# Engagement Recognition based on Multimodal Behaviors for Human-Robot Dialogue


**Koji Inoue**

Graduate School of Informatics
Kyoto University

# Abstract

It is becoming a reality for humans to interact with robots in daily life. Conversational robots are expected to be involved in human society in a symbiotic manner by playing a specific social role. The goal of this study is to realize natural and smooth dialogue between humans and robots. To this end, the robots need to understand a variety of information such as the interaction state during the dialogue. This understanding requires making use of not only linguistic information such as utterance content but also non-linguistic information that can be captured from multimodal sensors such as backchannels and eye gaze. Understanding the states leads to the generation of adaptive behaviors of the robots, which increases the quality of the user experience through the dialogue.

As one of the interaction states in human-robot dialogue, this thesis focuses on engagement that represents how much a user is interested in and willing to continue the current dialogue. Engagement is also related to other concepts such as involvement and rapport. Therefore, for conversational robots, engagement provides a cue to adaptively manage the dialogue with users. This thesis addresses engagement recognition based on multimodal listener behaviors such as backchannels, laughing, head nodding, and eye gaze. In this study, the problem of subjectivity on the perception of engagement is tackled. There is often disagreement on engagement labels among annotators due to the subjectivity of the annotation task. Novel engagement recognition models are introduced by taking into account the difference among the annotators.

Chapter 1 describes the background of this thesis in the context of spoken dialogue systems and human-robot interaction in order to clarify problems and approaches. Chapter 2 briefly reviews related works on engagement including its definitions, recognition methods, and adaptive behavior generation.

In Chapter 3, a hierarchical Bayesian model for engagement recognition is proposed. It is assumed that each annotator has a character that affects his/her perception of engagement. This character is modeled as latent variables. The proposed method estimates parameters for the engagement label and the character of each annotator from the training data. Thus, it can recognize each annotator's label precisely by considering the corresponding character.

Experimental results show that the proposed method achieved a higher recognition accuracy than other methods which do not consider the character such as majority voting.

Chapter 4 addresses automatic detection of the listener behaviors to realize real-time engagement recognition for live spoken dialogue systems. Backchannels and laughing are detected from the audio signal using bi-directional long short-term memory (LSTM) with connectionist temporal classification (BLSTM-CTC). Head nodding and eye gaze are detected from the head direction captured by the Kinect v2 sensor using LSTM and a heuristic rule, respectively. The outputs of the detection models are integrated as the input to the engagement recognition model in a statistical framework. Online engagement recognition is achieved without degrading accuracy.

Chapter 5 addresses an engagement recognition model using neural networks. The proposed neural network consists of two layers to consider both different annotators' labels and an integrated label. The first layer corresponds to each annotator's model that is trained with the corresponding labels independently. The second layer aggregates the different annotators' models to obtain one integrated label. After each layer is pre-trained, the whole network is fine-tuned through back-propagation of prediction errors. Experimental results show that the proposed network outperforms baseline models that directly recognize the integrated label without considering different annotations.

Chapter 6 addresses application of the engagement recognition to the android robot ERICA. ERICA has a physical mechanism to generate various behaviors such as facial expression and eye gaze, which leads to natural dialogue. ERICA is also designed to interact with users autonomously by using multimodal sensors. ERICA is given specific social roles such as attentive listening, job interview, speed dating, and laboratory guide. In the role of the laboratory guide, ERICA generates adaptive feedbacks toward the user according to the user engagement.

In Chapter 7, toward multi-party conversations, multimodal speaker diarization using both acoustic and eye-gaze information is proposed. Eye-gaze information plays an important role in turn-taking. Thus, it is also useful for predicting speech activity. The proposed method integrates a variety of eye-gaze features with acoustic information stochastically. Experiments in poster conversations show that eye-gaze information contributes to the improvement of diarization accuracy under noisy environments, and its weight is automatically determined according to the level of surrounding noise.

Finally, this thesis is concluded with future directions in Chapter 8.

# Acknowledgements

This work was accomplished at Kawahara Laboratory, Graduate School of Informatics, Kyoto University. I express my gratitude to all people who helped me in this work.

At first, I would like to express my special thanks and appreciation to my supervisor Prof. Tatsuya Kawahara. His comments were essential and insightful for advancing this work. This work would not have been completed without his continuing support.

I also express my special thanks and appreciation to the members of my dissertation committee, Prof. Toyoaki Nishida and Prof. Takayuki Kanda for their time and valuable comments and suggestions.

This thesis cannot be accomplished without Dr. Katsuya Takanashi, Dr. Shizuka Nakamura, and Dr. Divesh Lala in the spoken dialogue system group of Kawahara Lab. They supported me and gave much time to meaningful discussions. Furthermore, I appreciate their cooperation to build a spoken dialogue system for the autonomous android robot ERICA. I also deeply thank the other members of the spoken dialogue system group in Kawahara Lab. I appreciate discussions with Mr. Takashi Yamaguchi, Mr. Ryosuke Nakanishi, Mr. Masanari Ishida, Mr. Kenta Yamamoto, Mr. Kohei Hara, and Mr. Koki Tanaka.

I also express my thanks to members in the ERATO Ishiguro symbiotic human-robot interaction project. Prof. Hiroshi Ishiguro from Osaka University gave me insightful comments in monthly meetings. Dr. Takashi Minato and Dr. Kurima Sakai from ATR continuously supported me to implement the spoken dialogue system for the android robot ERICA.

I also deeply thankful to Dr. Milhorat Pierrick, Dr. Yukoh Wakabayashi, and Dr. Hiromasa Yoshimoto for their kind supports.

I would like to thank both current and past members in Kawahara Lab. I am grateful to comments and supports from Prof. Shinsuke Mori, Dr. Yuya Akita, Dr. Kazuyoshi Yoshii, Dr. Katsutoshi Itoyama, Mr. Masato Mimura, Dr. Yoshiaki Bando, and the other members.

This work was supported by the Japan Society for the Promotion and Science (JSPS) with their financial support as a Fellowship for Young Scientists (DC1).

Lastly, I am truly grateful to my family for their daily and continuous support.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

It is becoming a reality for humans to interact with robots in daily lives. Towards the symbiotic society between humans and robots, this thesis addresses engagement recognition for human-robot dialogue.

## 1.1 Background

Many conversational robots have been developed by researchers and companies, especially over the past decade. It is no longer unusual that robots are being deployed to have conversations with humans in real social occasions. The robots are designed to play social roles in our society by creating a social bond between the robots and humans. In the future, it is expected that the conversational robots will be more involved in human society in a symbiotic manner. To this end, the communication between the robots and humans is expected to be more smooth and natural.

In the symbiotic society of the robots and humans, spoken dialogue is an important interface between them. Human language technologies, such as automatic speech recognition (ASR) and natural language understanding (NLU), have been drastically advanced by machine learning techniques such as deep learning [1–9]. It has allowed researchers to develop various application systems that communicate with people by utilizing the spoken language. Besides conversational robots, other typical applications are smartphone applications and smart speakers. These systems provide not only web search but also many tasks by coordinating with third-party applications. Furthermore, the systems can also handle simple chatting to make the users more engaged in the interaction with the systems.

However, interaction with the systems, including conversational robots, is machine-specific and is based on the simple exchange that consists of one question and one answer. Users adapt their utterances and behaviors to make them more recognizable and understand-

able for the systems. For example, the users speak simple commands slowly and politely. For turn-taking, the systems assume that the users say a keyword to start the interaction, and the systems then use explicit signals such as lighting in order to express that the systems understand the users' initiation. This is much different from human-human conversations which are more natural and smooth. It is important for the systems to be more user-friendly and to behave like human beings to be more integrated into our daily lives. In this thesis, the ultimate goal is to realize the natural interaction between humans and robots, like that of human-human conversation.

## 1.2 Spoken Dialogue Technology

Spoken dialogue systems have been investigated for several decades. The origin of works on dialogue systems can be traced to ELIZA [10] and SHRDLU [11]. ELIZA played the role of a psychotherapist by having text chatting implemented by simple database retrieval and keyword rephrasing. SHRDLU handled the text command by natural language understanding based on the state of the blocks, which means that it made use of context of the dialogue. Following studies have proposed and implemented various functions for dialogue systems, such as *frame* [12] and *planning* [13–15] for flexible handling of the user input and the context of the dialogue. Several concepts were proposed such as *theory of grounding* for handling the error of automatic speech recognition [16–18]. Several architectures were also investigated such as *blackboard* [19] and *hub-spoke structure* [20, 21] to manage the communication between modules including automatic speech recognition and the dialogue manager. Afterwards, many practical systems have been developed. These systems were designed for task-oriented dialogues such as bus navigation and flight booking. For the last decade, statistical approaches were introduced such as partially observed Markov decision process (POMDP) [22] and RNNs [23]. Non-task-oriented dialogues like casual chatting has also been investigated [24–26]. Recently, attention has turned to text generation based on sequence-to-sequence learning with a large amount of conversational data extracted from such as social networking services [27–29].

Spoken dialogue systems have also been studied from the perspective of multimodal. Making use of multimodal features is helpful for a more precise understanding of the users. Using gesture and body-pose features improved the accuracy of utterance intention understanding such as dialogue act classification [30]. Emotion recognition has also been enhanced by using multimodal features [31–35]. It has also been found that eye-gaze and body movement features were informative for prediction of turn-taking behaviors [36–39]. On the other hand, how to generate multimodal behaviors of systems has also been

studied [40, 41]. The generation of multimodal behaviors is also important to interact with users smoothly and cooperatively. Toward natural human-robot interaction in social occasions, this thesis focuses on making use of multimodal behaviors for better user understanding.

## 1.3 Human-Robot Dialogue

Conversational robots have been widely studied and developed. Although conventional robots communicate with humans through simple commands in specific protocols, it is essential for conversational robots to interact with humans in natural language. Therefore, spoken dialogue is the inevitable interface for the communication robots. On the other hand, for spoken dialogue systems, the advantage of making use of robots is to be able to utilize multimodality. The robots can coordinately generate their multimodal behaviors in accordance with user behaviors. Furthermore, it is possible to refer to and operate on objects in the real world by such as symbol grounding [42–44]. Another notable thing is that the co-existence of robots in the real world makes human dialogue partners more engaged in the dialogue. Therefore, the human-robot dialogue has a potential to realize natural communication that is close to human-human dialogue.

The studies of human-robot dialogue were inspired by the development of intelligent robots. The origin of intelligent robots was Shakey [45], which was equipped with various sensors for recognition of the surrounding environment by moving itself. Afterwards, researchers started to investigate humanoid robots that imitated the shapes and features of humans. Kismet [46, 47] had a mechanical face that imitated the human face in order to generate facial expression and eye-gaze behaviors. COG [48] was a humanoid robot that was able to operate its own arms, neck, and torso. SIG [49] was designed to integrate audio and visual information for active audio-scene understanding that consists of tasks such as sound localization and sound source separation. Robovie [50] was directed for communication with humans in real environments by utilizing various sensors such as a microphone, color camera, omnidirectional camera, touch sensor, and an ultrasonic distance sensor. As the intermediate system between a robot and a virtual agent, Furhat [51, 52] has a virtual face projected on the real object shaped of a head in order to generate a variety of facial expressions and eye-gaze behaviors. In the context of the above intelligent robots, an android robot, name ERICA, that has an appearance similar to human beings has been developed. Fig 1.1 shows the snapshot of the android robot ERICA. ERICA is designed as a 23 year old woman. ERICA has 46 active joints inside its body to generate various behaviors such as facial expressions, eye gaze, head nodding, and arm gestures. For understanding the surrounding environment and the conversational behaviors including users' utterances, the system of ERICA integrates

Fig. 1.1 Android robot ERICA

information from sensors such as microphone arrays and cameras that are deployed around the robot. Therefore, conversation with ERICA is aimed to be more natural through various expressions and behaviors and a deep understanding of the conversation and user state.

Conversational robots have been studied in specific dialogue situations. A role of a museum guide was given to a robot to be operated in the long-term period [53]. A conversational robot was designed for a navigation guide in a real environment such as a shopping mall [54]. Other robots that attended to multi-party conversations were developed [55–59]. Small humanoid robots were implemented to handle open-domain conversation consisting of a wide range of topics [60–62]. The Furhat robot was applied to a map instruction task [63] and quiz games [64, 65]. Interactive agents were implemented as a counselor [66], a museum guide [67], and a job interviewer [68–70]. It is expected that conversational robots are deployed to a diversity of situations.

This thesis addresses the human-robot dialogue with the android robot ERICA. The goal of this study is to realize natural human-robot dialogue with ERICA like human-human dialogue. However, it is challenging to implement it from scratch without any preconditions. Therefore, specific dialogue situations are assumed for ERICA. In this study, dialogue tasks for ERICA are considered by taking into account the balance between listening and speaking roles. This thesis addresses several social roles: attentive listening, speed dating, job interview, and laboratory guide.

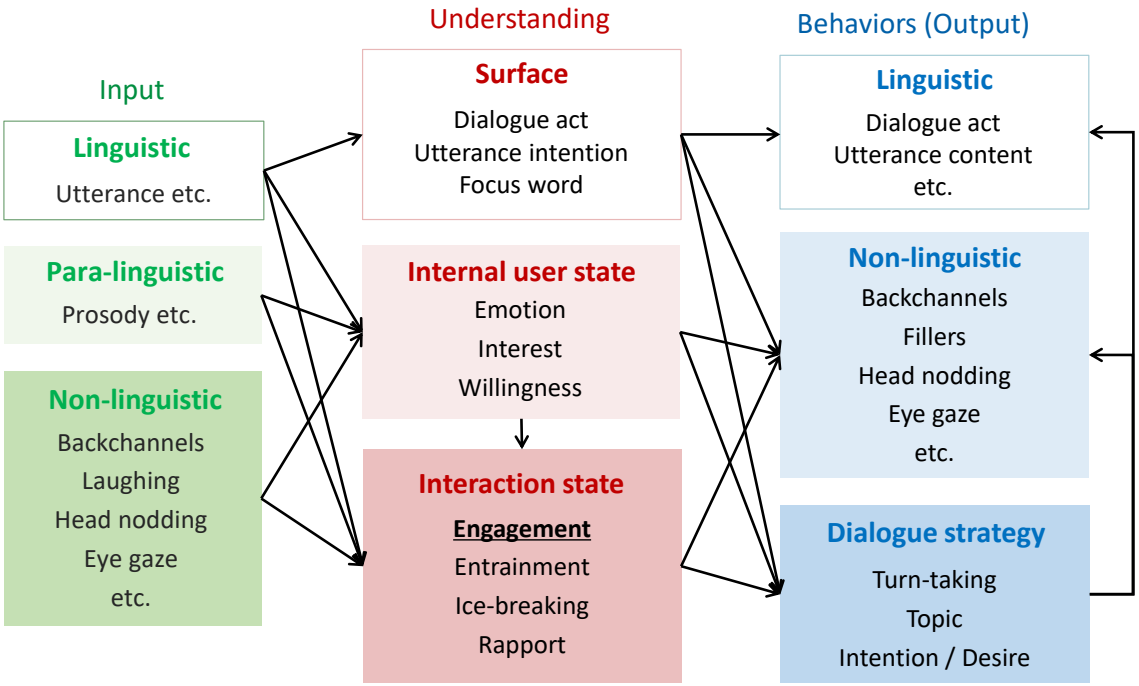


Fig. 1.2 Information which conversational robots should understand during the dialogue

## 1.4 Dialogue Understanding from Multimodal Behaviors

It is important for conversational robots to realize natural dialogue with humans that is close to those of human-human dialogue. For natural dialogue which is deep and long, the robots need to correctly recognize and understand a variety of information during the dialogue. Furthermore, the robots are expected to generate adaptive behaviors according to the understood information. Fig. 1.2 summarizes the information which the robots should understand during the dialogue. Conventional dialogue systems take linguistic information such as the ASR result. Then, the systems understand the input on the surface level such as the dialogue act and the utterance intention. Finally, the systems generate robot behaviors such as the utterance content based on the surface understanding. Moreover, the ideal conversational robots should use a variety of input. The input information includes not only linguistic features but also para-linguistic information such as prosody of user utterances. The input also includes non-linguistic information such as backchannels, laughing, head nodding, eye gaze, and so on. Based on the variety of input, the robots should understand more states of the dialogue. The understood states can be classified into two types: internal user states and interaction states.

The internal user states include emotion [71–73], interest [74, 75], willingness [76], and personality [77, 78]. Although it is important for the conversational robots to recognize these

internal states, it is difficult to know the real states of them during the dialogue. As the first step to realize social interaction, instead of the internal user states, it is also important for the robots to recognize the interaction states that describe the relationship between the robots and users. They are affected by not only the internal user states mentioned above but also social norms and dialogue situations. This thesis focuses on recognition of the interaction states. The interaction states include such phenomena as engagement, entrainment, ice-breaking, and rapport. Engagement represents the process by which dialogue participants establish, maintain, and end their perceived connection to one another [79]. Entrainment is a phenomenon where a dialogue participant adapts his/her behaviors to those of the dialogue partner [80–83]. Ice-breaking is a phenomenon in which dialogue participants ease the tension and become more friendly where they meet for the first time [84, 85]. Rapport is a close and harmonious relationship between participants and is more considered in long-term interactions [86–89].

The robots are expected to adaptively generate a variety of behaviors according to the understood states. The robot behaviors should contain not only the utterance content but also non-linguistic behaviors such as backchannels, fillers, laughing, head nodding, and eye gaze. Behind these explicit behaviors, the dialogue strategy should also be controlled according to the understood states. The strategy manages turn-taking, the dialogue topic, and the intention and desire of the robots. These adaptive behaviors are needed for realizing the natural and smooth interaction.

## 1.5 Challenges and Approaches

This thesis addresses engagement as one of the interaction states. Engagement has been massively studied in the research field of human-robot interaction. Engagement provides an indication of how much a user wants to start (or continue) the interaction with a robot. In the case of spoken dialogue systems, it is assumed that users are almost engaged in the dialogue because the interaction is limited to simple and short. Therefore, it is not needed to recognize engagement of the users. However, to deal with deep and long interaction, it would be important for spoken dialogue systems to recognize engagement during the dialogue.

It is expected that conversational robots can dynamically adapt their behaviors according to engagement, and increases the quality of the user experience through the dialogue. For example, some studies have made made to control turn-taking behaviors [90], dialogue topics [91, 92], and utterance style [93] based on automatically recognized engagement. Furthermore, engagement can be utilized as an evaluation metric for dialogue. The dialogue task is sometimes not evident in natural human-human dialogue so it is difficult to evaluate the

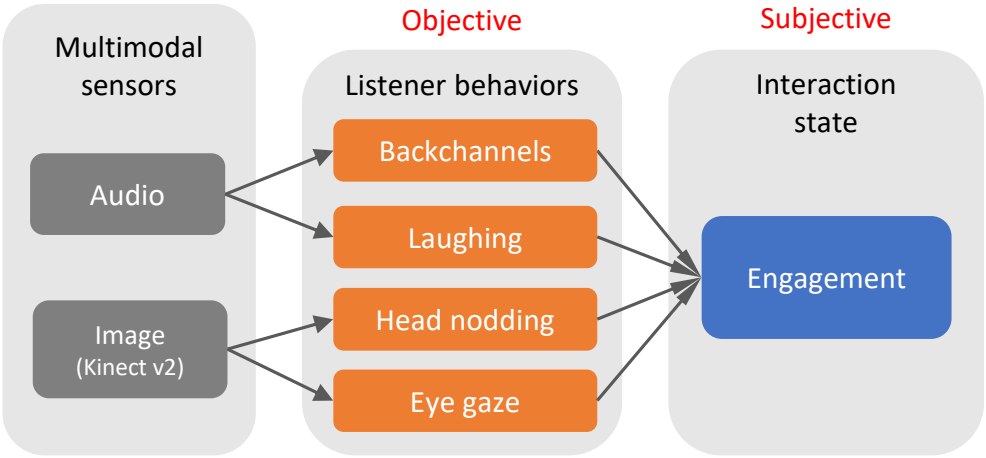


Fig. 1.3 Scheme of engagement recognition

dialogue. In this case, the evaluation of dialogue tends to be subjective. For example, studies on non-task oriented dialogue such as casual chatting used evaluation metrics including the length of dialogue, linguistic appropriateness, and human judgment. However, there is still no clear metric to evaluate this kind of dialogue. The measurement of engagement can be used for the evaluation of this kind of dialogue, thus dialogue models can be trained according to the maximization of engagement.

This thesis addresses engagement recognition based on multimodal listener behaviors such as backchannels and eye gaze. Engagement is an interaction state that depends on both dialogue participants including a speaker and a listener. Therefore, methods of engagement recognition should focus on not only speaker behaviors but also listener behaviors that are responses towards the speaker. Although the speaker behaviors would be informative for engagement recognition, these are affected by the error of automatic speech recognition, and the linguistic features depend on the dialogue domain. On the other hand, the listener behaviors are expected to be robust against linguistic errors. Fig. 1.3 depicts the scheme of engagement recognition in this study. At first, listener behaviors are detected from signals of multimodal sensors. The behaviors consist of backchannels, laughing, head nodding, and eye gaze. Detection of backchannels and laughing is based on audio signals, and head nodding and eye gaze are detected from image signals captured by the Kinect v2 sensor. Since a training data set is needed for machine learning, the annotation work is conducted to obtain reference labels by recruiting manual workers (called annotators). The listener behaviors can be defined objectively so that the consistency in the annotation work is kept and there are few disagreements among the annotators. Accordingly, it is not challenging to collect a large amount of training data. Machine learning models such as deep learning methods can be applied to this problem in order to realize enough accuracy [94]. Afterward, the

level of engagement is recognized according to the observations of the listener behaviors. However, the perception of engagement is subjective and tends to depend on each annotator. It is difficult to keep consistency between annotators. It is possible to give a description of the sample cases indicating the level of engagement to annotators, and it can also be restricted which kinds of behaviors or modalities the annotators focus on during the annotation work. Nevertheless, low agreement scores among annotators are still observed. This poses a difficulty of annotation of engagement. Consequently, it is also difficult to collect a large amount of training data.

This thesis tackles the issue caused by the subjectivity of engagement to realize a robust recognition model. In order to obtain the engagement labels, one possible way is to ask dialogue participants to evaluate their own engagement right after the dialogue. However, there is often evaluators' bias where they tend to evaluate themselves positively [95]. Therefore, other studies have asked third-party annotators to give labels of engagement by watching the dialogue video. As mentioned, a disagreement on engagement labels among annotators is often observed. To deal with this issue, other studies integrated the different engagement labels into one label. The most widely used approach was majority voting [96]. Another approach was to use the ratio of the number of annotators who gave the corresponding label [97]. For example, if three of five annotators gave the positive label (e.g. engaged), a regression model was trained to predict a value of 0.6 (=3/5). This study considers a different approach that focuses on the difference of annotators. A scheme of engagement recognition that models the difference of annotators is proposed. Here, it is hypothesized that the difference among annotators arises from the difference of the character (e.g., personality) of annotators. Each annotator has a character that affects his/her perception of engagement. The character represents not only the difference but also common points among annotators. The proposed model estimates the distribution of the character of each annotator. Based on each annotator's character, the engagement label of each annotator is recognized. In other words, different annotators' perceptions can be simulated. Since the character represents the relationship between annotators, a small amount of training data can be effectively used. An engagement recognition model is proposed to introduce the character as latent variables. The proposed model estimates the distribution of the engagement label by considering the corresponding characters. The proposed approach makes a contribution to studies on recognition tasks containing subjectivity such as emotion recognition in that the proposed model takes into account the difference of annotators. Furthermore, this thesis argue automatic detection models of the input listener behaviors. Machine learning approaches have shown to be accurate enough for these detection tasks. A deep learning method is utilized to conduct the model training without exact time alignment in the training data. Finally, online engagement

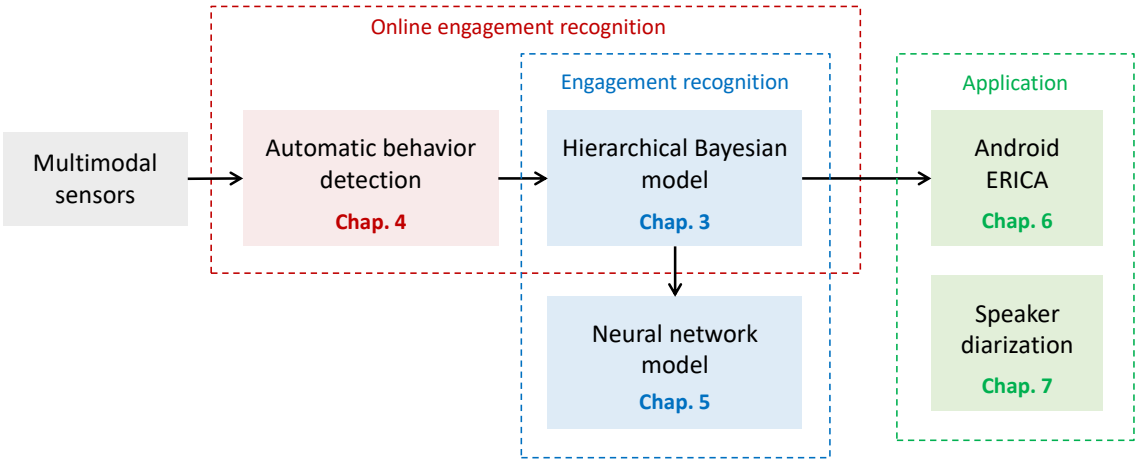


Fig. 1.4 Organization of this thesis

recognition is realized by integrating the automatic detection models of listener behaviors with the engagement recognition model. It is important for live spoken dialogue systems to implement a real-time engagement system. There is also an issue of how to generate actions and behaviors of robots based on the result of engagement recognition. In this thesis, as one of the applications for the android robot ERICA, feedback responses are generated according to the result of engagement recognition. These feedback responses are aimed at increasing user engagement during the dialogue.

## 1.6 Organization of this Thesis

The rest of this thesis is organized as follows. At first, Chapter 2 overviews related works. The definitions of engagement, engagement recognition models in previous studies, and some attempts on how to generate adaptive behaviors of systems according to engagement are introduced. Then, engagement recognition models for human-robot dialogue are proposed in the following chapters. Fig. 1.4 summarizes the organization of this thesis. In Chapter 3, a hierarchical Bayesian model is proposed to estimate the character as latent variables. This chapter also describes the human-robot dialogue corpus used in this thesis and how to annotate engagement. Chapter 4 addresses automatic detection models of the listener behaviors. Then, the detection models are integrated with the engagement recognition model in order to realize online engagement recognition fro live spoken dialogue systems. Chapter 5 addresses another engagement recognition model with a neural network. A two-step engagement recognition with a neural network that consists of two layers is proposed. In the first step, each annotator's recognition is modeled as a unit in the intermediate layer. In the second step, the different annotators' models are aggregated to recognize the integrated label in the

output layer. This chapter also addresses an efficient method that trains both each annotator's label and the integrated label by using the advantage of neural networks. In Chapter 6, as an application of engagement recognition, the real-time engagement recognition model is applied to a spoken dialogue system of the android robot ERICA. Several social roles given to ERICA are explained. As one of the social roles, a spoken dialogue system is implemented for a laboratory guide that utilizes engagement recognition. Chapter 7 addresses speaker diarization toward multi-party party conversations. A multimodal speaker diarization method is proposed by integrating acoustic and eye-gaze information. Finally, Chapter 8 concludes with future directions of human-robot dialogue with engagement recognition.

# Chapter 2

# Literature Review

This chapter overviews related works on engagement recognition in dialogue. At first, several definitions of engagement are presented. Next, previous studies on engagement recognition are reviewed. Finally, conventional systems to generate system behaviors according to user engagement are introduced.

## 2.1 Definition of Engagement

The origin of engagement studies was a sociology study by Erving Goffman [98]. He defined engagement through an observation and an analysis on real communication as

> Joining each other openly in maintaining a single focus of cognitive and visual attention

This definition is still used in some recent studies on human-robot interaction [99, 100]. Then, many studies have been made and there exists a variety of definitions of engagement [91]. Engagement was defined in the context of the communication process such as

> The process by which two (or more) participants establish, maintain, and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in interaction, evaluating staying involved, and deciding when to end connection. [79]

This definition was frequently used [97, 101, 102]. Another definition on the communication process was

> The process subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume or terminate an interaction [103]

In human-human conversations, a definition was made by focusing on each dialogue participant as

> The value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction [104].

In studies on human-agent interaction or human-robot interaction, many definitions were made. The definitions focused on the user state during the current dialogue so that those were called *user engagement*. A frequently used definition was

> User engagement describes how much a participant is interested in and attentive to a conversation [105]

Another definition was based on a description of user experience as

> Quality of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control [106]

Engagement can also be defined by referring to other concepts. The most related concept is *interest* during the dialogue. Engagement was defined in the relationship of interest as

> User's interest to continue the conversation [92]

Note that the static interest that corresponds to the fixed profile of each participant is distinguished from engagement. For example, even if a user is not interested in the current topic, there is a possibility that the user is engaged in the dialogue. *Involvement* is a concept that represents the process when a participant attends to the conversation and is closely related to engagement [107]. A definition of engagement referring to involvement was made as

> The degree of involvement a user chooses to have with a system over time [108]

User's *willingness* is also a proxy of engagement. A definition of engagement was made by using willingness for the situation on turn-taking as

> Engagement intention refers to a participant's willingness to start/end a conversation or to take over/release the conversational floor [90]

Empathy is also related to engagement and an extended definition was mentioned as

> Empathic engagement is the fostering of emotional involvement intending to create a coherent cognitive and emotional experience which results in empathic relations between a user and a synthetic character [109]

*Attention* to the robot leads to engagement when the user intends to start the dialogue. In this sense, attention is a vital indicator of engagement [110]. *Rapport* is a description of the relationship between a user and a system from the perspective of longer-term relationship. A study on virtual agents confirmed that an existing of rapport between an agent and a user is positively correlated to engagement [111].

Summary, in human-robot dialogue, the definitions of engagement can be classified into two types as follows. The first type focuses on cues to start and end the dialogue. For example, one of the tasks was to detect the potentially engaged person who wants to start the conversation with a situated robot [112, 113]. If the robot correctly recognizes this kind of engagement for each surrounding person, the robot can start the dialogue with the engaged person. This leads to the smooth communication and avoid the false action of the robot where the robot is ignored by the not-engaged person. The second type of the definitions is about the quality of the interaction between the robot and a user during the dialogue. For example, engagement can be interpreted as an indicator on how much the user is being interested in the current dialogue. If the robot correctly recognizes this kind of engagement for the current interlocutor, the robot can adaptively control the dialogue strategy in accordance with engagement. Both types of definitions are important for human-robot dialogue to accomplish a high quality of dialogue. This thesis focuses on the latter type of definitions, so that the recognition model understands how the user is being engaged during the dialogue. Furthermore, engagement can also be interpreted as the degree of agreement that the dialogue participants start and continue the dialogue. This interpretation means that engagement represents the superficial connection between the robots and users, but it leads to realizing the social interaction between them. Besides, engagement is affected by internal user states such as emotion and interest, and it is also influenced by social norms and dialogue situations.

## 2.2 Engagement Recognition

A considerable number of studies have been made on engagement recognition over the last decade. Engagement recognition has been generally formulated as a binary classification problem: engaged or not (dis-engaged). A category classification problem was also proposed [114]. For example, in the study on engagement in remote conversation, they defined the category as a set of *No interest*, *Following*, *Responding*, *Conversing (without active discourse management)*, *Influencing*, and *Governing/managing*.

The useful features for engagement recognition have been investigated and identified from a multimodal perspective. Non-linguistic behaviors are commonly used as the clue to

Table 2.1 Previous studies on the investigation of effective features for engagement recognition

| feature | | references |
|---|---|---|
| head | eye gaze | [115, 97, 90, 116–118, 114, 96] |
| | facial expression | [115, 117] |
| | head direction | [117] |
| | head nodding | [79, 96, 119] |
| conversational | voice activity | [90] |
| | backchannels | [96] |
| | vocal laughing | [120] |
| | adjacency pair | [116] |
| | turn length | [90] |
| | overlap timing | [121] |
| body | posture | [122, 119] |
| spatial | trajectory | [123] |
| | distance | [90] |
| | region space | [124, 113] |
| low-level signal | acoustic signal | [105, 125, 76, 126] |
| | image signal | [126] |

recognize engagement because verbal information like linguistic features is specific to the dialogue domain and content, and speech recognition is error-prone. Table 2.1 summarizes the feature set investigated in the previous studies.

A series of studies on human-agent/robot interaction has investigated the feature set for engagement recognition. The most widely used feature was eye-gaze behaviors [115, 97, 90, 116–118]. For example, it is informative for engagement recognition if a user is looking at the agent/robot and when both the user and the agent/robot are looking each other, called mutual gaze. Facial expression (facial action units) is also useful for engagement recognition [115, 117]. For example, positive smiling is an indicator that a user is engaged in the current dialogue. Head direction is investigated as an effective feature [117]. Head direction is often used as a proxy of eye-gaze direction. In order to track eye gaze, eye-tracking systems requires the calibration process that takes much time on the occasion of interaction. It is easier to track head direction than eye gaze. For example, making use of depth cameras such as the Kinect v2 sensor does not require any calibration process. This proxy is important for live conversational robots interacting with many people one after another. In human-robot interaction, since a user and a robot interact in a real space, physical information is also useful. For example, some studies used trajectory of the user, distance between the user and the robot, and region spaces where the user is standing [124, 123, 90,

113]. Conversational behaviors were also investigated. For example, voice activity and the turn length are speaker behaviors and reflective of engagement [90]. Head nodding was also examined as feedback towards the robot [79]. Besides, interaction behaviors were also investigated including backchannels and an adjacency pair that is a set of utterances such as a question and an answer [116]. Furthermore, the relationship between engagement and timing of overlapping (interrupting) was examined [121]. The study also confirmed that some interaction states including engagement and involvement were affected by the kind of system utterance (cooperative or not). Other used features include vocal laughing [120] and posture [122]. Empirical studies implemented engagement recognition modules utilizing the above multimodal features. In a human-robot interaction, a guidance robot recognized user engagement by measuring the above-mentioned multimodal features [90]. In a chatting bot system, user engagement was recognized by detecting the facial action units and head direction [117] with the Open FACE software [127].

Next, in human-human interaction, the effective features in dyadic conversations were investigated. In this case, researchers focused on low-level signals such as acoustic and image signals [105, 125, 76, 126]. Acoustic features include log voice power, fundamental frequency (F0), and mel-frequency cepstrum coefficients (MFCC). Image features include RGB raw data and local binary pattern [128], and so on. Furthermore, the study was extended to multi-party conversations, showing that eye-gaze behaviors were also informative in multi-party human-human dialogue[114, 96]. Verbal and visual backchannels (head nodding) were also examined in an experiment on engagement recognition [96, 119]. Posture of each participant was also investigated [119]. From the above, engagement is related to various non-linguistic behaviors Since the non-linguistic behaviors are independent of dialogue contents, these are expected to be robust for engagement recognition. Therefore, this thesis focuses on non-linguistic listener behaviors such as backchannels, laughing, head nodding and eye gaze.

The recognition models using the above multimodal features were initially based on heuristic rules. That uses the pattern of eye gaze direction, head pose direction, and space region [129, 107, 124]. These heuristic rules were robust, but it would take time to create additional rules for more complicated behaviors. Afterward, studies on data collection and annotation of engagement were made. This approach allowed us to train machine learning models with the annotation data. The most frequently used method was support vector machines (SVM) [90, 96, 119, 117]. Hidden Markov model (HMM) was used in handling time-wise signals such as acoustic features [105]. Recently, convolutional neural networks (CNN) was applied to engagement recognition [126]. Since the machine learning approach makes it possible to automatically learn the complicated relationship between behaviors and

engagement, this thesis also takes this approach. Concretely, a hierarchical Bayesian model is used to utilize latent variables for the character modeling.

How to annotate engagement is also important. Since the perception of engagement is subjective, it is difficult to define the annotation criteria objectively. Furthermore, there exists an evaluator bias so that it is also difficult to ask dialogue participants to evaluate self-engagement. In general, previous studies asked third-party annotators to evaluate user engagement by watching dialogue videos. In this case, it is important for the annotation work to keep consistency among annotators. Therefore, some studies trained a few annotators to keep the consistency [114, 90, 126, 117]. However, in the case of annotating natural dialogue, the annotation would become more complicated and challenging to be consistent among annotators. Another approach is based on "wisdom of crowds" where many annotators were recruited [97, 90, 96]. These studies integrated multiple labels given by many annotators into one label by such as majority voting. This thesis takes the latter approach to collect many labels without time-consuming training of annotators. In contrast to previous studies, the proposed method takes into account the different views of the annotators.

## 2.3 Adaptive Behavior Generation According to Engagement

It is also important for engagement studies to investigate how to generate system behaviors after the system recognizes user engagement. Studies on the adaptive behavior generation lead to clarifying the significance of engagement recognition. These generated behaviors should be coordinated to the users.

System behaviors include turn-taking behaviors. Since turn-taking behaviors are reflective of engagement, the system should control its own turn-taking behaviors according to user engagement. In a human-robot dialogue, the robot was designed to control the turn-taking behavior after the robot recognizes engagement of the user [90]. If the user was recognized as engaged based on non-linguistic behaviors, the robot started the dialogue by proactive calling from the robot. Besides, if the user was engaged during the dialogue, the robot gave the conversational floor to the user because engaging of the user means that the user wants to talk more. The relationship between turn-taking behaviors and engagement was investigated in detail [130]. The analysis indicated that if a participant was engaged, the duration of the participant's turn became longer than the case of not engaged. Furthermore, the frequency of backchannels given by the corresponding interlocutor was also higher. These results suggest that a person tends to hold the conversational floor when the person is engaged, and also that

the interlocutor acknowledges that the willingness of the person by verbal backchannels. It is desired that turn-taking models adopt the viewpoint of user engagement to realize smooth turn-taking.

Adaptive system actions in accordance with engagement were also investigated. Topic selection based on user engagement was proposed [131]. The system was designed to predict user engagement on each topic, and select the next topic that maximizes both user engagement and the system's preference. A chatbot system was implemented to select a dialogue module according to user engagement [92]. For example, when the user was not engaged in the conversation, the system switched the current dialogue module into another. The system also had a function that refers back to a previously engaged topic. Subject experiments revealed that the perceived appropriateness of the system utterance were improved by considering user engagement. Another chatting system was implemented to react to user disengagement [117]. In an interview dialogue, when the user (interviewee) was disengaged, the system gave positive feedbacks to elicit more self-disclosure from the user. A subject experiment showed that self-reported engagement and user experience were improved by the positive feedbacks. It was also reported that a size of the vocabulary of the user was increased in the condition with the positive feedbacks. Another research [93] investigated how to handle user engagement in a human-robot interaction where the user had a main task and the robot sometimes interrupt by chatting. The main task was data input by personal computers. At first, they defined engagement states by eye gaze behaviors and verbal backchannels. They also extracted action patterns according to the engagement states from human-human dialogue. Based on the observation, they implemented two kinds of system actions: *explicit* and *implicit*. The explicit action means that the robot explicitly interrupts the user by saying "*Are you listening?*" It is expected that the explicit system utterance would be more noticeable for the user. On the other hand, the implicit action contains inserting fillers between system utterances to gain attention from the user. It is expected that the implicit action is more natural and does not annoy the user. A subject experiment showed that the explicit action increased perceived annoyance, controlling, and aggression of the robot. On the other hand, the implicit action increased consideration, appropriateness, and politeness of the robot. These results suggest that conversational robots should take the implicit strategy, for cooperation with the user.

# Chapter 3

# Engagement Recognition based on Multimodal Behaviors

This chapter addresses an engagement recognition method based on multimodal listener behaviors. The method is formulated as a latent variable model to consider the subjectivity of the annotation of engagement.

## 3.1 Introduction

Engagement recognition has been widely studied for human-robot interaction and spoken dialogue systems. The used features were based on non-linguistic information such as backchannels and eye-gaze behaviors. To realize engagement recognition, it is necessary to collect reference labels of engagement. However, it is unrealistic to ask dialogue participants to evaluate their own engagement due to time constraint. Furthermore, there is an evaluation bias where people tend to evaluate their engagement positively. In general, it is common to ask third-party people, called annotators, to annotate the engagement by independently watching dialogue videos. Although annotation criteria are carefully considered and given to the annotators, disagreement among annotators are observed because the perception of engagement is subjective. Previous studies dealt with this problem by integrating the different labels by such majority voting [97, 96]. This chapter aims to recognize each annotator's label instead of recognizing the integrated label as depicted in Fig. 3.1. There are two ways to use the result of engagement recognition. The first way is to use the integrated label, which means that a dialogue system uses a common perception of engagement. Another way is to use a specific engagement label corresponding to one annotator, which means that the

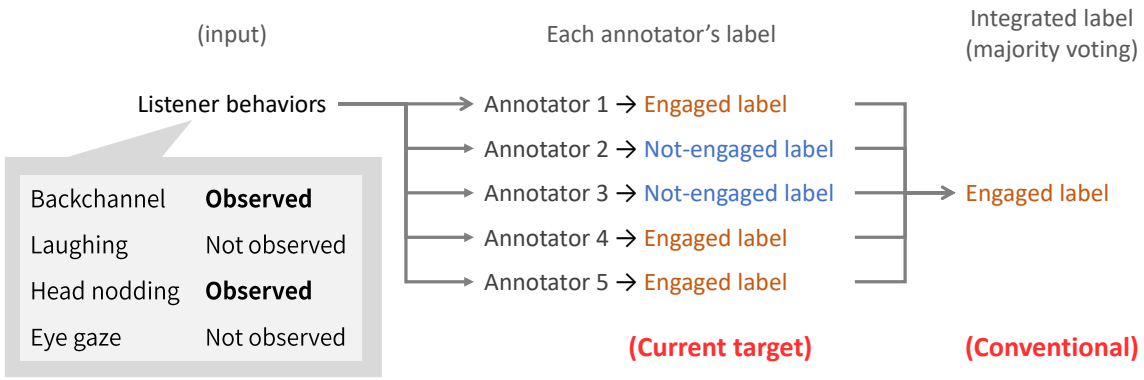


Fig. 3.1 Recognition of each annotator's label

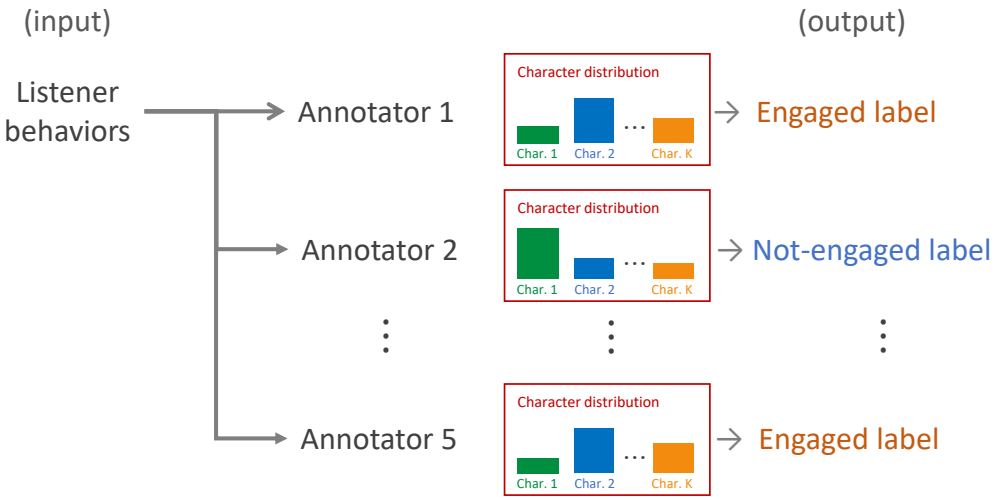


Fig. 3.2 Modeling of individual perception of engagement with different character distribution

dialogue system imitates the annotator's perception. An engagement recognition model is proposed for the latter use.

To recognize each annotator's label, this chapter addresses modeling of each annotator's perception of engagement. A recognition model can be trained for each annotator by using only each annotator's labels. However, if the amount of training data of each annotator is limited, the model training falls to overfitting due to data sparseness. To tackle this problem, it is assumed that perceptions of the annotators are clustered into several groups. To implement the assumption, this study introduces a latent variable called character. It is modeled that each annotator has its own character and the character affects the perception of engagement as illustrated in Fig 3.2. The character is represented in several dimensions on a latent space and is implemented by a hierarchical Bayesian model. It is expected that the proposed model can precisely recognize each annotator's label, avoiding the overfitting.

## 3.2 Human-Robot Dialogue Corpus

A human-robot dialogue corpus has been collected in which the humanoid robot ERICA [132, 133] interacted with human subjects. ERICA was operated by another human subject, called an operator, who was in a remote room. The dialogue was one-on-one, and the subject and ERICA sat on chairs facing each other. Fig. 3.3 shows a snapshot of the dialogue. The dialogue scenario was as follows. ERICA works in a laboratory as a secretary, and the subject visited the professor. Since the professor was absent for a while, the subject talked with ERICA until the professor would come back. The operator was asked to think of utterances and dialogue content by herself. The voice uttered by the operator was directly played with a speaker placed on ERICA in real time. When the operator spoke, the lip and head motions of ERICA were automatically generated from the prosodic information [134, 135]. The operator also manually controlled the head and eye-gaze motions of ERICA to express some behaviors such as eye-contact and head nodding. The dialogue was recorded with multimodal sensors, referring to other corpus studies [136, 137]. The used sensors were directed microphones, a 16-channel microphone array, RGB cameras, and the Kinect v2 sensor. After the recording, the manual annotation was conducted to build the conversation data including utterances, turn units, dialogue acts, backchannels, laughing, fillers, head nodding, and eye gaze (the object at which the participant is looking). Besides, both participants were asked to evaluate the level of self-engagement on the 7-point scale.

For an annotation of subject engagement, 20 dialogue sessions were used in this paper. The subjects were 12 females and 8 males, with ages ranging from teenagers to over 70 years old. The operators were six amateur actresses in their 20s and 30s. Whereas each subject participated in only one session, each operator was assigned several sessions. Each dialogue lasted about 10 minutes. All participants were native Japanese speakers.

## 3.3 Annotation of Engagement

The annotation work of the subject engagement in the human-robot dialogue corpus was conducted. This section explains how to annotate the subject engagement and an analysis of the annotation result below.

### 3.3.1 How to Annotate Engagement

There are several choices to annotate ground-truth labels of the subject engagement. The intuitive method is to ask the subject to evaluate his/her own engagement right after the dialogue session. However, in practice, it can be obtained only one evaluation on the

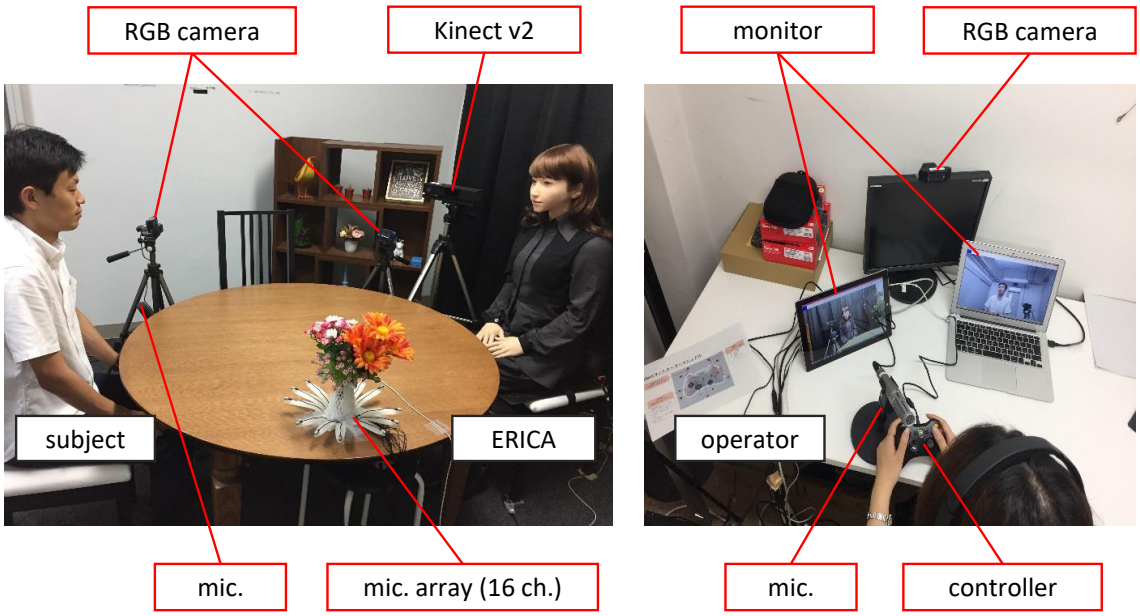


Fig. 3.3 Setup for dialogue collection

whole dialogue due to time constraints. It is difficult to obtain evaluations on fine-grained phenomena such as every conversational turn. Furthermore, there often is a bias where subjects tend to give positive evaluations of themselves. This kind of bias was observed in other works [95, 138]. Another method is to ask the ERICA's operators to evaluate the subject engagement. However, it was difficult to let the actresses participate in this annotation work due to time constraints. Similar to the first method, also in the current case, it was obtained only one evaluation on the whole dialogue, but this is not useful for the current recognition task. This problem often happens in other studies for building corpora because the dialogue recording and the annotation work are done separately. Most of the previous studies adopted a practical approach where they asked third-party people (annotators) to evaluate engagement. This approach is categorized into two types: training a small number of annotators [114, 90, 126, 117] and making use of the wisdom of crowds [97, 96]. The former type is valid when the annotation criterion is objective. On the other hand, the latter type is better when the criterion is subjective and when a large amount of data is needed. This study took the latter approach.

As annotators, 12 females who had not participated in the dialogue data collection were recruited. Their gender was set to be same as those of the ERICA's operators. The annotators were instructed to take the point of view of the operator. Each dialogue session was randomly assigned to 5 annotators. The definition of engagement was presented as "*How much the subject is interested in and willing to continue the current dialogue with ERICA*". The annotators were asked to annotate the subject engagement based on its behaviors while the

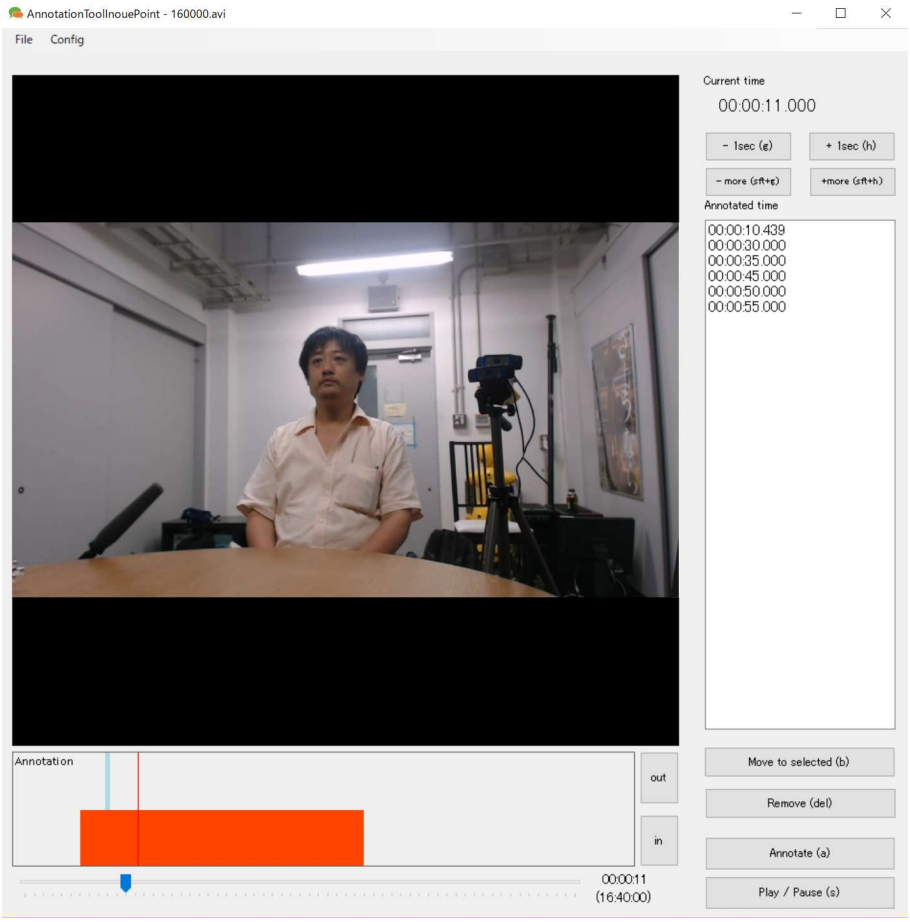


Fig. 3.4 Graphical user interface of annotation tool

subject was listening to ERICA's talk. Therefore, the subject engagement can be interpreted as *listener engagement*. It also means that the annotators observe *listener behaviors* expressed by the subject. The annotators were shown a list of listener behaviors that could be related to engagement, with example descriptions. This list included facial expression, laughing, eye gaze, backchannels, head nodding, body pose, moving of shoulders, and moving of arms or hands. The annotators were instructed to watch the dialogue video by standing in the viewpoint of the ERICA's operator, and to press a button when all the following conditions were being met:

(1) ERICA was holding the turn

(2) the subject was expressing any listener behaviors

(3) the behavior means the high level of subject engagement

For condition (1), the annotators were notified of auxiliary information which showed the timing of when the conversational turn was changed between the subject and ERICA. Fig. 3.4 illustrates an annotation tool implemented for this annotation work.

Before the annotation work, the annotators were given the practice time to understand what kind of behaviors could appear in this dialogue corpus. Each annotator was assigned other 5 dialogues for the practice, and they annotated only the listener's behaviors without considering engagement. The result of the practice was verified by the experimenter to make sure that each annotator had noticed the enough amount of the behaviors.

### 3.3.2 Analysis

Across all annotators and sessions, the average number of button presses per session was 18.13 with a standard deviation of 12.88. Since each annotator was assigned some of the 20 sessions randomly, one-way ANOVA was tested for both inter-annotator and inter-session, respectively. As a result, it was observed significant differences in the average numbers of button presses among both the annotators ($F(11,88) = 4.64$, $p = 1.51 \times 10^{-5}$) and the sessions ($F(19,80) = 2.56$, $p = 1.92 \times 10^{-3}$). There was a variation among not only sessions but also annotators.

In this study, each ERICA's conversational turn is used as a unit for engagement recognition. The conversational turn is useful for spoken dialogue systems to utilize the result of engagement recognition because the systems typically make a decision on their behaviors on a turn basis. If an annotator pressed the button more than once in a turn, it was regarded that the turn was annotated as *engaged* by the annotator. This dataset excluded short turns whose durations are smaller than 3 seconds, and also some turns corresponding to the greeting. As a result, the total number of ERICA's turns was 433 over 20 dialogue sessions. The numbers of engaged and not engaged turns from all annotators were 894 and 1,271, respectively.

The agreement level among the annotators is investigated as below. The average value of Cohen's kappa coefficients on every pair of two annotators was 0.291 with a standard deviation of 0.229. However, as Fig. 3.5 shows, some pairs showed coefficients higher than the moderate agreement (larger than 0.400). The result suggests that the annotators can be clustered into some groups based on their tendencies to annotate engagement similarly. Each group regarded different behaviors as important, and had different thresholds to accept the behaviors as engaged events.

The annotators were asked to complete a survey on which behaviors were essential to judge the subject engagement. For every session, the annotators were asked to select all the essential behaviors to judge the subject engagement. Table 3.1 lists the results of this survey. Note that this survey was conducted in total 100 times (5 annotators in 20 sessions). The

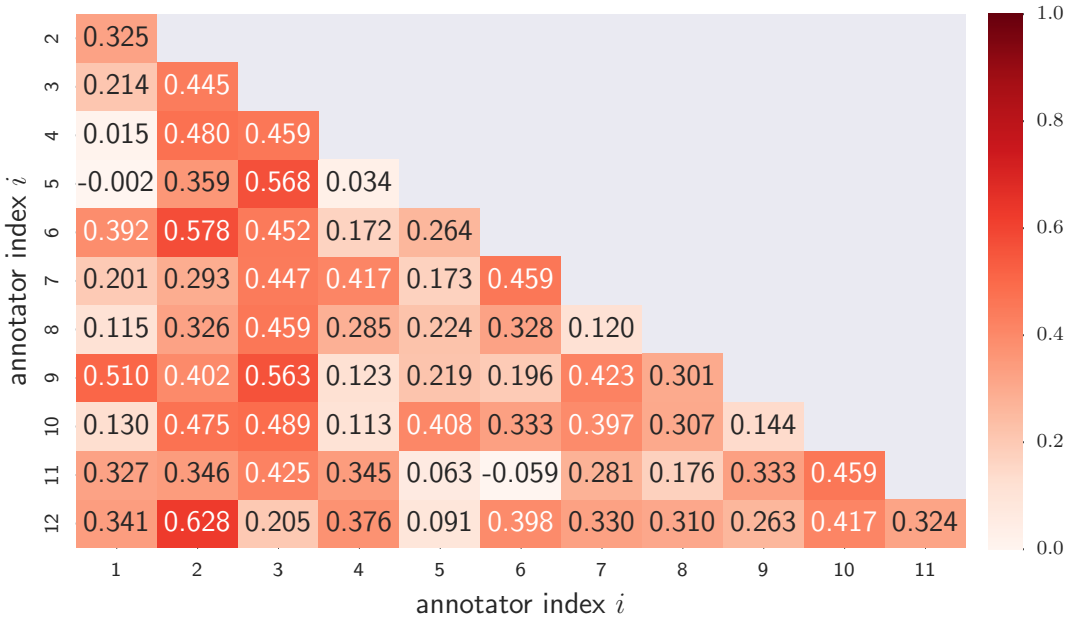


Fig. 3.5 Inter-annotator agreement score (Cohen's kappa) on each pair of annotators

Table 3.1 The number of times selected by annotators as essential behaviors

| listener behavior | #selected |
|---|---|
| facial expression | 77 |
| backchannels | 67 |
| head nodding | 65 |
| eye gaze | 40 |
| laughing | 39 |
| body pose | 32 |
| moving of shoulders | 3 |
| moving of arms or hands | 2 |
| others | 4 |

result indicates that engagement could be related to some listener behaviors such as facial expression, backchannels, head nodding, eye gaze, laughing, and body pose.

In following experiments, four behaviors are used: backchannels, laughing, head nodding, and eye gaze. As seen in Chapter 2, these behaviors have been identified as indicators of engagement. The occurrences of the four behaviors were manually annotated. The definition of backchannel in this annotation was responsive interjections (such as "*yeah*" in English and "*un*" in Japanese), and expressive interjections (such as "*oh*" in English and "*he-*" in Japanese) [139]. The laughing was defined as vocal laughing, not including just smiling without any vocal utterance. The occurrence of head nodding was annotated based on the

Table 3.2 Relationship between the occurrence of each behavior and the annotated engagement (*1*: occurred / engaged, *0*: not occurred / not engaged)

| behavior | | engagement | | accuracy |
|---|---|---|---|---|
| | | *1* | *0* | |
| backchannel | *1* | 790 | 830 | 0.569 |
| | *0* | 104 | 441 | |
| laughing | *1* | 139 | 46 | 0.630 |
| | *0* | 755 | 1,225 | |
| head nodding | *1* | 653 | 532 | 0.643 |
| | *0* | 241 | 739 | |
| eye gaze | *1* | 332 | 243 | 0.628 |
| | *0* | 562 | 1,028 | |

vertical movement of the head. The occurrence of eye-gaze behaviors is acknowledged when the subject was gazing at ERICA's face continuously for more than 10 seconds. This 10-second threshold was decided by confirming a reasonable balance between the accuracy in Table 3.2 and the recall of the engaged turns. It was challenging to annotate facial expression and body pose due to their ambiguity. The other behaviors not used in this study will be considered as additional features in the future work. The relationship between the occurrences of the four behaviors and the annotated engagement is summarized in Table 3.2. Note that the numbers are calculated from the engagement labels given by all individual annotators. The result suggests that these four behaviors are useful cues to recognize the subject engagement. The accuracy scores on each annotator is further analyzed in Table 3.3. The results show that each annotator has a different perspective on each behavior. For example, the engagement labels of the first annotator (index 1) are related to the labels of backchannels and head nodding. On the other hand, those of the second annotator (index 2) are related to those of laughing, head nodding, and eye gaze. This difference implies that it is needed to consider each annotator's different perspective on engagement.

## 3.4 Latent Character Model

A hierarchical Bayesian model for engagement recognition is proposed. The annotators can be clustered into some groups based on their perception manners. It is assumed that each annotator has a character which affects his/her perception of engagement. The character is a latent variable estimated from the annotation data. The proposed model is called as a latent character model. This model is inspired by latent Dirichlet allocation (LDA) [140]

Table 3.3 Accuracy scores of each annotator's engagement labels when the reference labels of each behavior are used

| annotator index | listener behavior | | | |
|---|---|---|---|---|
| | backchannel | laughing | head nodding | eye gaze |
| 1 | 0.751 | 0.348 | 0.736 | 0.488 |
| 2 | 0.588 | 0.690 | 0.722 | 0.717 |
| 3 | 0.584 | 0.753 | 0.695 | 0.689 |
| 4 | 0.487 | 0.702 | 0.607 | 0.696 |
| 5 | 0.301 | 0.850 | 0.566 | 0.659 |
| 6 | 0.553 | 0.648 | 0.620 | 0.581 |
| 7 | 0.647 | 0.479 | 0.647 | 0.599 |
| 8 | 0.441 | 0.752 | 0.466 | 0.596 |
| 9 | 0.749 | 0.518 | 0.769 | 0.559 |
| 10 | 0.576 | 0.589 | 0.576 | 0.677 |
| 11 | 0.542 | 0.651 | 0.657 | 0.627 |
| 12 | 0.553 | 0.619 | 0.604 | 0.660 |

and the latent class model which estimates annotators' abilities for a decision task like diagnosis [141].

### 3.4.1 Problem Formulation

At first, the problem formulation of this engagement recognition is defined as follows. Engagement recognition is done for each turn of the robot (ERICA). The input is based on listener behaviors of the user during the turn: laughing, backchannels, head nodding, and eye gaze. Each behavior is represented as binary: occur or not as defined in the previous section. The input feature is a combination of the occurrences of the four behaviors and referred to as *behavior pattern*. In this study, since the four listener behaviors are used, the possible number of the behavior patterns is 16 ($= 2^4$). Although the number of the behavior patterns would be massive if many behaviors are used, the observed patterns are limited so that the less-frequent patterns can be excluded. The output is also binary: engaged or not, as annotated in the previous section. Note that this ground-truth label differs for each annotator.

### 3.4.2 Generative Process

The latent character model is illustrated as the graphical model in Fig. 3.6. Table 3.4 describes symbols used below. The generative process is as follows. For each annotator, parameters of

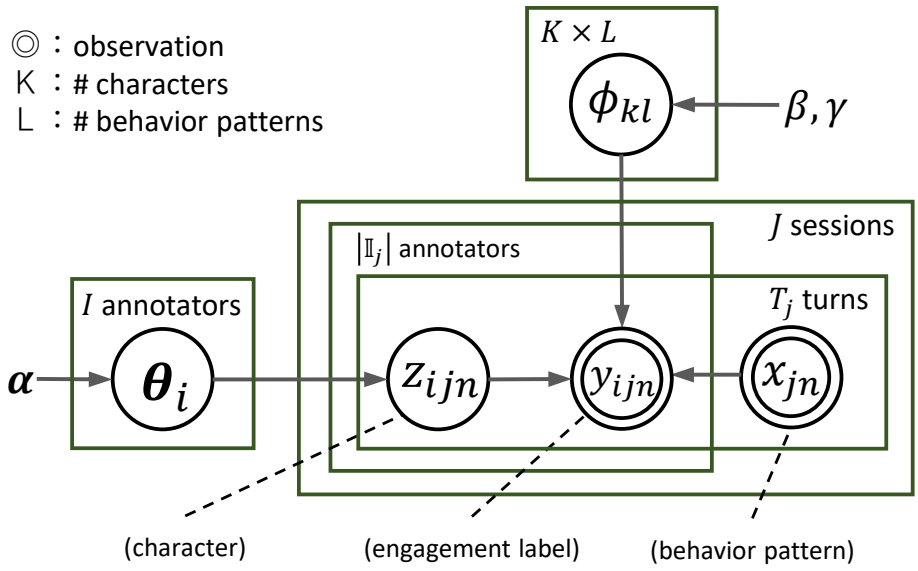


Fig. 3.6 Graphical model of latent character model

a character distribution are generated from the Dirichlet distribution as

$$\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{ik}, \cdots, \theta_{iK}) \sim \text{Dirichlet}(\boldsymbol{\alpha}) \ , \ 1 \leq i \leq I \ , \tag{3.1}$$

where $I$, $K$, $i$, $k$ denote the number of annotators, the number of characters, the annotator index, and the character index, respectively, and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k, \cdots, \alpha_K)$ is a hyperparameter. The parameter $\theta_{ik}$ represents the probability that the $i$-th annotator has the $k$-th character. For each combination of the $k$-th character and the $l$-th behavior pattern, a parameter of an engagement distribution is generated by the beta distribution as

$$\phi_{kl} \sim \text{Beta}(\beta, \gamma) \ , \ 1 \leq k \leq K \ , \ 1 \leq l \leq L \ , \tag{3.2}$$

where $L$ denotes the number of behavior patterns, and $\beta$ and $\gamma$ are hyperparameters. The parameter $\phi_{kl}$ represents the probability that annotators with the $k$-th character interpret the $l$-th behavior pattern as an *engaged* signal.

The number of the total dialogue sessions is represented as $J$. For the $j$-th session, the set of annotator indices who were assigned to this session is represented as $\mathbb{I}_j$. The number of conversational turns of the robot in the $j$-th session is represented as $T_j$. For each turn, a character is generated from the categorical distribution corresponding to the $i$-th annotator as

$$z_{ijn} \sim \text{Categorical}(\boldsymbol{\theta}_i) \ , \ 1 \leq j \leq J, \ 1 \leq n \leq T_j \ , \ i \in \mathbb{I}_j \ , \tag{3.3}$$

Table 3.4 Description of symbols

| symbol | description |
|---|---|
| $K$ | # of characters |
| $L$ | # of behavior patterns |
| $I$ | # of annotators |
| $J$ | # of dialogue sessions |
| $T_j$ | # of dialogue turns in $j$-th dialogue session |
| $\mathbb{I}_j$ | set of annotator indices who were assigned to $j$-th dialogue session |
| $\theta_{ik}$ | parameter of character distribution (probability that $i$-th annotator has $k$-th character) |
| $\phi_{kl}$ | parameter of engagement distribution (probability that annotators with $k$-th character interpret $l$-th behavior pattern as *engaged* signal) |
| $x_{jn}$ | behavior pattern in $n$-th turn of $j$-th dialogue session |
| $y_{ijn}$ | engagement label given by $i$-th annotator in $n$-th turn of $j$-th dialogue session |
| $z_{ijn}$ | character of $i$-th annotator in $n$-th turn of $j$-th dialogue session |
| $\boldsymbol{\Theta}$ | dataset of parameters of character distribution |
| $\boldsymbol{\Phi}$ | dataset of parameters of engagement distribution |
| $\boldsymbol{X}$ | dataset of behavior patterns |
| $\boldsymbol{Y}$ | dataset of engagement labels |
| $\boldsymbol{Z}$ | dataset of characters |
| $D_i$ | # of turns where $i$-th annotator was assigned |
| $D_{ik}$ | # of turns where $i$-th annotator had $k$-th character |
| $N_{kl}$ | # of engagement labels given by annotators with $k$-th character against $l$-th behavior pattern |
| $N_{kl1}$ | # of times when engaged labels are given in the condition of $N_{kl}$ |
| $N_{kl0}$ | # of times when not-engaged labels are given in the condition of $N_{kl}$ |
| $N_{ijnl}$ | binary variable indicating if $i$-th annotator observed $l$-th behavior pattern in $n$-th turn of $j$-th dialogue session |
| $N_{ijnl1}$ | binary variable indicating if $i$-th annotator gave an engaged label in the condition of $N_{ijnl}$ |
| $N_{ijnl0}$ | binary variable indicating if $i$-th annotator gave a not-engaged label in the condition of $N_{ijnl}$ |
| $D_{A\backslash B}$ | # of turns in the condition of $D_A$ without considering the index $B$ |
| $N_{A\backslash B}$ | # of engagement labels in the condition of $N_A$ without considering the index $B$ |

where $n$ denotes the turn index. The input behavior pattern in this turn is represented as $x_{jn} \in \{1, \cdots, L\}$. Note that the behavior pattern is independent from the annotator index $i$. The binary engagement label is generated from the Bernoulli distribution based on the

character and the input behavior pattern as

$$y_{ijn} \sim \text{Bernoulli}(\phi_{z_{ijn}x_{jn}}) \ . \tag{3.4}$$

Given the dataset of the above variables and parameters, the conditional distribution is represented as

$$p(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{\Theta},\boldsymbol{\Phi}|\boldsymbol{X}) = p(\boldsymbol{Z}|\boldsymbol{\Theta})p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Phi})p(\boldsymbol{\Theta})p(\boldsymbol{\Phi}) \ , \tag{3.5}$$

where the bold capital letters represent the datasets of the variables written by the corresponding small letters. Note that $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are the model parameters, and the dataset of the behavior patterns $\boldsymbol{X}$ is given and regarded as constant.

### 3.4.3 Training

In the training phase, the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are estimated. The training datasets of the behavior patterns $\boldsymbol{X}$ and the engagement labels $\boldsymbol{Y}$ are given. A collapsed Gibbs sampling is used to marginalize the model parameters and to efficiently sample only target variables. Here, each character is sampled alternately and iteratively from its conditional probability distribution as

$$z_{ijn} \sim p(z_{ijn}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}_{\backslash ijn},\boldsymbol{\alpha},\beta,\gamma) \ , \tag{3.6}$$

where the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are marginalized. Note that $\boldsymbol{Z}_{\backslash ijn}$ is the set of the characters without $z_{ijn}$. The conditional probability distribution is expanded in the same manner as the other work [142]. The distribution is proportionate to the product of two terms as

$$\begin{aligned} & p(z_{ijn}=k|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}_{\backslash ijn},\boldsymbol{\alpha},\beta,\gamma) \\ \propto \ & p(z_{ijn}=k|\boldsymbol{Z}_{\backslash ijn},\boldsymbol{\alpha})p(y_{ijn}|\boldsymbol{X},\boldsymbol{Y}_{\backslash ijn},z_{ijn}=k,\boldsymbol{Z}_{\backslash ijn},\beta,\gamma) \ , \end{aligned} \tag{3.7}$$

where $\boldsymbol{Y}_{\backslash ijn}$ is the dataset of the engagement labels without $y_{ijn}$.

To extend each term of Eq. (3.7), it is introduced the joint probability of the datasets of the characters $\boldsymbol{Z}$ and engagement labels $\boldsymbol{Y}$ where the datasets of the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are marginalized as

$$p(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\alpha},\beta,\gamma) = p(\boldsymbol{Z}|\boldsymbol{\alpha}) \ p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\beta,\gamma) \ . \tag{3.8}$$

The first term is expanded by marginalizing the parameter set $\boldsymbol{\Theta}$ as

$$p(\boldsymbol{Z}|\boldsymbol{\alpha}) = \int p(\boldsymbol{Z}|\boldsymbol{\Theta})\ p(\boldsymbol{\Theta}|\boldsymbol{\alpha})\ d\boldsymbol{\Theta} \tag{3.9}$$

$$= \int \prod_{i=1}^{I}\prod_{k=1}^{K} \theta_{ik}^{D_{ik}} \prod_{i=1}^{I} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{ik}^{\alpha_k - 1} d\boldsymbol{\Theta} \tag{3.10}$$

$$= \frac{\Gamma(\sum_k \alpha_k)^I}{\prod_k \Gamma(\alpha_k)^I} \int \prod_{i=1}^{I}\prod_{k=1}^{K} \theta_{ik}^{D_{ik}+\alpha_k - 1} d\boldsymbol{\Theta} \tag{3.11}$$

$$= \frac{\Gamma(\sum_k \alpha_k)^I}{\prod_k \Gamma(\alpha_k)^I} \prod_{i=1}^{I} \frac{\prod_k \Gamma(D_{ik}+\alpha_k)}{\Gamma(D_i + \sum_k \alpha_k)}\ . \tag{3.12}$$

Note that $D_i$ and $D_{ik}$ denote the number of turns where the $i$-th annotator was assigned, and the number of turns where the $i$-th annotator had the $k$-th character, respectively. Besides, $\Gamma(\cdot)$ is the gamma function. Eq. (3.10) is derived from the definitions of the categorical and Dirichlet distributions. The last expansion is based on the partition function of the Dirichlet distribution as

$$\int \prod_{k=1}^{K} \theta_{ik}^{\alpha_k - 1} d\boldsymbol{\theta}_i = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}\ . \tag{3.13}$$

Similarly, the second term of Eq. (3.8) is also expanded by marginalizing the parameter set $\boldsymbol{\Phi}$ as

$$p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\beta,\gamma)$$

$$= \int p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\Phi})\ p(\boldsymbol{\Phi}|\beta,\gamma)\ d\boldsymbol{\Phi} \tag{3.14}$$

$$= \int \prod_{k=1}^{K}\prod_{l=1}^{L} \phi_{kl}^{N_{kl1}} (1-\phi_{kl})^{N_{kl0}} \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{\Gamma(\beta+\gamma)}{\Gamma(\beta)\Gamma(\gamma)} \phi_{kl}^{\beta-1} (1-\phi_{kl})^{\gamma-1} d\boldsymbol{\Phi} \tag{3.15}$$

$$= \frac{\Gamma(\beta+\gamma)^{KL}}{\Gamma(\beta)^{KL}\Gamma(\gamma)^{KL}} \int \phi_{kl}^{N_{kl1}+\beta-1} (1-\phi_{kl})^{N_{kl0}+\gamma-1} d\boldsymbol{\Phi} \tag{3.16}$$

$$= \frac{\Gamma(\beta+\gamma)^{KL}}{\Gamma(\beta)^{KL}\Gamma(\gamma)^{KL}} \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{\Gamma(N_{kl1}+\beta)\Gamma(N_{kl0}+\gamma)}{\Gamma(N_{kl}+\beta+\gamma)}\ . \tag{3.17}$$

Note that $N_{kl}$ is the number of engagement labels given by annotators with $k$-th character against the $l$-th behavior pattern. Among them, $N_{kl1}$ and $N_{kl0}$ represent the numbers of engaged and not-engaged labels, respectively. Eq. (3.15) is derived from the definitions of the Bernoulli and beta distributions. The last expansion is also based on the partition function like Eq. (3.13).

Using Eq. (3.12), the first term of Eq. (3.7) is calculated as

$$
\begin{aligned}
& p(z_{ijn}=k|\boldsymbol{Z}_{\backslash ijn}\boldsymbol{\alpha}) \\
=& \frac{p(z_{ijn}=k,\boldsymbol{Z}_{\backslash ijn}|\boldsymbol{\alpha})}{p(\boldsymbol{Z}_{\backslash ijn}|\boldsymbol{\alpha})} && (3.18)\\
=& \frac{\Gamma(D_{ik\backslash ijn}+1+\alpha_k)\prod_{k'\neq k}\Gamma(D_{ik'\backslash ijn}+\alpha_{k'})}{\Gamma(D_i+\sum_{k'}\alpha_{k'})} \Big/ \frac{\prod_{k'}\Gamma(D_{ik'\backslash ijn}+\alpha_{k'})}{\Gamma(D_i-1+\sum_{k'}\alpha_{k'})} && (3.19)\\
=& \frac{\Gamma(D_{ik\backslash ijn}+1+\alpha_k)}{\Gamma(D_{ik\backslash ijn}+\alpha_k)} \frac{\Gamma(D_i-1+\sum_{k'}\alpha_{k'})}{\Gamma(D_i+\sum_{k'}\alpha'_k)} && (3.20)\\
=& \frac{D_{ik\backslash ijn}+\alpha_k}{D_i-1+\sum_{k'}\alpha_{k'}}\,. && (3.21)
\end{aligned}
$$

Note that a property of the gamma function is used as

$$\Gamma(x+1)=x\Gamma(x) \tag{3.22}$$

where $x$ is any complex number except zero and negative integers. Using Eq. (3.17), the second term of Eq. (3.7) is also developed as

$$
\begin{aligned}
& p(y_{ijn}|\boldsymbol{X},\boldsymbol{Y}_{\backslash ijn},z_{ijn}=k,\boldsymbol{Z}_{\backslash ijn},\beta,\gamma) \\
=& \frac{p(\boldsymbol{Y}|\boldsymbol{X},z_{ijn}=k,\boldsymbol{Z}_{\backslash ijn},\beta,\gamma)}{p(\boldsymbol{Y}_{\backslash ijn}|\boldsymbol{X},\boldsymbol{Z}_{\backslash ijn},\beta,\gamma)} && (3.23)\\
=& \prod_{l=1}^{L}\frac{\Gamma(N_{kl1\backslash ijn}+N_{ijnl1}+\beta)\Gamma(N_{kl0\backslash ijn}+N_{ijnl0}+\gamma)}{\Gamma(N_{kl\backslash ijn}+D_{ijnl}+\beta+\gamma)} \prod_{\substack{k'=1\\k'\neq k}}^{K}\prod_{l=1}^{L}\frac{\Gamma(N_{k'l1\backslash ijn}+\beta)\Gamma(N_{k'l0\backslash ijn}+\gamma)}{\Gamma(N_{k'l\backslash ijn}+\beta+\gamma)} \\
& \Big/ \prod_{k'=1}^{K}\prod_{l=1}^{L}\frac{\Gamma(N_{k'l1\backslash ijn}+\beta)\Gamma(N_{k'l0\backslash ijn}+\gamma)}{\Gamma(N_{k'l\backslash ijn}+\beta+\gamma)} && (3.24)\\
=& \prod_{l=1}^{L}\left\{\frac{\Gamma(N_{kl\backslash ijn}+\beta+\gamma)}{\Gamma(N_{kl\backslash ijn}+N_{ijnl}+\beta+\gamma)}\frac{\Gamma(N_{kl1\backslash ijn}+N_{ijnl1}+\beta)}{\Gamma(N_{kl1\backslash ijn}+\beta)}\frac{\Gamma(N_{kl0\backslash ijn}+N_{ijnl0}+\gamma)}{\Gamma(N_{kl0\backslash ijn}+\gamma)}\right\}\,. && (3.25)
\end{aligned}
$$

Note that $N_{kl\backslash ijn}$ represents the number of engagement labels given by annotators with the $k$-th character against the $l$-th behavior pattern without considering $x_{ijn}$. Among them, $N_{kl1\backslash ijn}$ and $N_{kl0\backslash ijn}$ are the numbers of engaged and not-engaged labels, respectively. Besides, $N_{ijnl}$ represents the binary variable indicating if the $i$-th annotator observed the $l$-th behavior pattern in the $n$-th turn of the $j$-th session. Among them, $N_{ijnl1}$ and $N_{ijnl0}$ are binary variables indicating if the annotator gave the engaged and not engaged labels, respectively. Therefore,

the sampling formulation in Eq. (3.7) is expanded as

$$
\begin{aligned}
&p(z_{ijn}=k|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}_{\setminus ijn},\boldsymbol{\alpha},\beta,\gamma) \\
&\propto p(z_{ijn}=k|\boldsymbol{Z}_{\setminus ijn},\boldsymbol{\alpha})p(y_{ijn}|\boldsymbol{X},\boldsymbol{Y}_{\setminus ijn},z_{ijn}=k,\boldsymbol{Z}_{\setminus ijn},\beta,\gamma) \qquad (3.26)\\
&=\frac{D_{ik\setminus ijn}+\alpha_k}{D_i-1+\sum_{k'}\alpha_{k'}} \\
&\quad\times\prod_{l=1}^{L}\left\{\frac{\Gamma(N_{kl\setminus ijn}+\beta+\gamma)}{\Gamma(N_{kl\setminus ijn}+N_{ijnl}+\beta+\gamma)}\frac{\Gamma(N_{kl1\setminus ijn}+N_{ijnl1}+\beta)}{\Gamma(N_{kl1\setminus ijn}+\beta)}\frac{\Gamma(N_{kl0\setminus ijn}+N_{ijnl0}+\gamma)}{\Gamma(N_{kl0\setminus ijn}+\gamma)}\right\}. \qquad (3.27)
\end{aligned}
$$

After sampling, one of the sampling results is selected where the joint probability of the variables is maximized as

$$
\boldsymbol{Z}^{*}=\underset{\boldsymbol{Z}^{(r)}}{\arg\max}\ p(\boldsymbol{Y},\boldsymbol{Z}^{(r)}|\boldsymbol{X},\boldsymbol{\alpha},\beta,\gamma)\ , \qquad (3.28)
$$

where $\boldsymbol{Z}^{(r)}$ represents the $r$-th sampling result. Using Eq. (3.12) and Eq. (3.17), the joint probability is expanded as

$$
\begin{aligned}
p(\boldsymbol{Y},\boldsymbol{Z}^{(r)}|\boldsymbol{X},\boldsymbol{\alpha},\beta,\gamma) &= p(\boldsymbol{Z}^{(r)}|\boldsymbol{\alpha})\ p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}^{(r)},\beta,\gamma) \qquad (3.29)\\
&\propto \prod_{i=1}^{I}\frac{\prod_k\Gamma(D_{ik}+\alpha_k)}{\Gamma(D_i+\sum_k\alpha_k)}\prod_{k=1}^{K}\prod_{l=1}^{L}\frac{\Gamma(N_{kl1}+\beta)\Gamma(N_{kl0}+\gamma)}{\Gamma(N_{kl}+\beta+\gamma)}\ , \qquad (3.30)
\end{aligned}
$$

where $D_i$, $D_{ik}$, $N_{kl}$, and $N_{kl1}$ are counted up among the $r$-th sampling result $\boldsymbol{Z}^{(r)}$. Finally, the model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ are estimated based on the sampling result $\boldsymbol{Z}^{*}$ as

$$
\theta_{ik}=\frac{D_{ik}+\alpha_k}{D_i+\sum_{k'=1}^{K}\alpha_{k'}}\ , \qquad (3.31)
$$

$$
\phi_{kl}=\frac{N_{kl1}+\beta}{N_{kl}+\beta+\gamma}\ , \qquad (3.32)
$$

where $D_i$, $D_{ik}$, $N_{kl}$, and $N_{kl1}$ are counted up among the sampling result $\boldsymbol{Z}^{*}$.

### 3.4.4 Inference

In the testing phase, the unseen engagement label given by a target annotator is predicted based on the estimated model parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$. The input data are the behavior pattern $x_t \in \{1,\cdots,L\}$ and the target annotator index $i \in \{1,\cdots,I\}$ . Note that $t$ represents the turn index in the test data. The probability that the target annotator gives the engaged label on

this turn is calculated by marginalizing the character as

$$p(y_{it} = 1|x_t, i, \mathbf{\Theta}, \mathbf{\Phi}) = \sum_{k=1}^{K} \theta_{ik}\ \phi_{kx_t}\ . \tag{3.33}$$

The $t$-th turn is recognized as *engaged* by the target annotator when this probability is higher than a threshold.

### 3.4.5 Related Models

Similar models considering the difference of annotators are summarized. In a task of backchannel prediction, a two-step CRF (conditional random fields) was proposed [143, 144]. They trained a prediction model per annotator. The final decision is based on voting by the individual models. The proposed method trains the model based on the character, not for each annotator, so that robust estimation is expected even if the amount of data for each annotator is not large, which is the case in many realistic applications. In a task of estimation of empathetic states among dialogue participants, a model classified annotators by considering both the annotators' tendencies of the estimation and their personalities [145]. The personalities correspond to the characters in the proposed model. Their model was able to estimate the empathetic state based on a specific personality. It assumed that the personality and the input features such as the behavior patterns are independent. The proposed model assumes that the character and the input features are dependent, meaning that how to perceive each behavior pattern is different for each character.

## 3.5 Evaluation

In this section, the latent character model is compared with other models that do not consider the difference of the annotators. Besides, the accuracy of the online implementation is evaluated. Furthermore, the effectiveness of each behavior is investigated in order to identify important behaviors in this recognition task. In this experiment, the task is to recognize each annotator's labels. Since the low agreement among the annotators was observed in Section 3.3.2, it does not make sense to recognize a single overall label such as majority voting. In real-life applications, a target system can select an annotator or a character appropriate for the system. The last part of this section suggests a method to select a character distribution for engagement recognition based on a personality trait expected for a system such as a humanoid robot.

### 3.5.1 Setup

The cross-validation with the 20 dialogue sessions was conducted: 19 for training and the rest for testing. In the proposed model, the number of sampling was 10,000, and all prior distributions were the uniform distribution. The evaluation was done for each annotator one by one where 5 annotators individually annotated each dialogue session. Given the annotator index $i$, the probability of the engaged label (Eq. (3.33) or Eq. (4.4)) was calculated for each turn. The accuracy score is calculated by setting the threshold at 0.5. Note that the accuracy score is a ratio of the number of the correctly recognized turns to the total number of turns. The final evaluation was made by averaging the accuracy scores for all annotators and also the cross-validation. The accuracy score of a majority baseline was 0.579 (=1,271 / 2,165).

### 3.5.2 Effectiveness of Character

At first, the proposed model was compared with two methods to see the effectiveness of the character. In this experiment, the input behavior patterns were manually annotated. For the proposed model, an appropriate number of characters ($K$) was explored by changing from 2 to 5 on a trial basis. The first compared model was the same as the proposed model other than a unique character ($K = 1$). The second compared models are based on other machine learning methods. Compared models were logistic regression, support vector machine (SVM), and a multilayer perceptron (neural network). For each model, two types of training are considered: *majority* and *individual*. In the *majority* type, the training labels of the five annotators was integrated by majority voting and trained a unique model which was independent of the annotators. In the *individual* type, an individual model was trained for each annotator with his/her data only and used each model according to the input annotator index $i$ in the test phase. Although the *individual* type can learn the different tendency of each annotator, the amount of training data decreases. Furthermore, the training data was divided into training and validation datasets on a session basis so that it corresponds to 9:1. Each model was trained with the training dataset, and then tuned the parameter of each model with the validation dataset. For logistic regression, the weight parameter of the $l2$-norm regularization was tuned. The range of the value of this weight was $\{10^n | n = -3, -2, -1,\ 0,\ 1,\ 2,\ 3\}$. For SVM, the radial basis function (RBF) kernel was used, and the penalty parameter of the error term was tuned. The range of the value of this parameter was same as the case of logistic regression. For the multilayer perceptron, the weight parameter of the $l2$-norm regularization was tuned for each unit. The range of the value of this weight was $\{10^n | n = -7, -6, -5, -4, -3, -2, -1, \}$. It also was needed to decide other settings for the multilayer perceptron such as the number of hidden layers,

Table 3.5 Engagement recognition accuracy ($K$ is the number of characters.)

| method | | accuracy |
|---|---|---|
| majority baseline | | 0.579 |
| logistic regression | *majority* | 0.670 |
| | *individual* | 0.681 |
| SVM | *majority* | 0.667 |
| | *individual* | 0.683 |
| multilayer perceptron | *majority* | 0.678 |
| | *individual* | 0.690 |
| latent character (proposed) | $K = 1$ (no character) | 0.674 |
| | $K = 2$ | 0.697 |
| | $K = 3$ | 0.703 |
| | $K = 4$ | 0.711 |
| | $K = 5$ | 0.703 |

the number of hidden units, the activation function, the optimization method, and the batch size. The weight parameter of the $l2$-norm regularization was tuned with the validation dataset, but it is not realistic to optimize the other settings in the same manner. Therefore, all combinations of the other settings were tried with the test dataset, and the best score is reported here. The range of the number of hidden layers was $\{1,2,3,4,5\}$. The range of the number of hidden units was $\{4,8,16,24,32\}$. The activation function was selected from hyperbolic tangent and rectified linear unit (ReLU) [146]. The optimization method was selected from Adam [147] and stochastic gradient descent. The range of the mini-batch size was $\{1,2,4,8,16,32\}$.

Table 3.5 summarizes the accuracy scores. Among the conventional machine learning methods, the multilayer perceptron showed the highest score on both of the *majority* and *individual* types. For the *majority* type, the best setting of the multilayer perceptron was 5 hidden layers, 16 hidden units, hyperbolic tangent, stochastic gradient descent, and 4 samples in mini-batch. For the *individual* type, the best setting was 1 hidden layer, 16 hidden units, ReLU, Adam, and 4 samples in mini-batch. The difference in the number of hidden layers can be explained by the number of available training data.

Considering the character ($K \geq 2$), the accuracy was improved than the w/o-character models including the multilayer perceptron. The highest accuracy was 0.711 with the four characters ($K = 4$). A paired t-test was conducted between the cases of the unique character ($K = 1$) and the four characters ($K = 4$) and found a significant difference between them ($t(99) = 2.55$, $p = 1.24 \times 10^{-2}$). Another paired t-test was conducted between the proposed model with the four characters and the multilayer perceptron models. There was a significant difference between the proposed model ($K = 4$) and the *majority* type of

Table 3.6 Recognition accuracy without each behavior of the proposed method

| used behavior | $K = 1$ | $K = 4$ |
|---|---|---|
| all | 0.674 | 0.711 |
| w/o backchannels | 0.669 | 0.699 |
| w/o laughing | 0.606 | 0.684 |
| w/o head nodding | 0.664 | 0.700 |
| w/o eye gaze | 0.650 | 0.681 |

the multilayer perceptron ($t(99) = 2.34$, $p = 2.15 \times 10^{-2}$). There also was a significant difference between the proposed model ($K = 4$) and the *individual* type of the multilayer perceptron ($t(99) = 2.55$, $p = 1.24 \times 10^{-2}$).

These results indicate that the proposed model simulates each annotator's perception of engagement more accurately than the others by considering the character. Apparently, the *majority* voting is not enough for this kind of recognition task that contains subjectivity. Although the *individual* model has the potential to simulate each annotator's perception, it fails to address the data sparseness problem in model training. This means that there was not enough training data for each annotator. This problem is often observed when data is collected by the wisdom of crowd approach where a large number of annotators are available, but the amount of data from each annotator is small.

### 3.5.3 Identifying Important Behaviors

The effectiveness of each behavior was examined by eliminating one of the four behaviors from the feature set. In this experiment, the number of characters ($K$) was set at only one and four. The results are reported in Table 3.6. From this table, laughing and eye-gaze behaviors are more useful for this engagement recognition task. This result is partly consistent with the analysis of Table 3.2. It is assumed that backchannels and head nodding were indicating engagement, but those can be used more frequently than the others. While backchannel and head nodding behaviors play a role to acknowledge that the turn-taking floor will be held by the current speaker, laughing and eye-gaze behaviors express the reaction towards the spoken content. Therefore, laughing and eye-gaze behaviors are more related to the high level of engagement. However, it is thought that some backchannels such as *expressive interjections* (such as "*oh*" in English and "*he-*" in Japanese) [139] are used to express the high level of engagement. From this perspective, there is room for further investigation to classify each behavior into some categories which are correlated with the level of engagement.

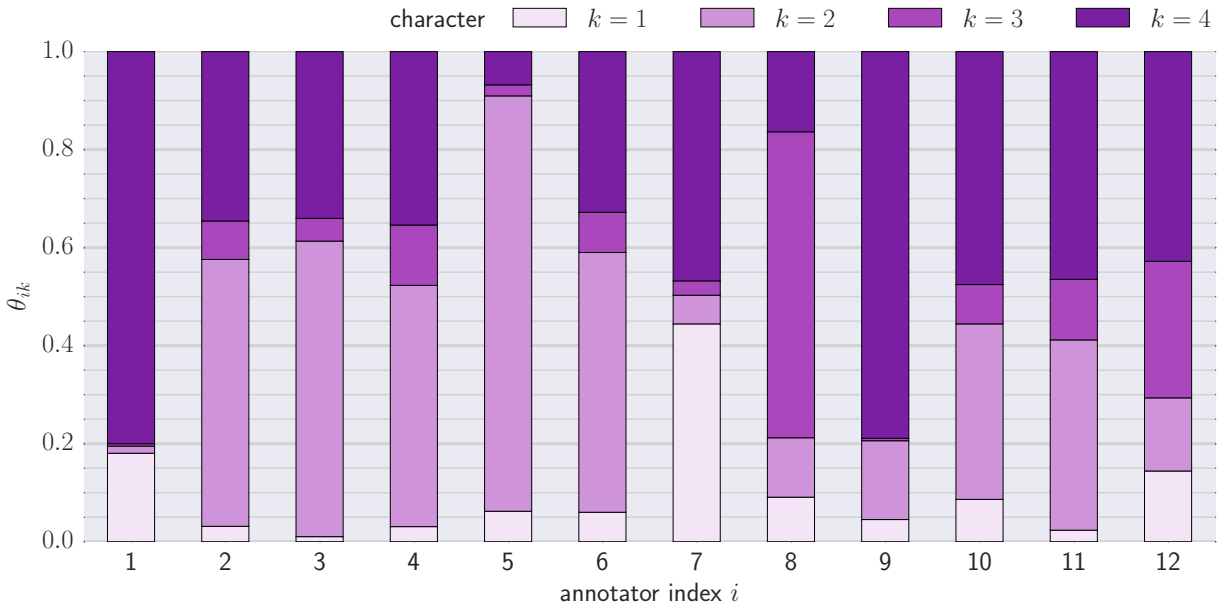


Fig. 3.7 Estimated parameter values of character distribution (Each value corresponds to the probability that each annotator has each character.)

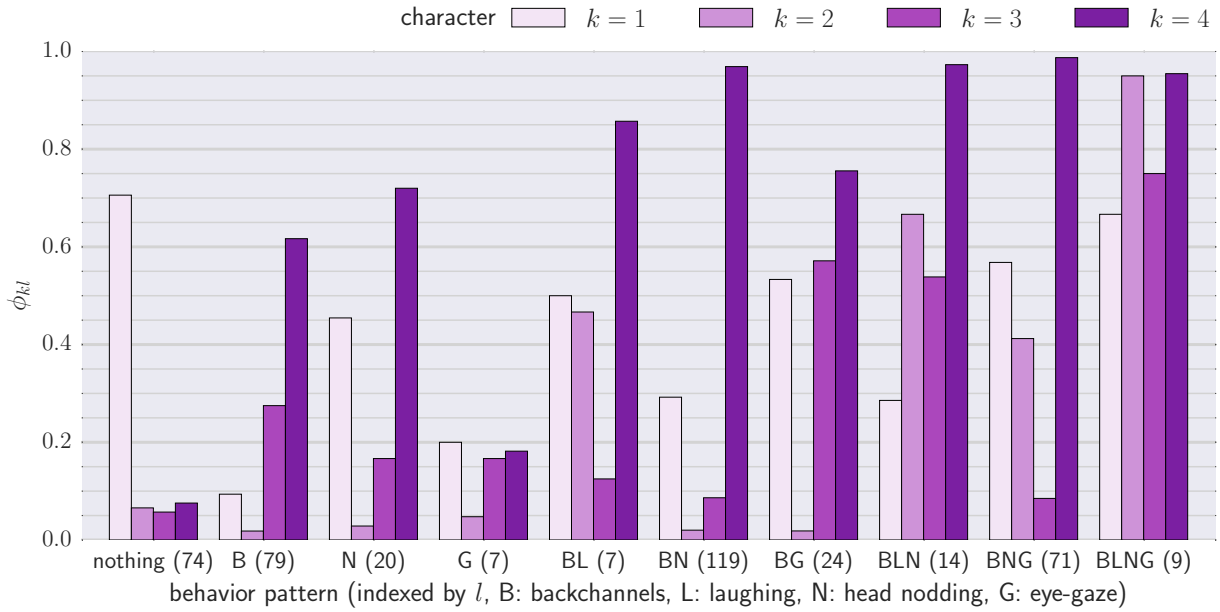


Fig. 3.8 Estimated parameter values of engagement distribution (Each value corresponds to the probability that each behavior pattern is recognized as engaged by each character. The number in parentheses next to each behavior pattern is the frequency of the behavior pattern in the corpus.)

Table 3.7 Clustered annotators based on character distribution and averaged in-cluster agreement scores

| cluster | annotator index | Cohen's kappa |
|---|---|---|
| A | 1, 9 | 0.510 |
| B | 2, 3, 4, 6 | 0.431 |
| C | 10, 11 | 0.459 |

Table 3.8 Averaged agreement scores between-clusters

| cluster pair | Cohen's kappa |
|---|---|
| A - B | 0.321 |
| A - C | 0.239 |
| B - C | 0.137 |

### 3.5.4 Example of Parameter Training

The result of parameter training was analyzed. The following parameters were trained using all 20 sessions. In this example, the number of characters was four ($K = 4$).

The parameters of the character distribution ($\theta_{ik}$) is shown in Fig. 3.7. The vertical axis represents the probability that each annotator has each character. It is observed that some annotators have common patterns. The annotators were clustered based on this distribution by using the hierarchical clustering with the unweighted pair-group method with arithmetic mean (UPGMA) algorithm. From the generated tree diagram, three clusters were extracted and reported in Table 3.7. Four annotators were independent (the annotators 5, 7, 8, 12). The table also shows the averaged agreement scores of the engagement labels among the annotators inside the same cluster. All scores are over the moderate agreement (larger than 0.400) and also higher than the whole averaged score (0.291) reported in Section 3.3.2. The averaged agreement scores between the clusters are analyzed. Table 3.8 reports the scores that are lower than the in-cluster agreement scores. For example, the annotators 1 and 9 share similar distributions. In Fig. 3.5, they showed the Cohen's kappa score of 0.510 which was higher than the average. Besides, the annotators 2, 3, 4, and 6 also share similar distributions, and they had the higher kappa scores among them.

The parameters of the engagement distribution ($\phi_{kl}$) is shown in Fig. 3.8. The vertical axis represents the probability that each behavior pattern is recognized as engaged by each character. Note that some behavior patterns were omitted due to less frequent (less than five in the whole corpus). The number of times when each behavior pattern is observed in the whole corpus is also reported in the figure. The proposed model can obtain the different

Table 3.9 Regression weights for mapping from Big Five scores to character distribution

| Big Five factor | character index ($k$) | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| *extroversion* | 4.11 | 1.95 | 0.00 | 4.00 |
| *neuroticism* | 0.71 | 0.00 | 0.00 | 1.94 |
| *openness to experience* | 1.52 | 0.00 | 5.10 | 0.00 |
| *conscientiousness* | 0.00 | 3.00 | 1.74 | 0.00 |
| *agreeableness* | 0.00 | 3.25 | 0.00 | 2.25 |

distribution for each character. Although the first character ($k = 1$) seems to be reactive to some behavior patterns, it is also reactive to behaviors other than the four behaviors because the high probability is estimated against the empty behavior pattern (nothing) where no behavior was observed. The second and third characters ($k = 2, 3$) show a similar tendency, but some are different (e.g., BL, BG, and BNG). The fourth character ($k = 4$) is reactive to all behavior patterns except the patterns of empty (nothing) and eye gaze only (G). Among all characters, the co-occurrence of multiple behaviors leads to higher probability (the right side of the figure). Especially, when all behaviors are observed (BLNG), the probability becomes very high in all characters. This tendency indicates that the co-occurrence of multiple behaviors expresses the high level of engagement.

### 3.5.5 How to Determine Character Distribution

The advantage of the proposed model is to simulate various kinds of perspectives for engagement recognition by changing the character distribution. However, when the proposed model is used in a spoken dialogue system, it is needed to determine one character distribution to be simulated. This experiment suggests a method based on the personality given to the dialogue system. Specifically, once the personality for the dialogue system is given, the character distribution for engagement recognition is decided. Here, as a proxy of the personality, the Big Five traits are used: *extroversion*, *neuroticism*, *openness to experience*, *conscientiousness*, and *agreeableness* [148]. For example, given a social role of a laboratory guide, the dialogue system is expected to be extroverted.

In the annotation work, the Big Five scores of each annotator were also measured [149]. Here, a softmax single-layer linear regression model is trained in order to map the Big Five scores to the character distribution as shown in Fig. 3.7. Note that the weight parameters of the regression are constrained by the $l1$-norm regularization and to be non-negative. The bias term was not added. Table 3.9 shows the regression weights and indicates that some characters are related to some personality traits. For example, *extroversion* is related to

the first and fourth characters, followed by the second character. This result can also be interpreted by relating to the engagement distribution in Fig. 3.8 The fourth character tends to give engaged labels on many behavior patterns and is related to *extroversion*. This result means that extrovert persons tend to give engaged labels than others. On the other hand, the second character tends to give not engaged labels on many behavior patterns and is related to *conscientiousness* and *agreeableness*. It can be said that persons with this personality tend to give not engaged labels than others.

The regressions with some social roles were tested. When a dialogue system plays a role of a laboratory guide, the input score on *extroversion* is set at the maximum value among the annotators, and the other scores are set to the average values. The output of the regression was $\boldsymbol{\theta} = (0.147, 0.226, 0.038, 0.589)$, which means the fourth character is weighted in this social role. For another social role such as a counselor, the input scores on *conscientiousness* and *agreeableness* are set at the maximum values among the annotators, and the other scores are set to the average values. The output was $\boldsymbol{\theta} = (0.068, 0.464, 0.109, 0.359)$, which means the second and fourth characters are weighted in this social role Further investigation is required on the effectiveness of this personality control.

## 3.6 Summary

This chapter has addressed engagement recognition using listener behaviors. Since the perception of engagement is subjective, the ground-truth labels depend on each annotator. It was assumed that each annotator has a character that affects his/her perception of engagement. The proposed model estimates not only user engagement but also the character of each annotator as latent variables. The model can simulate each annotator's perception by considering the character. In the experiment, the proposed model outperforms the other methods that use either the majority voting for label generation in training or individual training for each annotator. It was also confirmed that the proposed model could cluster the annotators based on their character distributions and that each character has a different perspective on behaviors for engagement recognition. The analysis result suggests the traits of each annotator or character. Therefore, a live spoken dialogue system can choose an annotator or a character that the system wants to imitate. Another method was also presented in order to select a character distribution based on the Big Five personality traits expected for the system.

# Chapter 4

# Engagement Recognition Integrating Automatic Behavior Detection

This chapter presents a method for integrating automatic behavior detection model with the engagement recognition model to realize online engagement recognition for live spoken dialogue systems. The method automatically detects the listener behaviors, and the posterior probabilities of the behaviors are used as the input of the engagement recognition model.

## 4.1 Introduction

Human conversation utilizes various non-verbal social signals for realizing smooth communication. The social signals convey internal states of participants, such as the level of engagement. This study focuses on four behaviors as social signals: backchannels, laughing, head nodding, and eye gaze. The series of previous studies have unveiled that these behaviors are related to engagement in both human-robot interaction and human-human conversation [79, 115, 97, 116, 114, 90, 96, 120, 119, 117, 118].

In the previous chapter, the behavior patterns were manually annotated and used as the input of the engagement recognition model. In order to use the engagement recognition model in spoken dialogue systems, this chapter addresses automatic detection of the behavior patterns. Fig. 4.1 illustrates the processing flow of the proposed method. Backchannels and laughing are detected with the 16-channel microphone array. Acoustic features were calculated from the audio signal enhanced by the microphone array processing. Head nodding and eye gaze are detected with the Kinect v2 sensor. Visual features consisted of the direction, velocity, and acceleration of the head estimated by the sensor. These devices make it possible to realize non-contact measurement of the behaviors. This factor is crucial for live spoken

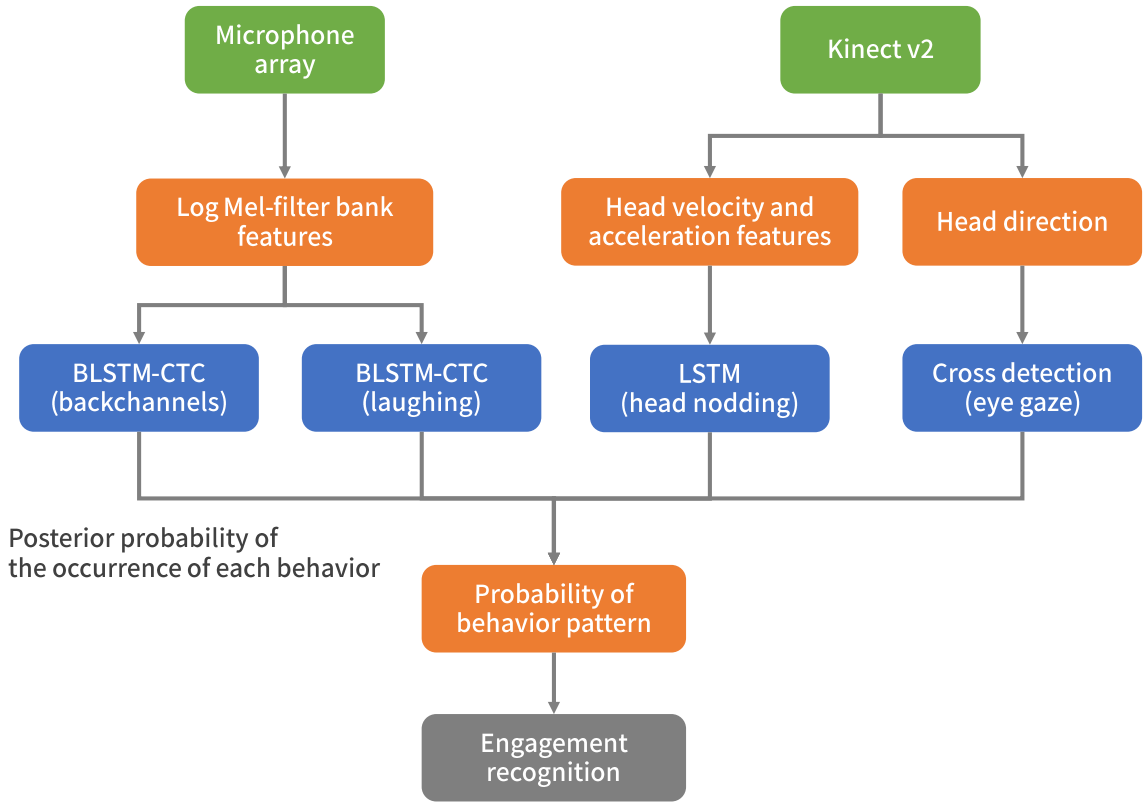


Fig. 4.1 Processing flow of automatic detection of listener behaviors

dialogue systems to be used in daily life. A detection model of each behavior outputs a posterior probability of that the behavior occurs. Finally, the posterior probabilities of the four behaviors are integrated and used as input of the engagement recognition model.

Machine learning approach is applied to detection of the listener behaviors. In contrast to the case of engagement, it is easy to collect a large amount of training data because the behaviors can be objectively defined. In early studies, the detection models were based on heuristic rules gained from analysis works. Recent studies on social signal detection have made use of machine learning methods [150, 151]. However, current engagement recognition systems have used only limited behaviors that can be obtained directly from sensors such as voice activity, distance, head direction, facial action units [90, 117]. This chapter reveals the feasibility of automatic detection of the listener behaviors for engagement recognition systems. Note that eye-gaze detection is done by a heuristic rule.

## 4.2 Automatic Behavior Detection

Automatic detection methods are described from one behavior to another. Bi-directional long short-term memory with connectionist temporal classification (BLSTM-CTC) is used for backchannels and laughing. An LSTM model is implemented for head nodding, and eye gaze is determined by a rule.

### 4.2.1 Backchannels

Backchannels are short vocal responses uttered by listeners towards the turn-holder [152]. The definition of backchannels in this study consists of responsive interjections (such as "*yeah*" in English and "*un*" in Japanese) and expressive interjections (such as "*oh*" in English and "*he-*" in Japanese) [139]. These responses are used by listeners to express acknowledgment that the current speakers can hold the turn, or the listeners are interested in the current talk. Therefore, backchannels are related to listener engagement.

The problem formulation in this study is as follows. On each user utterance, the log-Mel filterbank features are extracted as a 40-dimension vector and also a delta and a delta-delta of them. Therefore, the number of dimensions of the input features is 120. Note that each utterance is segmented as inter-pausal unit (IPU) [153] where every pause is longer than 200 milliseconds. The output is a binary label corresponding to whether the backchannel is contained in the utterance.

Earlier studies made use of heuristic rules for backchannel detection [152, 154]. Then, the Gaussian mixture model (GMM) was applied to this task [155]. Conditional random field (CRF) was applied to another task that predicts the timing of backchannels [38]. These machine learning models were trained as frame-wise classification. However, in the current case, event-wise detection is needed (e.g., whether the backchannel is uttered or not). This means that the frame-wise detection is not needed. Moreover, it is difficult to manually annotate frame-wise segmentation in training data.

Bi-directional long short-term memory with connectionist temporal classification (BLSTM-CTC) [156] is used for detection of backchannels. Recurrent neural networks (RNNs) can capture context information. Among them, LSTM can use longer context by introducing memory cells with switching mechanism [157]. To capture the context information more accurately, bi-directional LSTM consists of two separate forward and backward layers that are fed into the same next layer [158]. The advantage of using CTC is that it is not needed to conduct the manual segmentation [3]. Therefore, CTC has the potential for realizing robust sequence labeling.

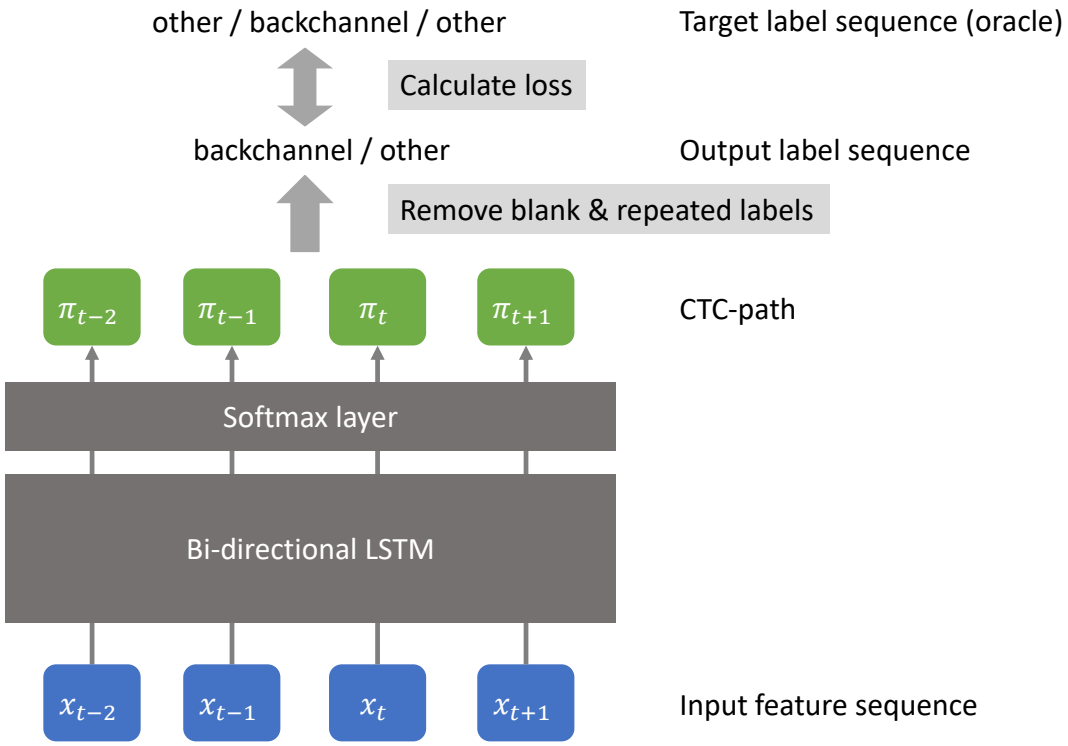


Fig. 4.2 Flowchart of bi-directional long short-term memory with connectionist temporal classification (BLSTM-CTC)

Fig. 4.2 describes the processing flow of BLSTM-CTC. CTC takes the outputs of BLSTM in frame-level, and then it generates an event-level sequence. CTC executes an optimization over all possible frame-level sequences. CTC uses a loss function for sequence labeling where the lengths of input and target sequences are different. For example, in the current case, the input sequence is frame-level audio features and the output sequence is the event-level behavior observations. CTC introduces the blank label in order to associate the input sequence to the target sequence. Let $\boldsymbol{X} = (x_1, x_2, \cdots, x_T)$ be the input sequence where the length is $T$. These values correspond to the outputs of the softmax layer of BLSTM. The target sequence is represented as $\boldsymbol{l} = (l_1, l_2, \cdots, l_U)$ where the length is $L$. CTC computes all the possible intermediate paths for the target sequence. The intermediate representation is represented as $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_T)$, called as CTC path, including the blank label and repeated labels. The posterior probability of the target sequence given the input sequence is represented as

$$p(\boldsymbol{l}|\boldsymbol{X}) = \sum_{\pi \in \Phi(\boldsymbol{l})} p(\boldsymbol{\pi}|\boldsymbol{X}) \,, \tag{4.1}$$

where $\Phi(\boldsymbol{l})$ is the set of CTC paths including the blank labels to represent the target sequence $\boldsymbol{l}$. The posterior probability of each CTC path is calculated based on independence assumption as

$$p(\boldsymbol{\pi}|\boldsymbol{X}) = \prod_{t=1}^{T} y_{t\pi_t} \ , \tag{4.2}$$

where $y_{tk}$ is the $k$-th output of the softmax layer at time $t$. Note that the output of the softmax layer contains a posterior probability for the blank label. The probability distribution of $p(\boldsymbol{l}|\boldsymbol{X})$ is efficiently computed by the forward-backward algorithm. When the model is tested, the maximum probability is selected from the outputs of the softmax layer by the greedy method.

In the current setting, it is assumed that the model is given an single utterance unit, then it detects whether the utterance contains a backchannel or not. Therefore, the target sequence consists of only two kinds of labels: *backchannel* and *other*. For example, if the person says a dialogue content after a backchannel in an utterance unit. The target sequence is *other backchannel*.

### 4.2.2 Laughing

Laughing displays a positive feeling towards the dialogue content. Therefore, it can be said that laughing is also related to engagement. Although smiling has the similar function, this study focuses on only vocal laughing because it is difficult to annotate smiling due to its ambiguity. Various machine learning methods were also applied to laughter detection [159–162]. Neural networks have also been applied to this problem, such as deep neural networks (DNNs) [163], convolutional neural networks (CNNs) [164], and bi-directional long short-term memory (BLSTM) [165]. The same architecture of BLSTM is used to detect vocal laughing. The target label sequence consists of *laughing* and *other*. Detection of backchannels and laughing are independently modeled because these can be uttered at the same time. The problem formulation is the same as detection of backchannels. The used input is also the same log-Mel filterbank features.

### 4.2.3 Head Nodding

Head nodding plays a role of backchannels without vocal utterance [166, 167] and is also related to engagement. Head nodding is defined as the vertical continuous movement of the head. Head nodding is detected from visual information captured by the Kinect v2 sensor. Detection of head nodding has also been widely studied in the field of computer vision. Since

detection of head nodding is formulated as the time series prediction, models that consider context information such as conditional random fields (CRF) have been proposed [154, 150]. This study uses LSTM for this task. The Kinect v2 sensor gives the head direction in the 3D space. A feature set contain the instantaneous speeds of the yaw, roll, and pitch of the head. Other features were the average speed, the average acceleration, and the range of the head pitch over the previous 500 milliseconds. An LSTM model is used with these features where the number of dimensions was 6. The occurrence probability of head nodding is estimated at every frame. The output sequence is smoothed with a Gaussian filter where the standard deviation for Gaussian kernel was 3.0.

### 4.2.4 Eye Gaze

Eye-gaze behavior has been studied as the indicator of engagement. In this study, the eye-gaze direction is approximated by the head orientation given by the Kinect v2 sensor. Precise eye-gaze direction would be measured if an eye tracker was used, but non-contact sensors such as the Kinect v2 sensor are preferable for spoken dialogue systems that interact with users on a daily basis. Eye gaze towards the robot is detected when the distance between the head-orientation vector and the location of the robot's head is smaller than a threshold. The threshold was set at 300 mm in the following experiment. The number of eye-gaze samples per second was about 30. To convert the estimated binary states (looking or not) to the occurrence probability of the eye-gaze behavior, a logistic function is used with a threshold of 10 seconds.

## 4.3 Integration with Engagement Recognition

The behavior detection models are used in the test phase of engagement recognition. On the detection of backchannels and laughing, the detection models output the posterior probability of the occurrence of each behavior on each IPU. On head nodding, the posterior probability of the occurrence of head nodding is calculated on each time-frame. On eye gaze, the posterior probability of eye-gaze behavior is calculated with the logistic regression on each on-set segment. For each behavior, the maximum posterior probability is used as the input of the engagement recognition model. The maximum probability of each behavior in the $t$-th turn of the test dataset is represented as $p_t(m)$ where $m$ is the behavior index.

The maximum posterior probabilities are converted into the occurrence probability of each behavior pattern for the engagement recognition model. The occurrence probability of

the $l$-th behavior pattern in the $t$-th turn is calculated as

$$p_t(l) = \prod_{m=1}^{M} p_t(m)^{b_m} (1 - p_t(m))^{(1-b_m)} , \tag{4.3}$$

where $M$, and $b_m \in \{0,1\}$ denote the number of behaviors and the occurrence of the behavior $m$, respectively. Note that $p_t(m)$ corresponds to the output of each behavior detection model, and the binary value $b_m$ is based on the given behavior pattern $l$. For example, when the given behavior pattern $l$ represents the case where both laughter ($m = 2$) and eye-gaze ($m = 4$) occur, the binary values are represented as $(b_1, b_2, b_3, b_4) = (0,1,0,1)$. As the behavior pattern $l$ is represented by the combination of the occurrences, $l = \sum_{m=1}^{M} b_m \cdot 2^{m-1}$. The probability of engaging (Eq. (3.33)) is reformulated by marginalizing not only the character but also the behavior pattern with its occurrence probability as

$$p(y_{it} = 1 | P_{tl}, i, \mathbf{\Theta}, \mathbf{\Phi}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \theta_{ik} \, \phi_{kl} \, p_t(l) , \tag{4.4}$$

where $P_{tl}$ denotes the set of the occurrence probabilities of all possible behavior patterns calculated by Eq. (4.3).

## 4.4 Evaluation

In this experiment, at first, the detection accuracy of each behavior is reported. Afterward, the detection result is integrated with engagement recognition, and the accuracy of online engagement recognition is evaluated.

### 4.4.1 Automatic Behavior Detection

A detection accuracy of each behavior is reported as below. The results are summarized in Table 4.1.

#### Backchannels

The BLSTM-CTC model was trained for backchannels. The number of hidden layers was 5 and the number of units on each layer was 256. The training data consisted of other 71 dialogue sessions recorded in the same manner as the dataset for engagement recognition. In the training set, the total number of user utterances was 14,704. Among them, the number of utterances containing backchannels was 3,931. Then, the detection model was tested with the

Table 4.1 Detection accuracy for each behavior

| behavior | evaluation type | precision | recall | F1 score |
|---|---|---|---|---|
| backchannels | event-based | 0.780 | 0.865 | 0.820 |
| laughing | event-based | 0.772 | 0.496 | 0.604 |
| head nodding | frame-based | 0.566 | 0.589 | 0.577 |
| | event-based | 0.608 | 0.763 | 0.677 |
| eye gaze | frame-based | 0.190 | 0.580 | 0.286 |
| | turn-based | 0.723 | 0.704 | 0.713 |

20 sessions which were used for engagement recognition. In the test dataset, the total number of the subject utterances was 3,517. Among them, the utterances containing backchannels was 1,045 Precision and recall of backchannels were 0.780 and 0.865, and the F1 score was 0.820. It can be said that the result is reasonable as the input for the engagement recognition.

### Laughing

The same BLSTM-CTC model was trained for laughing as well as the backchannel detection. The same 71 dialogue sessions were also used for training. In the training set, the number of utterances containing laughing was 1,003 of 14,704. In the test dataset of the 20 sessions, the utterances containing laughing was 240 of 3,517. Precision and recall scores were 0.772 and 0.496, and the F1 score was 0.604. The result suggests that detecting laughing is more difficult than those of backchannels. One possible reason is that the distribution of the data is more imbalanced. However, the precision score is still comparable, and the false alarm could be suppressed in engagement recognition.

### Head Nodding

An LSTM model that consists of a single hidden layer with 16 units is trained for the detection of head nodding. Since the annotation of head nodding was done by manual only for the same 20 sessions as the engagement recognition task, the 10-fold cross-validation was applied. The number of data frames per second was about 30, and a prediction was made every 5 frames. There were 29,560 prediction points in the whole dataset, and 3,152 of them were manually annotated as points where the user was nodding. Note that the data frames on the subject's turn were discarded. For prediction-point-wise evaluation, precision and recall of head nodding frames were 0.566 and 0.589, and the F1 score was 0.577. For event-wise detection, a continuous sequence of detected nodding was regarded as a head nodding event where the duration is longer than 300 milliseconds. If the sequence overlapped with a ground-truth event, the event was correctly detected. On an event basis, there were 855

head nodding events. Precision and recall of head nodding events were 0.608 and 0.763, and the F1 score was 0.677. The LSTM performance was compared with several other models such as SVM and DNN and found that the LSTM model had the best score [168].

#### Eye Gaze

The eye-gaze behavior was directly detected for the 20 sessions because the process is based on the rule. In the dataset of the 20 sessions, there were 300,426 eye-gaze samples in total, and 77,576 samples were manually annotated as *looking at the robot*. For frame-wise evaluation, precision and recall of the eye-gaze towards the robot were 0.190 and 0.580, and the F1 score was 0.286. This result implies that the ground-truth label of eye gaze based on the actual eye-gaze direction is sometimes different from the head direction. However, even if the frame-wise performance is low, it is enough if the eye-gaze behavior feature can be detected. The detection performance on a turn basis was also evaluated. There were 433 turns in the corpus, and the continuous eye-gaze behavior was observed in 115 turns of them. For this turn-wise evaluation, precision and recall of the eye-gaze behavior were 0.723 and 0.704, and the F1 score was 0.713. This result means that this method is sufficient to detect the eye-gaze behavior for engagement recognition. Note that some *not looking* states were ignored if the duration is smaller than 500 milliseconds.

### 4.4.2 Online Engagement Recognition

The online engagement recognition was evaluated. The same 20 dialogue sessions were used, and the cross-validation was conducted where 19 sessions were used for training and the rest was for testing. Compared methods were same as the previous chapter: logistic regression, support vector machine, multilayer perception, and the proposed model with a unique character ($K = 1$). Training ways are also same: training with majority labels (*majority*) and training for individual annotator (*individual*). Two types of the input features were compared in the test phase: manually annotated and automatically detected. Note that the manually annotated features were used for training in both cases. The number of characters ($K$) was tested from one to five. Table 4.2 shows the difference between the manual and automatic features. The accuracy is not degraded so much even when the input behavior features were automatically detected. Similar to the result in Section 3.5.2, the best number of characters ($K$) was 4 among the results of the automatic behavior detection. The paired t-test was performed on the proposed model with the four characters ($K = 4$), and there was no significant difference between the cases of the manual and automatic features ($t(99) = 1.45$, $p = 1.51 \times 10^{-1}$). This result indicates that the engagement recognition model

Table 4.2 Engagement recognition accuracy of online processing

| method | | behavior detection type | |
|---|---|---|---|
| | | manual | automatic |
| logistic regression | *majority* | 0.670 | 0.663 |
| | *individual* | 0.681 | 0.678 |
| SVM | *majority* | 0.667 | 0.668 |
| | *individual* | 0.683 | 0.673 |
| multilayer perceptron | *majority* | 0.678 | 0.653 |
| | *individual* | 0.690 | 0.681 |
| latent character (proposed) | $K = 1$ (no character) | 0.674 | 0.663 |
| | $K = 2$ | 0.697 | 0.700 |
| | $K = 3$ | 0.703 | 0.690 |
| | $K = 4$ | 0.711 | 0.700 |
| | $K = 5$ | 0.703 | 0.697 |

can be applied in live spoken dialogue systems. Note that all detection models can run in real time with short processing time which does not affect the decision-making process in spoken dialogue systems.

## 4.5 Summary

This chapter has addressed automatic detection of the listener behaviors and how to integrate them into the engagement recognition model in order to realize online engagement recognition for live spoken dialogue systems. For detection of backchannels and laughing, acoustic features were modeled by bi-directional long short-term memory with connectionist temporal classification (BLSTM-CTC). BLSTM can capture the context information of the input sequence feature. The advantage of using CTC is that it is not needed to have time-wise annotation of the listener behaviors. For head nodding, an LSTM model was used with visual features captured by the Kinect v2. Detection of eye gaze was implemented with a rule of cross detection. The experimental result on the detection of each behavior showed that each behavior can be detected by reasonable accuracy. Afterward, the detection results were integrated into the engagement recognition model. Experimental results showed that online engagement recognition was achieved without degrading accuracy. This real-time engagement recognition system will be used in a dialogue experiment explained in Chapter 7. A visualization tool for real-time engagement recognition was also implemented as shown in Fig. 4.3. This tool visualizes not only the current engagement recognition result but also the


Fig. 4.3 Visualization tool for online engagement recognition

recognition history so that it can show how and when the user engaged or not in the whole dialogue.

# Chapter 5

# Engagement Recognition Aggregating Different Annotators' Models

This chapter addresses a different model for engagement recognition implemented by a neural network. By taking advantages of neural networks, the proposed model considers both the difference among annotators and recognition of the integrated label.

## 5.1 Introduction

This chapter addresses engagement recognition using a neural network. In the annotation work done by third-party annotators, the ground truth label of engagement is sometimes inconsistent between the annotators because the perception of engagement is subjective and may depend on each annotator. Therefore, it is important to consider the difference between the annotators. However, it is sometimes needed to recognize the integrated label like majority voting in order to utilize the recognition result for applications such as spoken dialogue systems.

A neural network is proposed to predict the integrated label by considering the different annotators' labels. Figure 5.1 depicts the overview of the proposed method. The proposed neural network consists of two parts. The first part corresponds to recognition of each annotator's label. The second part aggregates the results of the first part to recognize the integrated label. Each part is pre-trained one by one, and then the whole network is fine-tuned. It is expected that the pre-training and fine-tuning lead to efficient training for the network so that the recognition accuracy is improved. The proposed network is more natural modeling in that each annotator's model is captured and aggregated to recognize the integrated labels,

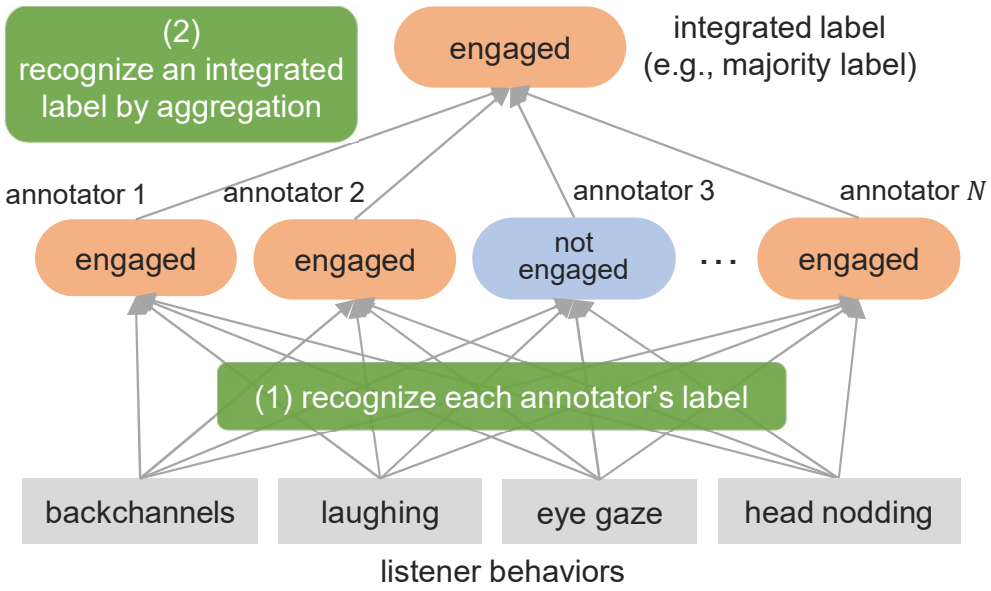


Fig. 5.1 Overview of the proposed neural network

and this study contributes to studies on recognition tasks with human subjectivity such as emotion recognition.

## 5.2 Neural Network Aggregating Different Annotators' Models

Each annotator's recognition should be individually modeled in the case where the recognition task has subjectivity. The proposed neural network consists of two parts where the first part corresponds to each annotator's recognition model and the second part aggregates the different annotators' models to recognize the majority label. Similar models that aggregate different annotators' models were proposed based on this two-step approach [143, 144]. These models used conditional random fields (CRF), while the proposed model uses neural networks, with the advantage being that the whole network is fine-tuned in the end-to-end manner.

### 5.2.1 Problem Formulation

Engagement recognition is done for each turn of the robot. The input is based on listener behaviors of the user during the turn: backchannels, laughing, eye-gaze, and head nodding. Table 5.1 summarizes the used feature set. Backchannels are distinguished into two types: responsive interjections (such as "*yeah*" in English and "*un*" in Japanese) and expressive

Table 5.1 Feature set of listener behaviors

| behavior type | feature |
|---|---|
| backchannels | (1) # of responsive interjections |
| | (2) # of expressive interjections |
| laughing | (3) # of laughing |
| eye-gaze | (4) time ratio of gaze toward robot |
| | (5) total time of gaze toward robot |
| | (6) # of gaze switching toward robot |
| head nodding | (7) # of head nodding |

interjections (such as "*oh*" in English and "*he-*" in Japanese) [139]. These features were manually annotated on the human-robot interaction corpus. The output is binary, engaged or not, for each turn. The reference label is based on the majority which was voted from labels of the five annotators.

### 5.2.2 Network Architecture

The proposed network consists of two parts as illustrated in Figure 5.2. Each part is realized by a linear interpolation of a GRU (gated recurrent unit) [169] and a linear transformation, inspired by the recurrent high-way network [170]. It is expected that the GRU captures context information through turns and the linear transformation captures local information of the current behavior input.

The network architecture is as follows. The input vector is represented as:

$$\boldsymbol{X}_t = (x_{1t}, \cdots, x_{bt}, \cdots, x_{Bt})^T \ , \tag{5.1}$$

where $x_{bt}$ represents the feature value of the $b$-th behavior in the $t$-th robot turn. $B$ is the number of behavior features ($B$=7 in this case). In the first part, a combination of a GRU and a linear transformation is prepared for each annotator, and it is trained to recognize each annotator's label. For the $n$-th annotator's model, the input vector is fed into the three functions as:

$$Z_{tn} = \sigma(\mathrm{GRU}_n(\boldsymbol{X}_t)) \ , \tag{5.2}$$

$$T_{tn} = \sigma(W_{Tn}\boldsymbol{X}_t + b_{Tn}) \ , \tag{5.3}$$

$$C_{tn} = \sigma(W_{Cn}\boldsymbol{X}_t + b_{Cn}) \ , \tag{5.4}$$

where $n \in \{1, \cdots, N\}$ represents the annotator index, $\sigma(\cdot)$ is the sigmoid function, and GRU$(\cdot)$ is the gated recurrent unit that stores a context hidden state inside it. $N$ is the number of total annotators in the training data ($N$=12 in this case). Afterwards, the outputs of the GRU and the linear transformation is linearly interpolated by the weight parameter as:

$$Y_{tn} = C_{tn} \cdot T_{tn} + (1 - C_{tn}) \cdot Z_{tn} \ . \tag{5.5}$$

This output corresponds to the recognition result of the $n$-th annotator in the $t$-th turn. In the second part, the outputs of all the annotators' models are concatenated as:

$$\boldsymbol{Y}_t = (Y_{t1}, \cdots, Y_{tn}, \cdots, Y_{tN})^T \ . \tag{5.6}$$

The same combination of a GRU and a linear transformation is applied to the concatenated vector to aggregate the different outputs as:

$$Z_t = \sigma(\mathrm{GRU}(\boldsymbol{Y}_t)) \ , \tag{5.7}$$

$$T_t = \sigma(W_T \boldsymbol{Y}_t + b_T) \ , \tag{5.8}$$

$$C_t = \sigma(W_C \boldsymbol{Y}_t + b_C) \ . \tag{5.9}$$

Finally, the probability of engaged in the turn is obtained in the similar way as the first part as:

$$E_t = C_t \cdot T_t + (1 - C_t) \cdot Z_t \ . \tag{5.10}$$

If the output $E_t$ is larger than a threshold, the $t$-th turn is recognized as *engaged* otherwise *not engaged*.

### 5.2.3 Pre-training and Fine-tuning

To make the first part of the proposed method realize each annotator's recognition and to make the second part aggregate different annotators' models, each part is pre-trained one by one. Although the whole network can be directly trained by using the integrated labels, it is expected that the pre-training leads to more efficient learning. Furthermore, the direct training sometimes falls into the local minima. Similar to the current case, it is often the case that each annotator partially gives ground-truth labels. In other words, some annotations of individual annotators are missing. In this case, the separated pre-training and fine-tuning would be a practical approach.

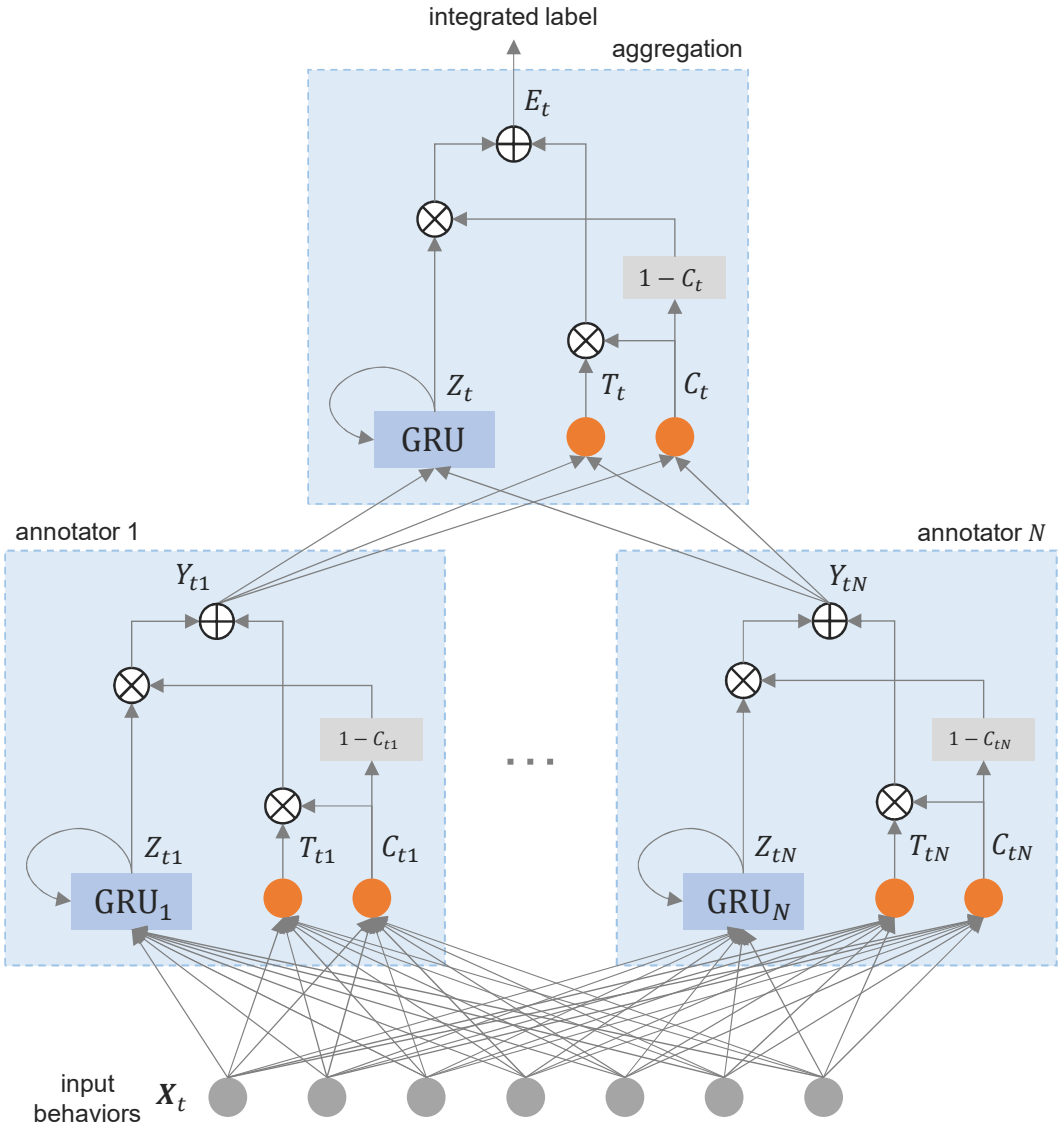


Fig. 5.2 Proposed network architecture

At first, the first part is trained by using each annotator's labels. Using the $n$-th annotator's labels, the $n$-th annotator's model is trained. The cost function is the squared error as

$$|Y_{nt} - Y'_{nt}|^2 , \tag{5.11}$$

where $Y'_{tn} \in \{0, 1\}$ is the reference label of the $n$-th annotator in the $t$-th turn. Second, after fixing the first-part parameters, only the second part parameters are trained by using the integrated labels. The cost function is also the squared error as

$$|E_t - E'_t|^2 , \tag{5.12}$$

where $E'_t \in \{0, 1\}$ is the integrated reference label (the majority label) in the $t$-th turn. Finally, the whole network is fine-tuned by using the integrated labels in the end-to-end manner. The cost function is same as the second part.

## 5.3 Evaluation

The proposed model is evaluated with cross validation of the 20 dialogue sessions, with 19 used for training and the rest for testing per fold. The input behavior features were manually annotated. The settings of the proposed model are as follows. The Adam algorithm was used as the optimizer [147]. The mini-batch size corresponded to one session. The number of epochs were 30 and 100 for the fine-tuning and the pre-training, respectively. The dropout rate was set at 0.2. The neural networks was implemented with Chainer 3.5.0[1] The reference labels were the majority labels where each session was annotated by five annotators. The probability of engaged (Eq. (5.10)) was calculated for each robot turn. The accuracy score was calculated by getting the decision threshold at 0.5. This score is a ratio of the number of the correctly recognized turns to the total number of the turns. The final evaluation was the averaged accuracy among all folds in the cross validation. The chance level was 0.617 (=267/433).

### 5.3.1 Effectiveness of Aggregating Different Annotators' Models

The proposed model is compared with some methods which do not consider the different annotators' models. The first method is a single-layer network using the combination of a GRU and a linear transformation, referred to as *one-layer*. This method was directly trained with the integrated labels. The second method is the same architecture of the proposed model but conducted only the fine-tuning directly without the pre-training, referred to as *two-layers (only fine-tuning)*. Additionally, a method without the fine-tuning was tested, referred to as *two-layers (only pre-training)*. This method partially corresponds to the earlier studies [143, 144].

The results are summarized in Table 5.2. The proposed method achieved the higher accuracy than the others. Comparing the one-layer model (B) with the two-layer models (C-E), the two-layer models had higher accuracies. This suggests that the more complicated network is effective to recognize the overall integrated label. Comparing the proposed method (E) with the fine-tuning only model (C), the pre-training is effective for this task and aggregating different annotators' models is important. Finally, the proposed method

[1] https://chainer.org

Table 5.2 Recognition accuracy

| method | accuracy |
|---|---|
| (A) chance level | 0.617 |
| (B) one-layer | 0.639 |
| (C) two-layers (only fine-tuning) | 0.686 |
| (D) two-layers (only pre-training) | 0.685 |
| (E) proposed | 0.728 |

Table 5.3 Recognition accuracy without each feature on the proposed method

| eliminated feature | accuracy |
|---|---|
| nothing | 0.728 |
| (1) # of responsive interjections | 0.730 (▲ 0.002) |
| (2) # of expressive interjections | 0.712 (▼ 0.016) |
| (3) # of laughing | 0.696 (▼ 0.032) |
| (4) time ratio of gaze toward robot | 0.724 (▼ 0.004) |
| (5) total time of gaze toward robot | 0.698 (▼ 0.030) |
| (6) # of gaze switching toward robot | 0.697 (▼ 0.031) |
| (7) # of head nodding | 0.713 (▼ 0.015) |

(E) showed higher accuracy than the pre-training only model (D). This gain is derived from the advantage of using neural networks where the whole network can be fine-tuned in an end-to-end manner.

### 5.3.2 Identifying Important Features

The effect of each behavior feature was also examined. The proposed model was tested by eliminating each of the features. The results are reported in Table 5.3. Note that the greater the decrease in accuracy, the more effective that feature is for the model. For backchannels, the responsive interjections (1), such as "*yeah*", was not so useful for the model. On the other hand, expressive interjections (2), such as "*oh*", are important because they are reactions toward what was said and better reflects the interest of the listener. laughing (3) is the most effective, so this behavior is important for engagement recognition. With regard to eye-gaze, while time ratio (4) was not effective, the total time duration (5) and number of gaze switching (6) were effective. Finally, head nodding (7) was also effective.

## 5.4 Summary

A neural network for engagement recognition has been proposed. Since perception of engagement is subjective, the ground-truth data depends on each annotator. A network aggregating different labels may help recognition of integrated labels (e.g., majority voting). The proposed method consists of two parts - recognition of each annotator's label, and aggregation of different annotators' models to recognize the overall integrated label. Each part is realized by the linear combination of the GRU and the linear transformation. The first part was pre-trained with the different annotators' labels, and the second part was also pre-trained with the integrated labels by fixing the parameters of the first part. Afterward, the whole network was fine-tuned. The experimental result shows that pre-training both each annotator's model and aggregation of different annotators' models is effective. The result also indicates that laughing and some eye-gaze features are informative in this engagement recognition task, followed by expressive interjections and head nodding.

# Chapter 6

# Implementation on Autonomous Android ERICA

This chapter addresses an application of engagement recognition. At first, the android robot ERICA together with specific social roles is introduced. Afterward, the real-time engagement recognition is implemented in a spoken dialogue system of laboratory guide by ERICA.

## 6.1 Introduction

It is important for conversational robots to adaptively generate actions and behaviors of robots in accordance with user engagement. As mentioned in Section 2.3, a few studies have been made on the adaptive generation. Here, the real-time engagement recognition model is applied to a live spoken dialogue system. This chapter argues the effectiveness of engagement recognition through subject experiments.

The research and development of the android robot ERICA have been conducted [132, 171]. ERICA is designed as a 23-year-old woman. Her design concept is to contain both the friendliness as an android robot and a sense of existence as a human being. The appearance of her face and body is artificially produced in reference to characteristics of beautiful ladies. ERICA mounts 46 active joints inside to move her face, head, shoulder, arms, hands, and back. The flexibility of her face has diversity, including eyebrow, eyelids, lip, eyeballs, and tongue. This enables the robot to generate various facial expressions. Therefore, ERICA is able to generate not only verbal responses but also non-verbal behaviors such as facial expression, eye gaze, head nodding, and gestures, in order to convey a variety of her internal states.

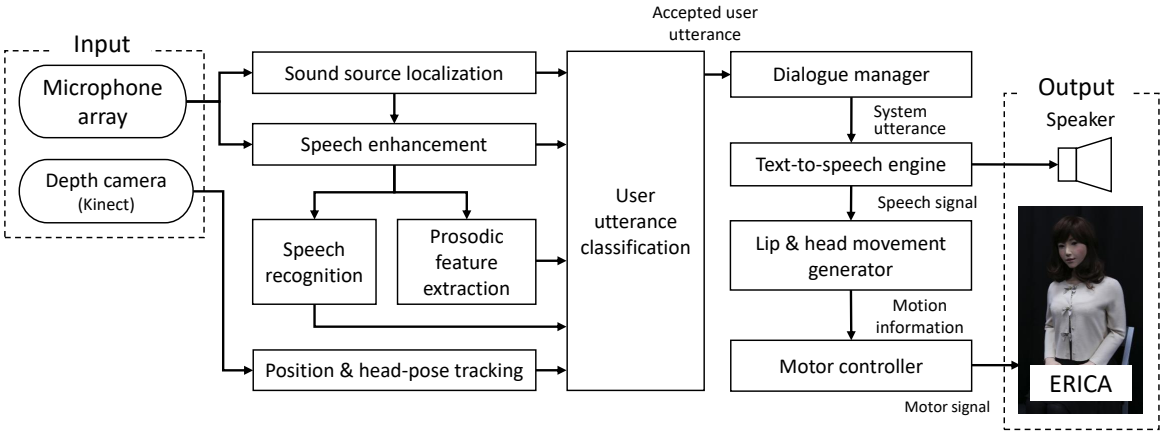


Fig. 6.1 System architecture of ERICA

It is expected for ERICA to realize the natural dialogue by integrating the human-like appearance with various kinds of behaviors. However, it is challenging to cover any types of conversations from scratch. Accordingly, at first, ERICA is given some specific social roles, aiming to realize the natural dialogue under each occasion. In the current setting, considered social roles are attentive listening, job interview, speed dating, and laboratory guide. Especially in the role of laboratory guide, ERICA mainly explains a laboratory to a user. In other words, ERICA holds the initiative of the dialogue and the ratio of utterance amount of ERICA is high. In this case, it is possible to utilize engagement recognition based on listener behaviors of the user so that ERICA can control the dialogue flow and her behaviors adaptively in order to realize the more user-friendly guidance. The advantage of using the android robot ERICA in this dialogue is to be able to generate various behaviors such as backchannels, head nodding, and eye gaze. These robot behaviors elicit user behaviors so that user engagement can be measured in the same manner as the case of the natural human-human dialogue.

## 6.2 System Architecture

A multimodal interactive system is implemented for ERICA. Figure 6.1 illustrates the entire configuration. The input sensors consist of a microphone array and a depth camera. These sensors are located around ERICA, not on the android, which increases the degree of freedom of sensor arrangement.

### 6.2.1 Speech Localization and Recognition with Microphone Array

The microphone array captures multi-channel audio signals and identifies which direction the acoustic signal comes from. Here, the sound source localization is realized by the multiple signal classification (MUSIC) method [172–174] with a 16-channel microphone array. Afterward, the input speech is enhanced by using the delay-and-sum beamforming. Prosodic information including fundamental frequency (F0) and power is calculated from the enhanced speech.

The automatic speech recognition (ASR) in ERICA is done by using the enhanced speech. Distant speech recognition elicits more natural human behavior because the user can use their arms and hands to show gestures. The used ASR is based on the acoustic-to-word end-to-end model that integrates the acoustic and language models into one neural network by utilizing the attention mechanism [175]. One of the advantages of making use of the above end-to-end model is that the processing time is quite fast so that the entire dialogue system can be stable even if the user talks a lot quickly. To realize the distant speech recognition, the training data consisted of a multi-condition speech audio where various noises and reverberations are contaminated.

### 6.2.2 Speaker Tracking with Depth Camera

To realize smooth interaction, it is essential for the system to correctly identify who talks to whom and if the user is giving his/her attention toward ERICA. The system tracks the location of users surrounding ERICA and also tracks those head orientations in the 3D space by using the Kinect v2 sensor. The user localization enables ERICA to spot if there is a person who wants to interact with her. ERICA identifies if the user is speaking to her by the head orientation. This enables ERICA not to respond to the talking between people, for example when a person introduces ERICA to a guest standing in front of ERICA. ERICA accepts user utterances when the following are met: the user is standing in front of ERICA and looking at ERICA's face, and the sound source is coming from the direction of the user. This function is needed when a demonstration is conducted for many people such as open laboratory events.

### 6.2.3 Text-To-Speech for ERICA

The speech of ERICA is generated by a text-to-speech engine that was developed for ERICA. It is based on the unit-selection framework from a database of many conversational-style utterances. It also contains many formulaic expressions and backchannels with a variety of

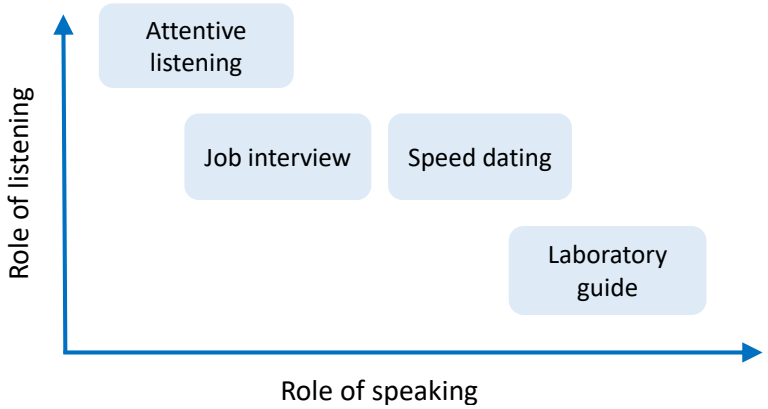


Fig. 6.2 Social roles covered by ERICA

prosodic patterns. At the same time, lip and head movements of ERICA are generated based on the prosodic information of the synthesized speech signals [134, 135].

## 6.3 Social Roles Played by ERICA

ERICA is aimed to realize natural and smooth dialogue with humans. To this end, some specific social roles are given to ERICA. Figure 6.2 illustrates the social roles on the two axes representing the trade-off relation: roles of listening and speaking. It is important for ERICA to cover the wide range of these axes comprehensively. As the case where the role of listening is dominant, attentive listening and job interview are considered. As the opposite case where the role of speaking is dominant, laboratory guide is addressed. As the mixed case where the roles of listening and speaking are balanced, speed dating is considered. Dialogue data also has been collected where the tele-operated ERICA was playing each social role by interacting with a human subject. The data will be used to train dialogue models for autonomous systems. In the long term, ERICA is expected to replace these human beings with comparable performance. The following explains the system and feasibility for each social role.

### 6.3.1 Attentive Listening

Attentive listening is to listen to the user talk carefully and to give feedbacks occasionally in order to encourage the user to talk more. A function of attentive listening is needed for counseling systems [66] and also casual conversation partner systems for elderlies. The attentive listening function consists of generating both bacchanals and additional listener feedbacks. To generate appropriate backchannels, it is needed to predict the timing and form

of the backchannel depending on the user utterance. Backchannel forms have a variety of different functions: one is to encourage the user to keep talking (*continuer*), and the other is to show a reaction to the user utterance (*assessment*) [176]. The automatic prediction on both the timing and form of backchannels have been studied and implemented [177, 178]. The additional listener feedbacks are short linguistic utterances toward the content uttered by the user, such as repeating a focus word, elaborating questions, evaluating responses, and lexical responses [179]. At first, a focus word of the user utterance is extracted by a heuristic rule. The focus word is directly used to generate a feedback of repeating the focus word. The elaborating question is generated by considering a pair of each interrogative and the focus word. For example, when the focus word is *movie* and an interrogative *which* is selected, a question is consequently generated like "*Which movie did you watch?*" The evaluating response is based on the sentiment polarity (positive or negative) of the user utterance. Each word of the user utterance is compared with sentiment word dictionaries[1] [180]. If the user utterance contained positive words, ERICA would say such as "*That is good.*" and "*That is nice.*" On the other hand, if the user utterance contained negative words, utterances of ERICA would be such as "*That is bad.*" The lexical responses are fixed responses such as "*I see.*" and "*It is right.*" Note that the uttered lexical response is selected based on the end form of the user utterance. Since each module generates each response independently, it is needed to select the suitable one. Here, it is selected by a rule based on the user silence. For example, if the system generates both repeating the focus word and an elaborating question, the current rule at first utters repeating the focus word. After that, if the user in silence for several seconds or if the user says only short utterances such as fillers, the system utters the elaborating question. Note that the above modules are implemented to process in real time.

A dialogue example is as follows. Note that **U** and **E** correspond to utterances by the user and ERICA, respectively. Besides, during the user utterances, ERICA generates backchannels at an appropriate timing.

**U1** I am going to give a talk about my traveling to Europe last year.

**E1** Europe. (Repeating the focus word)

**U2** Yeah. At first, I went to France.

**E2** France. (Repeating the focus word)

**U3** Yeah. There are a lot of places for sightseeing. Then, I went to a museum there.

**E3** A museum. (Repeating the focus word)

**E4** Which museum was it? (Elaborating question)

**U4** It was the Louvre museum.

[1] http://www.gsk.or.jp/en/catalog/gsk2011-c/

**E5** I see. (Lexical response)

**U5** I really enjoyed so many painting and sculptures there.

**E6** That is nice. (Evaluating response)

### 6.3.2 Job Interview

The job interview dialogue is organized and is system-initiative where the interviewer (ERICA) can control the dialogue flow. Therefore, the job interview dialogue can be implemented with such as the finite state transition. Although conventional systems for job interview could make questions that were defined in advance [69], the proposed system aims to dynamically generate elaborating questions according to interviewee (user) utterances. It is expected that the dynamic elaborating questions make the interviewee more diligent and engaged in the dialogue so that the system can elicit enough amount of information from the interviewee for the decision-making of the result of the job interview. A set of elaborating questions is prepared in advance, but the suitable one is automatically selected according to the preceding interviewee's response. For example, if the interviewee had already mentioned, the interviewer would not select an elaborating question that asks about the already-mentioned topic. Furthermore, to dynamically generate an elaborating question, a keyword is extracted from the interviewee's response. Then, the system asks the interviewee to explain about the keyword. The extracted keyword is the frequent and rare word measured by such as tf-idf [181]. Note that the above modules are independent of the kind of industry and company so that the system can be applied to any situations of the job interview.

A dialogue example is as follows.

**E1** What is the reason why you applied to our company?

**U1** I applied to this company because I studied informatics in my university.

**E2** I see. Do you have any experiences that are useful in this company? (Elaborating question)

**U2** My research theme was about deep learning. I have studied deep learning to solve some problems.

**E3** That is interesting. Could you explain deep learning? (Elaborating question with a keyword)

**U3** Well, it is a kind of machine learning method to train a non-linear prediction/discriminative model from a large amount of data.

**E4** I see. Thank you.

### 6.3.3 Speed Dating

Speed dating is a dialogue between a male and a female who meet each other for the first time. They talk about such as their hobbies and profiles to identify if they will be able to keep the relationship in the future. Some studies were made on analyses and prediction in a speed dating dialogue corpus [182, 183]. People are expected to be more engaged in the speed dating dialogue than the case of normal chatting because there exists a clear purpose of the dialogue. The range of conversational topics can be restricted to some that people use in the first-encounter dialogue. However, it is difficult to implement the speed dating dialogue system due to the mixed initiative where the initiative is frequently exchanged among them.

A preliminary dialogue system for speed dating was implemented with ERICA. At first, conversational topics were defined. These topics are thought to be frequently used in the first-encounter dialogue. Then, a response table for ERICA is prepared to respond to user questions. An ERICA's response was prepared for each pair of topic and interrogative. For example, for the pair of *hobby* and *what*, the ERICA's response is "*My hobby is reading books.*". A question sentence of ERICA was also prepared on each pair so that ERICA can make the question on the corresponding topic. Here, a dialogue system was implemented for ERICA making a question by taking the initiative when the user is silent for several seconds or that a configurable internal state of ERICA such as liking is higher. To handle the mix-initiative dialogue, a dialogue act classification was utilized in order to classify the user utterance into a question or others [184]. If the user utterance contains a question, the system makes a question with the response table. Otherwise, ERICA says a response towards the user statement, such as "That is nice". It is an issue of how to handle other topics where ERICA's utterances are not prepared. If ERICA says the same backup utterance such as "*Sorry. I do not know.*" every time for this case, the user would be disengaged in the dialogue. Therefore, it is needed to generate other backup utterances that can keep user engagement.

A dialogue example is as follows.

**U1** What is your hobby?

**E1** My hobby is watching movies.

**U2** What did you watch recently?

**E2** I watched *Star Wars* last week.

**E3** What movie did you watch?

**U3** Well, I watched *Alice in Wonderland* last month.

**E4** That is nice.

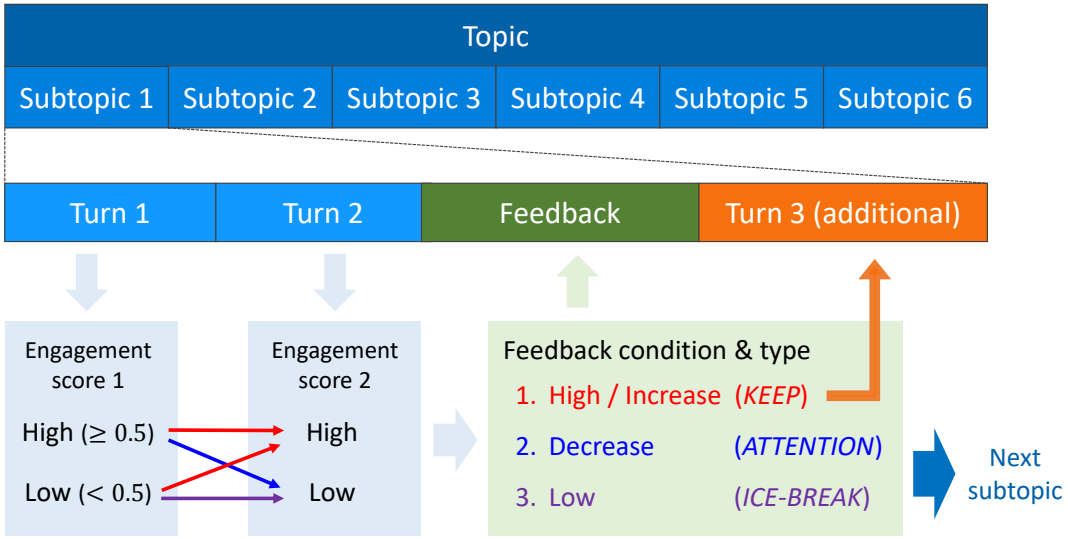


Fig. 6.3 Structure of dialogue and adaptive behaviors

### 6.3.4 Laboratory Guide

In the role of laboratory guide, ERICA holds the dialogue initiative and talks in most of the time. Some studies on information navigation have been made [67, 185]. Although it is easier to implement the autonomous system of a laboratory guide, users tend to be disengaged in the dialogue because the users say few utterances. In this study, a dialogue system that tracks user engagement is implemented by utilizing the real-time engagement recognition model. According to user engagement, the system adaptively generates a feedback towards the user in order to increase and keep user engagement. This system will be explained in the next section.

## 6.4 Adaptive Behaviors based on Engagement Recognition

A dialogue system for ERICA was implemented to play a role of the laboratory guide so that ERICA utilizes the real-time engagement recognition model. The dialogue content of the laboratory guide included several research topics. The dialogue structure of each topic is illustrated in Fig. 6.3. Each topic consisted of 6 subtopics where each subtopic corresponds to each research theme. Each subtopic included two ERICA's turns. In each turn, an engagement score was measured by the recognition model, and the result was regarded as a binary: high or low, by a threshold of 0.5. ERICA was implemented to generate feedback responses according to the combination of the two engagement scores from the continuous two turns. In this case, three kinds of feedbacks were prepared. When both scores were

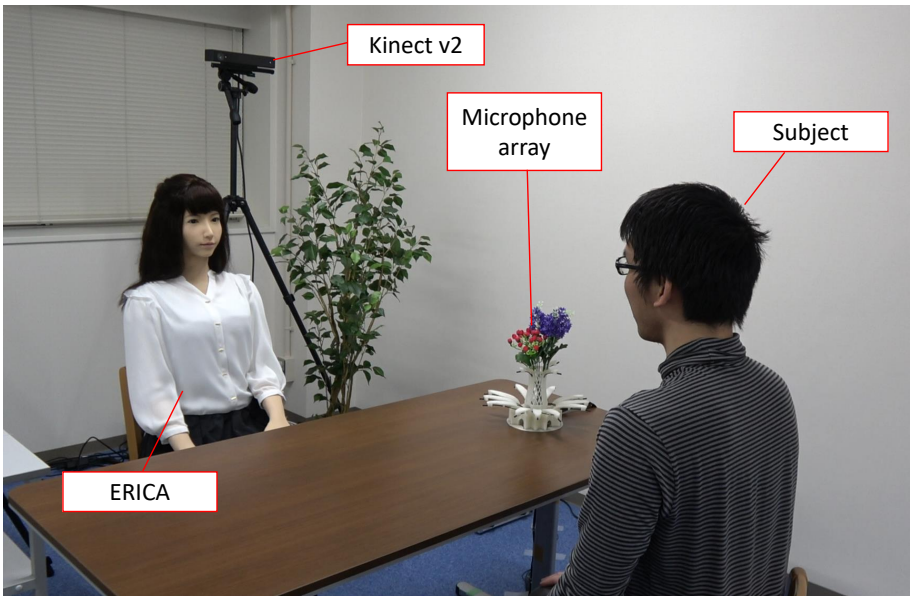


Fig. 6.4 Snapshot of dialogue experiment

high or the scores change from low to high (increase), ERICA said feedbacks like "*You seems to be interested in my talk. I am delighted.*" in order to keep the current engagement (*KEEP* feedbacks). When the scores changed from high to low (decrease), ERICA said a different type of feedbacks such as "*Are you tired? I will explain it in easier words.*" to gain attention from the user (*ATTENTION* feedbacks). When both scores were low, ERICA said another type of feedbacks such as "*Are you nervous? Please relax.*" to ease the tension of the user like ice-breaking (*ICE-BREAK* feedbacks). Furthermore, when after a *KEEP* feedback, ERICA gave an additional explanation to the user to elaborate the current subtopic (Turn 3). Note that when after an *ATTENTION* or *ICE-BREAK* feedback, ERICA shifted the dialogue to the next subtopic. It was expected that these adaptive behaviors including the feedbacks and the additional turns made the user more engaged in the laboratory guide.

A subject experiment was conducted in order to confirm the effectiveness of engagement recognition and the adaptive behaviors. Fig. 6.4 is the snapshot of the dialogue experiment. In this experiment, a subject and ERICA sat on each chair to face each other. To elicit listener behaviors of the subjects, ERICA generated backchannels and head nodding automatically during the subjects' turns [178]. The subjects were 11 persons (6 males and 5 females) who were recruited inside the university. The experiment procedure is as follows. At first, each subject had a practice dialogue to get used to talking with ERICA. They practiced a simple interaction consisting of several turns. Then, each subject talked about two research topics: automatic speech recognition and spoken dialogue systems. The order of dialogue topics was randomized among the subjects. Two experiment conditions were prepared: *engagement* and

Table 6.1 Evaluation scale for internal user states (* means invert scale.)

| internal state | question |
|---|---|
| interest | (1) I felt the dialogue was boring. * |
| | (2) I want to talk about other topics more. |
| | (3) I was fascinated by the dialogue content. |
| | (4) I was not concerned with the dialogue content. * |
| continue | (5) I wanted to quit the dialogue during that. * |
| | (6) I think the dialogue should finish earlier. * |
| | (7) I wanted to continue the dialogue more. |
| willingness | (8) I wanted to make the dialogue fulfilling. |
| | (9) I could not feel like talking. * |
| | (10) I participated in the dialogue by concentrating on that. |
| | (11) I felt that what I needed to do was just being there. * |
| | (12) I actively answered the questions from the robot. |
| | (13) I actively responded to the robot talk. |
| trust | (14) I liked the robot. |
| | (15) I felt the robot was friendly. |
| | (16) I was relieved when I was having the dialogue with the robot. |
| | (17) I could trust the dialogue content the robot talked. |
| empathy | (18) I could understand what emotion the robot was having. |
| | (19) I could agree with the idea the robot had. |
| | (20) I could advance the dialogue by considering the viewpoint from the robot. |
| | (21) I could understand the reason why the robot had that kind of emotion. |

*control*. In the *engagement* condition, ERICA measured engagement scores and generated adaptive behaviors mentioned above. In the *control* condition, ERICA did not generate any adaptive behaviors, which meant only the turn 1 and 2 were explained. The dialogue of the first topic was conducted with the *control* condition, and the second topic was done with the *engagement* condition. The order of these conditions was not randomized because it was thought that the adaptive behaviors in the *engagement* condition would affect the subjects' behaviors and impressions in the subsequent topic. Among the two conditions, measured engagement scores were compared. On the setting of the engagement recognition model, the character distribution was based on the mapping from the personality traits of the perceiver (ERICA), and *extroversion* was set at the maximum value, which was expected to the laboratory guide.

Additionally, subjective evaluation on internal user states was made after they talked about each topic. The internal user states included *interest*, *continue*, *willingness*, *trust*, and

*empathy*. To measure them, concrete question items were considered, as listed in Table 6.1. Two related researchers independently validated each question item by considering both the relevance to the internal state and also the correctness of the question sentence. In the experiment, each subject evaluated each item in the 7 point scale. Finally, the evaluated scores were averaged for each internal state. As described in Section 2.1, these internal states are related to engagement so that it was expected that the scores of them would be improved by the adaptive behaviors in the *engagement* condition.

Average engagement scores are reported in Fig.6.5. Since each topic consisted of 6 subtopics and each subtopic included two turns, the total number of turns was 12. Note that scores of the additional turns in the *engagement* condition are not included in this result. The difference between the two conditions was observed in the latter part of the dialogue. Paired t-tests were conducted on the scores of each turn between the two conditions. As a result, it turned out that the *engagement* condition significantly increased the engagement scores in 7-th and 9-th turns ($p < .01$). In addition, the tendencies of improvement were also observed in 8-th, 10-th, and 12-th turns ($p < .1$). This result suggests that the adaptive behaviors in the *engagement* condition made the subjects more engaged in the dialogue. Accordingly, engagement recognition is the important function in social human-robot dialogue such as the laboratory guide.

Average subjective scores on internal user states are reported in Fig. 6.6. For each internal state, a paired t-test was conducted between the two conditions. The results show that the scores of *interest* and *empathy* significantly increased in the *engagement* condition ($p < .05$ for *interest* and $p < .01$ for *empathy*). One possible reason is that the additional explanation made the subjects more be interested in the research topic. Besides, the feedback responses were perceived as emotional expressions of ERICA so that the perceived empathy scores increased.

## 6.5 Summary

This chapter has addressed applications of engagement recognition for the android robot ERICA. Specific social roles are set for ERICA as attentive listening, job interview, speed dating, and laboratory guide. The real-time engagement recognition model was applied to the dialogue of laboratory guide. In the laboratory guide, since ERICA talks in the most of the time, ERICA needs to track user engagement based on listener behaviors while ERICA is speaking. ERICA was implemented to adaptively generate feedback responses and additional explanations by measuring user engagement. The experimental results showed that the

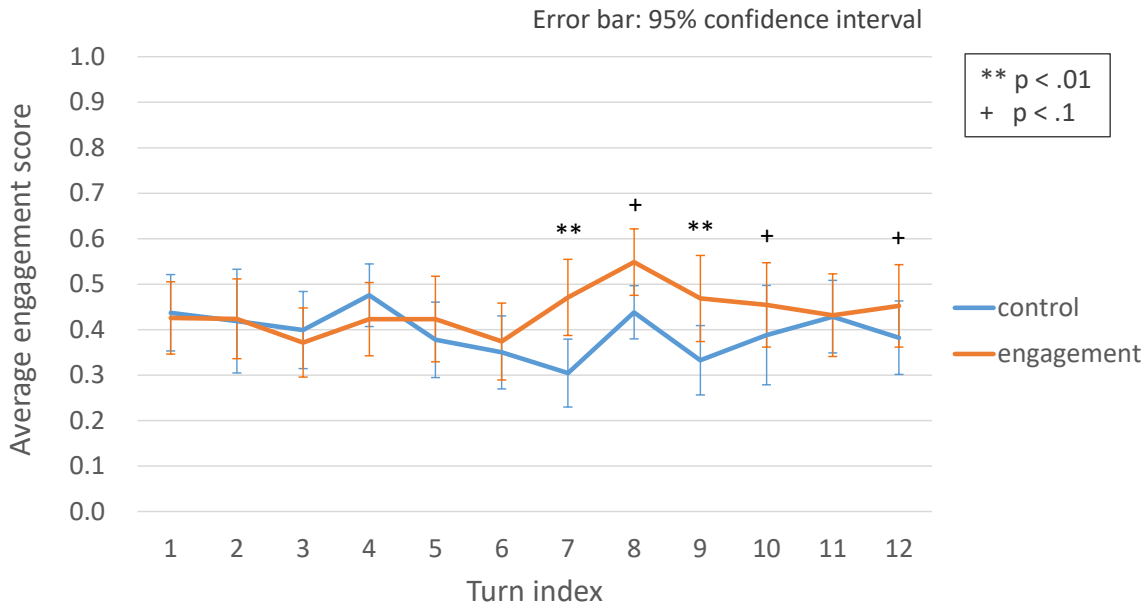


Fig. 6.5 Engagement scores during dialogue

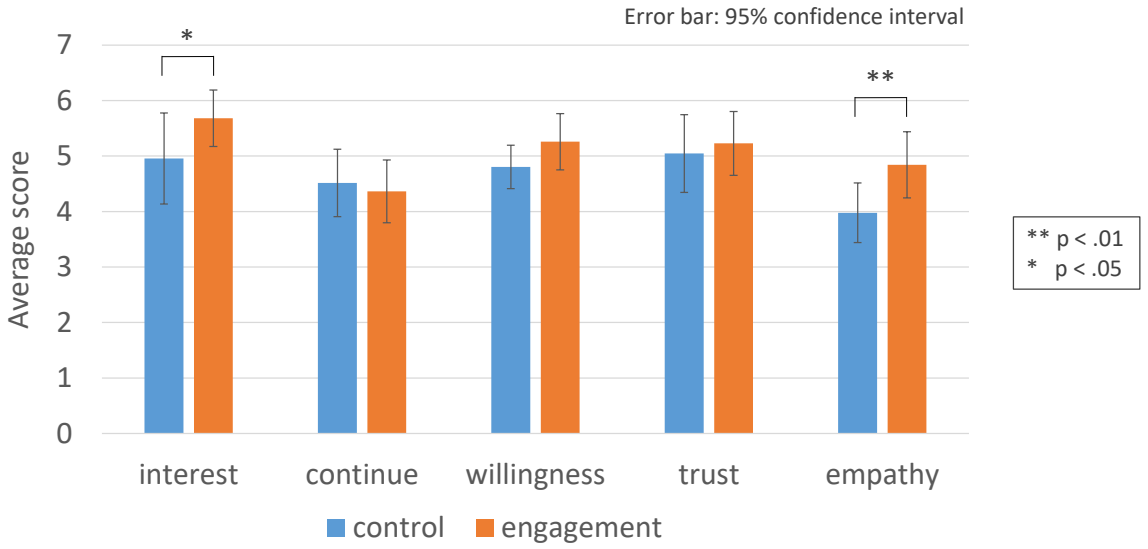


Fig. 6.6 Subjective evaluation on internal user states

adaptive behaviors increased the engagement scores and internal user states of interest and empathy.

# Chapter 7

# Speaker Diarization using Eye-gaze Information

This chapter addresses speaker diarization toward multi-party conversations. A multimodal speaker diarization method using eye-gaze information is proposed.

## 7.1 Introduction

Handling multi-party conversations is a challenging task for conversational robots. Speaker diarization is to identify "who spoke when" in multi-party conversations and is an important function for conversational robots that participate in multi-party conversations. A number of diarization methods have been investigated from the standpoint of acoustic signal processing [186–188]. In real multi-party conversations, the diarization performance is degraded by adversary acoustic conditions, such as background noise [186] and distant talking [189]. To solve the problem, some studies tried to incorporate multimodal data such as motion and gesture [190, 188, 191].

A multimodal diarization method is proposed in order to integrate eye-gaze information with acoustic information. The eye-gaze behavior plays an important role in turn-taking in multi-party conversations [36, 192]. For example, it is often observed that eye-gaze directions of participants tend to intersect each other when they change the conversational turn. Although it is known that eye-gaze information can be used to predict participants' utterances [37, 193, 39], eye-gaze information has not yet been integrated in speaker diarization tasks. The proposed method extracts acoustic and eye-gaze features, which are stochastically integrated to detect utterances. The findings of this study would be helpful

Table 7.1 Total utterance duration [sec.]

| session ID | *presenter* | audience 1 | audience 2 | total |
|---|---|---|---|---|
| 140206-01 | 1,251 | 19 | 227 | 1,497 |
| 140206-02 | 1,406 | 283 | 164 | 1,853 |
| 140206-03 | 1,333 | 328 | 170 | 1,831 |
| 140206-04 | 1,495 | 129 | 102 | 1,726 |
| 140207-01 | 1,343 | 164 | 123 | 1,630 |
| 140207-02 | 1,229 | 134 | 117 | 1,480 |
| 140207-03 | 1,205 | 106 | 267 | 1,578 |
| 140207-04 | 1,208 | 216 | 135 | 1,559 |
| total | 10,470 | 2,684 | | 13,154 |

to design multimodal spoken dialogue systems or conversational robots to realize smooth dialogues with a user.

## 7.2 Multimodal Corpus of Poster Conversations

Multimodal conversational data has been collected in poster sessions (= poster conversations), where a presenter made an interactive presentation to a small audience. To analyze poster conversations, "the smart posterboard" has been developed. This is a recording environment equipped with multimodal sensing devices such as a microphone array and cameras [194]. The smart posterboard system consists of a 19-channel microphone array, Kinect sensors, and HD cameras, which are attached to a large LCD, showed in Fig. 7.1. With this setting, eight poster sessions were recorded. In each session, one presenter made a poster presentation on his/her academic research, and there was an audience of two persons, standing in front of the poster and listening to the presentation. The duration of each session was 20 to 30 minutes. Speaker diarization and detection of backchannels are conducted with the sensors (the microphone array and Kinect) attached to the posterboard, so the participants do not have to wear any devices. For the ground truth annotation of the data used in this work, however, speech data were also recorded with a wireless headset microphone, and participants' head locations and orientations were captured by magnetometric sensors.

Table 7.1 summarizes the total utterance duration in the recorded sessions. While the presenter holds a majority of the turns, utterances by the audience (audience 1 and 2) are not frequent, which means it would be difficult to detect these utterances accurately. The audience's backchannels account for about 40 percent of their utterance duration.

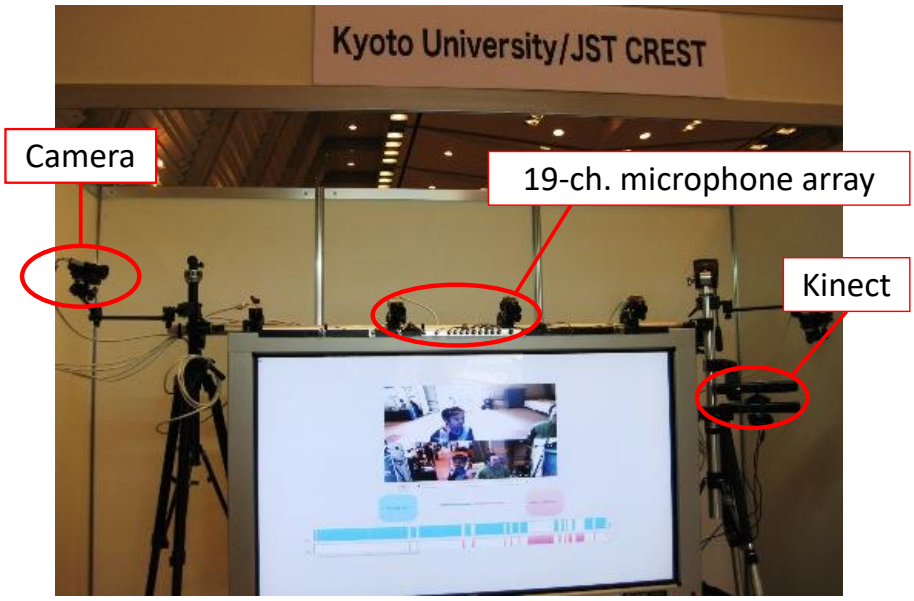


Fig. 7.1 Smart posterboard

## 7.3 Multimodal Speaker Diarization

This section describes the proposed speaker diarization method which integrates acoustic and eye-gaze information.

### 7.3.1 Baseline Acoustic Method

Conventional speaker diarization methods have used mel-frequency cepstral coefficients (MFCCs) [186, 188] and directions of arrival (DOA) of sound sources [195–198] as acoustic features. An acoustic baseline method in this study detects participants' utterances based on DOAs derived from the microphone array.

The multiple signal classification (MUSIC) method [172, 173] is used in order to detect multiple DOAs simultaneously. The MUSIC spectrum $M_t(\theta)$ is calculated based on the orthogonal property between an input acoustic signal and a noise subspace. Note that $\theta$ is an angle between the microphone array and the target of estimation, and $t$ represents the time frame. The MUSIC spectrum represents the DOA likelihood, and the large spectrum suggests that a participant makes an utterance from that angle. To calculate the spectrum, it is needed to determine the number of sound sources. In this study, the number of sound sources is predicted with an SVM using the eigenvalue distribution of a spatial correlation matrix [199].

### 7.3.2 Multimodal Method Integrating Eye-gaze Information

The proposed method incorporates eye-gaze information to speaker diarization. The method first extracts acoustic and eye-gaze features to compute a probability of speech activity respectively, then it combines the two probabilities for the frame-wise decision. The process is conducted independently on every time frame $t$ and for each participant $i$.

#### Acoustic Features

The acoustic features are calculated based on the MUSIC spectrum. The $i$-th participant's head location $\theta_{it}$ is tracked by the Kinect sensors. The possible location of the participant is constrained within a certain range ($\pm\theta_B$) from the detected location $\theta_{it}$. The acoustic features of the $i$-th participant in the time frame $t$ consist of the MUSIC spectra in this range as

$$\boldsymbol{a}_{it} = \left(M_t\left(\theta_{it}-\theta_B\right),\cdots,M_t\left(\theta_{it}\right),\cdots,M_t\left(\theta_{it}+\theta_B\right)\right) . \tag{7.1}$$

#### Eye-gaze Features

The eye-gaze direction used in this study is approximated by the head orientation estimated from RGB and depth images captured by the Kinect sensor. The head orientations are tracked by a particle filter [200]. The eye-gaze object is determined by the head-orientation vector and the location of the objects. In this work, the object is limited to the poster and the other participants.

The eye-gaze features are designed based on the eye-gaze objects of the $i$-th participant and a conversational partner. The eye-gaze feature vector $\boldsymbol{g}_{it}$ consists of the followings:

1. Eye-gaze object
   This feature represents which object the $i$-th participant looks at. For the presenter, ($P$) poster or ($I$) audience; For (anybody in) the audience, ($p$) poster or ($i$) presenter.

2. Joint eye-gaze event: "$Ii$", "$Ip$", "$Pi$", "$Pp$"
   This feature represents combination of the eye-gaze objects of both the $i$-th participant and the conversational partner. For example, "$Ii$" and "$Pp$" correspond to mutual gaze and joint attention, respectively. The definition of the event is summarized in Table 7.2.

3. Maximum duration of each state of the above 1.

4. Maximum duration of each state of the above 2.

5. Uni-gram and bi-gram of the above 1.
   This feature represents the transition of the eye-gaze objects.

Table 7.2 Definition of joint eye-gaze events

| | | presenter | |
|---|---|---|---|
| | gaze at | audience ($I$) | poster ($P$) |
| audience | presenter ($i$) | $Ii$ | $Pi$ |
| | poster ($p$) | $Ip$ | $Pp$ |

6. Uni-gram and Bi-gram of the conversational partner's eye-gaze objects

The first and the second features are calculated for the time frame $t$, and the others are measured within the preceding period (the preceding $C$ milliseconds).

**Multimodal Integration Model**

The acoustic features $\boldsymbol{a}_{it}$ are integrated with the eye-gaze features $\boldsymbol{g}_{it}$ to detect the $i$-th participant's speech activity $v_{it}$ in the time frame $t$. Note that the speech activity $v_{it}$ is binary: speaking ($v_{it} = 1$) and not speaking ($v_{it} = 0$). Here, the probabilities are linearly interpolated in the late fusion manner [188]:

$$f_{it}(\boldsymbol{a}_{it}, \boldsymbol{g}_{it}) = \alpha \; p(v_{it} = 1|\boldsymbol{a}_{it}) + \; (1-\alpha) \; p(v_{it} = 1|\boldsymbol{g}_{it}) \; , \tag{7.2}$$

The parameter $\alpha \in [0,1]$ is a weight coefficient. Each probability is computed by a logistic regression model. It is also possible to combine the two feature sets in the feature domain and directly compute a posterior probability $p(v_{it}|\boldsymbol{a}_{it}, \boldsymbol{g}_{it})$. Compared with this joint model, the linear interpolation model has a merit that training data does not need to be aligned between the acoustic and eye-gaze features because of independence of the two discriminative models. Furthermore, the weight coefficient $\alpha$ can be appropriately determined based on the acoustic environments such as signal-to-noise ratio (SNR). Here, it is estimated using an entropy $h$ of the acoustic posterior probability $p(v_{it}|\boldsymbol{a}_{it})$ [201, 202] as

$$\alpha = \alpha_c \times \frac{1-h}{1-h_c} \; , \tag{7.3}$$

where $h_c$ and $\alpha_c$ are an entropy and an ideal weight coefficient in a clean acoustic environment, respectively. When the estimated weight coefficient is larger than one or less than zero, the coefficient is set to one or zero, respectively. For online processing, the coefficient is updated every 15 seconds by using the preceding 15-second data.

## 7.4 Evaluation

The proposed method was compared with other methods that do not use eye-gaze information, using the corpus of the poster conversations described in Section 7.2.

### 7.4.1 Setup

Logistic regression models were trained separately for the presenter and the audience by cross-validation of the eight sessions. To evaluate one session, the other seven sessions were used for training. For the proposed multimodal method, the training data were also used for training of the SVM which determines the number of sound sources, and calculating the entropy $h_c$ of the clean acoustic environment. The constrained range of the MUSIC spectra of the acoustic features was 10 degrees ($\theta_B = 10^\circ$). This setting was intended to prevent overlapping between the participants. Since the MUSIC spectrum was calculated every 1 degree, the dimension of the acoustic features was 21. The scope to calculate the duration and bi- and tri-grams in the eye-gaze features was 1,000 milliseconds ($C = 1,000$). Frame rates per second of the acoustic and eye-gaze features are 62.5 and 29.5, respectively. Consequently, an interpolation using the nearest sample was done to the eye-gaze features. The ideal weight coefficient $\alpha_c$ (Eq. (7.3) ) in a clean environment was set to 0.9.

To evaluate performance under ambient noise, audio data are superimposed with a diffusive noise recorded in a crowded place on the 19-channel audio signals. SNRs were set to 20, 15, 10, 5, and 0 dB. In real poster conversations carried out in academic conventions, the SNRs are expected to be around 0 to 5 dB.

The proposed multimodal method was compared with other methods listed below:

1. *baseline MUSIC* [197]
   This method conducts peak tracking of the MUSIC spectrum and GMM-based clustering in the angle domain. Each cluster corresponds to each participant. This method does not use any cue from visual information.

2. *baseline with location constraint* [203]
   This method also performs peak tracking of the MUSIC spectrum, and compares the detected peak with the estimated head location within the $\theta_B$ range. If this constraint is not met, the hypothesis is discarded.

3. *acoustic-only model*
   This method fixes the weight coefficient $\alpha$ to 1 in Eq. (7.2), and uses only the acoustic information.

Table 7.3 Diarization error rate

| method | SNR [dB] | | | | | | ave. |
|---|---|---|---|---|---|---|---|
| | ∞ | 20 | 15 | 10 | 5 | 0 | |
| *baseline MUSIC* | 15.28 | 20.44 | 28.34 | 42.80 | 64.09 | 87.22 | 43.03 |
| *baseline with location constraint* | 7.76 | 13.66 | 21.18 | 35.49 | 55.27 | 77.81 | 35.20 |
| *acoustic-only model* | 6.52 | 7.60 | 9.63 | 14.20 | 22.33 | 34.34 | 15.77 |
| *multimodal model* | 7.35 | 8.55 | 10.73 | 14.23 | 18.21 | 21.22 | 13.38 |

For an evaluation measure, diarization error rate (DER) [204] was used in this experiment. DER consists of false acceptance (FA), false rejection (FR), and speaker error (SE) as

$$\mathrm{DER} = \frac{\#\mathrm{FA} + \#\mathrm{FR} + \#\mathrm{SE}}{\#\mathrm{S}} \times 100 \ ,$$

where #S is the number of speech frames in the reference data. This metric does not evaluate $\pm$ 250 milliseconds collars of the reference utterance segments. The minimum DER was evaluated by varying the threshold for Eq. (7.2) after smoothing (hangover) the detected utterance segments.

### 7.4.2 Results

Table 7.3 lists DERs for each SNR. The two baseline methods (*baseline MUSIC* and *baseline with location constraint*) showed lower accuracy than the stochastic methods (*acoustic-only model* and *multimodal model*) because the baseline methods are rule-based and not robust against dynamic changes of the MUSIC spectrum and the participants' locations. Compared with the acoustic-only model, the proposed multimodal model achieved higher performance under the noisy environments (SNR = 5, 0 dB). Meanwhile the acoustic-only model was slightly better than the multimodal model under clean environment because the acoustic information would be dominant in that case. Thus, the effect of the eye-gaze information can be seen under noisy environments expected in real poster sessions.

The weight coefficient $\alpha$ in Eq. (7.2) was also manually tuned with the step size of 0.1. In the clean environment (SNR was ∞ dB), the optimal weight was 1.0. On the other hand, in the noisy environments (SNRs were 5 and 0 dB), the optimal weights were 0.6 or 0.5. These results suggest that the weight of eye-gaze features is adequately increased in noisy environments. The average DER of the manual tuning was 11.78%, which is slightly better than the result (12.92%) by the automatic weight estimation (Eq. (7.3)). Therefore, the automatic weight estimation works reasonably according to the acoustic environment.

Table 7.4 Equal error rate on voice activity detection of presenter

| method | SNR [dB] | | | | | | ave. |
|---|---|---|---|---|---|---|---|
| | ∞ | 20 | 15 | 10 | 5 | 0 | |
| *baseline MUSIC* | 12.98 | 13.61 | 15.32 | 16.98 | 20.41 | 26.35 | 17.61 |
| *baseline with location constraint* | 20.41 | 19.84 | 20.17 | 21.50 | 24.19 | 28.14 | 22.38 |
| *acoustic-only model* | 9.81 | 10.11 | 11.40 | 14.31 | 19.03 | 25.19 | 14.98 |
| *multimodal model* | 10.47 | 10.19 | 11.46 | 14.47 | 19.71 | 25.61 | 15.32 |

Table 7.5 Equal error rate on voice activity detection of audience

| method | SNR [dB] | | | | | | ave. |
|---|---|---|---|---|---|---|---|
| | ∞ | 20 | 15 | 10 | 5 | 0 | |
| *baseline MUSIC* | 19.43 | 24.03 | 26.27 | 29.37 | 33.37 | 38.40 | 28.48 |
| *baseline with location constraint* | 22.90 | 25.55 | 28.46 | 29.52 | 33.77 | 38.57 | 29.80 |
| *acoustic-only model* | 19.82 | 23.07 | 25.35 | 28.47 | 32.66 | 37.76 | 27.86 |
| *multimodal model* | 19.86 | 23.08 | 25.25 | 28.02 | 31.84 | 36.12 | 27.36 |

Voice activity detection was separately evaluated on the presenter and the audience because there is a bias in utterance amount among them. The used evaluation metric was the equal error rate (EER) that is a value where the false acceptance ratio (FAR) equals the false rejection rate (FRR). FAR and FRR are calculated as

$$\mathrm{FAR} = \frac{\#\mathrm{FA}}{\#\mathrm{NS}} \times 100 \,, \tag{7.4}$$

$$\mathrm{FRR} = \frac{\#\mathrm{FR}}{\#\mathrm{S}} \times 100 \,. \tag{7.5}$$

Note that #NS is the number of non-speech frames in the reference data. For simplicity of the voice activity labels of the two audiences, a union set of them was used. For example, in the frame where if at least one audience is speaking, the reference label is positive. On the other hand, if no audience is speaking, the reference label is negative.

The results on EERs are listed in Table 7.4 for the presenter and in Table 7.5 for the audience. There is no improvement by the proposed method in the result of the presenter. Since the presenter speaks at most of the time in poster conversations, it is easier to detect the voice of the presenter. On the other hand, the proposed multimodal model improved the accuracy in noisy environments.

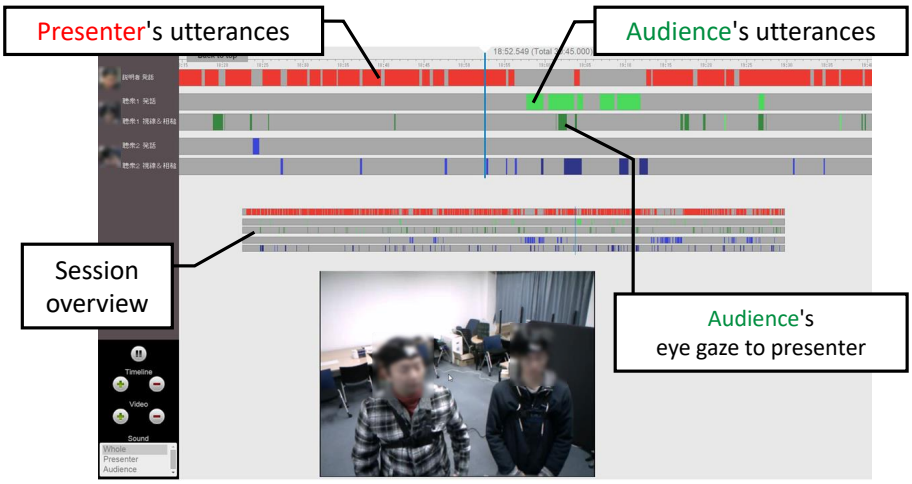


Fig. 7.2 Poster session browser

## 7.5 Summary

A multimodal speaker diarization method that integrates eye-gaze information with acoustic information has been proposed. The acoustic feature is based on sound source localization and the eye-gaze feature is calculated from the joint eye-gaze events. These features are stochastically integrated to detect the existence of utterances. The stochastic multimodal scheme improved the performance of speaker diarization and the effect of eye-gaze information is confirmed in the noisy environments which are expected in real poster sessions. It is expected that this multimodal speaker diarization will be implemented in the system of ERICA to handle multi-party conversations.

As an application to visualize the diarization results, a poster session browser was developed as showed in Fig. 7.2). This tool can be executed on Web browsers. This makes it possible to effectively review meaningful utterances such as questions and comments given by the audience, together with eye-gaze events in the poster conversation.

# Chapter 8

# Conclusion

This thesis has addressed engagement recognition for human-robot dialogue. This chapter reviews the contributions of this thesis and addresses the future directions.

## 8.1 Contribution

This thesis addressed engagement recognition based on multimodal listener behaviors such as backchannels, laughing, head nodding, and eye gaze. The goal of this study is to tackle the problem on the subjectivity of engagement that causes disagreement among annotators. A hierarchical Bayesian model for engagement recognition was proposed. It was assumed that each annotator has a different perspective on engagement. This was interpreted by a character that affects his/her perception of engagement. The proposed model represented the character as latent variables so as to recognize each annotator's labels precisely. This thesis also made a contribution to studies on recognition tasks that contain subjectivity such as emotion recognition in that the proposed model takes into account the difference of annotators.

Engagement recognition was also implemented by the neural network. The neural network has been applied little to the task of engagement recognition. In this thesis, the neural network was applied by utilizing the advantages of them. The proposed network represented the difference of annotators in the intermediate layer. The accuracy of engagement recognition was improved by both the representation of the difference of annotators and an effective training method. This neural network model is effective when a large amount of training data and a variety of input behavior features are used.

The automatic detection of the listener behaviors was also implemented. It is important for live conversational robots to recognize engagement in real time. Conventional engagement recognition systems used only limited behaviors that can be directly obtained from sensors like head direction, distance, and facial action units. In this thesis, backchannels, laughing,

and head nodding were detected by neural network models such as LSTM. These detection models were integrated with the engagement recognition model in order to realize real-time recognition. It was confirmed that the accuracy of engagement recognition was not degraded even the input listener behaviors were detected automatically. The real-time engagement recognition model was applied to the system of the android robot ERICA to generate adaptive behaviors in the dialogue of laboratory guide. It was also observed that the adaptive behaviors increased user engagement during the dialogue. This result suggests that engagement recognition is the important function for social human-robot dialogue.

In summary, in this thesis, several recognition models were proposed for the engagement recognition task that has subjectivity. Furthermore, it was also shown that engagement recognition was implemented into the live conversational robot. Therefore, the findings of this thesis contribute to the realization of symbiotic conversational robots that interact with humans in deep and long dialogue.

## 8.2 Future Work

In engagement recognition, since only four listener behaviors were used, it is expected to use a more variety of behaviors. For example, in this study, facial expression and body posture were not used. These were identified as effective features by previous studies. It is expected that this extension will lead to more accurate and robust engagement recognition. The number of training data should be increased by efficient annotation ways. In this case, neural network models would be more necessary to train from a large amount of data. Furthermore, it is also thought that shorter-term concepts such as entrainment can be used in order to recognize engagement more accurately. In contrast, engagement can be used to predict other longer-term concepts such as rapport. Another engagement recognition model can be considered. In this thesis, the models of engagement recognition and behavior detection were independently trained. It would be possible to train the integrated model by using neural networks in an end-to-end manner. However, it has to be solved that units of the recognition and the detection are different.

In adaptive behavior generation, although the heuristic rule was used in this thesis, it is expected to train a model from dialogue corpora. In this case, it is needed for the corpora to have reference labels of engagement. It is difficult to obtain the engagement labels by manual. If the corpora have the behavior data, the labels can be generated by the engagement recognition model. Then, the scheme of semi-supervised learning can be applied. Besides, the adaptively generated behaviors are not limited to linguistic features such as contents of robot utterances. It is expected that the robots control their listener behaviors such as

backchannels and head nodding according to user engagement. This leads to entrainment and natural interaction. Furthermore, it is also important to consider what the robots should do after the robots keep the users engaged in the dialogue. In this thesis, the goal of engagement recognition was to increase user engagement by the adaptive behaviors of the robot. This ultimate goal should be designed by considering the kind of dialogue task. For example, in the case of the laboratory guide, if a user keeps engaged in the dialogue, it would be possible that the robot can introduce researchers in the laboratory to the user.

Engagement can be used as an evaluation metric for dialogue. Recently, many studies have been made on machine learning to train dialogue models such as text generation [27–29]. These models use the whole dialogue corpus to train a model, and the objective function is how much the models generate the same system utterances compared to the training corpus. If engagement is used as the training metric for the dialogue model, the model would maximize engagement and is expected to realize more engaged dialogue. The automatic engagement recognition model can be also applied to this case as the scheme of semi-supervised learning. This idea can be applied to autonomous conversational robots. Engagement can be used to evaluate the current dialogue with real users. Then, the robot can update a rule of the dialogue strategy by using the engagement score as a reward in reinforcement learning. This framework allows the robot to train the ideal dialogue strategy by itself through interaction with real users.

# References


[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] L. Deng and D. Yu, "Deep learning: Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, pp. 1764–1772, 2014.

[4] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 577–585, 2015.

[5] L. Deng, G. Tur, X. He, and D. Hakkani-Tur, "Use of kernel deep convex networks and end-to-end learning for spoken language understanding," in *Workshop on Spoken Language Technology (SLT)*, pp. 210–215, 2012.

[6] A. Deoras and R. Sarikaya, "Deep belief network based semantic taggers for spoken language understanding.," in *Interspeech*, pp. 2713–2717, 2013.

[7] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding.," in *Interspeech*, pp. 3771–3775, 2013.

[8] K. Yao, G. Zweig, M.-Y. Hwang, Y. Shi, and D. Yu, "Recurrent neural networks for language understanding.," in *Interspeech*, pp. 2524–2528, 2013.

[9] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "Spoken language understanding using long short-term memory neural networks," in *Workshop on Spoken Language Technology (SLT)*, pp. 189–194, 2014.

[10] 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.

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

[12] 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.

[13] S. Carberry, *Plan recognition in natural language dialogue*. MIT press, 1990.

[14] A. Cawsey, *Explanation and interaction: The computer generation of explanatory dialogues*. MIT press, 1992.

[15] J. D. Moore, *Participating in explanatory dialogues*. The MIT Press, 1995.

[16] D. R. Traum, *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.

[17] J. F. Allen, L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Poesio, and D. R. Traum, "The TRAINS project: A case study in building a conversational planning agent," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 7, no. 1, pp. 7–48, 1995.

[18] G. Ferguson and J. F. Allen, "TRIPS: An integrated intelligent problem-solving assistant," in *Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pp. 567–572, 1998.

[19] 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*, vol. 12, no. 2, pp. 213–253, 1980.

[20] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: A reference architecture for conversational system development," in *International Conference on Spoken Language Processing (ICSLP)*, 1998.

[21] J. Polifroni and S. Seneff, "Galaxy-II as an architecture for spoken dialogue evaluation," in *International Conference on Language Resources and Evaluation (LREC)*, 2000.

[22] 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.

[23] J. Perez, Y. Boureau, and A. Bordes, "Dialog system and technology challenge 6 overview of track 1 – End-to-end goal-oriented dialog learning," in *Dialog System Technology Challenges 6*, 2017.

[24] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *International Conference on Computational Linguistics (COLING)*, pp. 928–939, 2014.

[25] R. S. Wallace, "The anatomy of A.L.I.C.E.," in *Parsing the Turing Test*, pp. 181–210, Springer, 2009.

[26] Z. Yu, Z. Xu, A. W. Black, and A. Rudnicky, "Chatbot evaluation and database expansion via crowdsourcing," in *Chatbot workshop of LREC*, pp. 15–19, 2016.

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

[28] M. Ghazvininejad, C. Brockett, M. Chang, B. Dolan, J. Gao, W. Yih, and M. Galley, "A knowledge-grounded neural conversation model," *arXiv preprint*, 2017. arXiv:1702.01932.

[29] R. T. Lowe, N. Pow, I. V. Serban, L. Charlin, C.-W. Liu, and J. Pineau, "Training end-to-end dialogue systems with the Ubuntu dialogue corpus," *Dialogue and Discourse*, vol. 8, no. 1, pp. 31–65, 2017.

[30] A. Ezen-Can, J. F. Grafsgaard, J. C. Lester, and K. E. Boyer, "Classifying student dialogue acts with multimodal learning analytics," in *International Conference on Learning Analytics and Knowledge (LAK)*, pp. 280–289, 2015.

[31] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *International Conference on Multimodal Interfaces (ICMI)*, pp. 205–211, 2004.

[32] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Multimodal approaches for emotion recognition: A survey," in *Internet Imaging VI*, vol. 5670, pp. 56–68, 2005.

[33] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, pp. 375–388, 2007.

[34] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: The multimodal emotion recognition test (MERT)," *Emotion*, vol. 9, no. 5, p. 691, 2009.

[35] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," in *International Conference on Affective Computing and Intelligent Interaction (ICACII)*, pp. 491–497, 2015.

[36] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta psychologica*, vol. 26, no. 1, pp. 22–63, 1967.

[37] K. Jokinen, K. Harada, M. Nishida, and S. Yamamoto, "Turn-alignment using eye-gaze and speech in conversational interaction," in *Interspeech*, pp. 2018–2021, 2010.

[38] L. P. Morency, I. D. Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, vol. 20, no. 1, pp. 70–84, 2010.

[39] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Analysis and modeling of next speaking start timing based on gaze behavior in multi-party meetings," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 694–698, 2014.

[40] J. Cassell, J. Sullivan, E. Churchill, and S. Prevost, *Embodied conversational agents*. MIT press, 2000.

[41] J. Gustafson and L. Bell, "Speech technology on trial: Experiences from the August system," *Natural Language Engineering*, vol. 6, no. 3-4, pp. 273–286, 2000.

[42] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura, "Embodied symbol emergence based on mimesis theory," *International Journal of Robotics Research*, vol. 23, no. 4-5, pp. 363–377, 2004.

[43] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. G. Okuno, "Two-way translation of compound sentences and arm motions by recurrent neural networks," in *International Conference on Intelligent Robots and Systems (IROS)*, pp. 1858–1863, 2007.

[44] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *International Conference on Human Robot Interaction (HRI)*, pp. 259–266, 2010.

[45] N. J. Nilsson, "Shakey the robot," tech. rep., SRI International, 1984.

[46] C. L. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1146–1151, 1999.

[47] C. L. Breazeal, *Designing sociable robots*. MIT press, 2002.

[48] B. Adams, C. Breazeal, R. A. Brooks, and B. Scassellati, "Humanoid robots: A new kind of tool," *IEEE Intelligent Systems*, vol. 15, no. 4, pp. 25–31, 2000.

[49] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pp. 832–839, 2000.

[50] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu, "Robovie: A robot generates episode chains in our daily life," in *International Symposium on Robotics (ISR)*, pp. 1356–1361, 2001.

[51] S. A. Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive behavioural systems*, pp. 114–130, Springer, 2012.

[52] S. A. Moubayed, G. Skantze, and J. Beskow, "The Furhat back-projected humanoid head–lip reading, gaze and multi-party interaction," *International Journal of Humanoid Robotics*, vol. 10, no. 1, pp. 1–25, 2013.

[53] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "The interactive museum tour-guide robot," in *Conference on Artificial Intelligence (AAAI)*, pp. 11–18, 1998.

[54] Y. Okuno, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Providing route directions: design of robot's utterance, gesture, and timing," in *International Conference on Human Robot Interaction (HRI)*, pp. 53–60, 2009.

[55] Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi, "Multi-person conversation via multi-modal interface-a robot who communicate with multi-user," in *European Conference on Speech Communication and Technology (EUROSPEECH)*, 1999.

[56] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech," in *International Conference on Multimodal Interfaces (ICMI)*, pp. 144–151, 2004.

[57] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing in human-robot conversations: How robots might shape participant roles using gaze cues," in *International Conference on Human Robot Interaction (HRI)*, pp. 61–68, 2009.

[58] D. Klotz, J. Wienke, J. Peltason, B. Wrede, S. Wrede, V. Khalidov, and J.-M. Odobez, "Engagement-based multi-party dialog with a humanoid robot," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 341–343, 2011.

[59] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. Petrick, "Two people walk into a bar: Dynamic multi-party social interaction with a robot agent," in *International Conference on Multimodal Interaction (ICMI)*, pp. 3–10, 2012.

[60] G. W. Lcock, "WikiTalk: A spoken wikipedia-based open-domain knowledge access system," in *COLING Workshop on Question Answering for Complex Domains*, pp. 57–69, 2012.

[61] K. Jokinen and G. Wilcock, "Multimodal open-domain conversations with the nao robot," in *Natural Interaction with Robots, Knowbots and Smartphones*, pp. 213–224, Springer, 2014.

[62] G. Wilcock and K. Jokinen, "Multilingual WikiTalk: Wikipedia-based talking robots that switch languages," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 162–164, 2015.

[63] G. Skantze, A. Hjalmarsson, and C. Oertel, "Turn-taking, feedback and joint attention in situated human–robot interaction," *Speech Communication*, vol. 65, pp. 50–66, 2014.

[64] G. Skantze and M. Johansson, "Modelling situated human-robot interaction using IrisTK," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 165–167, 2015.

[65] I. Leite, A. Pereira, A. Funkhouser, B. Li, and J. F. Lehman, "Semi-situated learning of verbal and nonverbal content for repeated human-robot interaction," in *International Conference on Multimodal Interaction (ICMI)*, pp. 13–20, 2016.

[66] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L. P. Morency, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 1061–1068, 2014.

[67] D. Traum, P. Aggarwal, R. Artstein, S. Foutz, J. Gerten, A. Katsamanis, A. Leuski, D. Noren, and W. Swartout, "Ada and Grace: Direct interaction with museum visitors," in *International Conference on Intelligent Virtual Agents (IVA)*, pp. 245–251, 2012.

[68] T. Baur, I. Damian, P. Gebhard, K. Porayska-Pomsta, and E. André, "A job interview simulation: Social cue-based interaction with a virtual character," in *International Conference on Social Computing (SocialCom)*, pp. 220–227, 2013.

[69] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "MACH: My automated conversation coach," in *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 697–706, 2013.

[70] K. Cofino, V. Ramanarayanan, P. Lange, D. Pautler, D. Suendermann-Oeft, and K. Evanini, "A modular, multimodal open-source virtual interviewer dialog agent," in *International Conference on Multimodal Interaction (ICMI)*, pp. 520–521, 2017.

[71] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech*, pp. 223–227, 2014.

[72] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, 2016.

[73] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[74] B. Schuller, N. Köhler, R. Müller, and G. Rigoll, "Recognition of interest in human conversational speech," in *Interspeech*, pp. 793–796, 2006.

[75] W. Y. Wang, F. Biadsy, A. Rosenberg, and J. Hirschberg, "Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification," *Computer Speech and Language*, vol. 27, no. 1, pp. 168–189, 2013.

[76] Y. Chiba, T. Nose, and A. Ito, "Analysis of efficient multimodal features for estimating user's willingness to talk: Comparison of human-machine and human-human dialog," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017.

[77] T. P. Costa Jr, "The NEO-PI-R professional manual: Revised NEO five-factor inventory (NEO-FFI)," *Psychological assessment resources*, 1992.

[78] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[79] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005.

[80] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL-HLT)*, pp. 169–172, 2008.

[81] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Interspeech*, pp. 3081–3084, 2011.

[82] J. Thomason, H. V. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," in *International Conference on Artificial Intelligence in Education (AIED)*, pp. 750–753, 2013.

[83] M. Mizukami, K. Yoshino, G. Neubig, D. Traum, and S. Nakamura, "Analyzing the effect of entrainment on dialogue acts," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 310–318, 2016.

[84] Y. Rogers and H. Brignull, "Subtle ice-breaking: Encouraging socializing and interaction around a large public display," in *OZCHI Workshop on Public, Community, and Situated Displays*, 2002.

[85] H. Inaguma, K. Inoue, S. Nakamura, K. Takanashi, and T. Kawahara, "Prediction of ice-breaking between participants using prosodic features in the first meeting dialogue," in *ICMI Workshop on Advancements in Social Signal Processing for Multimodal Interaction*, pp. 11–15, 2016.

[86] N. Lubold and H. Pon-Barry, "Acoustic-prosodic entrainment and rapport in collaborative learning dialogues," in *Multimodal Learning Analytics Workshop and Grand Challenges*, pp. 5–12, 2014.

[87] 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.

[88] R. Zhao, T. Sinha, A. W. Black, and J. Cassell, "Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior," in *International Conference on Intelligent Virtual Agents (IVA)*, pp. 218–233, 2016.

[89] P. Müller, M. X. Huang, and A. Bulling, "Detecting low rapport during natural interactions in small groups from non-verbal behaviour," in *Annual Meeting of the Intelligent Interfaces Community (IUI)*, 2018.

[90] Q. Xu, L. Li, and G. Wang, "Designing engagement-aware agents for multiparty conversations," in *Conference on Human Factors in Computing Systems (CHI)*, pp. 2233–2242, 2013.

[91] N. Glas, K. Prepin, and C. Pelachaud, "Engagement driven topic selection for an information-giving agent," in *Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, pp. 48–57, 2015.

[92] Z. Yu, L. Nicolich-Henkin, A. W. Black, and A. I. Rudnicky, "A Wizard-of-Oz study on a non-task-oriented dialog systems that reacts to user engagement," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 55–63, 2016.

[93] M. Sun, Z. Zhao, and X. Ma, "Sensing and handling engagement dynamics in human-robot interaction involving peripheral computing devices," in *Conference on Human Factors in Computing Systems (CHI)*, pp. 556–567, 2017.

[94] O. Rudovic, M. A. Nicolaou, and V. Pavlovic, "Machine learning methods for social signal processing," in *Social Signal Processing*, pp. 234–254, Cambridge University Press, 2017.

[95] V. Ramanarayanan, C. W. Leong, and D. Suendermann-Oeft, "Rushing to judgement: How do laypeople rate caller engagement in thin-slice videos of human-machine dialog?," in *Interspeech*, pp. 2526–2530, 2017.

[96] C. Oertel, K. A. F. Mora, J. Gustafson, and J.-M. Odobez, "Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions," in *International Conference on Multimodal Interaction (ICMI)*, 2015.

[97] Y. I. Nakano and R. Ishii, "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *Annual Meeting of the Intelligent Interfaces Community (IUI)*, pp. 139–148, 2010.

[98] E. Goffman, *Behavior in public places: Notes on the social organization of gatherings*. Simon and Schuster, 1966.

[99] A. Couture-Beil, R. T. Vaughan, and G. Mori, "Selecting and commanding individual robots in a vision-based multi-robot system," in *International Conference on Human Robot Interaction (HRI)*, pp. 355–356, 2010.

[100] J. Le Maitre and M. Chetouani, "Self-talk discrimination in human–robot interaction situations for supporting social awareness," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 277–289, 2013.

[101] A. Holroyd, C. Rich, C. L. Sidner, and B. Ponsler, "Generating connection events for human-robot collaboration," in *Inernational Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 241–246, 2011.

[102] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting engagement in HRI: An exploration of social and task-based context," in *International Conference on Social Computing (SocialCom)*, pp. 421–428, 2012.

[103] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 225–234, 2009.

[104] I. Poggi, *Mind, hands, face and body: A goal and belief view of multimodal communication*. Weidler, 2007.

[105] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," in *International Conference on Spoken Language Processing (ICSLP)*, pp. 1329–1332, 2004.

[106] H. L. O'Brien and E. G. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," *Journal of the American society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, 2008.

[107] C. Peters, "Direction of attention perception for conversation initiation in virtual environments," in *International Conference on Intelligent Virtual Agents (IVA)*, pp. 215–228, 2005.

[108] T. Bickmore, D. Schulman, and L. Yin, "Maintaining engagement in long-term interventions with relational agents," *Applied Artificial Intelligence*, vol. 24, no. 6, pp. 648–666, 2010.

[109] L. Hall, S. Woods, R. Aylett, L. Newall, and A. Paiva, "Achieving empathic engagement through affective interaction with synthetic characters," in *International Conference on Affective Computing and Intelligent Interaction (ICACII)*, pp. 731–738, 2005.

[110] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, I. Poggi, and U. R. Tre, "Engagement capabilities for ECAs," in *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2005.

[111] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in *International Conference on Intelligent Virtual Agents (IVA)*, pp. 125–138, 2007.

[112] Z. Yu, D. Bohus, and E. Horvitz, "Incremental coordination: Attention-centric speech production in a physically situated conversational agent," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 402–406, 2015.

[113] D. Bohus, S. Andrist, and E. Horvitz, "A study in scene shaping: Adjusting F-formations in the wild," in *AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*, 2017.

[114] R. Bednarik, S. Eivazi, and M. Hradis, "Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification of high and low levels of engagement," in *ICMI Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2012.

[115] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *International Conference on Multimodal Interfaces (ICMI)*, pp. 119–126, 2009.

[116] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *International Conference on Human Robot Interaction (HRI)*, pp. 375–382, 2010.

[117] Z. Yu, V. Ramanarayanan, P. Lange, and D. Suendermann-Oeft, "An open-source dialog system with real-time engagement tracking for job interview training applications," in *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2017.

[118] A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, and A. Lim, "UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions," in *International Conference on Multimodal Interaction (ICMI)*, pp. 464–472, 2017.

[119] M. Frank, G. Tofighi, H. Gu, and R. Fruchter, "Engagement detection in meetings," *arXiv preprint*, 2016. arXiv:1608.08711.

[120] B. B. Türker, Z. Buçinca, E. Erzin, Y. Yemez, and M. Sezgin, "Analysis of engagement and user experience with a laughter responsive social robot," in *Interspeech*, pp. 844–848, 2017.

[121] A. Cafaro, N. Glas, and C. Pelachaud, "The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions," in *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 911–920, 2016.

[122] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *International Conference on Human Robot Interaction (HRI)*, pp. 305–311, 2011.

[123] D. Bohus and E. Horvitz, "Learning to predict engagement with a spoken dialog system in open-world settings," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 244–252, 2009.

[124] M. P. Michalowski, S. Sabanovic, and R. Simmons, "A spatial model of engagement for a social robot," in *International Workshop on Advanced Motion Control (AMC)*, pp. 762–767, 2006.

[125] Y. Chiba and A. Ito, "Estimation of user's willingness to talk about the topic: Analysis of interviews between humans," in *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2016.

[126] Y. Huang, E. Gilmartin, and N. Campbell, "Conversational engagement recognition using auditory and visual cues," in *Interspeech*, 2016.

[127] T. Baltrušaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

[128] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[129] C. L. Sidner and C. Lee, "Engagement rules for human-robot collaborative interactions," in *International Conference on Systems, Man and Cybernetics (SMC)*, pp. 3957–3962, 2003.

[130] K. Inoue, D. Lala, S. Nakamura, K. Takanashi, and T. Kawahara, "Annotation and analysis of listener's engagement based on multi-modal behaviors," in *ICMI Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI)*, 2016.

[131] N. Glas, K. Prepin, and C. Pelachaud, "Engagement driven topic selection for an information-giving agent," in *Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, pp. 48–57, 2015.

[132] D. F. Glas, T. Minato, C. T. Ishi, T. Kawahara, and H. Ishiguro, "ERICA: The ERATO intelligent conversational android," in *Inernational Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 22–29, 2016.

[133] K. Inoue, P. Milhorat, D. Lala, T. Zhao, and T. Kawahara, "Talking with ERICA, an autonomous android," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 212–215, 2016.

[134] C. T. Ishi, H. Ishiguro, and N. Hagita, "Evaluation of formant-based lip motion generation in tele-operated humanoid robots," in *International Conference on Intelligent Robots and Systems (IROS)*, pp. 2377–2382, 2012.

[135] K. Sakai, C. T. Ishi, T. Minato, and H. Ishiguro, "Online speech-driven head motion generating system and evaluation on a tele-operated robot," in *Inernational Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 529–534, 2015.

[136] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, pp. 28–39, Springer, 2006.

[137] L. Chen, R. T. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. Huang, "VACE multimodal meeting corpus," in *Machine Learning for Multimodal Interaction*, pp. 40–51, Springer, 2006.

[138] V. Ramanarayanan, C. W. Leong, D. Suendermann-Oeft, and K. Evanini, "Rcrowdsourcing ratings of caller engagement in thin-slice videos of human-machine dialog: Benefits and pitfalls," in *International Conference on Multimodal Interaction (ICMI)*, pp. 281–287, 2017.

[139] Y. Den, N. Yoshida, K. Takanashi, and H. Koiso, "Annotation of Japanese response tokens and preliminary analysis on their distribution in three-party conversations," in *Conference of the Oriental COCOSDA*, pp. 168–173, 2011.

[140] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[141] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied statistics*, pp. 20–28, 1979.

[142] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl. 1, pp. 5228–5235, 2004.

[143] D. Ozkan, K. Sagae, and L. P. Morency, "Latent mixture of discriminative experts for multimodal prediction modeling," in *International Conference on Computational Linguistics (COLING)*, pp. 860–868, 2010.

[144] D. Ozkan and L. P. Morency, "Modeling wisdom of crowds using latent mixture of discriminative experts," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 335–340, 2011.

[145] S. Kumano, K. Otsuka, M. Matsuda, R. Ishii, and J. Yamato, "Using a probabilistic topic model to link observers' perception tendency to personality," in *International Conference on Affective Computing and Intelligent Interaction (ICACII)*, pp. 588–593, 2013.

[146] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 315–323, 2011.

[147] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[148] M. R. Barrick and M. K. Mount, "The Big Five personality dimensions and job performance: A meta-analysis," *Personnel Psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[149] S. Wada, "Construction of the Big Five scales of personality trait terms and concurrent validity with NPI," *The Japanese Journal of Psychology*, vol. 67, no. 1, pp. 61–67, 1996.

[150] L. P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[151] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech*, pp. 148–152, 2013.

[152] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[153] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs," *Language and speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[154] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi, "A conversation robot using head gesture recognition as para-linguistic information," in *Inernational Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 159–164, 2004.

[155] T. Kawahara, K. Sumi, Z.-Q. Chang, and K. Takanashi, "Detection of hot spots in poster conversations based on reactive tokens of audience," in *Interspeech*, 2010.

[156] A. Graves, M. Liwicki, H. Bunke, J. Schmidhuber, and S. Fernández, "Unconstrained on-line handwriting recognition with recurrent neural networks," in *Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 577–584, 2008.

[157] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[158] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[159] T. F. Krikke and K. P. Truong, "Detection of nonverbal vocalizations using gaussian mixture models: Looking for fillers and laughter in conversational speech," in *Interspeech*, pp. 163–167, 2013.

[160] G. Gosztolya, "Detecting laughter and filler events by time series smoothing with genetic algorithms," in *International Conference on Speech and Computer (ICSC)*, pp. 232–239, 2016.

[161] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting autism, emotions and social signals using AdaBoost," in *Interspeech*, pp. 220–224, 2013.

[162] H. Salamin, A. Polychroniou, and A. Vinciarelli, "Automatic detection of laughter and fillers in spontaneous mobile phone conversations," in *International Conference on Systems, Man and Cybernetics (SMC)*, pp. 4282–4287, 2013.

[163] R. Gupta, K. Audhkhasi, S. Lee, and S. Narayanan, "Paralinguistic event detection from speech using probabilistic time-series smoothing and masking," in *Interspeech*, pp. 173–177, 2013.

[164] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Laughter and filler detection in naturalistic audio," in *Interspeech*, pp. 2509–2513, 2015.

[165] R. Brueckner and B. Schulter, "Social signal classification using deep BLSTM recurrent neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4823–4827, 2014.

[166] S. K. Maynard, "Interactional functions of a nonverbal sign head movement in Japanese dyadic casual conversation," *Journal of pragmatics*, vol. 11, no. 5, pp. 589–606, 1987.

[167] C. Hanzawa, *Listening behaviors in Japanese: Aizuchi and head nod use by native speakers and second language learners*. PhD thesis, University of Iowa, 2012.

[168] D. Lala, K. Inoue, P. Milhorat, and T. Kawahara, "Detection of social signals for recognizing engagement in human-robot interaction," in *AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*, 2017.

[169] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint*, 2014. arXiv:1406.1078.

[170] G. Pundak and T. N. Sainath, "Highway-LSTM and recurrent highway networks for speech recognition," in *Interspeech*, pp. 1303–1307, 2017.

[171] T. Kawahara, "Spoken dialogue system for a human-like conversational robot ERICA," in *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2018.

[172] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transaction on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[173] Y. Huang, T. Otsuka, and H. G. Okuno, "A speaker diarization system with robust speaker localization and voice activity detection," in *Contemporary Challenges and Solutions in Applied Artificial Intelligence*, pp. 77–82, Springer, 2013.

[174] C. T. Ishi, C. Liu, J. Even, and N. Hagita, "Hearing support system using environment sensor network," in *International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[175] S. Ueno, H. Inaguma, M. Mimura, and T. Kawahara, "Acoustic-to-word attention-based model complemented with character-level CTC-based model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5804–5808, 2018.

[176] P. M. Clancy, S. A. Thompson, R. Suzuki, and H. Tao, "The conversational use of reactive tokens in English, Japanese, and Mandarin," *Journal of Pragmatics*, vol. 26, no. 3, pp. 355–387, 1996.

[177] T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. Ward, and T. Kawahara, "Analysis and prediction of morphological patterns of backchannels for attentive listening agents," in *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2016.

[178] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N. G. Ward, "Prediction and generation of backchannel form for attentive listening systems," in *Interspeech*, pp. 2890–2894, 2016.

[179] D. Lala, P. Milhorat, K. Inoue, M. Ishida, K. Takanashi, and T. Kawahara, "Attentive listening system with backchanneling, response generation and flexible turn-taking," in *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 127–136, 2017.

[180] N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima, "Collecting evaluative expressions for opinion extraction," in *International Conference on Natural Language Processing (ICON)*, pp. 596–605, 2004.

[181] G. Salton and M. J. McGill, *Introduction to modern information retrieval*. McGraw-Hill, 1986.

[182] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: Identifying interactional style in spoken conversation," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 638–646, 2009.

[183] R. Ranganath, D. Jurafsky, and D. McFarland, "It's not you, it's me: Detecting flirting and its misperception in speed-dates," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 334–342, 2009.

[184] T. Zhao and T. Kawahara, "Joint learning of dialog act segmentation and recognition in spoken dialog using neural networks," in *International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 704–712, 2017.

[185] K. Yoshino and T. Kawahara, "Conversational system for information navigation based on POMDP with user focus tracking," *Computer Speech and Language*, vol. 34, no. 1, pp. 275–291, 2015.

[186] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[187] D. A. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in *Interspeech*, pp. 1047–1050, 2009.

[188] G. Friedland, A. Janin, D. Imseng, X. A. Miro, L. Gottlieb, M. Huijbregts, M. T. Knox, and O. Vinyals, "The ICSI RT-09 speaker diarization system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 371–381, 2012.

[189] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Interspeech*, 2006.

[190] B. Xiao, V. Rozgic, A. Katsamanis, B. R. Baucom, P. G. Georgiou, and S. Narayanan, "Acoustic and visual cues of turn-taking dynamics in dyadic interactions," in *Interspeech*, pp. 2441–2444, 2011.

[191] B. G. Gebre, P. Wittenburg, S. Drude, M. Huijbregts, and T. Heskes, "Speaker diarization using gesture and speech," in *Interspeech*, pp. 582–586, 2014.

[192] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of personality and social psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[193] T. Kawahara, T. Iwatate, and K. Takanashi, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," in *Interspeech*, pp. 727–730, 2012.

[194] T. Kawahara, "Smart posterboard: Multi-modal sensing and analysis of poster conversations," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–5, 2013.

[195] D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, M. Wolfel, U. Klee, M. Omologo, A. Brutti, P. Svaizer, G. Potamianos, and S. M. Chu, "Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus," in *International Conference on Multimedia and Expo (ICME)*, pp. 876–879, 2005.

[196] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[197] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 29–32, 2008.

[198] K. Ishiguro, T. Yamada, S. Araki, T. Nakatani, and H. Sawada, "Probabilistic speaker diarization with bag-of-words representations of speaker angle information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 447–460, 2012.

[199] K. Yamamoto, F. Asano, T. Yamada, and N. Kitawaki, "Detection of overlapping speech in meetings using support vector machines and support vector regression," *IEICE Transactions on Fundamentals*, vol. 89, no. 8, pp. 2158–2165, 2006.

[200] H. Yoshimoto and Y. Nakamura, "Cubistic representation for real-time 3D shape and pose estimation of unknown rigid object," in *International Conference on Computer Vision Workshops*, pp. 522–529, 2013.

[201] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 741–744, 2003.

[202] K. Iwano, T. Matsuo, and S. Furui, "A study on unsupervised stream-weight estimation for multimodal speech recognition," in *The Special Interest Group Technical Reports of Information Processing Society of Japan*, SLP-76-24, pp. 1–6, 2009.

[203] Y. Wakabayashi, K. Inoue, H. Yoshimoto, and T. Kawahara, "Speaker diarization based on audio-visual integration for smart posterboard," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2014.

[204] J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo, *The rich transcription 2006 spring meeting recognition evaluation*. Springer, 2006.

# Appendix A

# List of Publications

## Journal Articles

1) K. Inoue, D. Lala, K. Takanashi, T. Kawahara, "Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue," *APSIPA Transactions on Signal & Information Processing*, 2018. **(Chapter 4)**

2) 井上昂治, Lala Divesh, 吉井和佳, 高梨克也, 河原達也, "潜在キャラクタモデルによる聞き手のふるまいに基づく対話エンゲージメントの推定," 人工知能学会論文誌, vol. 33, no. 1, pp. DSH-F_1–12, 2018. **(Chapter 3)**

3) 井上昂治, 若林佑幸, 吉本廣雅, 河原達也, "多人数会話における音響・視線情報を統合した話者区間検," 電子情報通信学会論文誌, vol. J99-D, no. 3, pp. 348–357, 2016. **(Chapter 7)**

## International Conferences

4) K. Inoue, D. Lala, K. Takanashi, T. Kawahara, "Engagement recognition in spoken dialogue via neural network by aggregating different annotators' models," In *Interspeech*, 2018. **(Chapter 5)**

5) K. Inoue, D. Lala, K. Takanashi, T. Kawahara, "Latent character model for engagement recognition based on multimodal behaviors," In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2018

6) D. Lala, K. Inoue, P. Milhorat, K. Takanashi, T. Kawahara, "Detection of social signals for recognizing engagement in human-robot interaction," In *AAAI Fall Symposium Natural Communication for Human-Robot Collaboration*, 2017.

7) K. Inoue, D. Lala, S. Nakamura, K. Takanashi, T. Kawahara, "Annotation and analysis of listener's engagement based on multi-modal behaviors," In *ICMI workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI)*, 2016.

8) D. Lala, P. Milhorat, K. Inoue, T. Zhao, T. Kawahara, "Multimodal interaction with the autonomous android ERICA," In *International Conference on Multimodal Interaction (ICMI)*, pp. 417–418, 2016.

9) K. Inoue, P. Milhorat, D. Lala, T. Zhao, T. Kawahara, "Talking with ERICA, an autonomous android," In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pp. 212–215, 2016. **(Chapter 6)**

10) K. Inoue, Y. Wakabayashi, H. Yoshimoto, K. Takanashi, T. Kawahara, "Enhanced speaker diarization with detection of backchannels using eye-gaze information in poster conversations," In *Interspeech*, pp. 3086–3090, 2015.

11) K. Inoue, Y. Wakabayashi, H. Yoshimoto, T. Kawahara, "Speaker diarization using eye-gaze information in multi-party conversations," In *Interspeech*, pages 562–566, 2014.

## Domestic Conferences

12) 井上昂治, Lala Divesh, 高梨克也, 河原達也, "自律型アンドロイドERICAにおけるエンゲージメント推定に基づく音声対話システム," 日本音響学会2018春季研究発表会, 2-8-9, 2018.

13) 井上昂治, Lala Divesh, Milhorat Pierrick, 高梨克也, 河原達也, "自律型アンドロイドERICAにおけるエンゲージメント推定に基づく音声対話システム," 人工知能学会研究会資料, SLUD-B508-18, 2017.

14) 井上昂治, Lala Divesh, Milhorat Pierrick, 石田真也, 趙天雨, 高梨克也, 河原達也, "自律型アンドロイドERICAにおける多様な聞き手応答を用いた傾聴対話," 人工知能学会研究会資料, SLUD-B508-11, 2017.

15) 井上昂治, Lala Divesh, 吉井和佳, 高梨克也, 河原達也, “潜在キャラクタモデルによる聞き手のふるまいに基づく対話エンゲージメントの推定,” 日本音響学会2017秋季研究発表会, 2-Q-12, 2017.

16) 井上昂治, Lala Divesh, 高梨克也, 河原達也, “聞き手の多様なふるまいに基づく対話エンゲージメントの推定,” 日本音響学会2017春季研究発表会, 3-5-1, 2017.

17) 井上昂治, 三村正人, 石井カルロス寿憲, 坂井信輔, 河原達也, “DAEを用いたリアルタイム遠隔音声認識,” 日本音響学会2017春季研究発表会, 1-Q-6, 2017.

18) 井上昂治, Lala Divesh, 高梨克也, 河原達也, “階層ベイズモデルを用いた聞き手の多様なふるまいに基づく対話エンゲージメントの推定,” 人工知能学会研究会資料, SLUD-B505-28, 2016.

19) 井上昂治, Milohorat Pierrick, Lala Divesh, 趙天雨, 河原達也, “自律型アンドロイドERICAによる社会的役割に則したインタラクション,” 人工知能学会研究会資料, SLUD-B505-7, 2016.

20) 井上昂治, 三村正人, 石井カルロス寿憲, 河原達也, “自律型アンドロイドERICAのための遠隔音声認識,” 日本音響学会2016春季研究発表会, 1-1-1, 2016.

21) 井上昂治, 河原達也, “自律型アンドロイドEricaのための音声対話システム. 人工知能学会研究会資料, SLUD-B502-5, 2015.

22) 井上昂治, 若林佑幸, 吉本廣雅, 高梨克也, 河原達也, “ポスター会話における音響・視線情報の確率的統合による話者区間及び相槌の検出,” 日本音響学会2015秋季研究発表会, 2-2-4, 2015.

23) 井上昂治, 若林佑幸, 吉本廣雅, 高梨克也, 河原達也, “スマートポスターボードにおける視線情報を用いた話者区間及び相槌の検出,” 情報処理学会研究報告, MUS-107-68, 2015.

24) 井上昂治, 若林佑幸, 吉本廣雅, 高梨克也, 河原達也, “スマートポスターボードにおける視線情報を用いた話者区間検出及び相槌の同定,” 第77回情報処理学会全国大会, 6P-09, 2015.

25) 井上昂治, 若林佑幸, 吉本廣雅, 高梨克也, 河原達也, “ポスター会話における音響・視線情報を統合した話者区間及び相槌の検出,” 情報処理学会研究報告, SLP-105-9, 2015.

26) 井上昂治, 若林佑幸, 吉本廣雅, 河原達也, “多人数会話における音響情報と視線情報の確率的統合による話者区間検出,” 日本音響学会2014秋季研究発表会, 2-8-4, 2014.

27) 井上昂治, 若林佑幸, 吉本廣雅, 河原達也, “多人数会話における視線情報を用いた話者区間検出,” 情報処理学会研究報告, SLP-102-1, 2014.