Title	非タスク型ダイアログシステムのための多モーダルユーザー印象認識に関する研究
Author(s)	魏, 文青
Citation	
Issue Date	2025-03
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/19926
Rights	
Description	Supervisor: 岡田 将吾, 先端科学技術研究科, 博士

Doctoral Dissertation

STUDY ON MULTIMODAL USER IMPRESSIONS RECOGNITION FOR NON-TASK DIALOGUE SYSTEMS

Wei Wenqing

Supervisor Okada Shogo

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March 2025

Abstract

With the continuous development of human-computer interaction technology, the application of dialogue systems in various fields is becoming increasingly widespread. Among them, non-task-oriented dialogue systems have shown great potential in fields such as chatbots and open-domain dialogue systems. While improving the quality of non-task-oriented dialogue systems is crucial for enhancing user interactions. A high-quality dialogue system must not only understand user intent but also generate accurate and natural responses. This necessitates a robust evaluation framework to assess the system's capabilities. Additionally, evaluation serves as the foundation for system improvement and optimization. Through evaluation of the dialogue system, weaknesses in the system can be identified, making it easier to fine-tune models, data, or algorithms to improve overall performance.

With the rise of multimodal dialogue systems, the demand for their evaluation has also increased. However, existing evaluation methods for dialogue systems often focus solely on text-to-text interactions, neglecting the importance of multimodal data in dialogue systems. In contrast, unimodal systems typically rely only on language content or speech intonation, which may lead to neglect or misinterpretation of users' emotions. Multimodal information, such as speech intonation, facial expressions, and body movements, can better capture users' emotional changes. Therefore, utilizing multimodal information to evaluate multimodal dialogue systems is crucial. Moreover, text-based evaluation metrics, such as BLEU and ROUGE, are insufficient for assessing multimodal dialogue systems. At the same time, existing multimodal databases face limitations in data collection, particularly in collecting speech and image data, resulting in incomplete and limited evaluation methods. Motivated by these challenges, this research aims to address data collection issues in the evaluation of multimodal non-task-oriented dialogue systems and propose innovative evaluation methods. By collecting, organizing, and utilizing multimodal data, we aim to evaluate dialogue system performance more comprehensively and accurately, thereby enhancing user experience and impressions. Therefore, this research has significant theoretical and practical implications and will make important contributions to the development of the field of multimodal dialogue system evaluation.

Above all, to establish an automated, robust, and accurate model for evaluating multimodal dialogue systems. Firstly, we introduce a method for identifying user satisfaction at the dialogue level, filling a gap in previous research. We use a method based on multimodal modeling, which comprehensively considers various information such as text, speech, and images to evaluate dialogue system performance more comprehensively.

Then, we utilize deep learning models to comprehensively analyze user satisfaction at the dialogue level and user impressions at the exchange level, enhancing the accuracy and reliability of the evaluation methods. Through experimental evaluation, we confirm the effectiveness and feasibility of the proposed methods in the field of multimodal dialogue system evaluation, providing new insights and methods for further research in this area.

The user impression can be analyzed and evaluated at two levels: the exchange level and the dialogue level. These two levels are closely interconnected but differ in focus, making them well-suited for capturing information at different hierarchical levels. While the relationship between user impressions at the dialogue level and user sentiments at the exchange level is secondly explored, which proposes a multi-task learning model that comprehensively considers information from both levels. By analyzing the relationship between 18 dialogue labels and user sentiment during dialogue exchanges and utilizing multi-task learning models, we successfully achieve accurate identification of user impressions at the dialogue level, bringing new insights and methods to the field of dialogue system evaluation.

Lastly, we address the issue of existing methods neglecting the influence of users' personal information which included age, gender, and personality on their impressions by proposing a model based on adversarial learning. By reversing the gradient direction during training, our network learns adversarial features that remain consistent across different users' personal information domains, effectively mitigating the influence of users' personal information and making the model applicable to evaluate non-task-oriented dialogue systems. Through experimental validation, we confirm the effectiveness and feasibility of the proposed method, providing new insights and methods for further research in the field of dialogue system evaluation.

In conclusion, this study proposes novel approaches to addressing the evaluation challenges encountered by multimodal non-task-oriented dialogue systems. The proposed methods improve the accuracy and comprehensiveness of dialogue system evaluation, offering valuable insights for enhancing user experience and satisfaction in various applications.

Keywords: Dialogue system, Multimodal, Evaluation, User impression, User traits adaptation.

Acknowledgment

I would like to express my deepest gratitude to my advisor, Professor Shogo Okada, for his meticulous guidance and selfless support throughout my doctoral studies. His patience and insightful feedback during our discussions have been invaluable. I also extend my sincere appreciation to Professor Kazunori Komatani for his invaluable collaboration on this study.

I am especially grateful to my family for their unwavering support and understanding throughout my academic journey. Their encouragement and love have consistently filled me with confidence and strength, and I am profoundly thankful for my parents' selfless dedication.

Additionally, I wish to thank all my lab mates and friends who have provided assistance and support during my research. Their collaboration, advice, and encouragement have been crucial in advancing my research.

Finally, I extend my heartfelt thanks to everyone who has offered support and assistance throughout my doctoral studies. Their care and encouragement have been vital in overcoming challenges and making progress. I am deeply grateful to all of you!

List of Figures

List of Tables

Contents

Chapter 1

Introduction

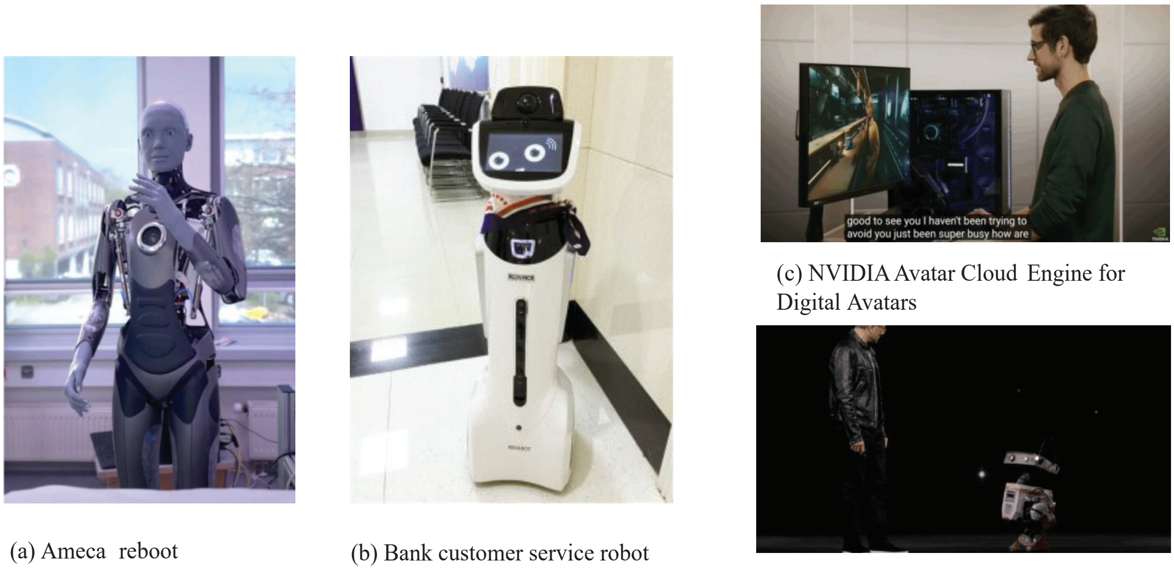
1.1 Backgorund

With the continuous advancement and widespread adoption of human-computer interaction technology, dialogue systems have become a key means of interaction across various fields. Among these, multimodal dialogue systems, as an advanced form, demonstrate significant potential, particularly in domains such as intelligent assistants and virtual customer service. Multimodal dialogue system ability to handle and integrate information from various modalities such as text, speech, images, and video. This integration allows for a richer and more nuanced understanding of user intentions, enhances the naturalness of human-computer interaction, and improves contextual comprehension and emotional sensitivity. By combining multiple signals, these systems achieve more accurate user intent recognition, better adapt to complex scenarios, and offer more personalized and robust services. This makes multimodal dialogue systems particularly valuable in applications requiring natural and sensitive interactions. Compared to traditional text-based dialogue systems, multimodal dialogue systems offer greater flexibility and adaptability, enabling a better understanding of user intent and emotions [2]. For instance, in the domain of intelligent assistants, multimodal dialogue systems can employ technologies like speech recognition and image processing to achieve more natural and efficient interactions shown in Figure 1.1(a)[1](b)(d)[2]. Similarly, in virtual customer service such as Figure 1.1(c), these systems can analyze user speech and facial expressions to more accurately discern user emotional states, thus providing more personalized service.

Non-task-oriented dialogue systems are dialogue systems designed not to accomplish specific tasks but to engage in natural and open-ended conversations. They provide a human-like interaction experience through casual conversation, emotional support, and social dialogue, thereby enhancing overall user satisfaction and trust. Non-task-oriented dialogue systems offer emotional support in fields such as mental health and caregiving, alleviating feelings of loneliness and stress, while also maintaining long-term user engagement through engaging and enjoyable dialogue. For instance, mental health

[1]https://engineeredarts.co.uk/robot/ameca/

[2]https://www.nvidia.com/en-us/geforce/news/nvidia-ace-for-games-generative-ai-npcs



(a) Ameca reboot

(b) Bank customer service robot

(c) NVIDIA Avatar Cloud Engine for Digital Avatars

(d) NVIDIA "WALL -E"

Figure 1.1: (a) Ameca is a chatbot created by Engineered Arts. (b) A banking service robot that interacts with users to answer questions. (c) An example released by NVIDIA of interacting with NPCs in a game. (d) An example of live interaction with "WALL-E" released by NVIDIA.

and emotional support, social interaction and entertainment, education and learning, customer service and user experience, as well as entertainment and creative domains. Consequently, evaluating such non-task-oriented dialogue systems becomes paramount.

Evaluating a non-task-oriented dialog system is a systematic evaluation of the performance, effectiveness, and user experience of a dialog system. This process aims to understand the actual performance of the system and identify its strengths and areas for improvement. The evaluation of dialogue systems is a critical step in ensuring their quality and effectiveness. With the widespread application of dialogue systems in areas such as intelligent assistants and virtual customer service, their performance directly impacts user experience and business outcomes.

Two fundamental issues exist in the current evaluation of dialogue systems. First, With the rise of multimodal dialogue systems [3, 4], the demand for their evaluation is increasing. Therefore, **exploring evaluation methods and metrics applicable to non-task-oriented multimodal dialogue systems represents a crucial research direction**. Previously, evaluation methods for dialogue systems primarily focused on text-to-text interactions [5–8] and failed consider the importance of fully multimodal data in dialogue systems. Although traditional text-to-text evaluation methods can to some extent reflect the performance of dialogue systems, they still have certain limitations and cannot comprehensively reflect the performance of multimodal dialogue systems and user impressions. Second **most evaluation methods based on word overlap [9–11] may not accurately reflect users' true intentions**, while methods based on embedding vectors may not fully consider contextual information. Word overlap methods rely on the similarity between generated responses and standard answers, but this approach has a low correlation with human judgment and makes it difficult to capture the fluency and naturalness of the dialogue. Although embedding vector methods can to some extent understand the semantics of words, they still prove inadequate for complex dialogue contexts and contextual variations in multimodal interactions. Therefore, evaluation methods need to more comprehensively consider the importance of multimodal data to ensure the effectiveness and reliability of dialogue systems in practical applications. Multimodal data, including text, speech, images, gestures, etc., can provide richer information, helping systems better understand user intentions and emotions. This comprehensive evaluation method that takes into account multimodal data not only improves the accuracy of system performance evaluation but also better reflects users' experiences in practical use.

In summary, non-task-oriented multimodal dialogue systems provide emotional support and social interaction through natural conversation and are widely used in fields such as mental health, education, and customer service. Evaluating the performance and user experience of these systems is crucial, however, traditional methods often fail to fully capture the effectiveness of multimodal systems. Therefore, the current research demand is to develop evaluation methods that effectively utilize multimodal data to comprehensively and accurately assess the performance of non-task-oriented multimodal dialogue systems. This research establish a solid foundation for enhancing user experience and interaction efficiency, and promote the widespread application and development of multimodal dialogue systems across various domains.

1.2 Research motivation

The motivation of this study is to evaluate non-task-oriented multimodal dialogue systems through user impressions, thereby addressing the research gap in the evaluation of non-task-oriented multimodal dialogue systems. This study aims to resolve data collection issues and explore innovative solutions. With the continuous advancement in social signal processing, multimodal methods have matured in identifying user emotions and intentions, as previous research indicates that these methods are better equipped to capture subtle changes in user emotions and intentions. Ultimately, the goal is to advance the field of multimodal dialogue system evaluation by leveraging multimodal data to identify user impressions, thereby promoting the development and progress of this field.

In the research of multimodal dialogue systems, a comprehensive and accurate evaluation of user impressions is crucial for enhancing system performance and user experience. Traditional evaluation methods often focus on individual dialogue exchanges, capturing only the performance at specific moments within the conversation. This approach is limited because it fails to fully capture the dynamic changes in user impression throughout the entire dialogue. User impression with a dialogue system is an accumulated and comprehensive experience, involving the interaction effects of multiple exchanges. Moreover, the contribution of each dialogue exchange to overall satisfaction is significant. Relying solely on dialogue-level evaluation may result in the omission of important information, thereby failing to accurately reflect the user's true experience.

Additionally, user characteristics such as personality and emotional state significantly impact user impressions. If these factors are not considered, evaluation models may exhibit bias towards different user groups, thus affecting the system's generality and robustness. Such bias not only leads to suboptimal performance in practical applications but also limits the system's applicability across diverse user groups. Therefore, our primary motivation is to systematically address these key challenges and develop a more comprehensive, accurate, and fair evaluation framework for multimodal dialogue systems.

1.3 Research contributions

Based on the above motivation, we have proposed the following three tasks to progressively achieve this goal. First, to solve the two fundamental issues mentioned in Section 1.1, we elevate the evaluation of user impression from the dialogue exchange level to the dialogue level. The core of this task lies in employing multimodal modeling methods, which comprehensively consider information from multiple modalities such as speech, text, and images throughout the dialogue process, to thoroughly assess the user's overall experience. Second, each dialogue consists of multiple conversational turns (exchange turn), and the experience of each exchange turn affects the overall impression of the dialogue. In the first task, the user experience at the exchange level was not considered. To address the relationship between the two levels, we aim to incorporate both aspects. We integrate user impressions at both the dialogue and exchange levels. Simultaneously processes and synthesizes information from both levels, capturing the overall impression of the entire dialogue while also attending to the specific performance of each exchange, thereby enhancing the precision and accuracy of the evaluation. Lastly, to ensure the robustness of dialogue systems, it is essential to consider user variability. Previous work has addressed issues related to modeling and dialogue structure information, but it has

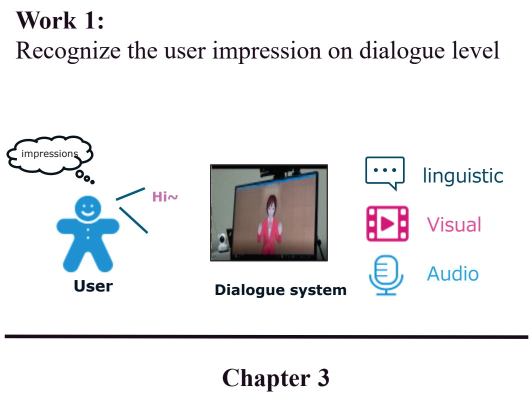


Figure 1.2: Overview of the work 1: Recognize the user impression on dialogue level.

not adequately considered the impact of users' personal information. We address the impact of user characteristics on user impressions by proposing an adversarial learning-based model. This model aims to mitigate the influence of user characteristics on evaluation results, thereby improving the fairness and generality of the system. Through this model, we enable the dialogue system to better adapt to the needs of diverse user groups, enhancing overall user experience and system robustness. Specifically, our research will address the following three main aspects:

1: Multimodal Molding:

Objective: We will investigate how to effectively integrate speech, facial, and text data to comprehensively evaluate the performance of dialogue systems. This objective includes developing new algorithms and models to capture and understand users' intentions and emotions in multimodal interactions. As shown in Figure 1.2 Multimodal Dialogue Systems (MDS) require perceiving users' multimodal behaviors to achieve natural language interaction with virtual agents. Evaluating the performance of dialogue systems is a crucial step in developing MDS. However, previous research has mainly focused on modeling user satisfaction in text-to-text dialogue systems, neglecting users' non-verbal behaviors (such as auditory and visual signals). Additionally, existing studies often identify satisfaction labels at the turn level rather than for the entire dialogue, posing challenges for system design.

Solution: Multimodal Recognition Method: A multimodal recognition method is proposed, utilizing sequence modeling algorithms (RNN, LSTM, and GRU) to identify user satisfaction at the dialogue level. Feature Extraction and Analysis: Language and non-language features are extracted from the exchange level (system and user utterance pairs), analyzing the contributions of multimodal and unimodal

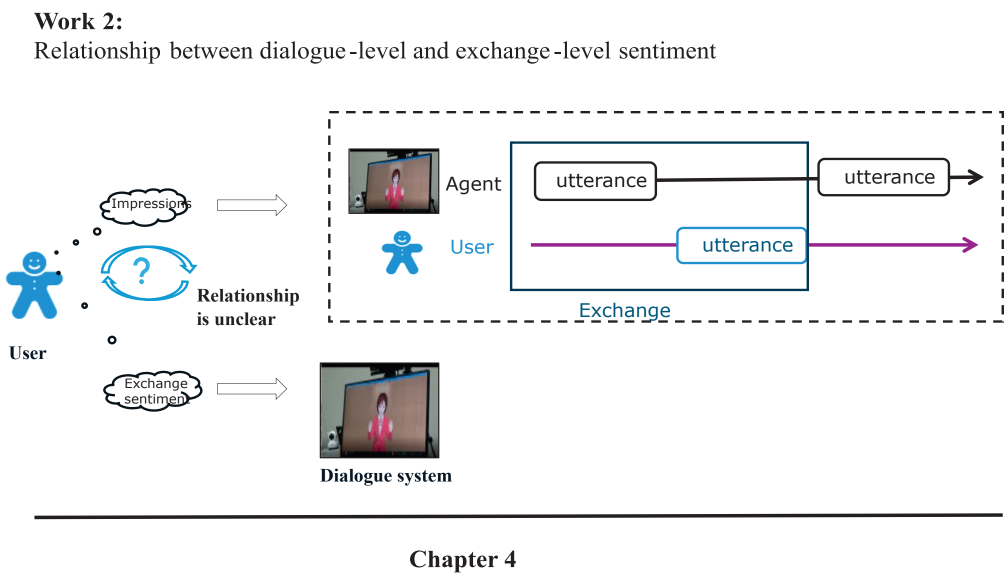


Figure 1.3: Overview of the work 2 Investigating the relationship between dialogue-level and exchange-level sentiment on multimodal human-system interaction.

features to identifying user satisfaction. The proposed method's identification accuracy is validated using a multimodal user-system dialogue data corpus with dialogue-level user satisfaction labels. The proposed model is compared with two types of human perception models based on external human coders and system operators, with results indicating the superior performance of the multimodal model in classification and regression tasks compared to human models.

2: Combine the Exchange Level and Dialogue Level Evaluation :

Objective: Multi-modal Dialogue Systems (MDS) have garnered increasing attention in recent years, particularly in the domain of automatically assessing user impressions. These systems facilitate natural language interaction with users by integrating speech, visual, and textual data. However, the majority of current research is predominantly focused on single-modal dialogue systems, which rely solely on textual data for modeling user satisfaction, thereby overlooking users' non-verbal behaviors such as auditory and visual cues. **Furthermore, existing studies often identify user satisfaction tags at the level of individual exchanges, neglecting the holistic satisfaction of the entire conversation.** In practical applications, users' overall impressions of dialogue systems are typically formed through their experiences in each interaction or exchange. Therefore, considering the emotional dynamics at the level of exchanges is crucial when identifying users' overall impressions of the dialogue. However, the relationship between exchange-level emotions and user impressions at the dialogue level remains unclear shown in Figure 1.3, and this relationship has not been adequately explored in impression analysis. Hence, there is a need for a method that comprehensively evaluates the performance of multi-modal dialogue systems to more accurately reflect users' overall satisfaction.

Solution: We conducted an in-depth investigation into the relationship between exchange-level emotions and 18 labels representing users' impressions at the dialogue level. These labels encompass various facets of users' perceptions of the dialogue, including naturalness, fluency, and emotional responses. Through rigorous analysis of the correlation between these labels and exchange-level emotions, we gained valuable insights into how users' emotional dynamics during the dialogue process influence their overall impression. To address the challenge of integrating these findings into a practical framework, we proposed a novel Multi-Task Learning (MTL) model. This model leverages annotations at the exchange level to identify labels at the dialogue level, enabling simultaneous learning of multiple related tasks. Our extensive experiments confirmed the effectiveness of the MTL model, demonstrating significant performance improvements over single-task models, with enhancements of up to 15.7% in dialogue system performance. Furthermore, comparison with human perception models highlighted the superior performance of our proposed approach in both classification and regression tasks. Overall, our findings contribute to a more nuanced understanding of multi-modal dialogue systems, facilitating their continued development and application by enhancing user satisfaction.

3: User Traits Adaption User Impression Recognition :

Objective: The existing research primarily focuses on exploring various algorithms, including traditional machine learning and deep learning, to effectively simulate the overall content of dialogues and strives to enhance the generalization of dialogue system content. However, these studies lack sufficient attention to the influence of user features on user impressions as shown in Figure 1.4, leading to the following issues: Firstly, there is insufficient attention and research on the influence of user features on user impressions, resulting in an inadequate understanding of the mechanisms behind user impression formation. Secondly, efforts in dialogue system content generalization are limited, with a lack of meticulous analysis and handling of user features, making it difficult to accurately understand and satisfy user needs. Lastly, data collection issues persist, especially regarding the relatively challenging collection of user impression data for multimodal dialogue systems, which hampers the performance evaluation and optimization of dialogue systems in multimodal environments.

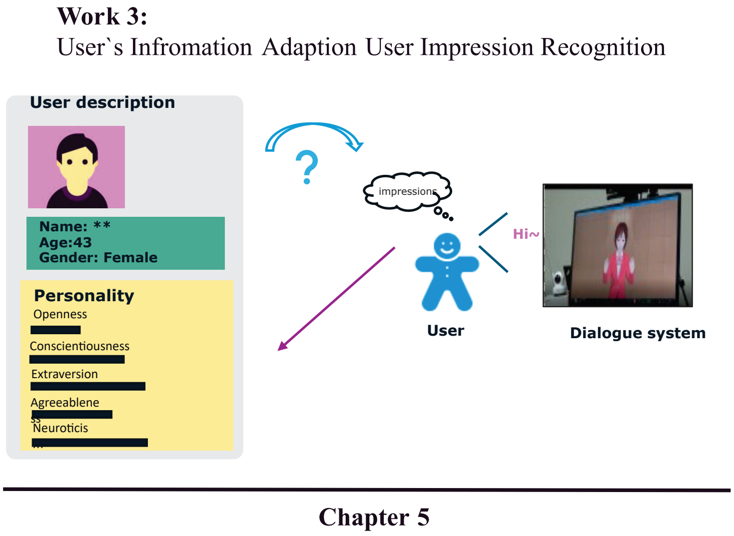


Figure 1.4: Overview of the work 3: User Traits Adaption User Impression Recognition.

Solution: To address the aforementioned issues, we have taken the following measures: Firstly, we collected a multimodal dialogue dataset comprising audio, body, visual, and text data, along with 18 types of user impression labels, to assess user impressions of the dialogue system. Additionally, we gathered user feature information such as age, gender, and personality to gain a more comprehensive understanding of user characteristics. Secondly, we proposed an adversarial adaptation method to mitigate the influence of user features. By leveraging gradient reversal layer (GRL) and adversarial discriminative (AD) models to invert the gradient flow of user feature classification, we were able to eliminate the user feature discrepancies learned from user impression embeddings, thereby improving the model's generalization performance. Lastly, we conducted comparisons between machine learning models and human models to validate the effectiveness of the proposed adversarial model. Through comparative analysis, we emphasized the importance of automatic multimodal recognition techniques in accurately estimating user impressions, providing valuable insights for further optimization of dialogue systems.

By progressively advancing through these three tasks, we aim to develop a more comprehensive, accurate, and fair evaluation framework for multimodal dialogue systems, thereby improving the overall performance and user experience of these systems. This approach not only facilitates the research and development of more efficient dialogue systems but also lays a solid foundation for their widespread application across various scenarios.

In this study, impression, satisfaction, and rapport are key concepts used to describe user experience and dialogue quality. Impression refers to the user's overall perception and evaluation of the dialogue system. It is derived from a comprehensive feedback encompassing eighteen labels, including "well-coordinated," "boring," "cooperative," "harmonious," "unsatisfying," "uncomfortably paced," "cold," "awkward," "engrossing," "unfocused," "involving," "intense," "unfriendly," "active," "positive," "dull," "worthwhile," and "slow." Impression provides a holistic assessment of the user's experience with the system. Satisfaction is a core component of impression, focusing on the user's specific evaluation of the system's performance, functionality, and interaction experience. It reflects the user's overall level of contentment. In our study, we define user satisfaction based on the scores of three labels: "well-coordinated," "awkward," and "unfriendly." Rapport emphasizes the quality of interaction between the user and the dialogue system, including the naturalness, friendliness, and emotional connection of the communication. We define user rapport using the scores of three labels: "well-coordinated," "awkward," and "engrossing." Therefore, a positive rapport can enhance user satisfaction and strengthen the overall impression of the dialogue system. These factors collectively contribute to the comprehensive impression of the dialogue system.

1.4 The organization of the dissertation

This dissertation consists of six chapters. Following Chapter 1, which This thesis first introduces the background of the research field and states the target research questions serve as the introduction, the contents of Chapters 2 through 6 are summarized as follows.

Chapter 2 reviews related work on dialogue systems and multimodal interactions. First, it summarizes the development of dialogue systems, covering both task-oriented and non-task-oriented approaches. It also discusses the evolution of dialogue systems from text-based to multimodal interactions. The importance of evaluation in dialogue systems is highlighted, with an introduction to previously used methods from both subjective and objective perspectives. Furthermore, this chapter emphasizes the limitations of current evaluation methods and underscores the necessity of incorporating multimodal data when assessing user satisfaction.

Chapter 3 addresses the first major task: enhancing user satisfaction evaluation at the dialogue level using multimodal modeling. This chapter presents the motivation for this shift and details a novel method for evaluating user satisfaction at the dialogue level. It discusses how this method leverages multimodal data to provide a more comprehensive assessment of user experience and system performance. The chapter also describes the experimental setup, the datasets used, and the results obtained, demonstrating the effectiveness of the proposed approach.

Chapter 4 introduces the second major task: integrating dialogue-level and exchange-level user impressions through a multi-task learning model. This chapter explains the motivation behind combining these two levels of information and describes the design of the multi-task learning model. It provides details on how the model simultaneously considers dialogue-level and exchange-level impressions to improve the overall evaluation accuracy. The chapter also covers the experiments conducted to validate the model and analyze the result, which demonstrates the advantages of simultaneously considering both dialogue-level and exchange-level approaches.

Chapter 5 discusses the third major task: mitigating the influence of user characteristics on user impressions using an adversarial learning model. This chapter outlines the challenges posed by user characteristics such as age, gender, and personality on the evaluation of user satisfaction. It introduces an adversarial learning model designed to reduce these biases, thus enhancing the generalizability and robustness of the dialogue system. The chapter details the model architecture, the training process, and the experiments carried out to test its effectiveness. It concludes with an analysis of the results, highlighting the improvements achieved through this approach.

Chapter 6 summarizes the dissertation and provides the conclusions of this research. This chapter recaps the key findings from each of the previous chapters, discussing their implications for the evaluation of multimodal dialogue systems. It also highlights the contributions of this study to the field and suggests potential directions for future research.

Chapter 2

Literature review

2.1 Dialogue system

Enabling machines to engage in communication with humans is a crucial task in the field of artificial intelligence, and it is also a highly challenging endeavor. In 1950, Turing proposed in his paper "Computing Machinery and Intelligence" the use of human-computer dialogue as a means to evaluate the level of machine intelligence, which sparked widespread interest among researchers. Early dialogue systems made significant contributions to natural language processing. ELIZA [12](1960s) was a rule-based system that responded by rearranging words or asking questions based on user input. LUNAR [13] (1970s) helped lunar geologists analyze lunar rock data using syntactic and semantic analysis. SHRDLU [14](1968–1970) allowed users to interact with a virtual environment of blocks, illustrating that understanding language requires contextual knowledge. Next, frame-based dialogue systems, such as the GUS system [15], were introduced. The GUS system aids users in completing tasks by understanding the semantic content of dialogues through a set of predefined frames representing specific meanings. In the 1980s, early speech dialogue systems, such as "Hearsay-II [16]," integrated both top-down and bottom-up processing approaches. By the 2000s, dialogue systems had developed fundamental interactive capabilities [17], including the ability to answer and ask questions, recognize interaction costs, manage interruptions, and address social and emotional needs. These early advancements laid the groundwork for the development of contemporary dialogue systems. Since the advent of deep learning, there have been significant advancements in dialogue systems. Initially, deep neural network-based models [18–22] brought notable improvements in system performance. Subsequently, deep learning techniques introduced large-scale pre-trained language models (such as GPT [23, 24] and BERT [25]), which further advanced dialogue systems' capabilities in understanding and generating natural language. These developments have enabled modern dialogue systems to handle complex conversational tasks more effectively and provide more natural and seamless user interactions.

Intelligent conversational agents are now commonly used in our daily lives. Dialogue systems are computer programs that can interact with humans using natural language. These systems can be classified into two categories: task-oriented [18–20] and non-task-oriented [21, 22]. Task-oriented dialogue systems are designed to assist humans in achieving their desired goals. These systems focus on specific tasks such as booking a flight, ordering food, or making a hotel reservation. The goal of these systems is to provide the user with the best possible solution to their problem by asking relevant questions and providing informative responses. On the other hand, non-task-oriented dialogue systems are designed for general conversation and do not have any specific goal or objective. These systems aim to engage users in casual conversation and provide them with an enjoyable experience.

In the domain of task-oriented systems, numerous products have emerged in the industry. Voice assistants such as Apple's Siri and Microsoft's Cortana have been widely utilized across various platforms such as smartphones and operating systems. Additionally, virtual assistant-style smart speakers like Amazon's Echo, Baidu's Duer, Google's Home, and Alibaba's Tmall Genie have also become ubiquitous household items. Meanwhile, in the realm of open-domain conversational systems, Microsoft has developed a range of chatbots tailored to different languages, such as Xiaoice, Rinna, Zo, and Ruuh, with user numbers reaching tens of millions. For an extended period, academic research on conversational systems has predominantly focused on specific tasks within constrained domains, constituting a significant branch of spoken dialogue system research. These systems typically impose strict definitions and limitations on inputs, designing corresponding rules, logic, and response statements tailored to specific tasks. Despite significant progress in the field of human-computer interaction achieved through these methods, their robustness, scalability, and domain adaptability are compromised by factors such as manual rule setting, rendering them unsuitable for open-domain conversational systems. In recent years, driven by the rapid growth of social data on the internet, data-driven open-domain conversational systems have gradually become a focal point of academic attention. The role of conversational systems has also shifted from mere service providers to emotional companions. Deep learning techniques have been widely applied in retrieval and generation methods, emerging as a crucial technical direction in open-domain conversational system research. In recent years, dialogue systems based on large-scale pre-trained language models (LMM) have gradually become mainstream, exemplified by ChatGPT, Gemini, and Claude. To engage in conversations with users. They can be viewed as a unique hybrid of multi-domain task-oriented spoken dialogue systems (SDS) and open-domain non-task-oriented SDS. Next, we will delve into task-oriented and non-task-oriented dialogue systems, exploring their different applications and characteristics in human-machine interaction.

2.1.1 Task-oriented dialogue system

Task-oriented dialogue systems aim to assist users in completing specific tasks, such as restaurant reservations and weather inquiries, with the core objective of providing

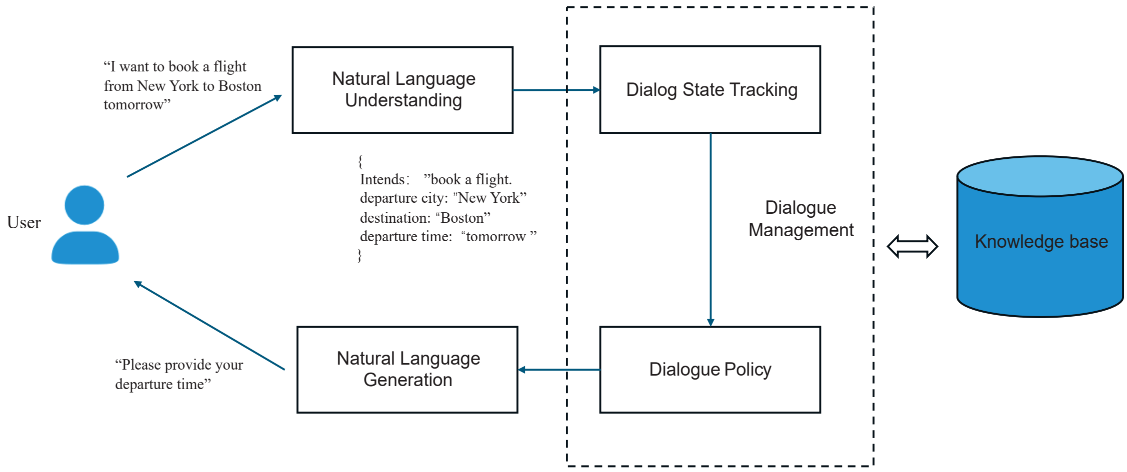


Figure 2.1: Overview of the pipeline task-oriented dialogue .

rapid and accurate service through precise information extraction and efficient task management. In contrast to open-domain dialogue systems, task-oriented dialogue systems typically feature more structured dialogue frameworks, rigorously defining and restricting user inputs, and designing rules, logic, and response statements specific to particular tasks. This structured dialogue framework helps the system efficiently process user requests within defined parameters, enhancing response accuracy and efficiency. The architecture of task-oriented dialogue systems can be broadly classified into two categories: pipeline [26] and end-to-end [27, 28]. In pipeline approaches, task-oriented dialogue systems typically rely on core modules such as Natural Language Understanding (NLU), Dialogue State Tracking (DST), and Dialogue Management (DM). The NLU module converts the user's natural language inputs into system-understandable intents and slots; the DST module tracks user states and contextual information during the dialogue process; and the DM module generates appropriate system responses based on the current dialogue state and predefined dialogue strategies. NLU, DST, and Natural Language Generation (NLG) components are typically trained independently in their respective training stages, while the dialogue policy component is trained after integrating these components into the complete systems, the structure for the pipeline of dialogue system shown in Figure 2.1.

The Natural Language Understanding (NLU) module performs semantic analysis in a task-oriented dialogue system. A prevalent pattern of semantic representation is dialogue acts, consisting of intents and slot values [19]. NLU tasks can be decomposed into intent detection and slot filling. Intent detection aims to identify the user's expressed intention or purpose [29–31], often viewed as a classification task. Methods based on RNNs or LSTMs are widely used due to their advantages in sequence encoding, allowing better consideration of context [30, 32]. For instance, given the input "I want to book a flight from New York to Boston tomorrow," the intent detection module needs to understand that the user intends to "book a flight."

Slot filling is responsible for extracting key information or parameters needed to complete specific tasks from user input. Slots can be understood as task-related variables. For example, in the sentence "I want to book a flight from New York to Boston tomorrow," slots include "departure city: New York," "destination: Boston," and "departure time: tomorrow." Sequence modeling methods [33, 34] are commonly used for slot filling. Recent advancements in natural language processing models, such as BERT and Transformer, have been widely applied to intent detection and slot-filling tasks [35, 36]. Following NLU, Dialogue State Tracking (DST) is a critical component in intelligent dialogue systems. It captures and manages state information in the conversation to ensure accurate understanding and response to user needs. DST dynamically updates dialogue state tables using various models and techniques, such as finite state machines, slot filling, instance-based methods, planning methods, and Bayesian networks, through supervised learning [37, 38]. DST plays a vital role in improving system response efficiency and user experience by ensuring dialogue coherence and context relevance, ultimately providing efficient and personalized services to users. Finally, the Natural Language Generation (NLG) module utilizes the state information from DST to generate natural language responses. NLG plays a primary role in dialogue systems by converting structured data into natural language output, enabling the system to produce fluent and natural replies. For example, given the input "I want to book a flight from New York to Boston tomorrow," potential responses generated by the NLG module could include "Your flight has been successfully booked for tomorrow from New York to Boston" or "Please provide your departure time." In practical applications, the NLG module is widely used in intelligent assistants (such as Siri and Alexa), customer service chatbots, and content generation systems (such as news summary generation). By transforming structured data into natural language output, NLG significantly enhances the naturalness and user experience of dialogue systems, allowing the system to provide personalized, fluent, and friendly services.

Template-based NLG methods are simple and effective but are constrained by predefined templates [39], limiting response flexibility. With the advancement of deep learning, improved methods such as RNN and LSTM-based generation approaches have been proposed to enhance coherence, while adding semantic conditions improves context relevance. Recently, pre-trained models like GPT, through fine-tuning, have notably improved user experience by generating more natural and fluent dialogues [40].

Typically, pipeline systems optimize each component separately, leading to complex model designs where improvements in individual components do not necessarily enhance

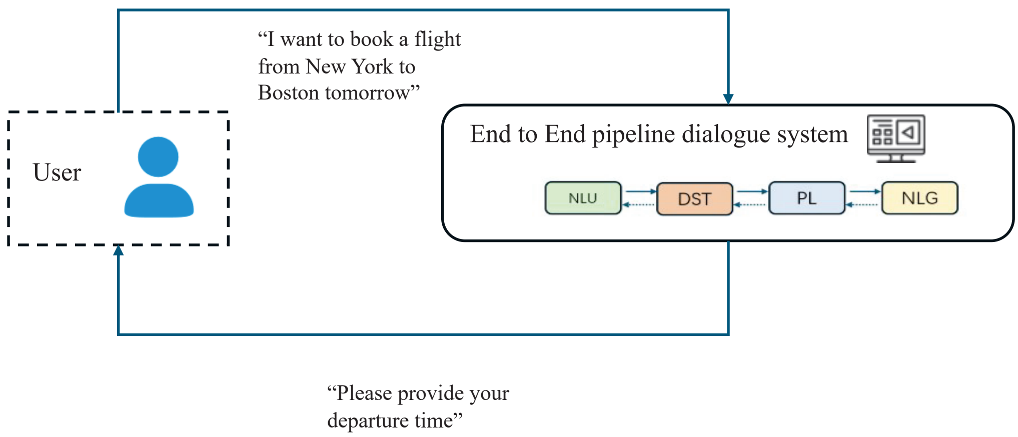


Figure 2.2: Overview of pipeline end-to-end dialogue system.

overall system performance. In contrast, task-oriented dialogue systems adopt end-to-end approaches inspired by open-domain dialogue systems, utilizing neural models to construct the system without the need for modular design. These methods often employ sequence-to-sequence models. For example, Wen et al. proposed a modular end-to-end model in which each component is modeled using neural networks [41] shown as Figure 2.3. Additionally, Lei et al. introduced a two-step seq2seq generation model that bypasses structured dialogue act representations, retaining only dialogue state representations [42].

However, a significant drawback of these end-to-end methods is their requirement for large amounts of training data, which is costly to acquire. Moreover, since the models only observe examples from the data, they cannot fully explore the state space. To address these issues, reinforcement learning methods have been introduced [42–46]. These end-to-end approaches enable task-oriented dialogue systems to achieve more efficient and coherent dialogue management without the need for manually designed complex modules. Nonetheless, they rely heavily on substantial data and computational resources to achieve high performance.

2.1.2 Non-task oriented dialogue system

In non-task-oriented dialogue systems, also known as open-domain dialogue systems, an end-to-end architecture is typically employed instead of the modular architecture used in task-oriented systems. Traditional dialogue systems used in task-oriented applications

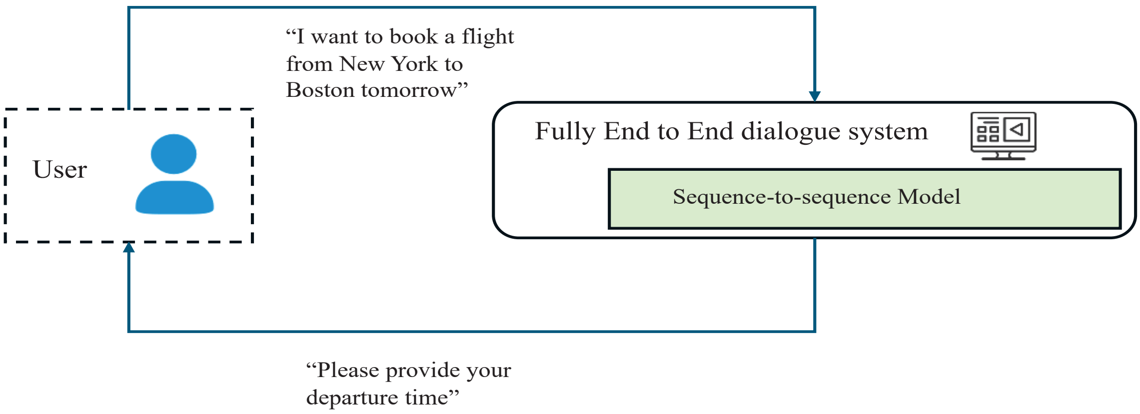


Figure 2.3: Overview of fully end-to-end dialogue system.

require extensive domain-specific manual crafting, hindering their scalability to new domains. End-to-end dialogue systems, where all components are trained from the dialogue data itself, alleviate this limitation.

Non-task-oriented dialogue systems aim to engage users in more spontaneous and open-ended conversations, prioritizing emotional engagement, entertainment, and knowledge sharing. The primary objective is to create natural, fluent, and meaningful dialogue experiences that make users feel understood and accompanied. Unlike task-oriented systems, non-task-oriented dialogue systems usually feature a more flexible dialogue structure, allowing users to express their thoughts, feelings, and interests freely. This flexibility enables the system to better adapt to user needs and preferences. It relies on natural language processing (NLP) and natural language generation (NLG) techniques to ensure understanding and generation of contextually appropriate dialogue content.

In terms of application domains, non-task-oriented dialogue systems demonstrate broad potential. In the entertainment domain, these systems can serve as chatbots to provide entertaining conversations and companionship during leisure time. In education, they can function as language learning assistants, offering dialogue practice to help students improve their speaking and comprehension skills. On social media platforms, non-task-oriented dialogue systems can facilitate user interaction and socialization. As virtual companions, these systems can provide emotional support and companionship, particularly for those feeling lonely or need of emotional communication.

According to the construction methods, non-task-oriented dialogue systems can be classified into retrieval-based, generation-based, and hybrid retrieval and generation

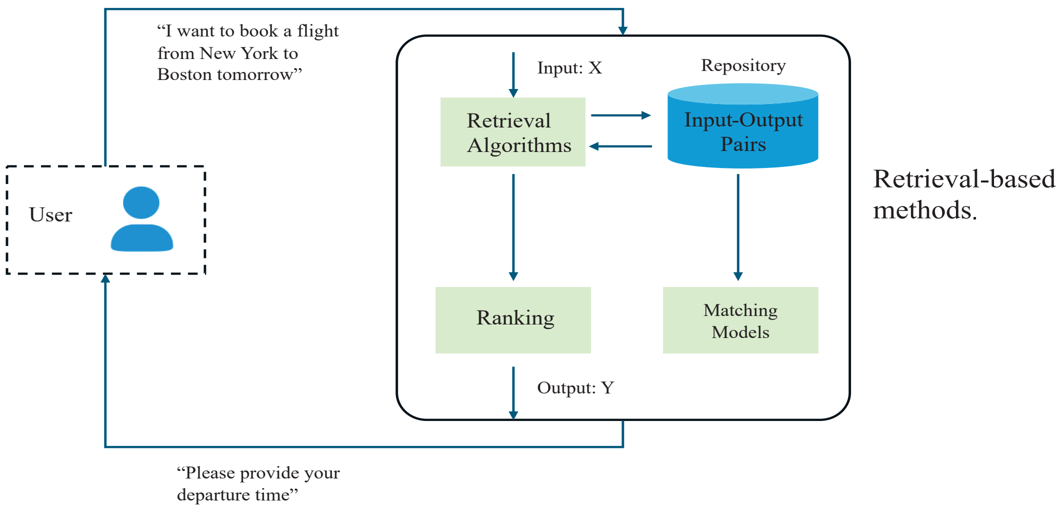


Figure 2.4: Overview of Retrieval-based-methods dialogue system.

approaches. Retrieval-based methods generate responses by retrieving the most relevant replies from predefined dialogue corpora or knowledge bases. These systems match user inputs against existing dialogue records or knowledge entries to find the best matches, which are then presented as system responses, shown in Figure 2.4. Unlike generation-based dialogue systems, retrieval-based systems do not generate new text but rely on existing content for matching and selection [47–49]. Ji et al. [50] introduced traditional learning-to-rank methods for selecting responses from large-scale response repositories. These methods first retrieve candidate lists from repositories containing many input-output pairs and then use matching functions to choose the highest-scoring candidates as output responses. Matching functions can be implemented using traditional ranking algorithms [51] or neural networks [52, 53]. Generative dialogue systems respond to user input by generating new content. These systems typically collect much conversational data as training data and build end-to-end dialogue models based on deep neural networks. Typical generative dialogue systems employ seq2seq models shown in Figure 2.3 or Encoder-Decoder models, which capture the conversational context and generate coherent responses [54–56]. Trained on extensive dialogue data, the models learn the mapping from input to output, thus generating natural language responses that fit the conversational context. Recently, large-scale language pre-training models like BERT and GPT have achieved remarkable results in generative dialogue [57, 58].

The combination of retrieval and generation methods integrates the advantages of both types of dialogue systems, aiming to construct more flexible and efficient dialogue systems. This approach can be divided into two stages: firstly, retrieval mechanisms to find the most relevant candidate responses from a large predefined dialogue repository, and then utilizing generation models to adjust or supplement these candidate responses to generate more natural and contextually relevant responses. Through this combination, the system ensures both the accuracy and relevance of responses, while providing diverse and flexible dialogue experiences [59, 60]. In [61], prototype responses are first retrieved from training data, providing a good starting point for the generator, and further enhancing the relevance and accuracy of generated responses. This method utilizes prototype responses from existing dialogue data as the basis for generating new responses, helping to ensure that the generated responses better fit the context and topic of the conversation. Additionally, Zhang et al. [62] proposed an adversarial learning framework to enhance retrieval-generation integrated models. This framework consists of three components: a generator for generating responses, a ranking generator for ranking candidate responses, and a ranking discriminator for filtering high-quality responses. By coordinating these three components, the framework aims to improve the quality of generated responses and ensure that the generated responses are superior to other candidate responses.

2.1.3 Multimodal dialogue system

Meanwhile, multimodal dialogue systems are computer programs that can interact with humans using multiple modes of communication such as text, speech, images, videos, etc. The development of multimodal dialogue systems has been a significant area of research in artificial intelligence. These systems have gained immense popularity due to their ability to bridge the gap between language and vision and facilitate better human-machine interaction.

These systems [52, 63] leverage complementary information from different modalities to build more robust and informative responses that better fulfill the user's needs. In a multimodal dialogue system, the user can interact with the system using different modes of communication. For example, a user can ask for restaurant recommendations by providing an image of a dish they like or describing it in text or speech. The system can then use this information and other relevant information such as location and cuisine preferences to provide personalized recommendations. Multimodal dialogue allows users to communicate with machines using multiple modes of communication. This approach enables machines to understand human needs more accurately and provide more informative responses. By leveraging diverse information from different modalities, these systems can build more comprehensive models of user needs and preferences. As conversational artificial intelligence continues to make significant strides, capturing the user's experience and impression has become increasingly crucial. This is especially true for open-domain dialogues that lack a clearly defined objective, which poses an even more significant challenge. Suppose the system can accurately predict how users feel about the conversation. In that case, it can better adapt to their interests or intentions or even manage the dialogue system more humanely to provide users with an enhanced experience.

The development of multimodal dialogue systems has evolved from single-modality to integrating multiple modalities, gradually enhancing the naturalness of human-machine interaction and user experience. The earliest systems primarily relied on text-based interactions. With the advancement of speech recognition and synthesis technologies, voice dialogue systems gained prominence. Subsequently, the introduction of computer vision enabled these systems to process image and video data. Initial multimodal systems attempted to integrate text, speech, and visual information, but the level of integration was relatively low. As deep learning and multimodal fusion technologies advanced, modern multimodal systems can more naturally understand and generate multimodal information, providing personalized and context-aware responses.

2.2 The development of evaluation of dialogue system

Dialogue system evaluation is a crucial task aimed at assessing and improving the system's performance, usability, and effectiveness [64]. By evaluating factors such as response accuracy, naturalness, and user satisfaction, developers can identify areas for improvement, enhancing the overall user experience. Additionally, evaluation helps optimize system performance across various aspects, such as language understanding, response generation, and context handling, thus aiding developers in refining algorithms and models. Moreover, evaluation ensures the reliability and stability of the system by assessing its performance in different contexts and reducing errors and interruptions during real-world applications. Furthermore, evaluation guides technological innovation by providing insights for research and development directions. Through analyzing user needs and feedback, evaluation assists in optimizing user services, ultimately leading to increased user satisfaction and loyalty. In summary, dialogue system evaluation is critical for optimizing user experience, improving system performance, ensuring system reliability, guiding technological innovation, and optimizing user services.

Evaluation methods for dialogue systems can be categorized into subjective evaluation and objective evaluation. Subjective evaluation methods primarily focus on collecting user feedback and opinions, including user satisfaction surveys, user experience testing, expert evaluations, and simulated dialogue assessments [5, 65–68]. These methods evaluate the system's user experience and interaction quality by designing standardized questionnaires, observing user behavior and listening to their feedback, or inviting experts to evaluate system performance. On the other hand, objective evaluation methods focus more on quantifying and analyzing system performance metrics, including automatic evaluation metrics, system performance indicators, dialogue content analysis, and task completion rates [69–71]. These methods quantify and evaluate the effectiveness and performance of the system using natural language processing techniques, performance monitoring tools, and text data analysis. By comprehensively applying subjective and objective evaluation methods, the performance and effectiveness of dialogue systems can be thoroughly evaluated, providing effective reference and guidance for system improvement and optimization.

2.2.1 Subjective evaluation

2.2.1.1 User Simulation

User simulation is an effective and straightforward evaluation strategy, which is also the most likely method to cover the broadest range of dialogue spaces. This is because simulating conversations in different scenarios enables efficient testing and evaluation across a wide spectrum [65–67].

As described in section 2.1.1 on pipeline task-oriented dialogue systems, in each interaction round, the dialogue system employs Natural Language Understanding (NLU) in each interaction round to parse user utterances into machine-understandable semantic labels. It then maintains an internal dialogue state through Dialogue State Tracking (DST), selects appropriate dialogue actions using Dialogue Policy (DP), and converts actions into natural language replies using Natural Language Generation (NLG). These interaction data and scores can be utilized for reinforcement learning training. However, the high cost and slow data feedback of fundamental user interactions hinder rapid iteration. Therefore, researchers construct a User Simulator (US) as an interactive environment for closed-loop training, generating a large volume of data to enable comprehensive exploration of state and action spaces in search of optimal strategies.

Nevertheless, the drawback of this approach lies in the potential disparity between honest user reactions and simulator responses, with the extent of this disparity depending on the quality of the user simulator. Despite this challenge, user simulation remains the most commonly employed method for evaluating task-oriented dialogue systems, extensively applied for assessing dialogue policies based on partially observable Markov decision processes (POMDPs) [5, 6]. There has been extensive research in modeling as well. It spans from foundational approaches like the bi-gram model [72], to practical and classic methods such as Agenda-based approaches [73], and more recently, to user models leveraging deep learning techniques [74, 75]. These advancements have significantly improved the effectiveness of user simulators and have also provided valuable methodologies for training dialogue models.

2.2.1.2 User satisfaction:

Research on modeling user satisfaction focuses primarily on measuring the impact of dialogue system attributes on user satisfaction to meet both interpretability and automation requirements. Typically, user satisfaction is modeled as either a regression or classification task, where system attributes serve as inputs and user judgments as target variables. Two key questions need to be addressed when assessing user satisfaction. Firstly, who evaluates the dialogue? Evaluation can be conducted by users themselves or by objective evaluators. Secondly, what is the level of evaluation, including dialogue-level and interaction-level aspects? The following will present relevant work about these two questions.

The PARADISE framework [76] is a proposed evaluation framework for task-oriented systems. Its main structure is outlined as follows: upon completion of user interaction with the dialogue system, satisfaction is assessed through questionnaire surveys, serving as the target variables. Subsequently, a linear regression model is trained on input variables extracted from recorded conversation data, either automatically or manually. Ultimately, user satisfaction for given input variables can be predicted by fitting the linear regression model. This evaluation methodology relies on user ratings at the dialogue level, thus classifying it as a dialogue-level evaluation approach. The framework amalgamates prior works, enabling the comparison of dialogue strategies through task representations, effectively separating the agent's task requirements from executing tasks via dialogue. PARADISE's performance measurement integrates considerations of task success and dialogue cost, boasting several advantages. It facilitates the evaluation of dialogue performance at any level and permits the comparison of agent executions across diverse tasks. Moreover, PARADISE accommodates task complexity, enabling partial task success measurement, and integrates objective and subjective cost metrics, ensuring a comprehensive evaluation perspective. Lastly, PARADISE utilizes user satisfaction to ascertain the weights of factors related to performance, enhancing the evaluation's comprehensiveness and accuracy. Subsequent studies (Walker et al., 2000) [77] have further delved into the PARADISE framework, primarily focusing on its generalizability across different systems and user cohorts, and its predictive capabilities.

In the assessment of user satisfaction at the exchange level, as opposed to evaluating the entire dialogue, the exchange-level evaluation focuses on individual interactions within each dialogue round to analyze the interaction process between users and the system in finer detail. Methods for exchange-level evaluation offer more specific and detailed feedback, aiding in pinpointing issues and areas for improvement at specific dialogue stages. Researchers have adopted two primary methodologies: one based on user feedback and the other on expert judgment. These methodologies leverage sequence modeling techniques, such as Hidden Markov Models (HMM), Conditional Random Fields (CRF), or Recurrent Neural Networks (RNN). Engelbrecht et al. [78] and

Higashinaka et al. [68] explored user-based and expert-based methodologies, employing digital keyboards or expert annotations to evaluate dialogues and train corresponding models. However, these approaches have inherent limitations, such as reliance on manual feature extraction or exclusive use of dialogue behavior as input features. To address these challenges, Schmitt and Ultes [79] introduced an" interaction quality" metric to identify issues in dialogues automatically. They rated dialogues at each time point and then computed the median rating as the interaction quality. This method employs Support Vector Machines (SVM) to fit input and target variables, which experts manually annotate. They found that interaction quality is an objectively effective approximation of user satisfaction and is more accessible. This holds particular significance for field dialogue system evaluations, given that users are typically not tasked with evaluating dialogue systems at the interaction level. Consequently, expert ratings at the exchange level are imperative for field dialogue evaluations.

2.2.1.3 Human evaluation:

anual evaluation methods play a crucial role in assessing dialogue systems, typically involving hiring testers to evaluate the dialogue outcomes generated by the system manually. The advantage of this approach lies in its ability to produce more authentic evaluation data, thereby enabling a more accurate performance assessment of the system. Testers usually assess the system within predefined task domains, interacting with the system through preset inquiries and scoring the system's performance based on the dialogue outcomes [6, 40, 80]. However, this evaluation method requires many testers and resources; thus, it is often found in research environments with ample resources, such as laboratories.

With the advent of outsourcing models and online media, such as the Amazon Mechanical Turk (AMT) service [81], researchers provide tasks and training instructions to the evaluators. Evaluators interact with the dialogue system via free telephone calls and provide feedback after each dialogue. This method can quickly generate a large amount of dialogue data, thereby aiding in the performance evaluation of dialogue systems. The evaluation process becomes more flexible but still faces challenges such as high costs and the representativeness of the evaluators. Therefore, in the future, reducing bias in the evaluation process, improving evaluation efficiency, and lowering costs are issues that require further research.

2.2.1.4 Advantages and Disadvantages of Subjective Evaluation

Subjective evaluation offers several key benefits in the assessment of dialogue systems.

1) In-depth Understanding of User Experience: Subjective evaluation can provide in-depth feedback on user experience, including user preferences, dissatisfaction,

and suggestions. This feedback helps development teams better understand user needs and expectations, improving conversational systems.

2) Identifying Potential Issues: Subjective evaluation can help identify potential issues and areas for improvement, including shortcomings in semantic understanding, response generation, and dialogue flow. Through users' subjective feedback, existing problems in the system can be recognized and addressed promptly.
3) Personalized Suggestions: Subjective evaluation can provide personalized questions and opinions, helping development teams understand the needs and preferences of different user groups. This customized feedback helps optimize system design and improve functionality, enhancing user satisfaction and experience quality.
4) Intuitive Effectiveness: Subjective evaluation is typically conducted through user surveys, focus group discussions, etc., and is characterized by its intuitive and effective nature. Rich and authentic feedback can be collected by directly communicating and interacting with users, providing valuable insights for system improvement.

However, subjective evaluation also has its drawbacks shown as follows:

1) Subjectivity: Subjective evaluation is influenced by subjective perceptions and viewpoints, leading to subjective preferences and personal emotional factors. Users may provide different subjective evaluations, resulting in subjectivity and inconsistency in evaluation results. Limitations: Subjective evaluation typically provides limited feedback information, making it challenging to comprehensively cover all aspects and performance indicators of the system. Evaluation results may be influenced by survey design, sample selection, and questionnaire design, leading to certain limitations.
2) Subjective Biases: Subjective evaluation may be influenced by users' biases and emotions, resulting in the subjectivity and distortion of evaluation results. Some users may provide subjective biased evaluations based on emotions, experiences, and expectations.
3) Difficulty in Quantification: Subjective evaluation results are typically unstructured textual information, making it difficult to quantify and analyze. Explaining and summarizing evaluation results requires considerable time and effort, making it less intuitive and straightforward than objective evaluation.

2.2.2 Objective evaluation

Objective evaluation of dialogue systems is crucial to assessing their performance and effectiveness. We can quantify the system's performance in various aspects through objective evaluation and derive objective metrics from the data to guide system improvement and optimization. This section will introduce the general principles and

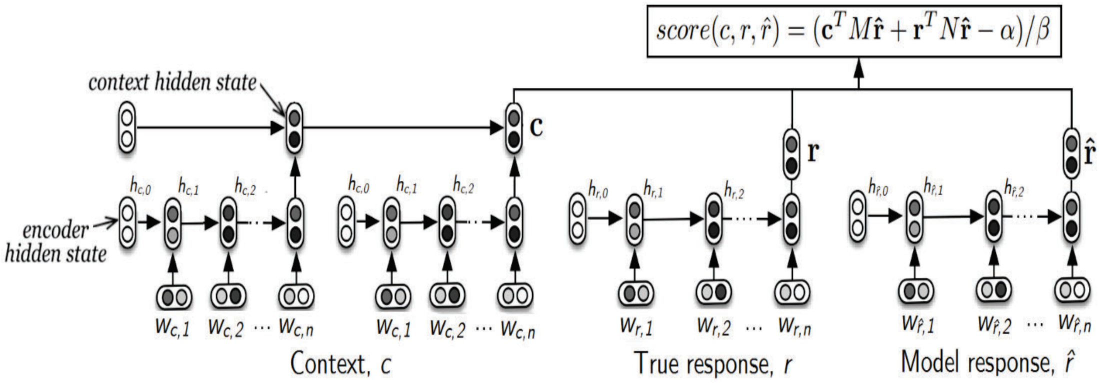


Figure 2.5: The basic structure of ADEM [1].

common methods of objective evaluation for dialogue systems and how to utilize objective evaluation to enhance system performance and user experience. In objective evaluation, commonly used metrics include word overlap evaluation metrics (such as BLEU, ROUGE, and METEOR) [9–11] and word vector evaluation metrics (such as Greedy Matching, Embedding Average, Vector Extrema, and perplexity) [69–71]. These metrics provide multiple perspectives to measure the performance of dialogue systems, ranging from word overlap to word vectors, offering effective tools and methods for comprehensive system performance assessment.

In enhancing dialogue systems, researchers have explored methodologies that simulate human evaluations to address the limitations of conventional objective assessment metrics. This research domain encompasses various approaches beyond the cited instances, including the utilization of recurrent neural network (RNN) methods [1], which gave rise to the Automatic Dialogue Evaluation Model (ADEM), aimed at predicting artificial ratings for dialogue responses.

Their feasibility was established through a comparison of ADEM ratings with traditional metrics. This method relies on manually annotated scores to train the network. To alleviate this issue, Zhang et al. [8] used a pre-training method to learn the encoder's parameters. In the original model, results are generated and used as inputs to an independent RNN. After training, the RNN can generate responses for specific texts under certain conditions. Concurrently, a structure similar to Generative Adversarial Networks (GANs) [7], comprising generators and discriminators, was employed to evaluate the similarity between responses generated by dialogue systems and humans. Both methodologies aim to simulate human evaluations through scoring, enhancing the quality of dialogue systems. At the same time, we can observe that most objective evaluation methods are based on text vectors. This means these methods primarily rely on converting the dialogue content into numerical vectors for analysis and comparison. As a result, these evaluation methods are particularly effective for assessing text-based dialogue systems. However, their applicability is limited, and they struggle to comprehensively cover dialogue systems involving multimodal inputs, such as speech, images, and videos. Consequently, existing objective evaluation methods may not provide accurate and comprehensive assessment results when dealing with more complex and diverse dialogue systems.

2.2.2.1 Advantages and disadvantages of objective evaluation:

Objective evaluation is based on objective standards and metrics, yielding relatively reliable results. Through objective data and metrics, the performance and effectiveness of conversational systems can be more accurately assessed such as:

1) Quantifiable Analysis: Results from the objective evaluation are typically quantifiable and analyzable, including performance metrics such as response time, accuracy, and fluency. These metrics facilitate development teams' quantitative analysis and comparison of system performance.
2) Comprehensive Coverage: Objective evaluation can comprehensively cover various aspects and functionalities of conversational systems, including semantic understanding, response generation, and dialogue flow. Through objective evaluation, the strengths and weaknesses of the system can be fully understood, and potential issues can be identified.
3) Objective Comparison: Objective evaluation can be used to objectively compare and analyze different systems, aiding users in selecting the system that best fits their needs. By comparing objective data and metrics, differences and advantages between systems can be better understood.

Objective evaluation may have limitations and not fully cover all users' needs and experiences. Some users' needs and preferences may not be fully reflected through objective metrics, resulting in incomplete evaluation results.

1) Fixed Thinking: Objective evaluation overly relies on data and metrics, potentially leading to fixed thinking and lack of innovation. Some system advantages and innovations may not be accurately assessed through objective evaluation, leading to neglect or underestimation.
2) 2) Unmeasurable Factors: Objective evaluation cannot fully capture users' emotions and subjective experiences, such as user satisfaction, preferences, and emotional responses. These factors may significantly affect user experience and system acceptance but are challenging to quantify and analyze through objective evaluation.

2.3 The Problem of evaluating dialogue systems

As described in Section 2.1 and Section 2.2, with the continuous advancement of hardware and the development of artificial intelligence technologies, multimodal systems demonstrate significant application potential across various industries in the future. In fields such as healthcare, education, smart homes, and intelligent customer service, multimodal dialogue systems are expected to play a crucial role. For instance, multimodal systems integrating speech, visual, and gesture recognition can provide more comprehensive assistance in medical diagnosis; in education, systems combining images, speech, and text can offer a more immersive learning experience. This cross-modal information fusion will further drive the intelligent development of various sectors.

However, existing evaluation methods for dialogue systems predominantly focus on text modalities, neglecting the influence of other modalities, such as speech and visual information. This leads to existing evaluation methods having many shortcomings in evaluating multimodal dialogue systems, which limit the comprehensive and practical evaluation of the performance of dialogue systems. Specific problems are shown as follows:

(1) This single-modal evaluation is inadequate in reflecting the overall performance of multimodal dialogue systems in real-world applications.

(2) Current mainstream automated evaluation metrics, such as BLEU and ROUGE in Section 2.2.2, although reflecting the quality of system-generated responses to some extent, do not consider information from multiple modalities such as speech, images, and video. This omission may lead to ambiguous dialogue understanding and consequently reduce user experience. This discrepancy can lead to poor system performance in practical applications despite high evaluation scores in laboratory tests.

(3) Consistency of emotional responses at the dialogue and turn levels: In multi-turn dialogues, overall user impression depends on the performance of individual exchanges and the emotional consistency throughout the entire conversation. However, existing evaluation methods typically focus only on the quality of single-turn dialogues, neglecting emotional changes in the context of the whole discussion. We believe that exchange-level evaluation and dialogue-level evaluation are complementary. Exchange-level evaluation helps identify and improve the quality of individual interaction turns, while dialogue-level evaluation provides a comprehensive perspective on the overall dialogue experience. Combining both approaches allows for a more thorough assessment of the dialogue system's performance.

(4) Ignoring users' personal information: Existing evaluation methods often fail to adequately consider individual user differences, such as age, gender, and personal-

ity. The experience and impression of different user groups using the same system can vary significantly, yet these differences are frequently overlooked in traditional evaluations.

These issues indicate that the current evaluation of multimodal dialogue systems is still insufficient, highlighting the need for evaluation methods that comprehensively consider multimodal data, users' personal information, and the overall dialogue experience.

To address issues (1) and (2) we evaluate multimodal dialogue systems in the Chapter 3 using user impressions. This work considers evaluation methods for multiple modalities, such as speech and vision, which provide a more comprehensive understanding of the system's performance in real-world applications. This ensures that multimodal systems can operate effectively across different inputs and outputs and offers a more accurate assessment. Additionally, improving evaluation methods to reflect better user experience aids in developing technically proficient and user-friendly systems.

To tackle the issue (3), in Chapter 4, we integrate user experience at both the exchange and the dialogue levels to create systems that maintain emotional consistency throughout interactions and provide more engaging experiences.

To tackle the issue (4), in Chapter 5, we take into account users' personal information to design systems that are more inclusive and adaptive, thereby enhancing the experience for diverse user groups.

In summary, this work's primary objective is to thoroughly and precisely evaluate the performance of multimodal, non-task-oriented dialogue systems based on user impressions. This approach aims to build a solid foundation for improving user experience and interaction efficiency. We seek to advance the application and development of multimodal dialogue systems across various industries by assessing user feedback and system effectiveness. This evaluation will significantly contribute to refining system design, optimizing user interactions, and promoting the integration of multimodal dialogue systems into diverse practical contexts.

Chapter 3

Multimodal User Satisfaction Recognition for Non-task Oriented Dialogue Systems

3.1 Overview

With natural language processing and speech recognition development, spoken dialogue systems, such as those of Amazon Alexa, SIRI, and Google Assistant, are used in many fields. There is great interest in developing non-task-oriented dialogue systems such as chatbots and open-domain dialogue systems [82] [83]. While improving the quality of the non-task-oriented dialogue system is vital for the user dialogue experience, it is not easy to evaluate how well the system works; therefore, an automatic evaluation of whether a user could be satisfied through the dialogue experience is crucial for developing and improving dialogue systems. Two unexplored problems exist in the current satisfaction recognition models. First, almost all previous research [84, 85] focused on user satisfaction modeling in text-to-text dialogue systems rather than multimodal ones. In evaluating user satisfaction in multimodal dialogue systems such as embodied conversational agents (ECAs) and social robots, the satisfaction level is observed from spoken dialogue contents and users' acoustic and visual nonverbal behaviors. Second, most previous work [78, 79, 85] recognized satisfaction labels at the turn level (per utterance or exchange) to ensure natural interactions. However, a user essentially feels satisfaction throughout the whole conversation; thus, the system designer needs to analyze satisfaction at the turn level and overall satisfaction concerning the entire conversation. We define overall satisfaction as "dialogue-level satisfaction".

This study presents a multimodal model to recognize user satisfaction at the dialogue level using multimodal features observed from users, which is suitable for evaluating non-task-oriented dialogue systems. The multimodal features extracted from each exchange (a pair of system and user utterances) are input to each unit of the sequence models (RNN, LSTM, and GRU). The output is the dialogue-level satisfaction annotated by the user who has talked with the system. We utilized a novel multimodal dialogue data corpus to construct these sequence models, including dialogue-level (overall) satisfaction labels, exchange-level sentiment annotation, and multimodal data, including spoken dialogue transcription, audio signals, face images, and body motion data. We used five feature sets to recognize user satisfaction. Meanwhile, this study analyzes the contributions of different features to user satisfaction.

To validate the proposed methods based on machine learning (ML models), we compared the performance of the proposed model with two types of human methods. The first method is a sequence model that recognizes dialogue-level user satisfaction from exchange-level impression annotations (**Human model (1)**). In the second method, the system operator (called "Wizard") directly recognizes the dialogue-level user satisfaction (**Human model (2)**). The main contributions of this study can be summarized in the following three aspects.

Multimodal user satisfaction recognition: This task is unexplored and a new challenge in the multimodal human-agent interaction domain. We proposed a multimodal approach utilizing sequence modeling algorithms to recognize user satisfaction at the dialogue level in multimodal interactions. This study combined audio, visual, and text features to recognize user satisfaction. We demonstrate that multimodal features performed better than unimodal features in Section 3.6.1.

Comparison between the contribution of multimodal features and exchange-level annotation: Many studies have focused on proposing multimodal models for recognizing the exchange-level sentiment label. However, how exchange-level labels correlate with dialogue-level satisfaction is still unclear. We first explored the relationship between exchange-level and dialogue-level annotations. Then, this study used exchange-level annotation scores as manual features to recognize user satisfaction and compared the results with those obtained multimodal (automatically obtained) features. The comparison between the two feature types is described in Section 3.6.2.

Comparison between the ML models and human model: To validate the effectiveness of the multimodal ML models, we compared the recognition result of the ML models with the user satisfaction score annotated by a system operator (Wizard). The comparative analysis in Section 3.6.3 demonstrates the challenging nature of the task and the contribution of the automatic multimodal recognition technique to estimating user satisfaction.

3.2 Related work

Intelligent conversational agents have become widely used in daily life. Measuring the performance quality of a dialogue system is a critical component during the development process. Initially, some researchers used dialogue efficiency and costs related to the length of the dialogue or task success to measure the performance [86]. However, when interacting with simulated or recruited users, there is no task success information in non-task oriented conversations (such as small talk and multi-domain dialogue).

To develop an appropriate and correct system, recent studies have focused on user-centered criteria that are defined based on human judgments to approximate the usability of dialogue systems. An annotated score, such as “user satisfaction”, is recognized by using machine learning techniques. For example, Engelbrecht et al. [78] used dialogue actions as input features to recognize user satisfaction at the exchange-level. Higashinaka et al. [68] used annotations by experts who observed the dialogue as target variables to model user satisfaction. Since most input features are annotated manually, this method is inconvenient and inefficient for online applications. Schmitt and Ultes [79] used dialog manager-related parameters, the semantic meanings of which were extracted automatically as input features, to recognize the median rating of several expert ratings at the exchange-level.

Regarding experimental methods, some researchers have regarded the user satisfaction recognition task as a sequence problem. Hara et al. proposed an N-gram model trained using sequences consisting of dialog acts to recognize user satisfaction [85]. A hidden Markov model (HMM) was also used to model user satisfaction transitions in dialogues [87]. However, the experiment has shown that Support Vector Machine methods that did not use sequence information were performed better than HMMs [79]. User satisfaction recognition is a temporal task that should benefit from time-series dialogue data. To investigate the effect of temporal information, Ultes et al. extended the set of temporal features to different levels, and the results showed that interaction parameters (e.g., ASR performance) at the window and dialogue levels that provide temporal information significantly affect interaction quality [88]. Recently, deep learning techniques have also been applied for user satisfaction recognition tasks. Ultes et al. [84] proposed a recurrent neural network (RNN) to improve recognition accuracy. To eliminate the heavy reliance on handcrafted temporal features, they presented a deep learning-based Interaction Quality (IQ) estimation model that utilizes recurrent neural networks’ capabilities to learn temporal information automatically.

Concerning features, all previous research focused on using linguistic features and dialogue content to recognize user satisfaction with a text-based dialogue system. In a multimodal dialogue system, it is well known that sensing multimodal information from the user helps recognize the inner states of the users, such as the sentiment level. Recently, many studies have proposed that multimodal information, including visual and acoustic features, improves sentiment recognition accuracy. The temporally selective attention model [89], multi-attention recurrent network [90], memory fusion network [91], and tensor fusion network [92] were proposed for multimodal sentiment analysis. Hirano et al. proposed a multitask deep learning neural network model (MT-DNN) using multimodal features to recognize the user exchange-level sentiment toward spoken dialogue systems [93].

The main difference between this work and previous works is summarized as follows.

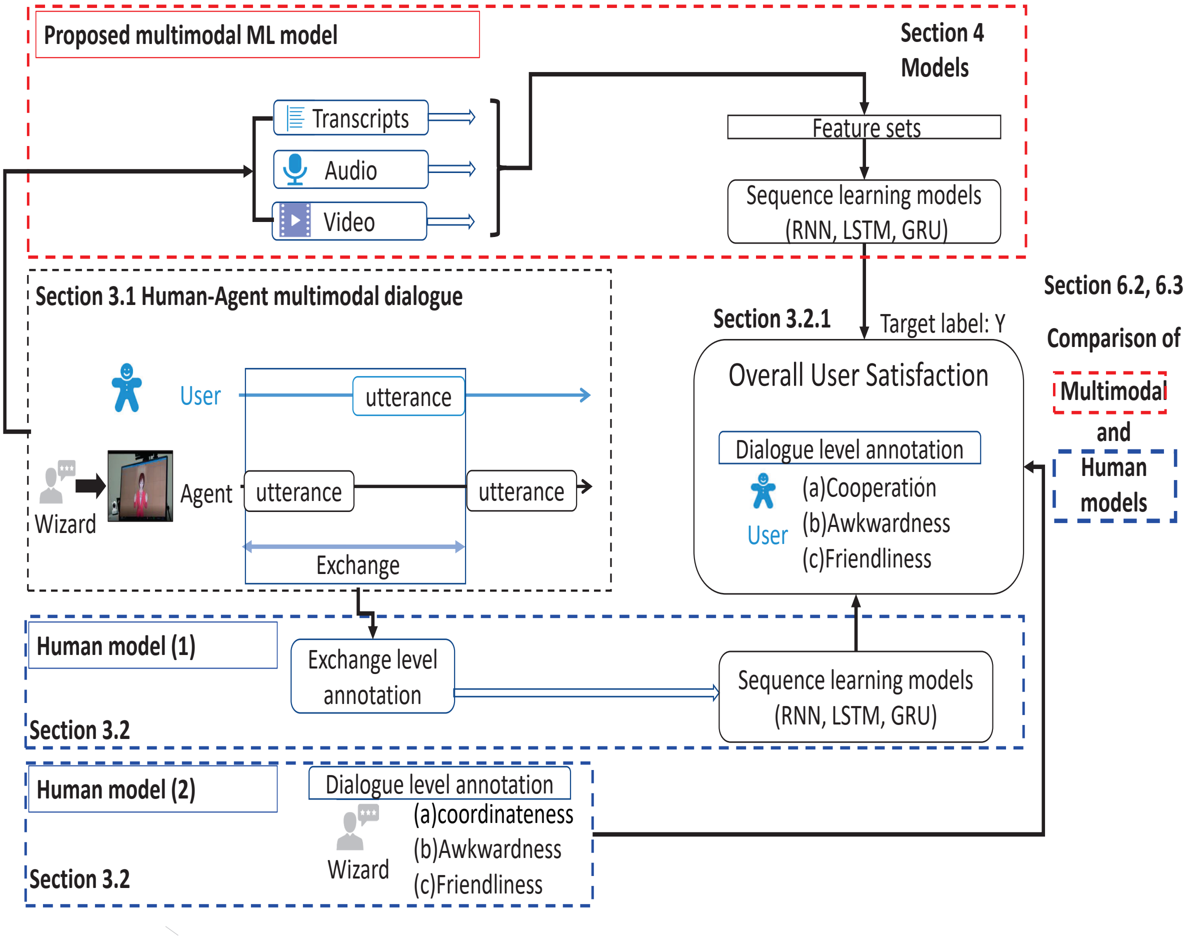


Figure 3.1: Overview of the estimation of the user's satisfaction at the dialogue level

Though most of the previous studies have focused on recognizing user satisfaction at the exchange level, the main focus of this study is to recognize user satisfaction at the dialogue level to evaluate non-task-oriented dialogue systems. We also present a multimodal modeling method based on the user's performance in the conversation.

3.3 Data and annotations

Figure 1 shows the overview of this research. In this section, we describe the multimodal dialogue dataset and its annotations.

3.3.1 Data recording

This study was conducted on two shared multimodal dialogue datasets [1] one named Hazumi1902 and one named Hazumi1911, in which recording settings were almost the same [94]. Both corpora were arranged to record facial videos, audio data, and upper body data, and a virtual agent called MMD-Agent was used as the interface to communicate with participants manipulated by an operator (Wizard) in another room. In this system, the operator could select a topic, utterances on the topic, and general responses used in conversation. To shorten the time interval before the machine responded, the operator was well-trained and had time to select the next utterance while the participant spoke (approximately 10 seconds).

Regarding acoustic signals and body posture, audio and posture of the upper body were recorded by a Kinect sensor. The posture information was recorded at 30 fps. Each participant's voice was recorded as a 16 kHz WAV file. The number of participants in the two corpora was 60, which included 25 males and 35 females. The participants' ages ranged from 20 to 70 years, and the public recruited them through a recruitment agency.

3.3.2 Annotations

The data corpus included two kinds of annotations. One was annotation at the dialogue level, and the other was annotation at the exchange-level. First, the dialogue-level annotations were used as the target labels in this study to develop the recognition model of user satisfaction. Second, the exchange-level annotations, including the user sentiment, indicate the user's perceptions of the system; therefore, these annotations were used as partial information for understanding the dialogue-level satisfaction.

[1] The doi is doi/10.32130/rdata.4.1

Table 3.1: Usage of different annotations

Unit	**Type**	**Annotator**	**Usage**
Exchange-level annotations	Topic continuance Third sentiment	Third-party coders	Input features for Human model (1)
	Self sentiment	Dialogue users	
Dialogue-level annotations	Coordinateness Awkwardness Friendliness	(i): Dialogue users (ii): Wizard	(i): Target label (ii): Using a human model to recognize (i) (Human model (2))

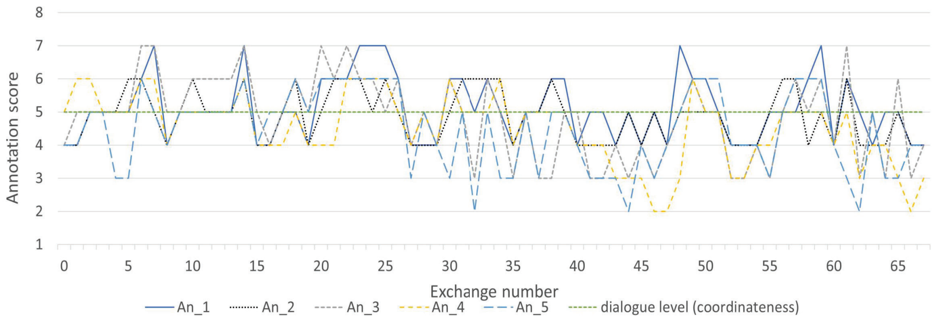


Figure 3.2: Example of annotation in a conversation. The An_1 to An_5 denote the topic continuance level annotated per each exchange by the five annotators.

3.3.2.1 Dialogue-level annotations

For dialogue-level annotations, defining user satisfaction based on one criterion is difficult. For this reason, this study used a questionnaire with 18 labels relating to the user's impression of the dialogue proposed in [95].[2] The questionnaire measured interpersonal communication cognition as a social skill. The 18 items were "well-coordinated", "boring", "cooperative", "harmonious", "unsatisfying", "uncomfortably paced", "cold", "awkward", "engrossing", "unfocused", "involving", "intense", "unfriendly", "active", "positive", "dull", "worthwhile", and "slow". Kimura et al. [96] analyzed the rapport in dyadic interactions (60 pairs, with 120 subjects). They reported that three labels ("well-coordinated", "awkward" and "unfriendly") were carefully extracted as representative labels from the 18 labels by conducting a factor analysis. Based on this finding, we defined the scores of the three labels as the user satisfaction level, and used these three values as grand-truth values for machine learning. We use the three labels ("well-coordinated", "awkward", and "unfriendly") as "coordinateness", "awkwardness" and "friendliness" in this study. Each label was evaluated on an eight-point scale from 1 to 8.

We asked the participants to annotate all 18 labels. Still, we asked the Wizard to annotate only the three labels ("well-coordinated", "awkward" and "unfriendly") to reduce the burden on the Wizard in annotating the rapport scores of all 60 participants after the dialogue[3].

3.3.2.2 Exchange-level annotations

The exchange was defined as the section beginnings from the start time of a system utterance and endings at the start time of the following system utterance. Exchange-

[2]We used the Japanese version [96] translated from the original questionnaire.

[3]Though we also conducted similar questionnaires before each dialogue, we did not use them

level annotations were collected to analyze the user's internal state in each exchange unit. Hirano et al. [93] and Katada et al. [97] presented multimodal models to recognize the exchange-level annotations. Three types of annotations were given at the exchange-level as follows:

Topic continuance: The topic continuance label was a degree indicating whether the topic should be changed. Five human coders assigned such labels depending on whether the system should have continued the current topic or changed the topic in the following system's utterance. The labels of the scores ranged from "strongly change the topic" 1 to "strongly continue the topic" 7, as shown in Figure 3.2.

External sentiment: When a participant communicated with the dialogue system, the participant had different sentiments during each turn. Human coders annotated the external sentiment level per exchange with scores ranging from 1 (the participants seemed bored with the dialogue) to 7 (participants seemed to enjoy the dialogue) while watching recorded videos of the dialogues.

Self-sentiment: This annotation was similar to the external sentiment annotation. Self-sentiment labels were assigned scores ranging from 1 to 7, divided into two categories. Positive sentiments included "enjoy talking" and "satisfied with the talk", and negative sentiments included "want to stop talking" and "confused about the system utterances".

Based on these definitions, 5373 exchanges obtained from 60 participants were annotated. The agreement scores of the annotators measured by Cronbach's alpha were 0.83 for the topic continuance and 0.86 for the external sentiment.

3.3.2.3 Usage of difference annotation

As described in Sections 3.3.2.1 and 3.3.2.2, two types of annotations were used in this study. As shown in Table 3.1, three of exchange-level labels were used as input features to recognize user satisfaction on the dialogue level. The details of the experiments are presented in Section 3.5.2. For dialogue-level annotation, both the user and Wizard annotated the user satisfaction labels at the dialogue level after the conversation. This study used user annotations at the dialogue level as target labels. We used the Wizard's annotation to evaluate the user satisfaction as a "human" model based on the Wizard's subjectivity. The results of estimations by the Wizard and models trained with multimodal features facilitated the comparison of the performances of humans and ML models.

Table 3.2: Pearson correlation coefficient between exchange annotations and the dialogue-level annotations

	Topic continuance	**External sentiment**	**Self-sentiment**
Coordinateness	0.17	0.24	0.30
Awkwardness	−0.16	−0.18	−0.36
Friendliness	0.07	0.05	0.29

3.3.3 Relation between exchange-level annotations and dialogue-level annotation

We used the Pearson correlation coefficient to calculate the correlations between dialogue-level annotation and the average value of all exchange-level annotations in one dialogue to explore the relationship between exchange labels and dialogue labels. Generally, it belongs to a weak correlation when the correlation coefficient is higher than 0.1; if the correlation is higher than 0.3, it is moderate. Table 2 shows the correlation matrix between the dialogue level of user satisfaction after dialogue annotations and the average value of all exchange-level annotations. Compared with the correlation between the third-party (topic continuance and external sentiment) annotations on the exchange and dialogue-level annotations, the self-sentiment annotations on the exchange had a higher correlation with dialogue-level annotations. We also found that all exchange annotations were positively correlated with the coordinateness and friendliness labels and negatively correlated with the awkwardness labels. The self-sentiment annotation had the highest correlation with the dialogue-level annotation.

However, we found that the correlation between exchange-level labels and dialogue-level labels was not strong, indicating that exchange-level annotation cannot accurately express user satisfaction at the dialogue level. For this reason, it is necessary to recognize user satisfaction directly at the dialogue level. In this study, the exchange-level annotation feature was used as a manual feature to identify the user's satisfaction at the dialogue level. We analyzed the exchange-level annotation feature results and compared them with the multimodal results in Section 3.6.2.

3.4 Multimodal user satisfaction modeling

3.4.1 Multimodal feature extraction

3.4.1.1 Audio feature

This study extracted acoustic features at the exchange level as the emotional information in speech using the speech feature extractor OPENSMILE [98]. The acoustic features corresponded to the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which achieves high performance in emotion-related fields. These features were extracted for each speaker turn and normalized by each speaker after extraction. Finally, we obtain a 88 dimensions vector.

3.4.1.2 Linguistic feature

We extracted two linguistic features from the manual transcription of spoken dialogue contents: **Part of speech:** The sentences were segmented into words and annotated with universal part-of-speech (POS) tags using Stanza NLP [4]. The PoS tag set was composed of 17 types: "adjective", "adposition", "adverb", "auxiliary", "coordinating conjunction", "determine", "interjection", "noun", "numeral", "particle", "pronoun", "proper noun", "punctuation", "subordinating conjunction", "symbol", "verb", "other". The PoS categories (nouns, verbs, etc.) in a user's utterance were counted.
BERT (Bidirectional Encoder Representations from Transformers [57]): In this study, we used a pre-trained model on only Japanese text (using Wikipedia) [99]. We used this model to extract features from the text at the exchange level, and finally, we obtained a 768-dimensional text representation vector.

3.4.1.3 Visual feature

This study used three-dimensional coordinate data for each upper body joint, estimated from a Microsoft Kinect V2 with a depth sensor.
Body activity features: This study used three-dimensional coordinate data for each joint of the upper body, which was estimated from a Microsoft Kinect v2, to extract motion features. We used five points of body motion, which included the left shoulder, right shoulder, left hand, right hand, and head. We denoted the three-dimensional coordinate data of each body point at tth-frame as $w(t) = x, y, z$. We calculated the absolute value of velocity between two frames as $|v(t)| = |w(t+1) - w(t)|$ and calculated the absolute value of acceleration between frames as $|a(t)| = |v(t) - v(t-1)|$. After $v(t)$ and $a(t)$ were calculated, we used the maximum value of acceleration, and the

[4] https://github.com/stanfordnlp/stanza

maximum, mean, and standard deviation of velocity in the user turn as body activity features. Finally, the body activity feature set included 20 dimensions in total.

Facial landmark feature: OpenFace [100] software outputs the three-dimensional coordinates of 68 facial landmarks in each frame. This study chose ten facial landmarks, including 2 points on each eye, 4 points around the mouth, and two on the eyebrows. We adopted the same method used for body feature tracking. We extracted the maximum acceleration value and the maximum, mean, and standard deviation of the velocity for each user exchange turn as facial features. Finally, we obtained a 40-dimensional vector.

Action units: Facial expressions display emotional states, which help regulate turn-taking during the conversation. This is often represented using facial action units (AUs), which objectively describe facial muscle activations [101]. In this study, we used OpenFace to obtain 18 types of AUs rated between 0 and 1, indicating absence and presence, respectively. Then we calculated the average of each AU in exchange for facial AU features (18 dim). Overall, 58 dimensions of facial features were used in this study.

3.4.2 Models

To recognize user satisfaction at the dialogue level, a machine learning model needs to capture dynamical change in multimodal behaviors while the user is talking with the system. To model the sequence of multimodal behaviors, we utilized the following three sequence models; Recurrent neural network (RNN), long short-term memory (LSTM) and gated recurrent unit (GRU) models.

The multimodal features extracted from each exchange were input to each RNN, LSTM, and GRU unit in the proposed multimodal model. This study used the early fusion method and the unimodal features (audio a_t: 88 dim., linguistic l_t: 785 dim., and visual v_t: 78 dim.) extracted from the t-th exchange were concatenated into one vector x_t (951 dim.). The input of these recurrent neural network models was x_t ($1 \leq t \leq T$). In all models, two-recurrent (hidden) layers with 128 units ($h_t^{(1)}$, $h_t^{(2)}$) are used to extract the features from the sequence input vector x_t with T exchanges. We obtained two final hidden states $h_T^{(1)}$, $h_T^{(2)}$ (2 (layers)*128 (units)) from the recurrent layers. A fully connected layer followed the recurrent layer to project the output (2*128) from the recurrent layers into the output layer. For a classification task, the output layer containing two units and the log-SoftMax function was used to output the probabilities of the different user satisfaction S_c. The sigmoid unit was used for a regression task to output the estimated value S_r of the user satisfaction level (1-8).

3.5 Experiments

The experiments aimed to recognize user satisfaction at the dialogue level. We evaluated user satisfaction recognition's accuracy through classification and regression tasks. Three

research questions were addressed, each corresponding to a subsection in Section 3.6.
RQ1: Do multimodal features contribute to improving user satisfaction recognition?
RQ2: Which features are more effective for user satisfaction recognition, multimodal or exchange-level annotation?
RQ3: How does the recognition of multimodal models perform compared with human subjective perception?

3.5.1 Experimental setting

3.5.1.1 Regression task setting

The regression tasks aimed to fit the labels of dialogue base on different feature sets. The mean squared error (MSE) was calculated using the square of the difference between the actual and estimated values, which were then summed and averaged. It was convenient to take the squared derivative of the results. This work used the MSE as the loss function for all regression tasks.

3.5.1.2 Classification task setting

The binary classification datasets were developed as follows. All dialogue-level label annotated scores (1-8) were converted into binary values (high and low) with a threshold of 4 (neutral state). The high/low data points for the three target labels at the dialogue level were 38/22 for the coordinateness label, 32/28 for the awkwardness label, and 49/11 for the friendliness label, respectively. We used the F1-score to evaluate the accuracy of imbalanced datasets in which the number of samples differed between the two classes.

3.5.1.3 Hyperparameter setting and evaluation

We used the same parameters in all models to evaluate the comparative models under equivalent conditions. We used the Adam optimizer, set the learning rate to 0.001, and set the total number of epochs to 30. Five-fold cross-validation was conducted, and their average F1-score was reported.

3.5.1.4 Combination of multimodal futures

According to the findings in previous works, linguistic features were the key descriptors in recognizing user satisfaction. For this reason, we set the unimodal model with a linguistic feature set as the baseline model. In addition to the baseline model, we prepare four combinations of unimodal features (audio, visual and linguistic) to analyze the effectiveness of the verbal-nonverbal and nonverbal multimodal models (without linguistic features).
(1) L: model trained with Linguistic features (baseline)

(2) A+V: model trained with Acoustic + Visual features
(3) A+L: model trained with Acoustic + Linguistic features
(4) V+L: model trained with Visual + Linguistic features
(5) ALL: model trained with Acoustic + Visual + Linguistic features

3.5.2 Comparative methods

We prepared two human models as comparative methods with the proposed multimodal models.

3.5.2.1 Human model (1) using exchange-level annotations

In this group experiment, five experts annotated two labels (external sentiment and topic continuance) in each exchange turn with scores ranging from 1 (the participants seemed bored with the dialogue) to 7. For the external sentiment and topic continuance labels, the averaged annotated scores were calculated per the t-th exchange and then combined with the self-sentiment annotation at t-th exchange as manually annotated features total three dimensions a_t $(1 \leq t \leq T)$. a_t is input features for the sequence models. The network architecture is the same as the multimodal models described in Section 3.4.2.

3.5.2.2 Human model (2) using subjective evaluation of Wizard

In the second group experiment, the Wizard's annotation was regarded as the result of human recognition. For the classification methods, similar to the annotation procedure described in Section 5.1, the Wizard's annotation results were divided into high and low satisfaction categories, then F1-score was calculated. We computed the MSE of the Wizard's annotations and user annotations for the regression.

Table 3.3: Regression results of each multimodal combination for three user satisfaction labels (Acoustic (A), Visual (V), Linguistic (L), and Exchange annotation (Exchange An). The accuracy denotes the mean squared error (MSE). The bold values indicate the best MSE for the performance index.)

		Proposed					Human models	
Labels	**Model**	**L**	**A+L**	**A+V**	**L+V**	**ALL**	**(1) Exchange An.**	**(2) Wizard**
Coordinateness	**RNN**	3.66	**3.14**	3.28	4.18	3.56	3.66	3.38
	LSTM	3.58	3.05	3.17	4.14	**2.93**	3.56	
	GRU	3.35	**3.21**	3.28	3.92	3.32	3.75	
Awkwardness	**RNN**	4.46	4.43	4.59	**4.31**	4.48	4.41	5.70
	LSTM	4.58	4.63	4.58	**4.00**	4.54	4.44	
	GRU	4.57	4.52	4.52	**4.22**	4.34	4.33	
Friendliness	**RNN**	**2.87**	3.23	3.26	3.14	3.45	3.13	4.15
	LSTM	**2.88**	3.06	3.29	3.03	3.13	3.09	
	GRU	3.02	3.25	3.28	**2.87**	3.14	3.32	

Table 3.4: Binary classification F1-score of each multimodal combination of three user satisfaction labels (Acoustic (A), Visual (V), Linguistic (L), and Exchange-level annotation (Exchange An). The bold values indicate the best F1-score)

		Proposed Method					Human models	
Labels	**Model**	**L**	**A+L**	**A+V**	**L+V**	**ALL**	**(1) Exchange An.**	**(2) Wizard**
Coordinateness	**RNN**	0.67	0.72	**0.73**	0.71	0.72	0.61	0.72
	LSTM	0.70	0.74	0.75	0.65	**0.76**	0.55	
	GRU	0.69	0.68	0.74	0.63	**0.75**	0.55	
Awkwardness	**RNN**	0.59	0.53	0.61	**0.72**	0.59	0.60	0.58
	LSTM	0.66	0.58	0.61	**0.68**	0.63	0.54	
	GRU	0.58	0.62	0.56	**0.63**	0.61	0.56	

3.6 Results

Tables 3.3 and 3.4 show the regression and binary classification results of satisfaction label recognition by the RNN, LSTM, and GRU, respectively, based on the five feature sets. In this section, we first compare the regression performance of these models' MSE and classification F1-score. Second, we compare the results of the proposed multimodal models with that of the model trained with the exchange-level annotation features. Finally, we compare the performance of the multimodal models to that of the "human model" based on subjective evaluation by the Wizard.

3.6.1 Comparison between unimodal and multimodal features (RQ1):

Columns 3 to 7 in Table 3 show the regression results: the MSE values of user satisfaction labels generated by training with different multimodal feature sets.

We observed that the multimodal models yielded the best performance for the well-coordinateness and awkwardness labels among all models. The ALL feature set (A+V+L) and L+V feature set produced the lowest MSE values (2.93 and 4.00) in the LSTM method for the coordinateness and awkwardness labels. Most feature fusion models (A+L and A+V) performed better regarding MSE than those using linguistic features. The unimodal set (L) achieved the best result for the friendliness label, with an MSE of 2.87. For the awkwardness label, L+V produced the lowest MSE for all methods.

Columns 3 to 7 in Table 4 present the F1-scores of the classification of user satisfaction labels (except friendliness) generated by training with different multimodal feature sets. Due to the imbalance in the friendliness label (49/11), all models overfitted this label. For the coordinateness dialogue label, in the LSTM and GRU methods, the ALL feature set (A+V+L) produced the best F1-scores (0.76 and 0.75) among all feature sets. Similarly, in the regression task, L+V yielded the best F1- score for all methods for the awkwardness label.

3.6.2 Comparison between multimodal and exchange-level-annotation features (RQ2):

Column 8 in Table 3 presents the regression results based on the features of the exchange-level annotation (annotation features). We report the lowest MSE of each label result for the annotation features as follows. The MSE was 3.56 for the coordinateness label with LSTM, 4.33 for awkwardness with the GRU, and 3.09 for friendliness with the LSTM. Although the performance of the annotation features was better in some cases, the multimodal feature sets achieved the best performance considering all the results.

Column 8 in Table 4 shows the user satisfaction classification results based on the annotation features. The results demonstrate that the RNN achieved the best performance for the coordinateness dialogue label, with an F1-score of 0.61. All multimodal features performed better than the annotation features for the coordinateness label. Similar to the coordinateness label, the RNN achieved the best performance (0.60) for the awkwardness label, which was better than that obtained by the RNN method trained using the A+V+L (0.59) feature set. The multimodal feature performance was better in all cases for the other methods (LSTM and GRU). Overall, the results show that multimodal features can improve recognition performance. In this evaluation, we used the two-layered RNN-based models for comparing models. However, the network architecture is not optimized for the Human model (1) with the low dimensional input (three dimensions), so the fair evaluation using the optimized network architecture per each model is future work.

3.6.3 Comparison of human model and ML models (RQ3):

Column 9 in Table 3 lists the regression results of the human model, in which the Wizard estimated user satisfaction. For the coordinateness label, the human model yielded an MSE of 3.38, which was better than some feature sets (A, L+V) but worse than the best result (2.93) achieved by LSTM. The regression results for the awkwardness and friendliness labels were worse than almost all ML model results.

Column 9 in Table 4 shows the classification results of the human model, in which the Wizard estimated user satisfaction. For the classification task, we calculated the F1-scores of binary classifications based on the annotations by the Wizard. Similar to the regression results, the human model obtained a better F1-score (0.72) than some feature sets (A, L+V) for the coordinateness label. In contrast, the result (0.76) of the LSTM model was higher than that of the human annotator. Most ML models performed better than the human models for the friendliness label. Overall, both regression and classification results indicate that the performance of the multimodal model was higher than that of the human model.

3.7 Discussion

3.7.1 Feature analysis

3.7.1.1 Contribution of each modality

To analyze the contribution of each modality to three satisfaction labels on classification tasks. We use ablation experiments, in which a GRU model was trained by removing feature sets individually. If the F1-score decreased, the removed feature set was effective

Table 3.5: Contribution of each modality feature to two labels in GRU (Diff denotes the difference in F1-scores for cases in which a specific modality was removed)

Modality	**Label**			
ALL (A+V+L)	**Coordinateness**		**Awkwardness**	
	0.75		0.61	
Remove modality	F1	Diff	F1	Diff
Acoustic	0.63	0.12	0.63	-0.02
Facial	0.71	0.04	0.61	0.00
Action Unit	0.68	0.07	0.60	-0.01
Body	0.77	-0.02	0.54	0.07
Linguistic	0.74	0.01	0.56	0.05

for the classification. On the contrary, if the F1-score improved, the removed feature set was ineffective for classification. Table 5 shows the binary classification recognition performance of the GRU model on user satisfaction labels (except friendliness) trained with feature sets after each feature set was excluded. This table shows that the acoustic feature set was the most effective (+0.12) for the coordinateness label. The second most effective feature set was facial features (+0.10). The linguistic features were less effective. The results indicated that non-linguistic features performed better than linguistic features in identifying the coordinateness label. For the awkwardness label, the body features (+0.07) and linguistic features (+0.05) yielded better values, meaning that body motion and linguistic features effectively recognized the awkwardness label. The acoustic features were less effective. For the friendliness label, the model with linguistic features performed better than the model with multimodal features (refer to Table 3). The results demonstrate that linguistic features can improve performance more than other modality features. However, the difference is 0.12 in the maximum case, which is insignificant. Therefore, analyzing the specific features or frames in the sequence data, which significantly improves accuracy, is essential for future work.

3.7.1.2 Comparison between unimodal and multimodal

The recognition performance was im- proved with multimodal features for the coordinateness and awkwardness labels. For the friendliness label, the difference in accuracy between the unimodal and the best multimodal models was insignificant (refer to Tables 3.3 and 3.5). We analyzed the results of Section 3.6.1 and the ablation experiments. For the coordinateness label, communication is a cooperative activity involving coordinated behaviors [102]. Participants in conversation spontaneously adjust facial expressions, postures, pronunciation and speech rates [103–105]. In this study, the multimodal fusion set (ALL) produced a better performance for the coordinateness label. Participants show

	Human model: Estimated low	Human model: Estimated High	Machine model: Estimated low	Machine model: Estimated High
Actual low	37%	10%	42%	5%
Actual high	31%	21%	28%	25%

Figure 3.3: Confusion matrix of the binary classification task for the awkwardness label (ML models: LSTM regression result using the L+V feature set), and human model (annotation by the Wizard))

a negative attitude when embarrassed in dialogue with the agent for the awkward label. They do not often physically express their feelings and communicate with the agent. At the same time, the degree of embarrassment has an important relationship with the participation attitude in the dialogue. Body features were the most effective (refer to Table 3.5). The result partially aligned with the finding that body features (hand and head movements) are closely related to embarrassment [106]. For the friendliness label, the linguistic feature performed better in most cases in this study.

3.7.2 Comparison between Multimodal recognition and human perception

As shown in Section 3.6.3, we found that the performance of our proposed model was better than that of the human model (2). To further analyze the difference in accuracy between the human and ML models, we evaluated the regression accuracy of 60 participants for the awkwardness label with the LSTM model using 5-fold cross-validation. This study divided the regression values into binary values (high and low). We considered the threshold of 4.5 because the regression result is continuous values. The evaluation using other threshold values is set as future work. We calculated the confusion matrix of the machine and human (Wizard) scores to observe the overall classification. As shown in Figure 3, the recognition performance of the ML models was better than that of the human model for the high and low awkwardness labels. However, both the ML and human models showed false low-level recognition results (true high embarrassment was mistaken for low embarrassment), accounting for 31 % and 28 % of the total samples, respectively. This result suggests that humans and machines have difficulty identifying high participant awkwardness at the dialogue level. In addition, compared to other labels, the regression MSE result for the awkwardness label was larger (refer to Table 3.3).

3.8 Chapter Summary

In this chapter, we introduced a multimodal user satisfaction recognition model explicitly designed for evaluating non-task-oriented dialogue systems at the dialog level. The model utilizes a unique multimodal user-system dialogue data corpus to capture the nuances of user satisfaction. Our study used LSTM, RNN, and GRU structures to incorporate contextual information. Our investigation yielded promising results, demonstrating that multimodal features outperformed unimodal, exchange annotation, and even human models in user satisfaction recognition. This underscores the reliability of our proposed model in identifying user satisfaction at the dialogue level. However, we acknowledge that there is still room for improvement in multimodal user satisfaction recognition.

In this study, we concatenated feature vectors from different modalities into a single feature vector for training. Moving forward, our focus will be refining the integration of multiple features to enhance performance in these tasks. Additionally, we aim to delve deeper into the relevance of specific multimodal features in user satisfaction recognition. By addressing these areas, we aim to further enhance the effectiveness and applicability of our multimodal user satisfaction recognition model in evaluating non-task-oriented dialogue systems.

Chapter 4

Investigating the relationship between dialogue and exchange-level impression

4.1 Overview

With the development of natural language processing (NLP) and speech recognition, dialog systems such as speech assistant systems, information guide systems in public locations, and intelligent customer service systems have become crucial in our lives. An important step in improving the quality of dialog systems is to evaluate the user's impression of the system. Many previous studies have used objective evaluation metrics to assess dialog systems. Previous research on user impressions can be divided into dialog-level and exchange-level user impression modeling. Exchange- level methods are designed to evaluate the user's impression at any point in a conversation. The main purpose of exchange-level user impression evaluations is to track the user's internal impressions (such as their sentiment) and to adapt verbal and nonverbal responses according to the user's impression. A good dialog system should be coherent, appropriate, and engaging [107]. The experience of each exchange also influences the user's overall impression of the conversation system, and the user's self-sentiment at the exchange level may help determine the user's overall evaluation of the dialog.

Previous studies [85, 87, 108] have recognized user interactions at the dialog and exchange levels separately, and few studies [109] have explored and utilized the relationship between the two. Bodigutla et al. [109] used multiple tasks to recognize turn- and dialog-level impression ratings for a task-oriented dialog, proving that dialog-level labels are beneficial for evaluating user satisfaction at the exchange level. However, this study focused on modeling user satisfaction in text-to-text dialog systems rather than multimodal systems. It did not explore the relationship between the dialog and exchange levels, which remains unclear. In particular, the type of user knowledge shared between dialog-level and exchange-level sentiment is important in developing multimodal dialog systems.

Based on this background, we use a publicly available multimodal dialog dataset that contains multimodal data [110], including audio, body, visual, and transcript data, as well as two types of user sentiment labels to evaluate user impressions of the system. The dataset was collected with a non-task-oriented dialogue system, and the dialogue level and exchange level sentiments were annotated. Therefore, this dataset allows us to investigate the relationship between the dialogue and exchange levels. Based on the findings of the correlation analysis between the two types of labels, we present a multitask model for recognizing dialog-level annotations that include two tasks: identifying the self-sentiment at the exchange level and recognizing the dialog-level label. To validate the effectiveness of the proposed multitask model, we compare the dialog-level labels of the single-task and multitask models. The comparative analysis in Section 4.6.2 indicates that considering user sentiment at the exchange level helped recognize the user impression at the dialog level.

Furthermore, we compared the results of our proposed model with those of other works using the same database. We demonstrate that our multitask model performed better in Section 4.7.2. The main contributions of this study can be summarized as follows:

Exploration of the relation between exchange-level and dialog-level labels: To explore the relationship between the exchange level and the dialog level, we use a dataset and annotate the user impressions at the exchange and dialog levels. We first analyze the relationship between user sentiment at the exchange and dialog levels in Section 5.3.3. Then, we investigate the effect of the correlation coefficient on the multitask performance in Section 4.7.2. The comparative analysis demonstrates that the correlation generally positively affects multitask performance.

Sequential multitask learning (MTL) of exchange-level and dialog-level labels: Sequential MTL [111] enables the model to utilize information from various tasks to learn important common information between different tasks. This characteristic allows MTL to train the model to handle one task while accounting for other sub-tasks. However, basic MTL assigns multiple labels to the same dimensional data input. Thus, we need to handle multiple labels assigned to different units (exchange and dialog) in this study. We utilize the sequence modeling method to train the model using labels assigned to different units. We utilize the sequence modeling method. We use long short-term memory (LSTM) and gated recurrent units (GRUs) as baselines. We show the impact of MTL in Section 4.6.2, and the results show that the MTL strategy improved the recognition accuracy on almost all dialog-level tasks.

4.2 Related works

With the development of social signal processing, human-computer interaction applications are increasingly used in people's daily lives. According to whether the dialog has a

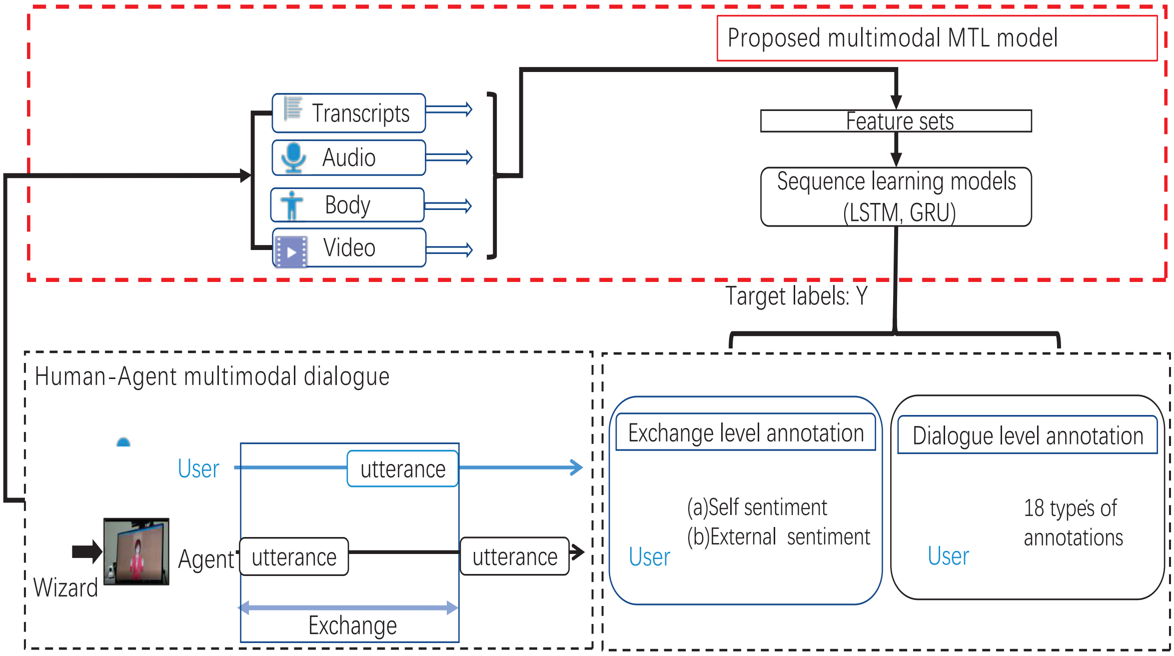


Figure 4.1: Overview of the MTL multimodal model for recognizing user impressions.

goal, the dialog system can be roughly divided into two categories: task-oriented dialog systems and open-domain non-task-oriented dialog systems. Task-oriented systems assist users in solving specific tasks as efficiently as possible [112]. In contrast, non-task-oriented systems do not have a specific task, and their main purpose is to entertain users with open-domain chats [22]. Non-task-oriented systems are used in many fields, such as assisting older adults and those learning a second language. In a non-task-oriented dialog system, it is more important to engage the user in the interaction as long as possible and to ensure that the users returns as often as possible rather than to respond to the user correctly. Previous studies have shown that people spontaneously adjust their facial expressions, postures, pronunciation, and speech rates during conversation [103–105], which demonstrates that dialog should be able to capture the unspoken intentions, attitudes, and emotions of interlocutors, especially in non-task-oriented dialog systems. [113] proposed a multimodal non-task-oriented dialogue system that improved user experience by assessing the multimodal behavior of users.

The evaluation of the performance of a dialog system is one crucial component of managing dialog. Non-task-oriented dialog systems, such as open-domain dialog systems, cannot set clear goals for the dialog; thus, evaluating task accomplishments is more difficult than evaluating task-oriented dialog systems. To address this issue, recent research has focused on recognizing user-centered criteria, such as satisfaction and interaction quality annotated by users [84,114]. In recent decades, many researchers have focused on evaluating user status. Most can be divided into two categories: exchange-level evaluations and dialog-level evaluations.

For exchange-level evaluation tasks, as the user's impressions can change dynamically during dialog exchanges, it is necessary to capture these dynamic changes in real-time so that the system's next action adopts a dialog strategy based on the user's impressions. Schmitt and Hara et al. [85, 108] used support vector machines and n-grams to predict the quality of interactions in ongoing dialogs at the exchange level. Engelbrecht et al. [78] used hidden Markov models (HMMs), and the user's opinion was regarded as a continuously evolving process. Historical context plays a crucial role in conversations and is beneficial for recognizing user satisfaction and considering temporal features at different levels. To overcome the limitation of handcrafting temporal features, Ultes et al. [84] developed a recurrent neural network for recognition of user satisfaction. To evaluate different aspects of the user's impressions, Hirano et al. [93] proposed a multitask deep learning neural network model that used multimodal features and deep neural networks (DNNs) to recognize the three exchange-level features: (1) the user's interest label, (2) the user's sentiment label, and (3) the topic continuance label toward the spoken dialog system.

The main purpose of the dialog level evaluation task is to learn dialog strategies to maximize the overall impression of the dialog, which is also helpful for identifying problematic conversation topics that led to user dissatisfaction. Higashinaka et al. [87] used overall dialog ratings to estimate dialog-level quality using HMM and overcame the limitation using task success [115] as dialog evaluation criteria. To estimate user satisfaction in conversations that span multiple domains, Bodigutla et al. [116] used new domain-independent feature sets (the aggregate topic popularity and the diversity of topics in a session) to estimate user satisfaction at both the turn and dialog levels. Wei et al. [114] proposed a multimodal user satisfaction recognition model to evaluate non-task-oriented dialog systems at the dialog level by using automatic multimodal features. Furthermore, they investigated the contribution of different modalities to user satisfaction at the dialog level.

All the above works recognized user impression on dialog and exchange levels separately. To utilize the relationship between the two, this study proposes a multimodal model for identifying the dialog level user impressions by considering the user's sentiment at the exchange level, which is suitable for evaluating non-task-oriented dialog systems. We first explore the relationship between 18 dialog labels and the user's sentiment at the exchange level and then utilize MTL models, which enable the model to recognize self-sentiment while considering the user's overall impression at the dialog level. Fig. 4.1 shows an overview of this research.

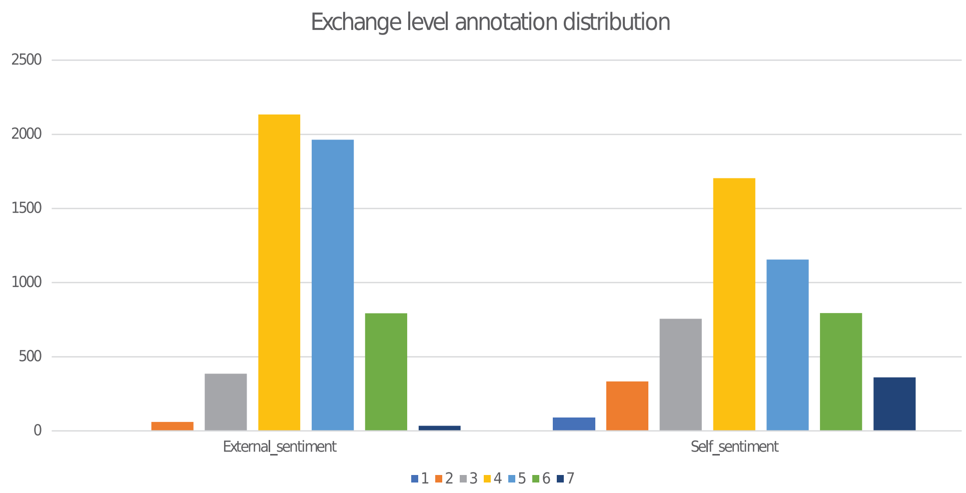


Figure 4.2: The rate distribution of the exchange-level annotations.

4.3 Data description

4.3.1 Data

Following [110, 114], the Hazumi1902 and Hazumi1911 data corpora were employed in this research. These two corpora include 60 participants (25 males/35 females aged 20-60). To reduce the effect of participants having different preferences on various topics, The behavior of the participants was recorded with a video camera and Microsoft Kinect V2 sensor.

4.3.2 Annotations

4.3.2.1 Dialog level annotations

The dataset used a questionnaire with 18 labels relating to the user's impression of the dialog, as proposed in [95]. The questionnaire measured cognition and rapport in interpersonal communication. The 18 items were "well-coordinated", "boring", "co-operative", "harmonious", "unsatisfying", "uncomfortably paced", "cold", "awkward", "engrossing", "unfocused", "involving", "intense", "friendly", "active", "positive", "dull", "worthwhile", and "slow". The users evaluated Each label on an eight-point scale from 1 to 8 after the dialogue.

4.3.2.2 Exchange level annotations

In this section, we describe the different exchange-level labels in detail.
External sentiment: In this work, an exchange was defined as the part that begins at the start time of a system utterance and ends at the start time of the next system

utterance. Human coders annotated the external sentiment according to participants performance during each exchange with a score ranging from 1 (participant seems bored with the dialog) to 7 (participants seem to enjoy the dialog), and five experts annotated the external sentiment labels. The distribution of the self-sentiment and external sentiment (average score annotated by the five experts) on the exchange-level labels are shown in Fig. 4.2.
Self-sentiment: This annotation was similar to the external sentiment annotation. Self-sentiment labels were assigned as scores ranging from 1 (want to stop talking, confused about the systems utterances) to 7 (enjoy talking, satisfied with the talk) and were annotated by the participants.

4.3.3 Data analysis

This study explores the relationship between the overall exchange sentiment within a dialog and the dialog-level sentiment. This relationship can reflect whether certain exchange sentiments lead to certain dialog sentiments from the viewpoint of the whole dialog. We computed the Pearson correlation coefficients between the average exchange labels (values) over the time-series exchanges and the dialog labels of all dialogues. We investigated the distribution of the coefficient values.

Moreover, as seen from the above definitions of dialog-level labels, the dialog-level labels describe the user impression with positive and negative annotations. Since different polarity labels describe opposing annotations, we divide the dialog-level labels into two categories to investigate the different annotation polarities precisely. The positive category includes the labels well-coordinated, cooperative, harmonious, engrossing, involving, friendly, active, positive, and worthwhile. The negative category consists of boring, unsatisfying, uncomfortably paced, cold, awkward, unfocused, intense, dull, and slow.

Table 4.1 (a) lists the coefficient value between exchange-level sentiments and positive dialog-level annotations, and Table 4.1 (b) lists the coefficient value between exchange-level sentiments and negative dialog-level annotations. Each row indicates a dialog-level annotation, and each column indicates an exchange-level sentiment. The intersection between a row and a column represents the coefficient value between an overall exchange-level sentiment and a dialog-level annotation. The average shows the average coefficient value of the coefficients of a given polarity corresponding to a given exchange-level sentiment.

As seen in the table, all coefficients between the exchange-level sentiment and the positive dialog-level annotations are positive, except the coefficient between the third-party sentiment and the cooperative label, which is negative but close to zero. The average coefficient between the third-party sentiment and the positive dialog-level annotations is 0.101. The average coefficient between the self-sentiment and the

Table 4.1: Pearson correlation coefficient and P-value (p) results between exchange and dialogue-level sentiments. (a) shows the coefficients between exchange-level sentiments and positive dialogue-level annotation; (b) shows the Pearson correlation coefficients between exchange-level sentiments and negative dialogue-level sentiments. (** represents $p < 0.001$, * represents $0.001 < p < 0.05$; if $p > 0.05$, no symbol is shown.)

(a) Pearson correlation coefficients of positive dialogue-level annotations

	Third-party sentiment	Self-sentiment
Well-coordinated	+0.222	+0.297 *
Cooperative	-0.002	+0.110
Harmonious	+0.058	+0.309 *
Engrossing	+0.179	+0.359 *
Involving	+0.154	+0.298 *
Friendly	+0.042	+0.283 *
Active	+0.068	+0.253 *
Positive	+0.154	+0.245
Worthwhile	+0.037	+0.531 **
Average	+0.101	+0.299

(b) Pearson correlation coefficients of negative dialogue-level annotations

	third-party sentiment	Self-sentiment
Boring	-0.135	-0.448 **
Unsatisfying	-0.123	-0.289 *
Uncomfortably paced	+0.049	-0.279 *
Cold	-0.270 *	-0.528 **
Awkward	-0.156	-0.362 *
Unfocused	-0.215	-0.261 *
Intense	-0.237	-0.164
Dull	-0.263	-0.453 **
Slow	-0.190 *	-0.386 *
Average	-0.171	-0.352

positive dialog-level annotations are 0.299. On the other hand, all coefficients between the exchange-level sentiment and the negative dialog-level annotations are negative. The average coefficient between the third-party sentiment and the negative dialog-level annotations is -0.171. The average coefficient between the self-sentiment and the negative dialog-level annotations is -0.352. These results demonstrate that dialog-level annotations are closely related to exchange-level sentiments. A higher overall exchange-level sentiment leads to a positive dialog-level annotation, while a lower overall sentiment leads to a negative dialog-level annotation. Almost all exchange-level self-sentiment and dialog-level annotation pairs correlate significantly, with p¡0.05; however, most third-party sentiment pairs are insignificant (p¿0.05). This result indicates that exchange-level self-sentiment labels correlate more and have more common information with dialog-level labels. Moreover, compared with the self-sentiment label annotated by the

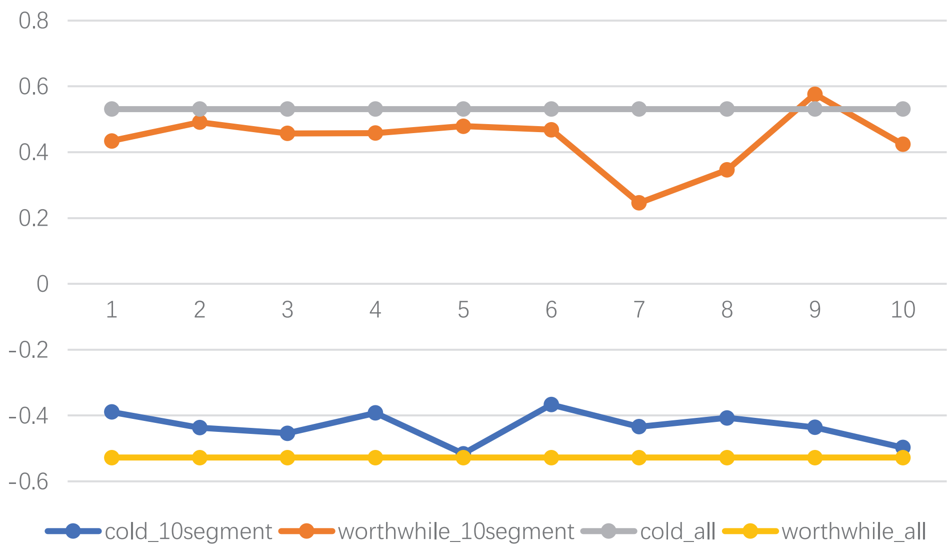


Figure 4.3: The Pearson correlation coefficient between each segment and dialog-level label (worthwhile and cold).

users themselves, the external sentiment annotated by the five experts was more time-consuming and expensive. Thus, our study used the self-sentiment as a sub-task target label.

Table 4.1 shows that the "worthwhile" label obtained the highest correlation with the self-sentiment among all positive annotations. In contrast, the "cold" label had the highest correlation with the self-sentiment label among all negative annotations. To explore the correlation between different conversation segments and dialog-level annotations, we divided the dialog into 10 segments. We used Pearson correlations to compute the average coefficient value of each segment. Fig. 4.3 shows the correlation coefficient of each segment's cold and worthwhile labels and the overall conversation with the self-sentiment label. We observed that different segments had distinct relations at the dialog label, and few segments (worthwhile-10th) had higher correlations at the dialog level. Compared with the correlation between the average value of each conversation segment and the self-sentiment, the average value of the overall dialog had a higher correlation with dialog-level annotations, which indicated that considering all exchanges is better for recognizing the dialog label. For this reason, all exchange-level information was utilized in all experiments.

We confirm that the dialog-level annotations are closely correlated with the exchange-level self-sentiment and suggest that considering the exchange-level sentiment can improve dialog-level annotation recognition. For this reason, in this study, we recognize user impressions by considering user self-sentiment at the exchange level.

4.4 Methods

4.4.1 Feature extraction

4.4.1.1 Audio features

For the acoustic modality, we use the speech feature extractor OpenSMILE [98] to extract acoustic features at the exchange level. The acoustic features correspond to the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which achieves high performance in emotion-related fields. The features of each speaker are extracted and normalized. Because these acoustic features represent the performance of an entire exchange, we apply the same acoustic features to utterances with different labels in one exchange.

4.4.1.2 Linguistic features

We extracted two linguistic features from transcriptions of spoken dialog contents for the linguistic modality.
Part of speech: The sentences were segmented into words and annotated with universal part-of-speech (POS) tags using Stanza NLP [1]. The number of different POS tags for each sentence is counted. We use a 17-dimensional vector as a sparse representation of 17 POS tags.
Bidirectional Encoder Representations from Transformers (BERT) [57]: Language model pretraining has proven useful for learning universal language representations. A model pretrained on Japanese text (using Wikipedia) [117] was employed in this work. We use this model to extract features from text at the exchange level, yielding a 768-dimensional text representation vector. Thus, we obtain a 785-dimensional linguistic feature vector.

4.4.1.3 Body features

This work uses three-dimensional coordinates of each joint in the upper body for body features, which were estimated with a Microsoft Kinect v2 sensor. Five points of body motion are employed: the left shoulder, right shoulder, left hand, right hand, and head. We denote the three-dimensional coordinate of each body point in frame t as $w(t) = x, y, z$ and the time between frames as t_1. We calculate the absolute value of the velocity between two frames as $|v(t)| = |w(t+1) - w(t)|$ and the absolute value of the acceleration between two frames as $|a(t)| = |v(t) - v(t-1)|$. We calculate the velocity and acceleration to coordinate the data of the five body points in all frames. After $v(t)$ and $a(t)$ are calculated, we use the maximum value of the acceleration and the maximum, mean,

[1] https://github.com/stanfordnlp/stanza

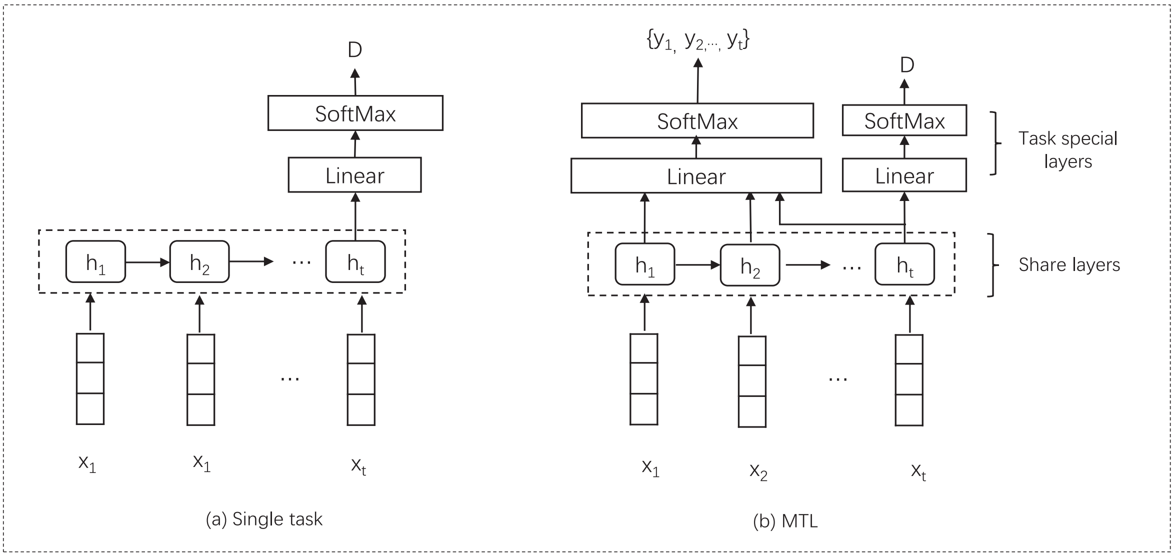


Figure 4.4: The structures of the signal task model and multitask model.

and standard deviation of the velocity in each exchange turn as body activity features. Thus, the body activity feature set has a total of 20 dimensions.

4.4.1.4 Visual features

To extract visual features, we used OpenFace [100] software.

Facial landmark features: OpenFace outputs three-dimensional coordinates of 68 facial landmarks in each frame. This study selected ten facial landmarks: two on each eye, four on the mouth, and two on the eyebrow. We adopted the same method for tracking body features. The maximum acceleration value and the maximum, mean, and standard deviation of the velocity were extracted for each user exchange and used as visual features. Thus, we obtained a 40-dimensional vector.

Action units: Facial expressions display emotional states that objectively describe facial muscle activation [101]. To extract the facial expression, this study used OpenFace software to obtain 18 action units (AUs)rated between 0 and 1, indicating absence and presence, respectively. Then, we calculated the average of each AU during the exchange to obtain the facial AU features (18-dimensional). Overall, 58 dimensions of visual features were used in this study.

4.4.2 Baseline and multitask model

4.4.2.1 Single-task deep learning neural network (baseline)

User multimodal behavior dynamically changes during a conversation. We use the LSTM and GRU methods to preserve the sequential information to recognize the user's impression. As described in Section 5.4, different unimodal features (audio a_t: 88-dim., linguistic l_t: 785-dim., body b_t 20-dim. and video v_t: 58-dim.) were extracted from the t-th exchange. We use the early fusion method to concatenate different unimodal features, generating the exchange-level multimodal feature $x_t = [a_t,\ v_t,\ b_t,\ v_t]$. The multimodal feature X = $(x_1,\ x_2,\ ...,\ x_t)$ was used as the input of these neural network models. In all models with one recurrent layer and 128 units, we obtained a 128-dimensional hidden state from the recurrent layer. The recurrent layer was followed by a fully connected layer, which projected the output (128-dimensional). At the end of the model output layer containing two units, the log- softmax function was used to output the probabilities of different user impressions. As shown in Fig. 4.4.1.4 (a), the output at the final moment ht can represent the whole sequence, which uses a fully connected layer followed by a softmax nonlinear layer to predict the probability distribution over different classes.

4.4.2.2 Multitask deep learning neural network (proposed model)

MTL is a machine learning approach that simultaneously solves multiple learning tasks by exploiting commonalities and differences across tasks [111]. An advantage of the multitask model is that it utilizes correlations among dialog-level and exchange-level tasks, improving the classification performance by learning several tasks in parallel. The key factor of MTL is the sharing scheme in the latent feature space. In a neural network-based model, the latent features can be regarded as the states of the hidden neurons. In general, MTL models are composed of two parts: shared layers and task-specific layers. The lower layers are shared across all tasks, with several task-specific layers. In the multitask model, we use a single recurrent layer with 128 units as the shared layers. These layers extract exchange and dialog-level features for the tasks shown in Fig. 4.4.1.4 (b). In the task-specific layers, for the exchange sentiment task, we obtained 128-dimensional hidden states H = $(h_1,\ h_2,\ ...,\ h_t)$ from the recurrent layer, which was followed by a fully connected layer that projected the output (128-dimensional). The output layer contains two units, and the log-softmax function of each hidden unit outputs the exchange-level task-specific layer at time step t, which are the probabilities for different exchange self-sentiments x_t. For the dialog-level user impression recognition task, the structure is the same as in the single task, and the mathematical formula of the model can be described as follows:

$$Share\ layer: h_t = LSTM(x_t W_e, h_{t-1}) \tag{4.1}$$

$$Exchange\ level\ task\ special\ layer: y_t = Softmax(h_t W_y + b_y) \tag{4.2}$$

$$Dialogue\ level\ task\ special\ layer: D = Softmax(h_t W_D + b_D) \tag{4.3}$$

Equation 4 shows the multitask loss function of the multitask model. L_e and L_d are the mean square error losses computed for the exchange-level and dialog-level label ratings, respectively. The values λ and (1-λ) are interpreted as the loss weights of the dialog- and exchange-level tasks, which were set manually.

$$L = \lambda * L_d + (1 - \lambda) * L_e \tag{4.4}$$

4.5 Experiment

The dialog-level user impression recognition task and exchange-level sentiment recognition task are both time-series tasks. Thus, considering the time-series information is beneficial for improving model performance. Recurrent neural networks are mainly used for tasks that involve sequential inputs, such as time-series predictions. This work uses LSTM and GRUs as baselines to model the sequence of multimodal behaviors. To eliminate the influence of unbalanced data, we adopt five-fold cross-validation to train the models, yielding five groups of evaluation results. The mean of the five groups was computed and used as the final result, and the F1-score of the label weights was used as the evaluation metric. According to previous works, linguistic features are key descriptors in recognizing user satisfaction. In this work, we used an unimodal model with a linguistic feature set as the baseline model. We compare the accuracy with the following five feature sets to analyze the contribution of each modality to the recognition of the dialog-level label:

(1) L: model trained with linguistic features (baseline)

(2) L+A: model trained with linguistic features + acoustic features

(3) L+B: model trained with linguistic features + body features

(4) L+V: model trained with linguistic features + visual features

(5) ALL: model trained with acoustic features + body features + visual features + linguistic features

Table 4.2: Binary classification F1-score of different multimodal combinations of LSTM base models on a dialogue-level (worthwhile) label (Acoustic (A), Body (B), Visual (V), and Linguistic (L)).

	GRU (1-128)	**LSTM(1-128)**	**LSTM(2-128)**	**LSTM(1-64)**
L	0.677	0.691	0.66	0.618
L+A	0.601	0.67	0.664	0.56
L+B	0.597	0.605	0.664	0.563
L+V	0.683	**0.738**	**0.7**	0.694
ALL	**0.692**	0.728	0.618	**0.697**

Table 4.3: Binary classification F1-score of different multimodal combinations of the LSTM (1-128) base model on a dialogue-level (worthwhile) label (Acoustic (A), Body (B),Visual (V), Linguistic (L)).

	Multitask									**Single task**
Loss weight (λ)	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	
L	0.648	0.603	0.668	0.702	0.615	0.677	0.722	0.668	0.629	0.691
L+A	0.697	**0.753**	**0.762**	0.639	0.677	0.707	0.715	0.663	**0.753**	0.67
L+B	0.639	0.715	0.692	0.625	0.657	0.563	0.59	0.643	0.601	0.605
L+V	0.615	0.7	0.715	0.643	0.677	**0.728**	**0.817**	0.713	0.707	**0.738**
ALL	**0.702**	0.744	0.677	**0.775**	**0.722**	0.629	0.775	**0.722**	0.692	0.728

4.5.1 Experimental settings

The experiment was performed for the different modalities based on the (1) dialog-level labels and (2) exchange-level user self-sentiment labels. The binary classification datasets were developed as follows. The dialog-level label annotated scores (1-8) were converted to binary values (high and low) with a threshold of 4 (neutral state). The self-sentiment, rated between 1 and 7, was converted to binary values (high/low) with a threshold of 4. The number of high/low points for the self-sentiment label on the exchange level was 2882/2311.

4.5.1.1 Comparative experiment settings (single task on the dialog level):

To investigate suitable hyperparameters, we first design comparative experiments to recognize dialog-level annotations.
GRU (1-128): A GRU layer with 128 hidden units is applied.
LSTM (1-128): An LSTM layer with 128 hidden units is applied.
LSTM (1-64)) An LSTM layer with 64 hidden units is applied.
LSTM (2-128): Two LSTM layers with 128 hidden units are applied.

For all the experiments, the number of epochs was 60, and we used the Adam optimizer with a learning rate of 0.001. Table 4.1 shows that the worthwhile label has the highest correlation coefficient (0.531) with self-sentiment. Therefore, we choose the worthwhile label as the target label. The high/low data points for the worthwhile labels at the dialog level were 38/22. The results of the comparisons are described in Section 4.6.1.

4.5.1.2 Multitask experiment settings

A 1-layer recurrent layer with 128 units was applied. The number of epochs was set to 60. Our experiments used the adaptive moment estimation (Adam) optimizer with a learning rate of 0.001. The information associated with various modes plays different roles in recognizing dialog-level labels and exchange-level sentiment. To determine the appropriate relationship between the two tasks in the multitask model, we used different values of λ $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

4.6 Results

4.6.1 Comparison of different methods

Table 4.2 shows the results of 4 comparative experiments with the multimodal features. To obtain a stable model, we first use the GRU and LSTM models with the same parameters (1 layer with 128 units) to recognize the worthwhile label. Columns 2 and

3 shows the results of the GRU (1-128) and LSTM (1-128) models, respectively. The L+V feature set achieved the best result (0.738) with the LSTM(1-128) model, which is better than the best result (0.683) achieved by the GRU (1-128) model with the ALL feature set. For this reason, we use the LSTM model as the base model. Then, to obtain appropriate parameters, different parameter settings were applied in the LSTM model, and columns 3 to 5 present the LSTM (1-128), LSTM (2-128), and LSTM (1-64) model results. The best results of the LSTM (1-128), LSTM (2-128), and LSTM (1-64) models were 0.738, 0.7, and 0.697, respectively. For the LSTM (1-128) model, the best result (0.738) was obtained with the L+V feature set. Thus, the LSTM (1-128) model was used as the baseline. Moreover, we found that in most cases, the ALL and L+V feature sets performed better than the unimodal feature set (L), while the L+A and L+B feature sets performed worse than the unimodal feature set (L) in all cases.

4.6.2 Comparison of the single-task and multitask models

Section 4.6.1 shows that the LSTM (1-128) model achieves the best F1-score with the L+V feature set. According to this result, we applied the same setting, namely, an LSTM layer with 128 hidden units, in the multitask model, and the structure is described in Section 4.4.2.2. Columns 1 to 10 in Table 4.3 show the binary classification results of the multitask model with different modality features and loss weight values (λ). Column 11 in Table 4.3 presents the binary classification results of a single task with different modality features. For the single-task model, the results with the ALL and L+V feature sets were better than those achieved with the unimodal feature set (L), and the L+V feature set produced the best result (0.738). For the multitask LSTM model, the L+V feature set achieved the best F1-score (0.817) by using MTL with loss weight loss ($\lambda = 0.7$). The best results of all feature sets (L, L+A, L+B, L+V, ALL) with the multitask model performed better than those with the single-task model, with results of 0.702, 0762, 0.715, 0.817, and 0.775, respectively, which represent improvements of 0.011, 0.092, 0.113, 0.079, and 0.047, respectively. The recognition performance resulted in a large improvement, demonstrating that our multitask model can learn the relation between exchange-level sentiment and a dialog-level label (worthwhile) and is thus useful and effective for recognizing dialog-level labels. Meanwhile, for most experiments, the L+B feature set performed worse than the unimodal featur e(L), and we suspect that the body feature does not predict the worthwhile label well. This also explains why the L+V feature set obtains the best result in Tables 4.2 and 4.3. On the other hand, in some cases ($\lambda = 0.1$), the multitask model performed worse than the single-task model, which indicates that a suitable weight loss is important.

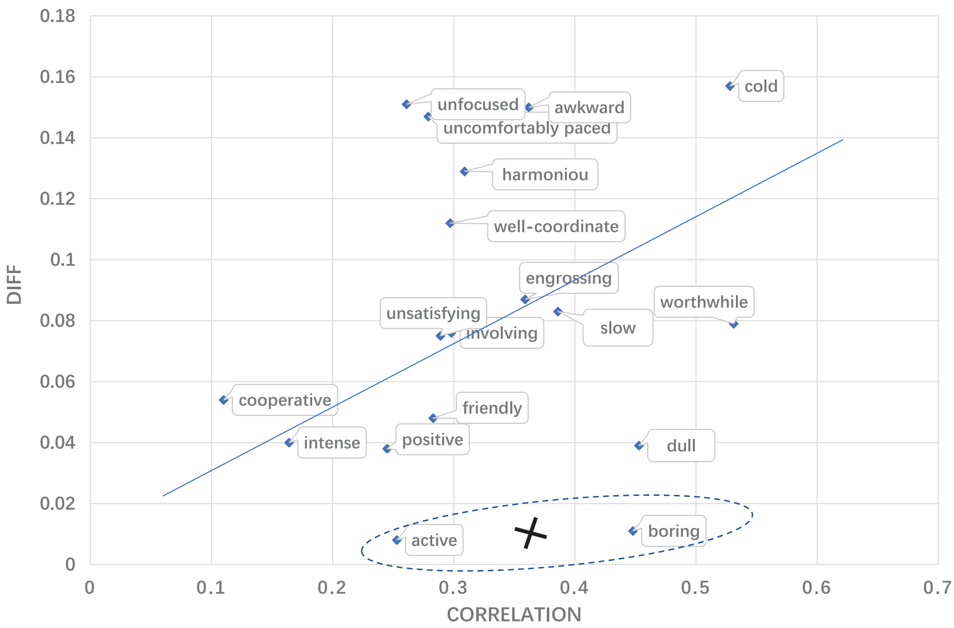


Figure 4.5: Analysis of the effect of correlations. (Diff denotes the difference in F1-scores between the single task and multitask models, while Correlation shows the Pearson correlation coefficient between self-sentiment and dialog-level annotations.)

4.6.3 Results of 18 types of annotations

Table 4.4 shows the best binary classification result of 18 types of annotations at the dialog level with the LSTM (1-128) model. Columns 3 and 4 present the best F1-score for a single task and a multitask with five types of comparative multimodal feature combinations, respectively. We observed that the multitask model performed better than the single-task model in all annotations. Compared with the average positive dialog annotation, the average negative dialog annotation was worse on the single-task model, with results of 0.701 and 0.685, respectively. After MTL was applied, the average negative dialog-level annotation (0.780) performed better than the average positive dialog-level annotation (0.767). Among all annotations, the cold label obtained the minimum F1-score (0.583) at the exchange level, and this label achieved the highest improvement (0.157) by applying MTL. The awkward label obtained the best F1-score (0.833) with the multitask model.

Table 4.4: Binary classification F1-scores of 18 annotations at the dialogue level. (a) shows positive dialog-level annotations; (b) shows negative dialog-level annotations. "diff" denotes the difference in F1-scores between the single task and multitask models.

(a)

	High/Low	Best MTL	Best Single	Diff
Well-coordinated	38/22	0.78	0.668	+0.112
Cooperative	46/14	0.737	0.683	+0.054
Harmonious	34/26	0.782	0.653	+0.129
Engrossing	27/33	0.798	0.71	+0.087
Involving	34/26	0.729	0.653	+0.076
Friendly	11/49	0.781	0.733	+0.048
Active	39/21	0.715	0.707	+0.008
Positive	46/14	0.798	0.76	+0.038
Worthwhile	38/22	0.817	0.738	+0.079
Average	/	0.767	0.701	+0.066

(b)

	High/Low	Best MTL	Best Single	Diff
Boring	16/44	0.778	0.767	+0.011
Unsatisfying	14/46	0.758	0.683	+0.075
Uncomfortably paced	35/25	0.826	0.679	+0.147
Cold	15/45	0.74	0.583	+0.157
Awkward	32/28	0.833	0.683	+0.150
Unfocused	18/42	0.795	0.644	+0.151
Intense	23/37	0.762	0.722	+0.040
Dull	15/45	0.744	0.705	+0.039
Slow	33/27	0.783	0.7	+0.083
Average	/	0.780	0.685	+0.095

Table 4.5: Binary classification F1-scores of the awkward and well-coordinated labels.

Label	Model	L	A+L	B+L	F+L	ALL	Human model(Wizard)
cWell_coordinated	MTL	0.738	0.791	0.733	0.733	0.791	0.72
	Single	0.697	0.722	0.597	0.753	0.764	
	Multi-other [114]	0.7	0.74	0.65 (B+V+L)		0.76	
Awkward	MTL	0.733	0.744	0.744	0.766	0.783	0.58
	Single	0.649	0.649	0.648	0.75	0.633	
	Multi-other [114]	0.66	0.58	0.68 (B+V+L)		0.63	

4.7 Discussion

4.7.1 Comparisons with previous works

To our knowledge, [114] used the same dataset as this work, which used LSTM (2-128) to recognize the awkward and well-coordinated labels. We used the LSTM (2-128) model to compare our results with the multi-other model and applied MTL to acknowledge the awkward and well-coordinated labels. Table 4.5 shows the binary classification results of the awkward and well-coordinated labels. For the well-coordinated label, the single task achieves a similar result to that presented in [114]. All models achieved the best F1-score with the ALL feature set, and the multitask model produced the best result (0.783). For the awkward label, when comparing the single-task model with the multi-other model presented in [114], the results with most feature sets are similar, except for the A+L feature set. The multitask model achieved a better F1-score than all other feature sets (A, L+V) for the awkward label. Overall, the multitask model proposed in this work performed better on all feature sets than the results presented in [114], demonstrating that our proposed method better utilizes exchange information and improves model performance. Column 8 shows the results of the human model proposed in [114]. Wizard annotated the user satisfaction label score, and the users were divided into high and low categories before the F1-score with the original annotation was calculated. The MTL models achieved better F1-scores on both the well-coordinated and awkward labels than the human model

4.7.2 Analysis of the effect of correlations

Combining the correlation coefficients in Table 4.1 and the improvement results in Table 4.4, we found that the correlation coefficients and improved performance are not significantly related to some labels. The friendly label has a correlation coefficient of 0.253 for positive labels, while the multitask and single-task models produce almost the same result. For negative labels, the boring label has a coefficient of -0.448, only achieving a slight improvement(0.011) by using MTL. Meanwhile, we found that the absolute value of the average correlation coefficient of the negative labels (0.352) was higher than that of the positive labels (0.299). The average improvement in the negative labels (0.095) by using MTL was higher than the average improvement in the positive labels (0.066), which indicates that although the performance improvement was not related to the correlation coefficient for every label, the correlation coefficient has a positive relationship with the overall improvement by comparing the positive labels and negative labels performances. Furthermore, we drew the scatter plot of the correlation coefficient between 18 types of labels and exchange- level sentiments, as well as the different improvements in the F1-score by using multiple tasks, as shown in Fig.4.5. We observed that the correlation coefficient is positively correlated with the performance improvement for overall user impression.

4.8 Chapter summary

In this Chapter, we first investigate the relationship between the exchange level and 18 dialog-level annotations. The worthwhile label has the highest correlation with user self-sentiment. We propose a multitask model to capture these correlations and learn the relevant information. By comparing our proposed multitask model with a single-task model and other relevant research, we show that the multitask model achieved the best performance, with a 15.7% performance improvement over the signal-task model with cold labels. Thus, our results demonstrate that our model can utilize this relation to achieve better performance. However, there is still room for improvement. [118] indicates that sex information is beneficial for recognizing emotions. This study used only user modal information to identify the users' impression of the dialog system, while other user characteristics, such as age and sex, were not considered. In future work, we will utilize the effects of user characteristics on user impressions to evaluate dialog systems.

Chapter 5

Influence of Personality Traits and Demographics on Rapport Recognition Using Adversarial Learning

5.1 Overview

With the recent advancements in natural language processing and speech recognition technologies, spoken dialogue systems such as Amazon Alexa, Siri, and Google Assistant have become widely popular across various domains. In recent years, LLMs-based dialogue systems, such as ChatGPT, Gemini, and Claude, have become mainstream in research and applications. Their advanced language generation and natural interaction capabilities have further driven the progress of research and development in non-task-oriented dialogue systems [82, 119–121].

Performance evaluation of dialogue systems plays a crucial role in optimizing data-driven dialogue systems and has been an active area of research. Rapport, a widely-used evaluation aspect for dialogue systems, has been defined as mutual attention, positivity, and coordination [122]. It refers to a harmonious, understanding, and trusting relationship established between individuals or between humans and machines. While rapport in human interactions [123–125] has been extensively studied, research on building and maintaining rapport in human-computer interactions [126] is relatively recent. Establishing good rapport in human-computer interaction can make interactions more natural and effective, increasing user satisfaction and experience. By evaluating user rapport during or after interacting with a dialogue system, developers can gain insights into how users respond to the system.

According to this background, In this study, we first investigate the impact of users' traits, such as age, gender, and personality, on user rapport recognition. We employed a dataset containing 18 types of user rapport with personal information such as age, gender, and personality traits as shown in Figure 5.1. It incorporates multiple modalities, including audio, body motion, visual cues, and transcript data, providing a comprehensive basis for evaluating the user rapport of dialogue systems. Following related research [127, 128], we used the Big Five source as the user personalities in this dataset. In this way, this dataset allowed us to explore the impacts of users' personal information on user rapport recognition. Through the analysis in Section 5.3.3, we confirmed that users' personal traits significantly impact rapport recognition.

While these traits may offer some relevance, an over-reliance on such features can lead to potential biases in the model, which may affect its fairness and performance across different user groups. Therefore, to mitigate the influence of personal information on the model, we employ an adversarial learning method. This approach utilizes gradient reversal techniques, to reduce the model's dependence on user traits, allowing it to focus on more general and crucial features. As a result, this method enhances the robustness and accuracy of user rapport recognition, ensuring that the model remains adaptable to various user profiles and scenarios. To validate the effectiveness of the proposed ADVER-based model Section 5.7.2 and 5.7.3 compare results between the adversarial approach, the baseline, human model, and instruction-based LLMs methods. The results demonstrate that the adversarial approach achieves superior performance. In summary, the main contributions of this study can be summarized as follows:

We first addressed the research question (RQ1): "Does adversarially learning users' personal information contribute to rapport recognition performance?" The effectiveness of the specifically utilized adversarial learning method is discussed in Section 5.7.1 and Section 5.8.2.

Age, gender, and personality influence users in human-computer dialogue in different ways. Age and gender often result in significant expression differences but typically do not directly impact user interaction with the employed system. Personality, however, can influence how users interact. For example, extroverted individuals may prefer open-ended conversations and more frequent interactions, whereas introverted individuals may favor direct and concise communication. Our second research question (RQ2) was as follows: "Are there differences between the impacts of demographic data (age and gender) and personality on user rapport recognition?" The relevant research findings are discussed in Section 5.7.2.

Additionally, the small size of our target dataset, with a total of 125 dialogues, might affect the performance of the developed model. To validate the effectiveness of the baseline and proposed models' effectiveness, we compared the machine learning models' recognition results with user rapport scores annotated by third-party experts and instruction-based LLMs. Our third research question (RQ3) was: "Do machine learning models outperform the results derived from multiple human observations and instruction-based LLMs in estimation accuracy?" The comparison results are detailed in Section 5.7.3.

SYS: Since last year, due to the impact of self–restraint on going out, the opportunity for dining out has decreased. Have you been creative in your meals?
(去年から外出自粛の影響で、外食があまりできないですが、食事で何か工夫されていることはありますか？)

USER: I've tried to add some variations to the dishes at home.
家で作る料理のバリエーションを増やしました

SYS: Have you tried adapting recipes or something similar?
(アレンジレシピとかはありますか？)

USER: Basically, I find recipes online and make various dishes, so I follow the recipe instructions without making too many adjustments.
(基本的にはネットで探して |変わった料理を作っているので アレンジもレシピ通りに作ってて自分では何もしていません)

SYS: I see. I tried a recipe using canned mackerel, and I was able to make a delicious Aqua Pazza. You can find it if you search online.
(なるほど、私はサバ缶を使ったレシピで、アクアパッツァがなかなかおいしくできました。ググったら出てきます。)

USER: Oh, I see. It looks very delicious.
(あーそっかとてもおいしそうですね)

SYS: It was easy to do while watching the video.
(動画をみながら作ると、簡単でした)

USER: Mm-hmm, it seems simple and like a good recipe.
(簡単そうでいいレシピですね)

User traits :{
"age" : " 20s",
" gender" : "female",
"Big-five": "4, 3, 3, 4, 5",
}

Figure 5.1: An example dialogue session contained in our dataset.

5.2 Related work

Generally, dialogue systems can be classified into task-oriented [18–20] and non-task-oriented [21, 22] systems. Task-oriented dialogue systems are designed to assist humans in achieving their desired goals. These systems focus on booking flights, ordering food, or making hotel reservations. The aim to offer the best possible solutions by asking relevant questions and providing informative responses. Non-task-oriented dialogue systems are designed for general conversations and do not have any specific goals or objectives. These systems aim to engage users in casual conversation and provide enjoyable experiences.

Given the diversity and complexity of dialogue systems, evaluating intelligent assistants has become a challenging task and an active research field. Most statistical

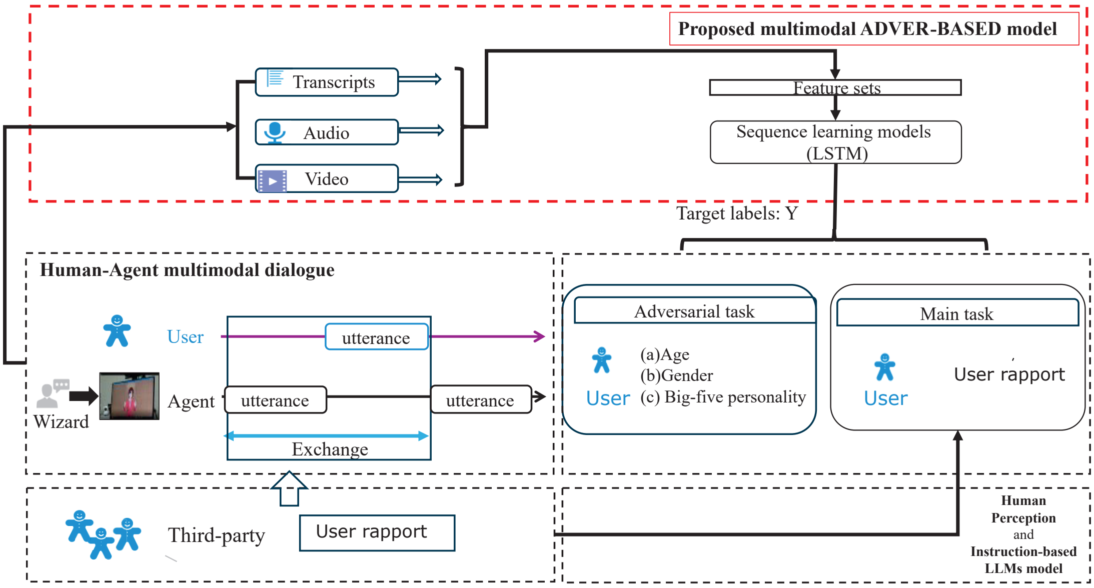


Figure 5.2: Overview of the multimodal model for adapting users' personal information to recognize user rapport.

approaches for spoken dialogue assessment consider objective criteria such as dialogue lengths or task success rates [86]. However, these metrics do not necessarily correspond to the target conversation's subjective and immediate user rapport. Especially for non-task-oriented conversations such as small talk and multidomain dialogues, no task success information is available when interacting with simulated or recruited users. This lack of information makes it difficult to evaluate non-task-oriented dialogue systems.

To address this problem, researchers have recently focused on more user-centered criteria, such as measuring user rapport levels during or after interacting with a dialogue system. Engelbrecht et al. and Klaus-Peter et al. [78]used hidden Markov models (HMMs) to recognize user satisfaction at each dialog step. With the continuous advancement of neural networks, numerous researchers are utilizing deep-learning techniques to predict user satisfaction. For example, Ultes et al. [84] proposed a bidirectional long short-term memory (BiLSTM) to assess the quality of interactions and achieve improved performance. To capture the different aspects of user satisfaction, [93] proposed a multitask deep learning-based neural network model that predicts user sentiment, user interest, and user topic continue based on the exchange level. T.E.Kim et al. [129] proposed a model that combines the user-utterance generation task with the user satisfaction scoring and action prediction tasks by applying a deep multitask neural model to achieve good user satisfaction prediction performance. A good dialogue system should provide coherent and appropriate responses and sufficiently engage to leave an overall rapport with the user. Therefore, it is essential to analyze the user rapport of the dialogue system at the exchange level and the dialogue level. The primary objective of the dialogue-level user satisfaction evaluation task is to learn dialogue strategies that maximize impressions in an overall conversation, which also helps identify problematic conversation topics that lead to user dissatisfaction. [79] used a statistical classification method with support vector machines to predict interaction quality on dialogue level with field and laboratory data, thus overcoming the limitation of using task success.

To improve the non-task-oriented multimodal dialogue system, Wei et al. [114] used automatic multimodal features to evaluate such systems at the dialogue level. Furthermore, to utilize the relationship between user impression at the dialogue level and exchange levels, Bodigutla et al. [109] proposed a multitask base model that jointly predicts turn-level annotation labels and user impression level for dialogue. With LLMs demonstrating impressive reasoning and dialogue-understanding capabilities, researchers have also employed them to evaluate user performance at the exchange level [130, 131].

Above all, most existing methods focus primarily on algorithmic improvements without considering the impact of users' information on user rapport. Several notable works have recently demonstrated that adversarial methods successfully enhance the robustness and generalizability of models in various tasks. Meng et al. [132] utilized an adversarial speaker adaptation method to achieve improved speech recognition Microsoft short message task by aligning the features of speaker-dependent models with a reference model, achieving significant word error rate gains Gao et al. [133] used an adversarial domain adaptation and a center loss to enhance the generalization capabilities of cross-corpus speech emotion recognition systems.

This study aims to investigate and mitigate the impacts that may lead to potential biases in the model of user rapport recognition for dialogue systems. Inspired by previous works [114, 132], we examined the relationships between 18 rapport labels and users' personal information. We subsequently employed an adversarial-based model to adapt more effectively to these personal information variations. An overview of this study is presented in Figure 5.2.

5.3 Dataset

5.3.1 Data

To develop a user-adaptive multimodal dialog system, Komatani et al [110] collected the multimodal Hazumi dataset. Most previous works [114, 135, 136] implemented based on this dataset involved laboratory settings. To better reflect real-world conditions, this study used three multimodal dialogue datasets Hazumi2105[1], Hazumi2012[2], and

[1]https://github.com/ouktlab/Hazumi2105/
[2]https://github.com/ouktlab/Hazumi2012/

Table 5.1: Data summary

Dataset	**Hazumi datasets** [134]		
Version name	Hazumi2010	Hazumi2012	Hazumi2015
Overview	Dialogues between human participants and virtual agents operated by a human Wizard for approximately 15 to 20 minutes per dialogue		
Instructions to the Wizard Participants	Chit-chat involving any topics to make the participants enjoy the dialogue		
Participants	33 (17 per male, 16 female)	63 (29 per male, 34 per female)	29 (14 perr male, 15 per female)
	Aged 20 to 70 (27 per: age<50, 6 per: age>50)	Aged 20 to 70 (54 per: age<50, 9 per: age>50)	Aged 20 to 70 (27 per: age<50, 6 per: age>50)
Sensors	Videos and voices of the participants		
	Video of agent		
Manual annotations	Third-party sentiment concerning the exchange turn level (on a 7-point scale) provided by five annotators Topic continuance conerning the exchange turn level (on a7-point scale) provided by five annotators		
	18 types of user rapport labels on dialogue level(out of 8-point scale), by five annotators and the participant		
	3 types of user rapport(coordinateness, awkwardness, and friendly-ness) at the dialog level (on an 8-point scale) provided by the Wizard		

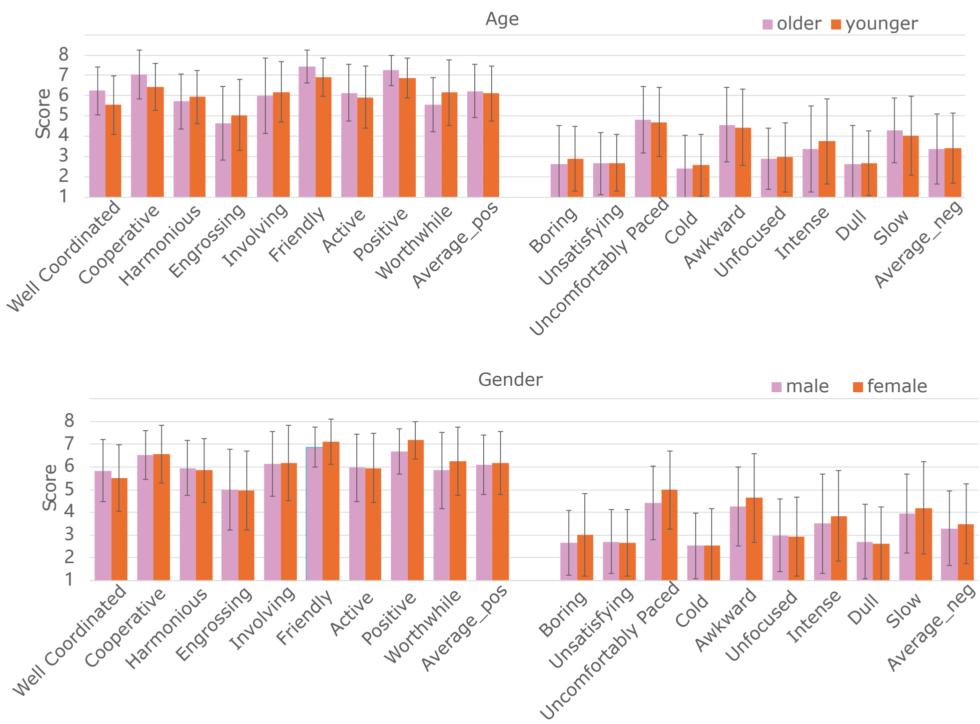


Figure 5.3: The average scores for 18 types of annotations for age (top) and gender (bottom). The average score for each positive rapport label is on the left side of each figure. Average_pos represents the average of all positive rapport labels. The average score for each negative rapport label is on the right side of each figure. Average_neg represents the average of all negative rapport labels.

Hazumi2010[3]. Each participant was recorded in their home. The MMD-Agent platform [137] was used as the interface for interacting with the participants, with their responses being controlled by an operator (Wizard). All corpora were arranged to record facial videos, and audio data via microphones and cameras through the Zoom software platform. Throughout the interactions, if participants displayed disinterest, the Wizard would proactively change the topic to rekindle their engagement. Conversely, if the participants appeared interested and actively participated in the conversation, the Wizard would listen and respond. The specific details of the database are provided in Table 5.1.

Table 5.2: The numbers of high/low data (4 as the threshold) for 18 types of annotations

User rapport	Hazumi 2010	Hazumi 2012	Hazumi 2105	High/Low	Level
well coordinated	0.804	0.774	0.631	102/23	pos
boring	0.856	0.793	0.667	23/102	neg
cooperative	0.844	0.731	0.575	119/6	pos
harmonious	0.823	0.711	0.643	105/20	pos
unsatisfying	0.781	0.757	0.726	17/108	neg
uncomfortably paced	0.577	0.710	0.511	72/53	neg
cold	0.716	0.546	0.386	17/108	neg
awkward	0.640	0.717	0.599	67/58	neg
engrossing	0.833	0.795	0.741	75/50	pos
unfocused	0.701	0.557	0.342	21/104	neg
involving	0.823	0.717	0.640	110/15	pos
intense	0.402	0.704	0.495	53/72	neg
friendly	0.854	0.721	0.667	125/0	pos
active	0.879	0.807	0.770	105/20	pos
positive	0.833	0.737	0.654	124/1	pos
dull	0.826	0.746	0.616	20/105	neg
worthwhile	0.794	0.716	0.612	106/19	pos
slow	0.820	0.696	0.506	50/75	neg

5.3.2 Annotations

5.3.2.1 User rapport

This dataset employed a questionnaire comprising 18 labels to measure the rapport of each user with the dialogue, as described in [95]. A questionnaire comprising 18 labels was employed to capture the user rapport with the dialogue in datasets. The questionnaire measured cognition and rapport in interpersonal communications. The 18 items were "well coordinated", "boring", "cooperative", "harmonious", "unsatisfying", "uncomfortably paced", "cold", "awkward", "engrossing", "unfocused", "involving", "intense", "friendly", "active", "positive", "dull", "worthwhile", and "slow". Three third-party experts evaluated each label on an eight-point scale from 1 to 8 after the dialogue finished. The agreement scores of the annotators measured by Cronbach's alpha are shown in Table 5.2. The dialog-level label annotated scores (1-8) were converted to binary values (high and low) with a threshold of 4. We found a high degree of consensus among users for some rapport labels such as "cooperative" (119/6), "friendly" (125/0), and "positive" (124/1), indicating that these system rapport labels were uniform and did not require recognition. In contrast, rapport labels that lack consensus among users could be beneficial for improving the dialogue system. Following the previous studies [114,135], we select rapport labels that are more contentious among users, such as "well coordinated" (102/23), "awkward"(67/58), and "engrossing"(75/50), as the prediction targets.

5.3.2.2 Personality

Among all inventories, the Ten Item Personality Inventory (TIPI) is a well-validated and brief version. Research [138] has shown that this questionnaire has high reliability and validity for measuring personality. The scale can be completed in approximately 1 minute and each of its 10 items is shown as follows:

I think I am:

- Q1. Energetic and outgoing
- Q2. Easily dissatisfied and prone to conflict
- Q3. Self-demanding and strict
- Q4. Anxious and worried
- Q5. Enjoys novelty
- Q6. Modest and shy
- Q7. Caring and kind
- Q8. Careless about details
- Q9. Calm and stable

[3] https://github.com/ouktlab/Hazumi2010/

- Q10. Lacks creativity, ordinary

According to [134], two items represent each Big Five factor. Manually assessed personality scores could be obtained from the participants; ratings of item scores could be obtained through simple calculations (with Qi representing the rating score for item i), such as subtracting the rating score for Q6 from that of Q1 to derive the extraversion score. The calculation method for each personality score can be calculated as follows:

- 1) Extraversion: (Q1 + 8 - Q6)/2,
- 2) Agreeableness:(Q2 + 8 - Q7)/2,
- 3) Conscientiousness: (Q3 + 8 - Q8)/2,
- 4) Neuroticism:(Q4 + 8 - Q9)/2,
- 5) Openness:(Q5 + 8 - Q10)/2.

5.3.3 data analysis

Owing to the influence of user age, gender, and personality on dialogues, this section aimed to investigate the impact of users' personal information on the rapport of dialogue systems. We examine the differences among user annotations based on age, gender, and personality. Furthermore, the dialogue-level labels depict both positive and negative annotations. Given that the labels in distinct polarities represent opposing annotations, we partition the dialogue-level labels into two categories to enable the presentation of different annotations. The positive category contains well-coordinated, cooperative, harmonious, engrossing, involving, friendly, active, positive, and worthwhile annotations. Conversely, the negative category includes boring, unsatisfying, uncomfortably paced, cold, awkward, unfocused, intense, dull, and slow annotations.

5.3.3.1 The relationship between user rapport labels and user gender and age.

Figure 5.3 separately shows the average ratings of 125 participants for the post-questionnaires consisting of 18 items based on age (old/young) with the boundary set at 50 years old and gender (male/female) separately. The vertical axis denotes the mean user rapport rating. The horizontal axis represents the 18 questionnaire items, with the first nine being positive labels. The overall average of the positive labels is displayed in the tenth position. Conversely, the remaining nine items are negative labels represented by tags ranging from the eleventh to nineteenth positions. The overall average of the negatively labeled items is shown at the right end of the horizontal axis. Older participants tended to be more positive for age than younger participants regarding sentiment-type rapport, such as friendly, active, positive, bored, and cold. For gender, the female data were more sensitive and had higher standard deviations and means than the male data for most labels, including positive and negative user rapport. Therefore,

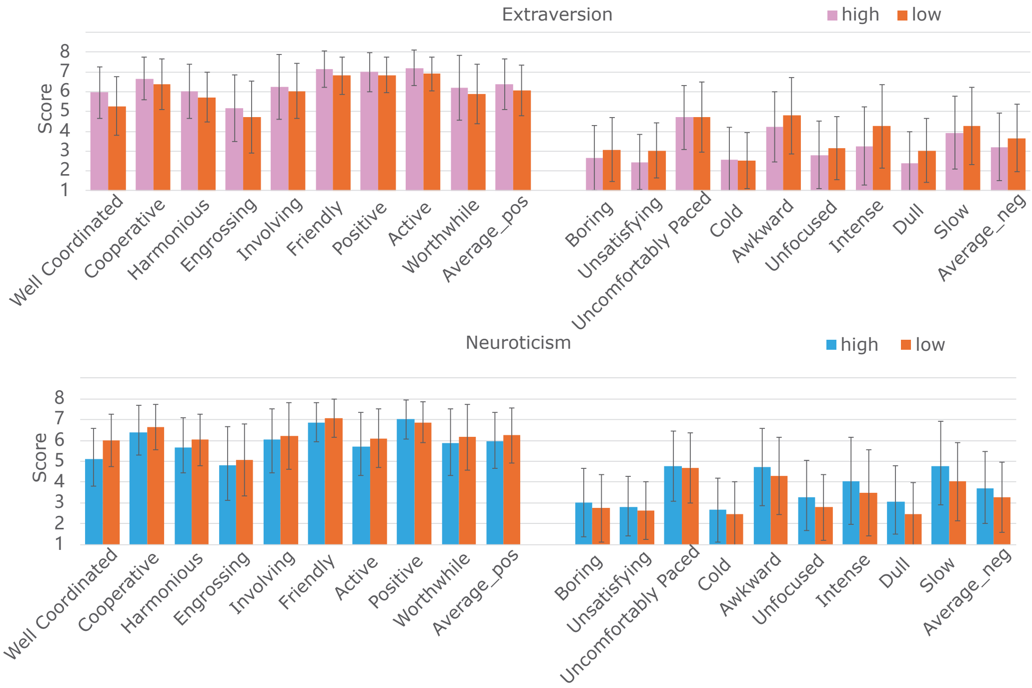


Figure 5.4: The average scores for 18 types of annotations for Big Five personality, On the left side of each figure is the average score for each positive rapport label. Average_pos represents the average of all positive rapport labels. The average score for each negative rapport label is on the right side of each figure. Average_neg represents the average of all negative rapport labels.

it is important to consider age and gender when recognizing a user's rapport at the dialogue level.

5.3.3.2 Relationships between user satisfaction and user personalities.

In this section, our main focus was exploring the relationships between user personalities and user rapport. We calculated the mean performance of different user rapport types across user personalities. In this section, we select Extraversion and Neuroticism, as representatives for discussion. Figure 5.4 illustrates the average ratings provided by 125 participants in the post-questionnaires comprising 18 items related to the Big Five. The personality scores were converted to two types (high and low personality) with a threshold of 4.

In Figure 5.4, blue represents the mean user rapport scores of high-personality users, whereas orange represents the mean scores of low-personality users. In terms of extroversion, we observed that for positive user rapport (columns 1 to 9), high-personality users tended to have higher mean rapport label scores than those of low-personality users. Conversely, for negative user rapport labels (columns 10 to 18), high-personality users tended to have lower mean rapport scores than low-personality users. This indicated a positive relationship between extraversion and user rapport and a negative relationship with negative user rapport labels. Similarly, for neuroticism, we found that for positive user rapport labels, high-personality users have lower mean user rapport scores compared to low-personality users. Conversely, high-personality users tend to have higher mean rapport scores for negative user rapport labels than low- personality users. This suggests a negative relationship between neuroticism and positive user rapport labels and a positive relationship with negative user rapport labels.

Table 5.3 (a) lists the Pearson correlations between the Big Five personality traits and the positive dialog-level user rapport scores. Generally, a correlation coefficient above 0.1 signifies a weak correlation, whereas a correlation above 0.3 indicates a moderate correlation. Table 5.3 (b) presents the coefficient values between the Big Five personalities and the negative dialogue-level user rapport labels. Each row represents a personality, and each column displays a dialogue-level user rapport label. The intersection of a row and a column indicates the coefficient value between the personality and rapport of the user. As illustrated in Table (a), all the coefficients between extraversion, conscientiousness, and openness personality and the positive dialogue-level user rapport labels are positive. However, the coefficient between openness and the positive label is an exception, being negative but close to zero. The average coefficients between the extraversion, conscientiousness, and openness positive dialogue-level annotations are 0.205 for extraversion, 0.108 for conscientiousness, and 0.123 for openness, respectively. Conversely, the correlation coefficients between agreeableness and neuroticism personality and the positive user rapport labels are negative. The average coefficients are -0.129 for agreeableness and -0.144 for neuroticism.

On the other hand, all the coefficients between extraversion, conscientiousness, and openness personalities and the negative dialogue-level user rapport labels are negative. This is consistent with the results shown in Figure 5.4. The above analysis indicates stronger correlations between user personalities and dialogue-level user rapport labels.

5.4 Features Extraction

5.4.1 audio feature

For audio features, we use OpenSMILE [98] to extract exchange-level acoustic features. These features corresponded to the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which excels in emotion-related fields.

Table 5.3: Pearson correlation coefficients results between user personality and dialogue-level user rapport. (a) shows the coefficients between user personality and positive dialogue-level user rapport labels; (b) shows the Pearson correlation coefficients between user personality and negative dialogue-level user rapport labels.

(a) Pearson correlation coefficients of positive user rapport labels

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Well Coordinated	0.314	-0.164	0.118	-0.347	0.146
Cooperative	0.157	-0.112	0.099	-0.148	0.057
Harmonious	0.193	-0.084	0.158	-0.134	0.098
Engrossing	0.229	-0.031	0.089	-0.117	0.223
Involving	0.117	-0.186	0.077	-0.052	0.048
Friendly	0.228	-0.230	0.154	-0.173	0.158
Active	0.349	-0.076	0.220	-0.155	0.277
Positive	0.092	-0.260	-0.058	0.006	-0.019
Worthwhile	0.169	-0.015	0.117	-0.179	0.115
Average_pos	0.205	-0.129	0.108	-0.144	0.123

(b) Pearson correlation coefficients of negative user rapport labels

Boring	-0.166	0.056	-0.046	0.080	-0.151
Unsatisfying	-0.241	0.133	-0.016	0.073	-0.155
Uncomfortably Paced	-0.077	-0.042	-0.118	0.076	-0.088
Cold	0.053	0.141	0.044	0.086	0.013
Awkward	-0.230	0.034	-0.098	0.171	-0.138
Unfocused	-0.178	0.015	-0.057	0.086	0.027
Intense	-0.239	0.002	0.039	0.202	-0.087
Dull	-0.237	0.110	-0.093	0.206	-0.140
Slow	-0.105	0.123	-0.281	0.192	-0.206
Average_neg	-0.158	0.063	-0.070	0.130	-0.103

5.4.2 Linguistic feature

The study extracted linguistic features from the participants' utterances and dialogue log data. We extracted two linguistic features from the manual transcription of spoken dialogue contents: **Part of speech:** The sentences were segmented into words and annotated with universal part-of-speech (POS) tags via Stanza NLP [4].

The PoS tag set was composed of 17 types: "adjective", "adposition", "adverb", "auxiliary", "coordinating conjunction", "determine", "interjection", "noun", "numeral", "particle", "pronoun", "proper noun", "punctuation", "subordinating conjunction", "symbol", "verb", "other". The PoS categories (nouns, verbs, etc.) in a user's utterance were counted. The frequencies of the PoS categories, such as nouns and verbs, were calculated in each user's utterance. We utilized a 17-dimensional vector to represent the 17 POS tags.

BERT (bidirectional encoder representations from transformers [57]): In this study, we employed a pre-trained model that specifically focused on Japanese text (trained using Wikipedia) [117]. This model was utilized to extract features from the text at the exchange level. Consequently, we obtained a 768-dimensional vector representing the text representation.

5.4.3 Visual feature

We extracted facial features as visual features via an RGB camera. **Facial landmark feature:** OpenFace [100] software outputs the three-dimensional coordinates of 68 facial landmarks in each frame. This study chose ten facial landmarks, including 2 points on each eye, 4 points around the mouth, and 2 points on the eyebrows. We utilized the same methodology for tracking body features to track facial features. For each user exchange, we extracted the maximum acceleration value and the maximum, mean, and standard deviation of the velocity value, resulting in facial features. Ultimately, we obtained a 40-dimensional vector representing these features.

Action units: Facial expressions are crucial in displaying emotional states and facilitating conversation turn-taking. These expressions are typically represented by facial action units (AUs), which provide objective descriptions of facial muscle activations [101]. This study employed OpenFace software to extract 18 types of AUs, each rated as 0 and 1 to indicate their absence or presence, respectively. The average value of each AU within an exchange was then calculated to derive facial AU features (18 dimensions). Consequently, a total of 58 dimensions of facial features were utilized in this study.

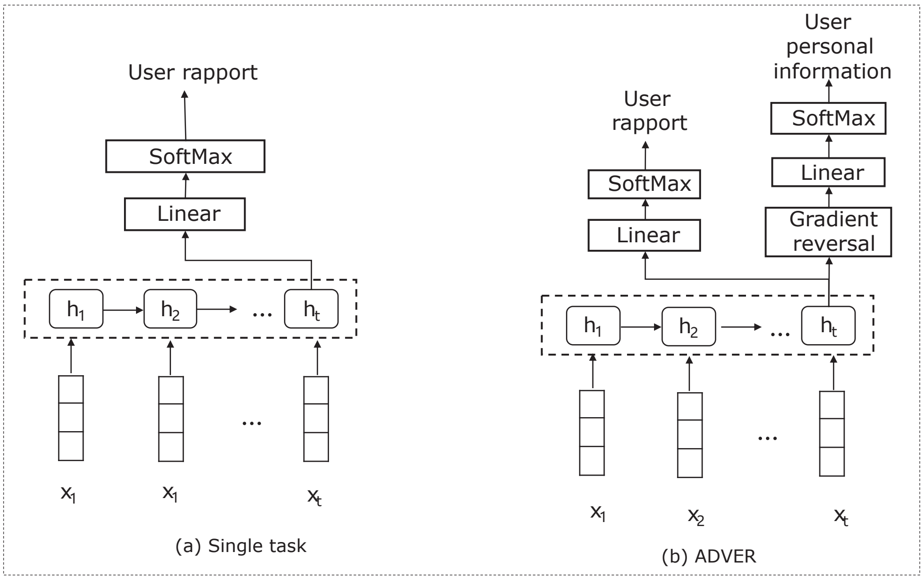


Figure 5.5: The structure of single-task model and adversarial model

5.5 Methods

5.5.1 Baselines and adversarial models

5.5.1.1 Single-task deep learning neural network (baseline)

To capture the dynamic changes in the multimodal behaviors of users during a conversation, based on previous work [135], we utilized LSTM methods to evaluate dialogue-level user impressions. As described in Section 5.4, different unimodal features (audio a_t: 88-dim., linguistic l_t: 785-dim.,and video v_t: 58-dim.) were extracted from the t-th exchange. We use the early fusion method to concatenate different unimodal features, generating the exchange-level multimodal feature $x_t = [a_t,\ v_t,\ v_t]$. The multimodal feature X = (x_1, x_2, ..., x_t) was used as the input of the neural network models. For all the models with one LSTM layer and 128 units, we obtained a 128-dimensional hidden state from the recurrent layer. The recurrent layer was followed by a fully connected layer, which projected the (128-dimensional) output. At the end of the model output layer containing two units, the log-Softmax function was used to output the probabilities of different user rapport labels.

As shown in Figure 5.5 (a), the output at the final moment ht can represent the whole sequence, which uses a fully connected layer followed by a softmax nonlinear layer to predict the probability distribution over different classes. As outlined in Section 5.4,

[4]https://github.com/stanfordnlp/stanza

various unimodal features (audio a_t: 88-dim., linguistic l_t: 785-dim., and video v_t: 58-dim.) were extracted from the t-th exchange. To combine these unimodal features, we employed the early fusion technique, concatenating them to create the exchange-level multimodal feature $x_t = [a_t,\ v_t,\ v_t]$. The multimodal feature X = (x_1, x_2, ..., x_t) was utilized as the input for the neural network models.

5.5.2 Domain adversarial neural network for user rapport (proposed model)

Gradient reversal is a technique inspired by multitasking learning that aims to address domain adaptation problems. In this approach, a neural network model is trained to perform two tasks simultaneously: a primary task and a domain adaptation task. The key idea behind gradient reversal is to force the utilized neural network to learn domain- invariant features during training. During the forward pass of the primary task, the shared layers of the network extract features from the input data. These shared layers capture general patterns and features across both tasks. However, the gradients are reversed for the domain adaptation task during the backward pass. This means that the gradients flowing through the shared layers are multiplied by a negative scalar, effectively reversing their direction. As a result, the shared layers are encouraged to learn domain-invariant features, making them less sensitive to variations between different domains. In summary, gradient reversal allows a neural network to learn task-specific representations while simultaneously learning domain- invariant features. This helps improve the model's performance in the primary task by reducing the influence of domain-specific characteristics.

The ADVER-based model used single LSTM layers with 128 units as the shared layers. These layers extracted features at the user personal information level and the dialog level of user rapport for the tasks shown in Figure. 5.5 (b). For the domain adaptation task, we obtained 128-dimensional hidden states H = (h_1, h_2, ..., h_t) from the LSTM layers, and the output at the final moment h_t could be regarded as a representation of the whole sequence. Subsequently, the h_t serves as the input of a gradient reversal layer, yielding an output G, which was then used as the input of a fully connected layer followed by a Softmax nonlinear layer for predicting the probability distribution over different classes of user‘ personal information. For the dialog-level user rapport recognition task, the structure was the same as that used for the single task, and the mathematical formula of the model can be described as follows:

$$\text{Share layer} : \mathrm{h_t} = \mathrm{LSTM}(\mathrm{x_t W_e}, \mathrm{h_{t-1}}) \tag{5.1}$$

$$\text{Gradient reversal layer} : \mathrm{G} = \mathrm{GRL}(\mathrm{h_t W_g} + \mathrm{b_g}) \tag{5.2}$$

$$\text{Adversarial task layer} : \mathrm{U} = \mathrm{Softmax}(\mathrm{G W_u} + \mathrm{b_u}) \tag{5.3}$$

Table 5.4: Binary classification F1 score of different unimodal user rapport

User rapport	Personal information	Unimodal		
		A	V	L
Well coordinated	Baseline	**0.526**	0.45	0.482
	Age	**0.685**	0.607	0.672
	Gender	**0.642**	0.449	0.583
	Extraversion	**0.613**	0.523	0.583
	Agreeableness	**0.764**	0.656	0.618
	Conscientiousness	**0.613**	0.5	0.552
	Neuroticism	0.505	0.549	**0.549**
	Openness	**0.662**	0.625	0.631
Awkward	Baseline	0.666	0.675	**0.7**
	Age	**0.709**	0.597	0.694
	Gender	0.648	0.665	**0.7**
	Extraversion	**0.719**	0.615	0.713
	Agreeableness	0.686	**0.725**	0.696
	Conscientiousness	0.64	0.642	**0.713**
	Neuroticism	0.608	0.587	0.587
	Openness	0.611	0.675	**0.714**
Engrossing	Baseline	0.659	0.671	**0.705**
	Age	0.658	0.67	**0.683**
	Gender	**0.708**	0.677	0.695
	Extraversion	0.684	0.681	**0.708**
	Agreeableness	0.699	0.711	**0.717**
	Conscientiousness	0.712	0.682	**0.72**
	Neuroticism	0.591	**0.681**	**0.681**
	Openness	0.693	0.679	**0.736**

Table 5.5: Binary classification F1-score of different multimodal for user rapport

User rapport	Personal information	Multimodal A+V	A+L	V+L	ALL	Human model	Instruction-based LMMs model
Well coordinated	Baseline	0.505	0.555	0.45	**0.611**	0.517	0.531
	Age	**0.723**	0.654	0.715	0.649		
	Gender	**0.643**	0.62	0.628	0.54		
	Extraversion	0.598	**0.663**	0.603	0.634		
	Agreeableness	0.676	0.704	0.669	**0.717**		
	Conscientiousness	0.558	0.55	**0.581**	0.576		
	Neuroticism	0.471	0.52	0.471	**0.531**		
	Openness	0.692	0.706	0.664	**0.722**		
Awkward	Baseline	0.682	0.655	**0.696**	0.673	0.607	0.365
	Age	0.684	0.674	**0.744**	0.711		
	Gender	0.678	0.619	**0.742**	0.729		
	Extraversion	0.679	0.662	**0.703**	0.675		
	Agreeableness	0.68	0.66	**0.719**	0.671		
	Conscientiousness	0.67	**0.703**	0.689	0.667		
	Neuroticism	0.649	0.593	0.649	**0.653**		
	Openness	0.641	**0.705**	0.698	0.675		
Engrossing	Baseline	**0.71**	0.69	0.665	0.655	0.571	0.540
	Age	0.647	**0.735**	0.713	0.666		
	Gender	0.683	**0.732**	0.674	0.665		
	Extraversion	0.62	0.693	**0.705**	0.685		
	Agreeableness	0.681	0.647	**0.729**	0.652		
	Conscientiousness	**0.725**	0.639	0.688	0.621		
	Neuroticism	**0.648**	0.632	0.648	0.627		
	Openness	**0.718**	0.653	0.691	0.698		

$$\text{User rapport task layer : D} = \mathrm{Softmax}(\mathrm{h_t W_d} + \mathrm{b_d}) \tag{5.4}$$

The loss of the ADVER base model can be defined as shown in Equation 5.5 where L_d is the cross-entropy loss for the dialogue user rapport classifier, and L_u is the cross-entropy loss for the users' personal information classifier. These two classifiers were adversarial trained. Specifically, the model parameters for user rapport classification were adjusted to minimize the L_d, and the users' personal information classification was adjusted to maximize L_u. A minimax competition enhances the discriminability of user rapport. It suppresses the discriminability of user users' personal information, leading the model to converge where the embeddings we extracted could recognize user rapport but cannot correctly classify them. Therefore, under ideal conditions, the embeddings we obtained were not influenced by users' personal information.

$$\mathrm{L} = \mathrm{L_d} - \lambda * \mathrm{L_u} \tag{5.5}$$

5.6 Experiments

User rapport recognition is a time series task that requires time series information to achieve improved model performance. Therefore, this study used recurrent neural networks suitable for handling sequential inputs of time series information. Based on previous research findings, we used long short-term memory (LSTM) as the baseline model to capture sequences of multimodal behaviors. Additionally, we employed adversarial models to eliminate the impact of users' personal information. In particular, as previously indicated, we aimed to answer the following research questions.

- **(RQ1):** Does adversarially learning users' personal information contribute to rapport recognition?
- **(RQ2):** Are there differences between the impacts of demographic data (age and gender) and personality on user rapport recognition?
- **(RQ3):** Do machine learning models outperform the results of multiple human observations regarding estimation accuracy?

5.6.1 Experimental settings

Given the relatively small database used in this study, a 5-fold cross-validation approach was employed to address potential issues that may arise from improper dataset partitioning. 5-fold cross-validation produces five sets of evaluation results, which are then averaged to produce a final result. The F1-score weighted by the label was used as the evaluation criterion. By utilizing this advanced methodology, highly accurate and reliable results were achieved in the analysis.

You are an expert dialogue evaluator. The following is a dialogue between a user and a dialogue system ;
{
User : ………
Agent : ………
User : ………
Agent : ………
………
}
Based on this dialogue, please assign scores for the user's rapport using the following 18 labels:
•"well coordinated", boring", cooperative", harmonious", unsatisfying", uncomfortably paced", cold", awkward", engrossing", unfocused", involving", intense", friendly", active", positive", dull", worthwhile", and slow". .
Scoring Guidelines:
Each label should be scored with 0 (low) or 1 (high).

•The scores must be presented in the format: coordinated: 1, boring: 0, using commas to separate each score.
After providing the scores, include a brief explanation for each score to justify the evaluation.

(a) Prompt words

well coordinated: 1, boring: 0, cooperative: 1, harmonious: 1, unsatisfying: 0, uncomfortably paced: 0, cold: 0, awkward: 0, engrossing: 1, unfocused: 0, involving: 1, intense: 0, friendly: 1, active: 1, positive: 1, dull: 0, worthwhile: 1, slow: 0

well coordinated: 1 - The dialogue flows smoothly with questions and responses that are relevant and connected, indicating good coordination.
boring: 0 - The conversation covers a variety of interesting topics, such as food, travel, and technology, keeping it engaging.
cooperative: 1 - Both the user and the system contribute actively to the dialogue, maintaining cooperation throughout.
harmonious: 1 - There is a sense of agreement and mutual understanding in the dialogue, making it harmonious.
unsatisfying: 0 - The dialogue covers several topics and provides informative exchanges, making it satisfying.
uncomfortably paced: 0 - The pacing of the conversation seems comfortable and natural, without rushed segments.
cold: 0 - The dialogue has a warm and friendly tone, with personal anecdotes and shared experiences.
awkward: 0 - The interaction does not exhibit any awkwardness; it is smooth and coherent.
engrossing: 1 - The dialogue includes engaging topics and the user appears interested, making it engrossing.
unfocused: 0 - The conversation stays on track with specific topics, indicating focus and direction.
involving: 1 - The user is actively participating and sharing personal experiences, making the dialogue involving.
intense: 0 - The conversation is relaxed and friendly rather than intense.
friendly: 1 - The tone of the conversation is friendly, with expressions of interest and shared laughter.
active: 1 - Both the user and the system maintain an active participation throughout the dialogue.
positive: 1 - The dialogue maintains a positive tone, with expressions of enjoyment and interest.
dull: 0 - The variety of topics and the dynamic interaction prevent the dialogue from being dull.
worthwhile: 1 - The dialogue is meaningful and informative, making it a worthwhile interaction.
slow: 0 - The dialogue progresses at a steady and engaging pace, not slow.

(b) Example of Chagpt4-o response

Figure 5.6: Instruction Templates and Evaluation Responses for Instruction-Based LLMs

We experimented with one baseline model and seven ADVER models for three selected rapport tasks. All machine learning models were trained with seven combinations of unimodal features (audio, visual, and linguistic features) to analyze the effectiveness of unimodal and multimodal features. The seven combinations of automatic feature sets:
(1) A: Model trained with acoustic features
(2) V: Model trained with visual features
(3) L: Model trained with linguistic features
(4) A+V: Model trained with acoustic + visual features
(5) A+L: Model trained with acoustic + linguistic features
(6) V+L: Model trained with visual + linguistic features
(7) ALL: Model trained with acoustic + visual + linguistic features

5.6.2 Comparative methods

5.6.2.1 Human model

We prepared a human model using third-party annotations to evaluate the user rapport levels of dialogues. These third-party annotations were considered the outcomes of human perception. The average of the third-party annotation results was classified into high- and low-satisfaction categories based on a predefined threshold, and the F1 score was computed to evaluate the corresponding performance.

Table 5.6: Binary classification of user rapport results, “Diff” denotes the difference in F1-scores between the single task and ADVER_base models

	Well Coordinated		Awkward		Engrossing	
	Best	**Diff**	**Best**	**Diff**	**Best**	**Diff**
Baseline	0.611	/	0.7	/	0.71	/
Age	0.723	0.112	0.744	0.044	0.735	0.025
Gender	0.643	0.032	0.742	0.042	0.732	0.022
Extraversion	0.663	0.052	0.719	0.019	0.708	-0.002
Agreeableness	0.764	0.153	0.725	0.025	0.729	0.019
Conscientiousness	0.613	0.002	0.713	0.013	0.725	0.015
Neuroticism	0.549	-0.062	0.653	-0.047	0.681	-0.029
Openness	0.722	0.111	0.714	0.014	0.736	0.026

5.6.2.2 Instruction-based LLMs model

We use GPT-4o as a dialogue expert and evaluate the overall dialogue by assessing the user's rapport. We carefully designed the prompt instruction using [130] to output stable evaluation results. Specifically, we input the complete dialogue content and a carefully designed prompt into the model to obtain user rapport scores across different dimensions. The specific requirements of response are as follows:
Scoring: Assign a score for each rapport label using 0 (low) or 1 (high).
Explanation: Provide a brief justification for each score, explaining the reasoning behind the evaluation.

ChatGPT-4o is a cross-lingual model. Our testing revealed that English prompts can effectively evaluate Japanese dialogues. For clarity, the full prompt used to generate the evaluation is shown in Figure 5.6 (a), and Figure 5.6 (b) presents an example of a response from ChatGPT-4o. As shown in the table, each score is accompanied by a rationale that explains the reasoning behind the model's assessment. Aligned with the human model, the F1 score was computed to evaluate the corresponding performance.

5.7 Results

Table 5.4 shows the F1 scores of the three unimodal [A, V, L] models across the three binary classification tasks. Table 5.5 presents the F1 scores obtained for the four multimodal feature sets [A+V, A+L, V+L, A+V+L] and the human model across the same three tasks. All the tasks are listed as eight sub-rows, which consist of one baseline model and seven ADVER-based models (age, gender, extraversion, agreeableness, conscientiousness, neuroticism, and openness).

5.7.1 Efficacy of Adversarial for user Rapport recognition (ANSWER TO RQ1)

5.7.1.1 Unimodal feature comparison

Table 5.4 shows the three-task classification results obtained based on the unimodal features. The table shows the following.

- Well_coordinated: The acoustic features achieved the best performance in the baseline. The age, gender, extraversion, agreeableness, and openness models improved upon the baseline, with the best result obtained by ADVER-Openness (0.526 to 0.764) with acoustic features.
- Awkward: The acoustic features yielded the best performance for the baseline (0.666). ADVER-Extraversion (0.719) achieved the best results, with a 0.53 improvement in the acoustic features. Among all the models, ADVER-Agreeableness (0.725) achieved the highest score with visual features.
- Engrossing: As shown in the table, the linguistic features produced the best results in the unimodal baseline (0.705). ADVER-Openness (0.736) achieved the highest score with linguistic features among all the adversarial models.

5.7.1.2 Multimodal feature comparison

Table 5 presents the classification results of the three tasks based on multimodal features. The results show the following.

- Well_coordianted: In the multimodal experiments, the All feature set yielded the best performance in the baseline (0.611), representing an improvement of 0.085 over the best unimodal feature. ADVER-Gender (0.723) achieved the highest score with the A+V feature set, closely followed by ADVER-Openness (0.722) with the All feature set.
- Awkward: The V+L feature set produced the best performance for the baseline (0.696) among the multimodal feature sets. For the adversarial task model, the V+L feature set yielded the best result with ADVER-AGE (0.744), with ADVER-Gender (0.742) following closely behind.
- Engrossing: The A+V feature set yielded the best performance in the baseline (0.71), slightly improving upon that achieved with the unimodal features (0.705). In the adversarial experiments, the age, gender, agreeableness, conscientiousness, and openness models performed better. The A+L feature set yielded the best result with ADVER-AGE (0.735), closely followed by ADVER-GENDER with the A+L feature set (0.732).

Overall, the results of the adversarial experiments showed improvements in both unimodal and multimodal settings .

5.7.2 Impact of Demographic Data vs. Personality on User Rapport Recognition (ANSWER TO RQ2)

To investigate the impact of personality and demographic data on user rapport in human-computer dialogue. We present the best F1 scores attained by the baseline and adversarial-based models with seven different multimodal feature sets, which are based on age, gender, and the Big Five personality in Table 5.6. Overall, we observed improvements in user rapport recognition effects achieved with the ADVER-based methods. As indicated in Table 5.6, for demographic data (age and gender), the ADVER-based models generally demonstrated enhancements in all tasks. Most adversarial experiments showed improvements for the Big Five personality traits, except for the engrossed task, where ADVER-Extraversion experienced a slight decrease (0.002) compared with the baseline result. Moreover, across all the tasks, the performance of the ADVER-Neuroticism model decreased, suggesting a lack of useful information for predicting user attitudes within the neuroticism personality trait. Overall, ADVER-Agreeableness achieved the greatest improvement (0.153) in the well-coordinated task, whereas ADVER-Age achieved the greatest improvement (0.044) in the awkward task. ADVER-Openness attained the greatest increase of 0.026 in the engrossed task.

In summary, the adversarial learning models generally improved the handling of demographic data (age and gender). While overall performance enhancements were observed when addressing the Big Five personality traits, notable variations were exhibited across different personalities. Some personalities, such as neuroticism, did not yield the expected improvements, whereas others, such as agreeableness and openness, demonstrated significant performance gains.

5.7.3 Validate the reliability of the overall system (ANSWER TO RQ3)

The human model and instruction-based LLMs columns of Table 5.5 present the results for the human model and the instruction-based LLMs model, respectively. When compared to the human model, the instruction-based LLMs model shows a +0.014 improvement in the ”well coordinated” label. However, the instruction-based LLMs model experiences a -0.031 decline in the ”Engrossing” label. Additionally, the performance on the ”Awkward” label is notably poor, with an F1 score of only 0.365, significantly lower than that of the other models. The human model and instruction-based LLMs scores were considerably inferior to those of the machine learning baseline model for every task, and the proposed method further improved the baseline score for every task. More specifically, the proposed method significantly improved the well-coordinated task with +0.153 improvement and increased the scores achieved in the other rapport tasks to various degrees.

The proposed system outperformed the existing models in terms of performance and demonstrated its reliability and effectiveness through empirical evidence, making it suitable for practical user rapport recognition applications..

5.8 Discussion

5.8.1 Feature analysis

Combining the results shown in Tables 5.4 and 5.5, for the well_coordinated label, we found that the All feature set yielded the best performance in the baseline, consistent with the results in [114]. This aligns with the understanding that communication is a cooperative activity involving coordinated behaviors [102]. Furthermore, some studies have shown that dialogue participants spontaneously adjust their facial expressions, postures, pronunciation, and speech rates [103–105]. Among the adversarial models, ADVER-Agreeableness achieved the best result (0.764). The agreeableness personality is significantly related to coordination in dialogue. Individuals with high agreeableness are generally more cooperative and easier to work with, leading to more coordinated interactions in conversations. For the awkward label, the V+L feature set yielded the best result in the baseline, consistent with [114, 135]. We found that ADVER-Age achieved the best results in both the awkward and engrossing tasks, indicating that age significantly impacts rapport in conversations. Therefore, it is important to consider the user's age in the decision-making process of dialogue systems.

Neuroticism, as mentioned in the previous section 5.7.2, did not achieve the expected improvements in most of the tasks. Previous work [134] reported that neuroticism significantly differs from self-reported measures even when annotated by humans. Therefore, annotators find it challenging to accurately judge a user's neuroticism based solely on a single dialogue. As [139]also indicates, neuroticism predictions are less accurate than conscientiousness and extroversion. In our study, the poor performance of Neuroticism related model may be due to the fact that the data in the dataset is based on casual conversations, where users typically engage in relaxed and humorous interactions. Such communication styles may not exhibit significant emotional fluctuations, especially concerning neuroticism, which is typically associated with emotional instability and anxiety. As a result, the system may not capture enough emotional variation to identify these traits. In summary, this may be because our proposed model did not learn relevant information about neuroticism from a single dialogue, thus failing to achieve the expected improvements.

In future data collection efforts, we will focus on enabling the system to gather user interaction data across multiple conversational scenarios. For instance, discussions around stress, anxiety, or emotional topics can provide contrasting contexts that will

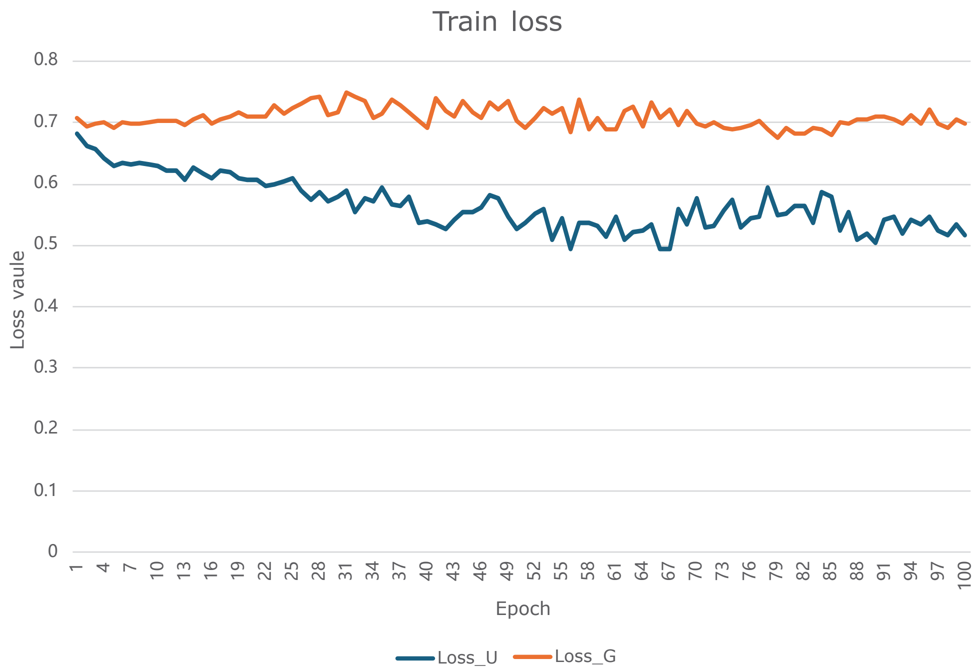


Figure 5.7: Train loss of ”Engrossing” label based on gender adversarial training in the A+L feature set. Loss_U: train loss of the use rapport (engrossing), Loss_G: train loss of gender task.

help the system observe and identify personality traits like neuroticism more effectively. By collecting data in various scenarios, we can more accurately assess the manifestation of personality traits, thereby improving the system's accuracy and adaptability in recognizing personality characteristics and enhancing its performance in identifying user rapport.

5.8.2 Effects of adversarial learning

To investigate whether the model genuinely learned features related to the task objective independently of users' personal information, we utilized an ADVER-based approach to recognize relevant users' personal information. Figure 5.7 presents the training loss of the ”engrossing” label based on adversarial gender training with the A+L feature set. The graph illustrates the training losses for gender and the ”engrossed” label across the different epochs in a single fold.

As shown in Figure 5.7:

1): Main task ”engrossing” label loss: As training progressed, the performance achieved by the model in the main task improved, with the main task loss gradually decreasing until convergence was reached.

Baseline

	Estimated High	Estimated Low
Actual High	42.4%	17.6%
Actual Low	12.8%	27.2%

Human model

	Estimated High	Estimated Low
Actual High	53.6%	6.4%
Actual Low	22.8%	11.2%

LLMs model

	Estimated High	Estimated Low
Actual High	38.4%	21.6%
Actual Low	22.4%	17.6%

ADVER_Gender

	Estimated High	Estimated Low
Actual High	45.6%	14.4%
Actual Low	12.0%	28.0%

Figure 5.8: The confusion matrix of Baseline, human model, Instruction-based LLMs model, and ADVER-Gender model for the engrossing label in the A+L feature set.

2): Complexity of the adversarial gender loss: the adversarial task loss exhibited a more complex pattern, initially decreasing during the early stages of training. However, owing to the reversal effect of the GRL, the loss experienced fluctuations, reflecting the ongoing adaptation of the feature extractor to the requirements of domain-adversarial training.

To gain further insights into the impact of engrossing labels, we separately compiled the results of different models for engrossing labels. The confusion matrix depicting these results is presented in Figure 5.8. The figure shows that the human model achieved the best performance in high engrossing with a result of 53.6%. Instruction- based LLMs performed poorly in both low-engrossing and high-engrossing categories. Additionally, both the human model and the instruction-based LLMs showed incorrect low-level recognition results (misclassifying true low engrossing as high engrossing), accounting for 22.8% and 22.4% of the total samples, respectively, which is significantly higher than the baseline and ADVER-Gender (12.8% and 12.0%). The human model's and instruction-based LLMs' poor performance may be attributed to the lack of domain-specific training and fine-tuning, hindering their ability to capture task-related information effectively.

Additionally, for the instruction-based LLMs model, the biases in pre-trained models and their limited understanding of the entire dialogue context also affect their performance. ADVER-Gender demonstrates improvements in both high and low-engrossing labels compared to the baseline, with respective increases of 3.2% and 0.8%. In conclusion, these results indicate that the model has learned gender-independent features while enhancing the label recognition performance. These results demonstrate that the model has indeed learned gender-independent features while enhancing the label recognition performance.

5.9 Chapter Summary

In this work, we first investigated the relationship between users' personal information and their rapport. We found that age, gender, and personality differences do influence user rapport. To address this influence, we proposed a domain-adversarial model that reduces the impact of user trials by learning adversarial features that are unrelated to users' personal information. The results indicate that our proposed adversarial learning model achieved a significant performance improvement. Moreover, our system consistently demonstrated its superior performance to the annotations by the human and instruction-based LLM models, thereby confirming the reliability and effectiveness of our system. In addition to the limitations imposed by the size of the utilized dataset, our model still has room for improvement. We continue to collect relevant data across multiple conversational scenarios and closely monitor the release of new datasets suitable for our research. As the data at the sentence level are sufficient, we will try to conduct analyses at the sentence level to explore the impact of users' personal information on users during human-computer interactions.

Chapter 6

Conclusion

6.1 Summary

In conclusion, this dissertation has explored three aspects of multimodal dialogue system evaluation, aiming to enhance user satisfaction and system performance through innovative methodologies. In Chapter 1, the background of the research field is introduced, and the target research questions are stated as part of the introduction. Through a comprehensive review of related work in Chapter 2, we discussed the evolution of dialogue systems and the limitations of existing evaluation methods in considering multimodal information, processing user data at the exchange and dialogue levels, and accounting for individual user information. Additionally, we emphasized the importance of incorporating multimodal data for a more comprehensive evaluation

Chapter 3 used a multimodal modeling approach to evaluate dialogue-level user satisfaction, addressing the evaluation challenges of non-task-oriented multimodal dialogue systems. By leveraging multimodal data, we achieved a more comprehensive understanding of user experience, as demonstrated by the experimental results.

In Chapter 4, to simultaneously consider exchange-level and dialogue-level information, we proposed a multi-task learning model that integrates user impressions at both levels, thereby improving evaluation accuracy by accounting for information from both perspectives and simultaneously considering both levels of information.

Chapter 5 addressed the challenge of mitigating the influence of user characteristics on user impressions through an adversarial learning model. This model effectively reduced biases introduced by user characteristics, enhancing the generalizability and robustness of the dialogue system.

The main findings and contributions of this research can be summarized as follows:

Multimodal Data Integration Enhances Evaluation Accuracy: his study demonstrates that integrating multimodal data into the evaluation of dialogue systems significantly enhances the comprehensiveness and accuracy of the evaluation. By combining information from various modalities, such as speech, text, gestures, and facial expressions, we can better understand user satisfaction and system performance. Compared to traditional unimodal evaluation methods, multimodal data fusion provides richer feedback on user experience, capturing users' complex emotions and interaction states. This finding highlights the necessity of integrating cross-modal information and the importance of modality differences in recognizing user impressions across different modalities.

Comprehensive Evaluation of Dialogue-Level and Exchange-Level Information: In Chapter 4, we introduced a multi-task learning model demonstrating the importance of simultaneously considering dialogue-level and exchange-level information to improve evaluation accuracy. We show that traditional approaches, which evaluate only one level of interaction (either dialogue-level or exchange-level), often overlook the holistic impact on user experience. Combining both levels offers more comprehensive and accurate user feedback. This finding reveals the complexity of dialogue interactions. It supports the need for incorporating multi-level information into evaluation frameworks, thereby improving the effectiveness and practical applicability of dialogue system evaluations.

of User Personality Traits Impact on Evaluation: In Chapter 5, we proposed an adversarial learning model that effectively reduces biases introduced by user characteristics (such as age, gender, etc.). This model allows the system to achieve greater universality and stability in evaluation results, ensuring that individual user differences do not skew the overall evaluation objectivity. This contribution is significant for promoting the universality and fairness of dialogue system evaluations, particularly in diverse user group applications, ensuring that the evaluation results are more representative and widely applicable.

Multidmolites Integrated Framework for Dialogue System Evaluation: The innovative approach proposed in this research enhances the depth and breadth of dialogue system evaluations at multiple levels. By integrating multimodal data, multi-task learning, and adversarial learning techniques, we designed a comprehensive framework that reflects user experience. This framework overcomes the limitations of traditional unimodal or single-level evaluations. It provides theoretical and practical insights for evaluating more complex dialogue systems in the future, advancing the development of dialogue system evaluation methodologies.

Improved Accuracy of User Satisfaction Prediction: By combining multimodal data with a multi-task learning model, we successfully improved the accuracy of user satisfaction prediction. When applied to non-task-oriented dialogue systems, this model allows for a better understanding and prediction of users' emotional shifts and satisfaction. This achievement provides a new methodological foundation for optimizing non-task-oriented dialogue systems, particularly in practical applications such as intelligent assistants and social robots.

Overall, this dissertation contributes to advancing multimodal dialogue system evaluation by introducing innovative methodologies and addressing key challenges in the field. By providing valuable insights and empirical evidence, this research lays the groundwork for future advancements and opens up new avenues for exploration.

6.2 Future work

The current work has the following limitations: First, the research primarily relies on the Hamauzi dataset, which involves a single type of dialogue system with limited diversity in dialogue styles and strategies. This restricts the ability of the evaluation method to differentiate between various dialogue systems, thereby limiting its generalizability and applicability across broader scenarios. Second, the evaluation method lacks sufficient explainability, as it does not identify which specific features of the dialogue system influence dialogue quality or how these features exert their impact. Although certain metrics or scores are provided, the connection between key factors—such as speech recognition accuracy or natural language understanding performance—and overall dialogue quality remains unclear.

Future work can be expanded into two main areas. First, the diversity of dialogue systems can be enhanced by introducing and evaluating various dialogue systems encompassing different dialogue styles and strategies. This will improve the generalizability of the evaluation methods, allowing them to differentiate and assess the performance of various dialogue systems in practical applications, thus broadening the applicability of the methods. Second, addressing the issue of explainability. The challenge arises because subjective evaluation results are often unstructured textual information, which is difficult to quantify and analyze, requiring considerable time and effort to interpret and summarize. To tackle this, we plan to integrate subjective and objective evaluations to develop more interpretable evaluation methods. These methods will reveal which features of the dialogue system impact dialogue quality and how they exert their effects. By conducting in-depth analysis and providing visualizations of the mechanisms behind these key factors, we aim to help developers understand the evaluation results better and make targeted improvements.

Future work can further expand into large models' application and interpretability research. With the rapid development of large models, particularly in emotion recognition and sentiment analysis, these models have demonstrated strong potential in understanding and predicting user emotions. By integrating deep learning and multimodal information, these models can more accurately identify and respond to users' emotional states during conversations, thereby providing more precise data support for evaluating and optimizing dialogue systems.

Appendix A1

ALL results of 18 type annotations

This section will present the prediction results for eighteen labels for a more comprehensive discussion on user performance in multimodal human-computer interaction. Based on the data in Table A1.1, the LSTM model yields relatively good results. Therefore, we use LSTM to identify these eighteen labels.

Table A1.1 presents the F1-scores for classifying all user labels using various multimodal feature sets. These results provide insights into the effectiveness of individual modalities (audio, video, and linguistic features) and their combinations for predicting different user rapport labels. This section delves into the performance trends and contrasts across feature sets to highlight their contributions.

Boring: For the **"boring"** label, linguistic-only (**L**) features achieved an F1-score of 0.74, outperforming all multimodal combinations. This result highlights that textual information, such as word choice, phrasing, and sentiment, is the most critical factor in detecting boredom. Interestingly, the adding audio (**A+L**) or video (**L+V**) features led to slight performance drops, suggesting that non-verbal cues might introduce noise rather than complementary information for this label.

Engrossing: The **A+V** feature set (audio and video) yielded the highest F1-score (0.71) for the **"engrossing"** label. This underscores the importance of non-verbal cues in capturing user engrossment, such as vocal prosody (tone, pitch, rhythm) and visual engagement (e.g., eye contact, gestures). By contrast, the linguistic-only (**L**) feature set achieved an F1-score of 0.63, showing that verbal content alone cannot represent the multimodal nature of engrossment.

Intense: For the **"intense"** label, the **A+L** feature set (audio and linguistic) achieved the highest F1-score (0.72), emphasizing the importance of prosody and speech content. This label seems to rely heavily on tone, volume, and verbal emphasis changes to convey intensity. The **ALL** feature set produced slightly lower performance (F1 = 0.71), suggesting that while video features can complement audio and linguistic information, their contribution to intensity detection may be limited.

Friendly: All models were overfitted due to the dataset imbalance for the "friendly" label, leading to unreliable results. Consequently, it is challenging to derive meaningful conclusions about the contribution of different feature sets for this label.

Worthwhile: The **A+V** feature set outperformed all others for the **"worthwhile"**

Table A1.1: Binary classification F1-score of each multimodal combination of three user satisfaction labels (Acoustic (A), Visual (V), Linguistic (L) with BiLSTM model

	L	**A+L**	**A+V**	**L+V**	**ALL**
well coordinated	0.70	0.74	0.75	0.65	0.76
boring	0.74	0.66	0.69	0.66	0.53
cooperative	0.68	0.67	0.68	0.66	0.61
harmonious	0.59	0.64	0.65	0.61	0.62
unsatisfying	0.69	0.67	0.57	0.68	0.54
uncomfortably paced	0.68	0.67	0.62	0.58	0.67
cold	0.58	0.50	0.65	0.48	0.56
awkward	0.65	0.58	0.61	0.68	0.63
engrossing	0.63	0.70	0.71	0.62	0.62
unfocused	0.63	0.64	0.44	0.60	0.58
involving	0.60	0.63	0.65	0.59	0.61
intense	0.64	0.72	0.70	0.66	0.64
friendly	0.73	0.60	0.72	0.72	0.73
active	0.71	0.59	0.64	0.58	0.71
positive	0.62	0.63	0.76	0.72	0.68
dull	0.66	0.47	0.63	0.70	0.59
worthwhile	0.69	0.67	0.74	0.64	0.73
slow	0.62	0.57	0.65	0.70	0.60

label, achieving an F1-score of 0.74. This suggests that non-verbal cues primarily influence users' perceptions of worthwhile interactions. Interestingly, the **ALL** feature set performed similarly (F1 = 0.73), indicating that including linguistic features offers only marginal improvements for this label.

A1.1 Discussion

Modality Dominance: From the results, specific user rapport labels, such as "boring" and "awkward," are primarily dominated by particular modalities. This suggests that different emotions or user reactions manifest in various forms during communication. For instance, the high accuracy of the "boring" label depends largely on linguistic modalities, such as word choice and tone, while "awkwardness" is more easily recognized through visual cues, such as body language and facial expressions. These findings emphasize the necessity of selecting the most relevant modal features when conducting multimodal analysis for different emotional states. For emotions that exhibit clear signals in specific modalities, a single modality might be sufficient to capture key features. However, multimodal integration becomes increasingly important for more complex or multidimensional emotional responses.

Multimodal Synergy: The high efficiency of the combined feature set (ALL) across most labels, especially when dealing with complex user reactions, highlights the significant advantages of multimodal synergy. By incorporating audio, video, and linguistic features, Multimodal systems can more comprehensively capture the multidimensional information inherent in human interactions, thereby improving the accuracy of emotional recognition. This also underscores the role of multimodal systems in addressing key information that a single modality might miss. For example, during emotional expression, the video modality might capture micro-expressions and gestures, which, combined with tonal and volume variations from the audio modality, help the system more accurately identify subtle emotional states. Future research could consider further optimizing multimodal feature fusion strategies and exploring the interactions between different modalities to enhance the system's emotional understanding capabilities.

Audio Contributions: Audio features, particularly prosody (such as pitch, tone, speech rate, and volume), play a crucial role in recognizing the "intense" label. Emotional fluctuations in speech can effectively convey strong emotions or tension during interactions, which is essential for accurately capturing emotional labels such as "intense." However, for labels like "boring," the contribution of audio features is relatively small, possibly because "boring" is more often expressed through monotonous text or a lack of emotional expression rather than audio features. This finding suggests that in constructing emotion analysis models, the role of audio might not be universally applicable, and the decision to include audio features should depend on the specific characteristics of the emotional label.

A1.2 Limitation

Dataset Imbalance: The imbalance in the dataset, particularly the extreme imbalance in the "friendly" label, severely impacts the model's training effectiveness.Imbalanced samples typically lead to overfitting, where the model performs well on training data but may fail to predict minority class labels in real-world applications reliably. This phenomenon highlights the need to apply specific data processing techniques, such as oversampling, undersampling, or class weight adjustment when dealing with emotion recognition tasks, especially when multimodal features are involved. Balancing the sample distribution across labels is crucial for improving the model's generalization capability.

A1.3 Summary

These results show the importance of modality selection and feature fusion in multimodal emotion recognition systems. The characteristics of different labels determine which modalities are most effective, so it is crucial to tailor the feature selection strategy for each label. The system can gain a more comprehensive understanding and assessment of user emotional states by effectively integrating multimodal information. Furthermore, addressing dataset imbalance through appropriate data processing techniques is crucial for enhancing the reliability and accuracy of the model.

Appendix A2

ALL results of other types annotations

A2.1 Analysis of Results for Uncomfortablypaced, Intense, and Slow Labels

In this part we add the more three "Uncomfortably paced," "Intense," and "Slow" as experimental prediction targets. From the perspective of data balance, the selection of "well coordinated," "awkward," "engrossing," "Uncomfortably paced," "Intense," and "Slow" as experimental prediction targets is primarily based on their diversity and representativeness in sample distribution. These six labels encompass a range of distribution types, from highly imbalanced to nearly balanced. For instance, "well coordinated" exhibits a high proportion of positive samples (81.6%), while "awkward" presents a near-balanced distribution (53.6% positive vs. 46.4% negative). This variation comprehensively evaluates the model's performance and robustness under different data balance conditions.

Moreover, these labels avoid extreme imbalance scenarios, such as "friendly" (125/0) or "cold" (17/108), where severe skewness in sample distribution could lead to overfitting on the majority class and hinder the model's ability to effectively learn minority class features. Additionally, the selected labels align with the experimental objectives, covering interaction-related attributes (e.g., "Uncomfortably paced" and "Slow") and emotion-related dimensions (e.g., "Intense" and "engrossing"). This ensures a comprehensive assessment of the model's capability in multimodal dialogue analysis, particularly in capturing emotional and interactional nuances.

The experimental results of "Uncomfortably paced," "Intense," and "Slow" in Table A2.1 provide insights into the performance of various unimodal and multimodal feature sets in binary classification tasks for the *Uncomfortablypaced*, *Intense*, and *Slow* user satisfaction labels. Additionally, the impact of demographics (age and gender) and personality traits (Big Five dimensions) on model performance is analyzed.

Table A2.1: Binary classification of ML model for the User impressions result

		Unimodal			Multimodal				Human model
		A	**V**	**L**	**A+V**	**A+L**	**V+L**	**A+V+L**	
Uncomfortably-paced	**Baseline**	0.655	0.65	**0.716**	0.674	0.703	0.701	0.707	0.590
	Age	0.617	0.639	**0.684**	0.632	0.683	0.632	0.642	
	Gender	0.68	0.615	0.669	0.632	0.607	**0.69**	0.59	
	Extraversion	0.676	0.636	0.707	0.648	**0.708**	0.642	0.602	
	Agreeableness	0.659	0.663	**0.722**	0.681	0.659	0.691	0.622	
	Conscientious-ness	0.659	0.642	**0.727**	0.678	0.701	0.654	0.663	
	Neuroticism	0.6	0.627	**0.627**	0.594	0.627	0.594	0.628	
	Openness	0.677	0.69	**0.703**	0.625	0.659	0.648	0.648	
Intense	**Baseline**	0.684	0.67	0.639	0.696	0.707	0.657	**0.726**	0.400
	Age	0.698	0.602	0.638	0.667	0.698	0.643	0.714	
	Gender	0.642	0.449	0.583	0.643	0.62	0.628	0.54	
	Extraversion	**0.742**	0.652	0.656	0.691	**0.744**	0.636	0.734	
	Agreeableness	**0.738**	0.698	0.7	0.724	0.688	0.722	0.671	
	Conscientious-ness	0.683	0.622	0.676	0.717	**0.723**	0.686	0.729	
	Neuroticism	0.684	0.69	0.69	**0.716**	0.678	**0.716**	0.712	
	Openness	0.697	0.701	0.697	**0.728**	0.691	0.672	0.685	
Slow	**Baseline**	0.67	0.67	0.655	0.67	0.663	0.616	**0.674**	0.563
	Age	0.621	0.705	0.653	0.662	**0.706**	0.648	0.664	
	Gender	0.607	0.708	0.656	0.698	**0.715**	0.691	0.631	
	Extraversion	0.681	0.67	0.646	0.66	**0.688**	0.653	0.684	
	Agreeableness	**0.705**	0.677	0.66	0.677	0.656	0.625	0.656	
	Conscientious-ness	0.688	0.674	0.646	0.694	0.642	0.632	**0.712**	
	Neuroticism	**0.674**	0.612	0.612	0.647	0.641	0.647	0.648	
	Openness	0.68	0.68	0.659	0.655	0.685	0.612	**0.691**	

A2.1.1 Uncomfortablypaced Label

Unimodal vs. Multimodal Comparison: Among unimodal features, the linguistic feature set (L) achieved the highest F1-score (0.716) for the baseline, outperforming audio (A: 0.655) and video (V: 0.650). Multimodal combinations, especially A+L and V+L, performed comparably (A+L: 0.703, V+L: 0.701), but did not surpass the linguistic-only model. The best-performing multimodal combination (A+V+L) achieved an F1-score of 0.707, indicating limited improvements from adding audio and video features.

Impact of Age and Gender: The age-based model's performance was slightly lower for L (0.684), with multimodal combinations like A+L and A+V failing to exceed unimodal results. Gender-based analysis showed an interesting reversal: the video feature (V: 0.615) underperformed compared to its combination with linguistic features (V+L: 0.690), suggesting that gender-related visual behaviors contributed to dialogue pacing classification.

Personality Analysis: The **Agreeableness** dimension yielded the best performance using linguistic features (L: 0.722). Similarly, Conscientiousness achieved strong results (L: 0.727). Multimodal combinations provided limited improvements for traits like Extraversion (A+L: 0.708) but performed comparably with unimodal setups.

A2.1.2 Intense Label

Unimodal vs. Multimodal Comparison: Multimodal models exhibited clear advantages for the Intense label. The A+V+L feature set achieved the best performance (0.726), significantly outperforming unimodal setups (A: 0.684, V: 0.670, L: 0.639). Audio combined with linguistic features (A+L) also performed well, with an F1-score of 0.707, indicating the importance of prosody in combination with verbal content.

Impact of Age and Gender: Age-related models showed robust performance improvements when using multimodal features (A+V+L: 0.714), suggesting that multimodal signals captured age-related nuances effectively. For gender, multimodal combinations provided inconsistent results, with A+V+L underperforming (0.540) compared to simpler setups like V+L (0.628). This may indicate gender-specific challenges in audio-visual integration for intense dialogue classification.

Personality Analysis: For personality traits, Extraversion and Agreeableness showed the strongest results with multimodal feature sets (Extraversion, A+L: 0.744, Agreeableness, A+V: 0.738), emphasizing the role of expressive non-verbal behaviors in intense dialogue scenarios. Other traits like Neuroticism and Openness also benefited from multimodal combinations, with A+V+L providing consistent improvements.

A2.1.3 Slow Label

Unimodal vs. Multimodal Comparison: For the Slow label, multimodal setups again outperformed unimodal ones. The A+V+L feature set achieved the highest F1-score (0.674), while unimodal results were lower (A: 0.670, V: 0.670, L: 0.655). Combining audio and linguistic features (A+L: 0.663) or video and linguistic features (V+L: 0.616) showed moderate improvements over unimodal setups.

Impact of Age and Gender: For age, the best performance was achieved with A+L (0.706), suggesting that prosody and verbal features were more indicative of age-related differences in dialogue pacing. Gender-based models showed similar trends, with the best results coming from A+L (0.715). This highlights the importance of audio-linguistic interactions in capturing gender-specific behaviors related to slow dialogue.

Personality Analysis: For personality traits, **Agreeableness** achieved the best performance using audio features (A: 0.705), followed by multimodal combinations like A+V+L (0.656). and Openness also exhibited strong results with A+V+L, emphasizing that pacing-related judgments rely on a combination of verbal and non-verbal cues.

A2.1.4 Discussion

Modality: Linguistic features consistently performed well across all labels, highlighting the importance of verbal content. However, multimodal combinations like A+V+L often provided slight improvements, especially for nuanced labels like Intense. **Impact of Demographics:** Age and gender played a significant role in dialogue behavior, with multimodal models showing enhanced sensitivity to these attributes. **Personality and Dialogue Labels:** Personality traits exhibited diverse patterns of interaction with dialogue labels. Traits like Agreeableness and Extraversion benefited the most from multimodal setups, whereas others like Neuroticism showed less variation across feature sets.

These results underscore the importance of multimodal feature selection to the specific nuances of user impression labels and highlight the challenges posed by demographic and personality-related variability.

References

[1] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, “Towards an automatic turing test: Learning to evaluate dialogue responses,” *arXiv preprint arXiv:1708.07149*, 2017.

[2] L. Liao, Y. Ma, X. He, R. Hong, and T.-s. Chua, “Knowledge-aware multimodal dialogue systems,” in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 801–809.

[3] W. Li, D. Jiang, W. Zou, and X. Li, “Tmt: A transformer-based modal translator for improving multimodal sequence representations in audio visual scene-aware dialog,” *arXiv preprint arXiv:2010.10839*, 2020.

[4] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” *Advances in neural information processing systems*, vol. 32, 2019.

[5] J. D. Williams, “Applying pomdps to dialog systems in the troubleshooting domain,” in *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*, 2007, pp. 1–8.

[6] B. Thomson and S. Young, “Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems,” *Computer Speech & Language*, vol. 24, no. 4, pp. 562–588, 2010.

[7] A. Kannan and O. Vinyals, “Adversarial evaluation of dialogue models,” *arXiv preprint arXiv:1701.08198*, 2017.

[8] Y. Zhang, W. Zhang, and T. Liu, “Survey of evaluation methods for dialogue systems,” *SCIE NTIA SINICA Informationis*, vol. 47, no. 8, pp. 953–966, 2017.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[10] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[11] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[12] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[13] W. A. Woods, "Progress in natural language understanding: an application to lunar geology," in *Proceedings of the June 4-8, 1973, national computer conference and exposition*, 1973, pp. 441–450.

[14] T. Winograd, "Understanding natural language," *Cognitive psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[15] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd, "Gus, a frame-driven dialog system," *Artificial intelligence*, vol. 8, no. 2, pp. 155–173, 1977.

[16] L. D. Erman, F. Hayes-Roth, V. R. Lesser, and D. R. Reddy, "The hearsay-ii speech-understanding system: Integrating knowledge to resolve uncertainty," *ACM Computing Surveys (CSUR)*, vol. 12, no. 2, pp. 213–253, 1980.

[17] R. W. Picard, "Toward computers that recognize and respond to user emotion," *IBM systems journal*, vol. 39, no. 3.4, pp. 705–719, 2000.

[18] Z. Wei, Q. Liu, B. Peng, H. Tou, T. Chen, X.-J. Huang, K.-F. Wong, and X. Dai, "Task-oriented dialogue system for automatic diagnosis," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 201–207.

[19] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Science China Technological Sciences*, vol. 63, no. 10, pp. 2011–2027, 2020.

[20] M. Zhao, L. Wang, Z. Jiang, R. Li, X. Lu, and Z. Hu, "Multi-task learning with graph attention networks for multi-domain task-oriented dialogue systems," *Knowledge-Based Systems*, vol. 259, p. 110069, 2023.

[21] S. Kobyashi and M. Hagiwara, "Non-task-oriented dialogue system considering user's preference and human relations," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 31, no. 1, 2016.

[22] M. Inaba, N. Iwata, F. Toriumi, T. Hirayama, Y. Enokibori, K. Takahashi, and K. Mase, "Constructing a non-task-oriented dialogue agent using statistical

response method and gamification," in *International Conference on Agents and Artificial Intelligence*, vol. 2. SCITEPRESS, 2014, pp. 14–21.

[23] D. Ham, J.-G. Lee, Y. Jang, and K.-E. Kim, "End-to-end neural pipeline for goal-oriented dialogue systems using gpt-2," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 583–592.

[24] Y. Jang, J. Lee, and K.-E. Kim, "Gpt-critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems," in *International Conference on Learning Representations*, 2022.

[25] S. Wang, S. Wang, Z. Liu, and Q. Zhang, "A role distinguishing bert model for medical dialogue system in sustainable smart city," *Sustainable Energy Technologies and Assessments*, vol. 55, p. 102896, 2023.

[26] B. Peng, C. Li, Z. Zhang, C. Zhu, J. Li, and J. Gao, "Raddle: An evaluation benchmark and analysis platform for robust task-oriented dialog systems," *arXiv preprint arXiv:2012.14666*, 2020.

[27] Y. Yang, Y. Li, and X. Quan, "Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 230–14 238.

[28] Y. Zhang, Z. Ou, and Z. Yu, "Task-oriented dialog systems that consider multiple appropriate responses under the same context," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9604–9611.

[29] G. Tur, D. Hakkani-Tür, and L. Heck, "What is left to be understood in atis?" in *2010 IEEE Spoken Language Technology Workshop*. IEEE, 2010, pp. 19–24.

[30] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," *arXiv preprint arXiv:1609.01454*, 2016.

[31] Y. Wang, Y. Shen, and H. Jin, "A bi-model based rnn semantic frame parsing model for intent detection and slot filling," *arXiv preprint arXiv:1812.10235*, 2018.

[32] P. Niu, Z. Chen, M. Song *et al.*, "A novel bi-directional interrelated model for joint intent detection and slot filling," *arXiv preprint arXiv:1907.00390*, 2019.

[33] D. Guo, G. Tur, W.-t. Yih, and G. Zweig, "Joint semantic utterance classification and slot filling with recursive neural networks," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 554–559.

[34] G. Kurata, B. Xiang, B. Zhou, and M. Yu, "Leveraging sentence-level information with encoder lstm for semantic slot filling," *arXiv preprint arXiv:1601.01530*, 2016.

[35] C.-S. Wu, S. Hoi, R. Socher, and C. Xiong, "Tod-bert: Pre-trained natural language understanding for task-oriented dialogue," *arXiv preprint arXiv:2004.06871*, 2020.

[36] C. Liu, S. Zhu, Z. Zhao, R. Cao, L. Chen, and K. Yu, "Jointly encoding word confusion network and dialogue context with bert for spoken language understanding," *arXiv preprint arXiv:2005.11640*, 2020.

[37] S. Lee, "Structured discriminative model for dialog state tracking," in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 442–451.

[38] K. Sun, L. Chen, S. Zhu, and K. Yu, "The sjtu system for dialog state tracking challenge 2," in *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2014, pp. 318–326.

[39] Q. Zhu, Z. Zhang, Y. Fang, X. Li, R. Takanobu, J. Li, B. Peng, J. Gao, X. Zhu, and M. Huang, "Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems," *arXiv preprint arXiv:2002.04793*, 2020.

[40] P. Budzianowski, I. Vulic, and I. Hello, "Gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems," *arXiv preprint arXiv:1907.05774*.

[41] T.-H. Wen, D. Vandyke, N. Mrksic, M. Gasic, L. M. Rojas-Barahona, P.-H. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," *arXiv preprint arXiv:1604.04562*, 2016.

[42] W. Lei, X. Jin, M.-Y. Kan, Z. Ren, X. He, and D. Yin, "Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1437–1447.

[43] T. Zhao and M. Eskenazi, "Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning," *arXiv preprint arXiv:1606.02560*, 2016.

[44] J. D. Williams, K. Asadi, and G. Zweig, "Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning," *arXiv preprint arXiv:1702.03274*, 2017.

[45] B. Dhingra, L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng, "Towards end-to-end reinforcement learning of dialogue agents for information access," *arXiv preprint arXiv:1609.00777*, 2016.

[46] B. Liu and I. Lane, "Iterative policy learning in end-to-end trainable task-oriented neural dialog models," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 482–489.

[47] R. Carpenter, "Cleverbot," *Cleverbot. com*, 2011.

[48] S. Jafarpour, C. J. Burges, and A. Ritter, "Filter, rank, and transfer the knowledge: Learning to chat," *Advances in Ranking*, vol. 10, no. 2329-9290, p. 17, 2010.

[49] A. Leuski and D. Traum, "Npceditor: Creating virtual human dialogue using information retrieval techniques," *Ai Magazine*, vol. 32, no. 2, pp. 42–56, 2011.

[50] Z. Ji, Z. Lu, and H. Li, "An information retrieval approach to short text conversation," *arXiv preprint arXiv:1408.6988*, 2014.

[51] T.-Y. Liu *et al.*, "Learning to rank for information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[52] C. Cui, W. Wang, X. Song, M. Huang, X.-S. Xu, and L. Nie, "User attention-guided multimodal dialog systems," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 445–454.

[53] Z. Lu and H. Li, "A deep architecture for matching short texts," *Advances in neural information processing systems*, vol. 26, 2013.

[54] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *arXiv preprint arXiv:1503.02364*, 2015.

[55] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," *arXiv preprint arXiv:1506.06714*, 2015.

[56] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[57] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," pp. 4171—4 186, 2018.

[58] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[59] J. Weston, E. Dinan, and A. H. Miller, "Retrieve and refine: Improved sequence generation models for dialogue," *arXiv preprint arXiv:1808.04776*, 2018.

[60] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[61] Y. Wu, F. Wei, S. Huang, Y. Wang, Z. Li, and M. Zhou, "Response generation by context-aware prototype editing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7281–7288.

[62] J. Zhang, C. Tao, Z. Xu, Q. Xie, W. Chen, and R. Yan, "Ensemblegan: Adversarial learning for retrieval-generation ensemble model on short-text conversation," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 435–444.

[63] M. Firdaus, N. Thakur, and A. Ekbal, "Aspect-aware response generation for multimodal dialogue system," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 2, pp. 1–33, 2021.

[64] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak, "Survey on evaluation methods for dialogue systems," *Artificial Intelligence Review*, vol. 54, pp. 755–810, 2021.

[65] T. Watanabe, M. Araki, and S. Doshita, "Evaluating dialogue strategies under communication errors using computer-to-computer simulation," *IEICE transactions on information and systems*, vol. 81, no. 9, pp. 1025–1033, 1998.

[66] H. Ai and F. Weng, "User simulation as testing for spoken dialog systems," in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008, pp. 164–171.

[67] J. Schatzmann, "Statistical user and error modelling for spoken dialogue systems," Ph.D. dissertation, University of Cambridge, 2008.

[68] R. Higashinaka, Y. Minami, K. Dohsaka, and T. Meguro, "Issues in predicting user satisfaction transitions in dialogues: Individual differences, evaluation criteria, and prediction models," in *Proc. International Workshop on Spoken Dialogue Systems (IWSDS)*, 2010, pp. 48–60.

[69] V. Rus and M. Lintean, "An optimal assessment of natural language student input using word-to-word similarity metrics," in *Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, 2012. Proceedings 11.* Springer, 2012, pp. 675–676.

[70] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," *arXiv preprint arXiv:1511.08198*, 2015.

[71] G. Forgues, J. Pineau, J.-M. Larchevêque, and R. Tremblay, “Bootstrapping dialog systems with word embeddings,” in *Nips, modern machine learning and natural language processing workshop*, vol. 2, 2014, p. 168.

[72] E. Levin and R. Pieraccini, “User modelling for spoken dialogue system evaluation,” in *Proc. IEEE ASR Workshop*, 1997.

[73] J. Schatzmann and S. Young, “The hidden agenda user simulation model,” *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 733–747, 2009.

[74] L. E. Asri, J. He, and K. Suleman, “A sequence-to-sequence model for user simulation in spoken dialogue systems,” *arXiv preprint arXiv:1607.00070*, 2016.

[75] F. Kreyssig, I. Casanueva, P. Budzianowski, and M. Gasic, “Neural user simulation for corpus-based policy optimisation for spoken dialogue systems,” *arXiv preprint arXiv:1805.06966*, 2018.

[76] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, “PARADISE: A framework for evaluating spoken dialogue agents,” in *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain: Association for Computational Linguistics, Jul. 1997, pp. 271–280. [Online]. Available: https://aclanthology.org/P97-1035

[77] M. Walker, C. Kamm, and D. Litman, “Towards developing general models of usability with paradise. natural language engineering,” *Special Issue on Best Practice in Spoken Dialogue systems*, 2000.

[78] K.-P. Engelbrecht, F. Gödde, F. Hartard, H. Ketabdar, and S. Möller, “Modeling user satisfaction with hidden markov models,” in *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2009, pp. 170–177.

[79] A. Schmitt and S. Ultes, “Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction,” *Speech Communication*, vol. 74, pp. 12–36, 2015.

[80] J. Henderson, O. Lemon, and K. Georgila, “Hybrid reinforcement/supervised learning for dialogue policies from communicator data,” in *IJCAI workshop on knowledge and reasoning in practical dialogue systems*, 2005.

[81] F. Jurcıcek, S. Keizer, M. Gašic, F. Mairesse, B. Thomson, K. Yu, and S. Young, “Real user evaluation of spoken dialogue systems using amazon mechanical turk,” in *Proceedings of INTERSPEECH*, vol. 11, 2011.

[82] M. Huang, X. Zhu, and J. Gao, "Challenges in building intelligent open-domain dialog systems," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 3, pp. 1–32, 2020.

[83] D. Peng, M. Zhou, C. Liu, and J. Ai, "Human–machine dialogue modelling with the fusion of word-and sentence-level emotions," *Knowledge-Based Systems*, vol. 192, p. 105319, 2020.

[84] S. Ultes, "Improving interaction quality estimation with bilstms and the impact on dialogue policy learning," in *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2019, pp. 11–20.

[85] S. Hara, N. Kitaoka, and K. Takeda, "Estimation method of user satisfaction using n-gram-based dialog history model for spoken dialog system," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, 2010.

[86] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation (Signals and Communication Technology).* Berlin, Heidelberg: Springer-Verlag, 2005.

[87] R. Higashinaka, Y. Minami, K. Dohsaka, and T. Meguro, "Modeling user satisfaction transitions in dialogues from overall ratings," in *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2010, pp. 18–27.

[88] S. Ultes, A. Schmitt, and W. Minker, "Analysis of temporal features for interaction quality estimation," in *Dialogues with Social Robots*, 2017, pp. 367–379.

[89] H. Yu, L. Gui, M. Madaio, A. Ogan, J. Cassell, and L.-P. Morency, "Temporally selective attention model for social and affective state recognition in multimedia content," in *Proc. ACM International Conference on Multimedia*, 2017, pp. 1743–1751.

[90] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 2018, 2018, p. 5642.

[91] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 32, no. 1, 2018.

[92] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017, p. 1114–1125.

[93] Y. Hirano, S. Okada, H. Nishimoto, and K. Komatani, "Multitask prediction of exchange-level annotations for multimodal dialogue systems," in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, 2019, pp. 85–94.

[94] K. Komatani and S. Okada, "Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels," in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2021.

[95] F. J. Bernieri, J. S. Gillis, J. M. Davis, and J. E. Grahe, "Dyad rapport and the accuracy of its judgment across situations: a lens model analysis." *Journal of Personality and Social Psychology*, vol. 71, no. 1, p. 110, 1996.

[96] M. Kimura, M. Yogo, and I. Daibo, "Expressivity halo effect in the conversation about emotional episodes," *Japanese Journal of Research on Emotions*, vol. 12, no. 1, pp. 12–23, 2005.

[97] S. Katada, S. Okada, Y. Hirano, and K. Komatani, "Is she truly enjoying the conversation? analysis of physiological signals toward adaptive dialogue systems," in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, 2020, pp. 315–323.

[98] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[99] Y. Kikuta, "Bert pretrained model trained on japanese wikipedia articles," 2019.

[100] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2018, pp. 59–66.

[101] J. F. Cohn, Z. Ambadar, and P. Ekman, "Observer-based measurement of facial expression with the facial action coding system," *The handbook of emotion elicitation and assessment*, vol. 1, no. 3, pp. 203–221, 2007.

[102] N. Latif, A. V. Barbosa, E. Vatiokiotis-Bateson, M. S. Castelhano, and K. Munhall, "Movement coordination during conversation," *PLoS one*, vol. 9, no. 8, p. e105036, 2014.

[103] J. N. Cappella and S. Planalp, "Talk and silence sequences in informal conversations iii: Interspeaker influence," *Human Communication Research*, vol. 7, no. 2, pp. 117–132, 1981.

[104] G. J. McHugo, J. T. Lanzetta, D. G. Sullivan, R. D. Masters, and B. G. Englis, "Emotional reactions to a political leader's expressive displays." *Journal of Personality and Social Psychology*, vol. 49, no. 6, p. 1513, 1985.

[105] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[106] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame." *Journal of personality and social psychology*, vol. 68, no. 3, p. 441, 1995.

[107] Z. Yu, Z. Xu, A. W. Black, and A. Rudnicky, "Strategy and policy learning for non-task-oriented conversational systems," in *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, 2016, pp. 404–412.

[108] A. Schmitt, B. Schatz, and W. Minker, "Modeling and predicting quality in spoken human-computer interaction," in *Proceedings of the SIGDIAL 2011 Conference*, 2011, pp. 173–184.

[109] P. K. Bodigutla, A. Tiwari, J. V. Vargas, L. Polymenakos, and S. Matsoukas, "Joint turn and dialogue level user satisfaction estimation on multi-domain conversations," *arXiv preprint arXiv:2010.02495*, 2020.

[110] K. Komatani and S. Okada, "Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021, pp. 1–8.

[111] Y. S. Abu-Mostafa, "Learning from hints in neural networks," *Journal of complexity*, vol. 6, no. 2, pp. 192–198, 1990.

[112] S. Young, M. Gašić, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1160–1179, 2013.

[113] Z. Yu, L. Nicolich-Henkin, A. W. Black, and A. Rudnicky, "A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement," in *Proceedings of the 17th annual meeting of the Special Interest Group on Discourse and Dialogue*, 2016, pp. 55–63.

[114] W. Wei, S. Li, S. Okada, and K. Komatani, "Multimodal user satisfaction recognition for non-task oriented dialogue systems," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 586–594.

[115] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young, "Agenda-based user simulation for bootstrapping a pomdp dialogue system," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 2007, pp. 149–152.

[116] P. K. Bodigutla, L. Polymenakos, and S. Matsoukas, "Multi-domain conversation quality evaluation via user satisfaction estimation," *arXiv preprint arXiv:1911.08567*, 2019.

[117] Y. Kikuta, "Bert pretrained model trained on japanese wikipedia articles," 2019.

[118] D. V. Sang, L. T. B. Cuong, and V. Van Thieu, "Multi-task learning for smile detection, emotion recognition and gender classification," in *Proceedings of the Eighth International Symposium on Information and Communication Technology*, 2017, pp. 340–347.

[119] Y. Ling, Z. Liang, T. Wang, F. Cai, and H. Chen, "Sequential or jumping: context-adaptive response generation for open-domain dialogue systems," *Applied Intelligence*, pp. 1–16, 2022.

[120] T. Young, F. Xing, V. Pandelea, J. Ni, and E. Cambria, "Fusing task-oriented and open-domain dialogues in conversational agents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 622–11 629.

[121] S. Iizuka, S. Mochizuki, A. Ohashi, S. Yamashita, A. Guo, and R. Higashinaka, "Clarifying the dialogue-level performance of gpt-3.5 and gpt-4 in task-oriented and non-task-oriented dialogue systems," in *Proceedings of the AAAI Symposium Series*, vol. 2, no. 1, 2023, pp. 182–186.

[122] L. Degnen and R. Rosenthal, "The nature of rapport and its nonverval correlates," *Psychol Inq*, vol. 1, pp. 285–293, 1990.

[123] P. Müller, M. X. Huang, and A. Bulling, "Detecting low rapport during natural interactions in small groups from non-verbal behaviour," in *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 2018, pp. 153–164.

[124] A. Abbe and S. E. Brandon, "The role of rapport in investigative interviewing: A review," *Journal of investigative psychology and offender profiling*, vol. 10, no. 3, pp. 237–249, 2013.

[125] T. Hayashi, C. O. Mawalim, R. Ishii, A. Morikawa, A. Fukayama, T. Nakamura, and S. Okada, "A ranking model for evaluation of conversation partners based on rapport levels," *IEEE Access*, vol. 11, pp. 73 024–73 035, 2023.

[126] A. Cerekovic, O. Aran, and D. Gatica-Perez, "Rapport with virtual agents: What do human social cues and personality explain?" *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 382–395, 2017.

[127] M. N. Shiota, D. Keltner, and O. P. John, "Positive emotion dispositions differentially associated with Big Five personality and attachment style," *The journal of positive psychology*, vol. 1, no. 2, pp. 61–71, 2006.

[128] J. A. Penley and J. Tomaka, "Associations among the big five, emotional responses, and coping with acute stress," *Personality and individual differences*, vol. 32, no. 7, pp. 1215–1228, 2002.

[129] T. E. Kim and A. Lipani, "A multi-task based neural model to simulate users in goal oriented dialogue systems," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2115–2119.

[130] J. Mendonça, I. Trancoso, and A. Lavie, "Soda-eval: Open-domain dialogue evaluation in the age of llms," *arXiv preprint arXiv:2408.10902*, 2024.

[131] ——, "Ecoh: Turn-level coherence evaluation for multilingual dialogues," *arXiv preprint arXiv:2407.11660*, 2024.

[132] Z. Meng, J. Li, and Y. Gong, "Adversarial speaker adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5721–5725.

[133] Y. Gao, S. Okada, L. Wang, J. Liu, and J. Dang, "Domain-invariant feature learning for cross corpus speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6427–6431.

[134] K. Komatani, R. Takeda, and S. Okada, "Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus," in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2023, pp. 104–113.

[135] W. Wei, S. Li, and S. Okada, "Investigating the relationship between dialogue and exchange-level impression," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 359–367.

[136] S. Katada, S. Okada, and K. Komatani, "Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 349–358.

[137] A. Lee, K. Oura, and K. Tokuda, "Mmdagent—a fully open-source toolkit for voice interaction systems," in *2013 IEEE international conference on acoustics, speech and signal processing.* IEEE, 2013, pp. 8382–8385.

[138] A. Oshio, A. Shingo, and P. Cutrone, "Development, reliability, and validity of the japanese version of ten item personality inventory (tipi-j)." *Japanese Journal of Personality/Pasonariti Kenkyu*, vol. 21, no. 1, 2012.

[139] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Twelfth annual conference of the international speech communication association*, 2011.

Publications

Main works peer review

[1] Wenqing Wei, Sixia Li, Candy Olivia Mawalim, Xiguang Li, Kazunori Komatani, and Shogo Okada, ”User Traits Adaptation for Dialogue-Level User Rapport Recognition Using Online Data.” MDPI MTI 2025

[2] Wenqing Wei, Sixia Li and Shogo Okada. Investigating the relationship between dialogue and exchange-level impression. ICMI 2022

[3] Wenqing Wei, Sixia Li, Shogo Okada, Kazunori Komatani. Multimodal User Satisfaction Recognition for Non-task Oriented Dialogue Systems. 23th ACM International Conference on Multimodal Interaction (ICMI) 2021

Other works peer review

[1] Wenqing Wei, Yuan Gao, Zhengdong Yang2, Jiyi Li, Chenhui Chu, Shogo Okada, Sheng Li . FedCPC: an Effective Federated Contrastive Learning Method for Privacy Preserving Early-Stage Alzheimer's Speech Detection. ASRU 2023

[2] Longfei Yang, Wenqing Wei, Sheng Li, Jiyi Li and Takahiro Shinozaki. Augmented Adversarial Self-Supervised Learning for Early-Stage Alzheimer's Speech Detection. Interspeech 2022

Non-peer review

[1] Yao LiuWenqing WeiSixia LiJianwu Dang. A Study of Punctuation Prediction by Considering Syntactic Information. 2022

[2] Wenqing Wei, Rui Wang, Sheng Li, Yachao Guo and Hao Huang, “System Description of Alzheimer's Disease Recognition Evaluation 2021 (NICT-Xinjiang Univ. team, short speech track)” NCMMSC 2021