What is Word Error Rate? Measuring the WER of Machine-Generated Transcripts and Its Limitations." 2019. [Online]. Available: https://medium.com/@bethworthy/what-is-word-error-rate-measuring-the-wer-of-machine-generated-transcripts-and-its-limitations-1457be914f3b

[67] N. Peters, G. Romigh, G. Bradley, and B. Raj, "A Comparative Analysis of Human-Mediated and System-Mediated Interruptions for Multi-user, Multitasking

Interactions," in *Advances in Human Factors and Systems Interaction*, 2018, pp. 339–347.

[68] "Developers | Audacity ®," 2019. https://www.audacityteam.org/community/developers/ (accessed Jan. 29, 2020).

[69] J. Sutherland, "What is word error rate and who is winning?," *Medium*, Nov. 13, 2017. https://medium.com/@sutherlandjamie/what-is-word-error-rate-and-who-is-winning-e623db5d7913 (accessed Jan. 29, 2020).

[70] "Steepest ascent method for multivariate optimization - Application Center." https://www.maplesoft.com/applications/view.aspx?SID=4194&view=html (accessed Jan. 29, 2020).

[71] A. C. Thomas, "What is computational complexity?," *Medium*, Jan. 23, 2020. https://towardsdatascience.com/what-is-computational-complexity-66722cd5f8dd (accessed Jan. 30, 2020).

[72] "Watson Speech to Text." https://www.ibm.com/cloud/watson-speech-to-text

[73] "Speech to Text." https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/#overview

[74] "Amazon Transcribe." https://aws.amazon.com/transcribe/

[75] "Application Programming Interface." 2020. [Online]. Available: https://www.ibm.com/cloud/learn/api

[76] "Android Operating System." 2021. [Online]. Available: https://www.android.com/

[77] R. W. Lent, “Longitudinal Relations of Self-efficacy to Outcome Expectations, Interests, and Major Choice Goals in Engineering Students,” *Journal of Vocational Behavior*, vol. 73, 2008.

[78] D. R. Garrison, T. Anderson, and W. Archer, “Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education,” *The Internet and Higher Education*, vol. 2, no. 2, pp. 87–105, 1999, doi: https://doi.org/10.1016/S1096-7516(00)00016-6.

[79] T. H. Fatani, “Student satisfaction with videoconferencing teaching quality during the COVID-19 pandemic,” *BMC Medical Education*, vol. 20, no. 1, p. 396, 2020, doi: 10.1186/s12909-020-02310-2.

[80] “Amazon EC2: Secure and resizable compute capacity to support virtually any workload.” 2021. [Online]. Available: https://aws.amazon.com/ec2/?ec2-whats-new.sort-by=item.additionalFields.postDateTime&ec2-whats-new.sort-order=desc

[81] “What is a URL?” 2005. [Online]. Available: https://developer.mozilla.org/en-US/docs/Learn/Common_questions/What_is_a_URL

[82] C. C. Aggarwal, *Neural Networks and Deep Learning*. Springer, 2018.

[83] C. Chiu *et al.*, “State-of-the-art speech recognition with sequence-to-sequence models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4774–4778.

[84] “Speech-To-Text basics.” 2022. [Online]. Available: https://cloud.google.com/speech-to-text/docs/basics

[85] S. Rollnick, W. R. Miller, and C. C. Butler, *Motivational Interviewing in Health Care*. The Guilford Press, 2008.

[86] Wren and Martin, *High School English Grammar and Composition*. Blackie Elt Books, 2016.

[87] "Open NLP." [Online]. Available: https://opennlp.apache.org/

[88] "Alphabetical list of part-of-speech tags used in the Penn Treebank Project." [Online]. Available: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

[89] "Understanding Cohen's Kappa." [Online]. Available: https://www.statisticshowto.com/cohens-kappa-statistic

[90] J. E. Arnold and D. G. Watson, "Synthesising meaning and processing approaches to prosody: performance matters," *Language, Cognition and Neuroscience*, vol. 30, no. 1–2, pp. 88–102, 2015, doi: 10.1080/01690965.2013.840733.

[91] R. Dieng, *Designing Cooperative Systems: The Use of Theories and Models : Proceedings of the 5th International Conference on the Design of Cooperative Systems (COOP'2000)*. IOS Press, 2000. [Online]. Available: https://books.google.com/books?id=6buG05d2h3MC

[92] S. Cocchimiglio, "What Is Behavior Modification? Psychology, Definition, Techniques & Applications," *Behavior*. https://www.betterhelp.com/advice/behavior/what-is-behavior-modification-psychology-definition-techniques-applications/

[93] B. Paltridge, *Discourse Analysis: An Introduction (Bloomsbury Discourse)*, 2nd ed. Continuum, 2012.

[94] M. Hasan *et al.*, "Identifying Effective Motivational Interviewing Communication Sequences Using Automated Pattern Analysis," *Journal of Healthcare Informatics Research*, vol. 3, no. 1, pp. 86–106, 2019, doi: 10.1007/s41666-018-0037-6.

[95] W. Sun, K. Zhang, S.-K. Chen, X. Zhang, and H. Liang, "Software as a Service: An Integration Perspective," in *Service-Oriented Computing – ICSOC 2007*, Berlin, Heidelberg, 2007, pp. 558–569.