Title	Situation Understanding for Turn-Taking in Human-Robot Dialogue
Author(s)	杉山, 貴昭
Citation	大阪大学, 2017, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/61711
rights	
Note	

Doctoral Dissertation

Situation Understanding for Turn-Taking in Human-Robot Dialogue

Takaaki Sugiyama

January 2017

Graduate School of Engineering,
Osaka University

Preface

Humanoid robots, which are autonomous human-shaped robots, have potential applications in public spaces. People expect the robots to communicate by using their human-like eye gazes, gestures, and postures, in the same manner as humans. Many research institutes have tried to develop humanoid robots that can talk like human beings. Thanks to their developments, the motor ability of these robots is similar to that of human beings. In contrast, their communication ability is still far from that of human beings.

Humanoid robots in public spaces need to use the on-board microphones to obtain sounds because the number of dialogue participants cannot be controlled by them. Thus, we cannot assume that all human participants wear microphones because this is not realistic. Moreover, the robot does not always detect sounds clearly, and receives not only utterances but also surrounding noises. Therefore, humanoid robots need to be able to determine whether they should take a turn when they detect a sound, i.e., turn-taking decision.

The main contribution of the dissertation is the implementation of the turn-taking decision module by using social norms, which are implicit rules that govern human behaviors in a group. This means that a robot infers the dialogue situation from several observations and selects an appropriate action according to social norms. The observations can be divided into two types: observations about the robot and observations about users. The former is the robot states, such as its postures, motion, and utterance; the latter is the users' states. This dissertation presents two methods for situation understanding based on each observation. The first method is to predict *ease of speaking*, i.e., when a user is likely to begin speaking to a robot based on the robot's states. The second one is to estimate *response obligation*, i.e., whether a robot should respond to its detected sounds based on the users' states. These methods can make the robot estimate the current dialogue situation and behave according to social norms.

Chapter 1 provides this study's background, and introduces a new architecture of spoken dialogue systems for humanoid robots in public spaces. Chapter 2 describes the first method, i.e., predicting when a user is likely to begin speaking to a robot. The problem is cast as a binary machine-learning task with the input consisting of features describing the robot's state. Chapter 3 shows that the first method can be used for spoken dialogue systems for humanoid robots. Before incorporating the method in a spoken dialogue system, this chapter discusses two issues. This chapter confirms whether *ease of speaking* is influenced by the specific participants and whether the robot can handle individual differences and influences of users' urgency. Moreover, an actual spoken dialogue system for a humanoid robot is introduced. Chapter 4 describes the second method, i.e., determining whether a robot should respond to its detected sounds in public spaces. Here, typical user behaviors in human-robot dialogue are exploited for making the estimation. Finally, Chapter 5 concludes this dissertation and points to directions that future studies can take.

This dissertation summarizes my study on human-robot dialogue in Doctor's Course of Department of Information and Communications Technology, Graduate School of Engineering, Osaka University. Contents of every chapter in this dissertation except Chapter 1 and 5 are based on some publications shown in the publication list as follows:

Chapter 2: Publications A-3, B-5, C-2, D-7, and D-8,
Chapter 3: Publications A-2, B-2, B-3, B-4, D-3, D-4 and C-5,
Chapter 4: Publications A-1, B-1, C-1, and D-1.

Acknowledgements

First and foremost, I offer my sincerest gratitude to my supervisor, Professor Noboru Babaguchi, for his valuable and insightful advices on my career and this dissertation. I have had the support and encouragement of him. Without his persistent help, this dissertation would not have been possible.

I am also grateful to Professor Satoshi Sato of Nagoya University for his supervision and continuing encouragement through the course of my research in Nagoya University and Osaka University. It was a great pleasure to work with him, and I have learned a lot of things from him. His profound insights always stimulated me, and his professional, diligent, and rigorous style of work has had a great influence on me. I would like to follow his gentle manners in the future.

I would like to express my appreciation to my doctoral committee members, Professor Takashi Washio and Associate Professor Naoko Nitta for their insightful comments and suggestions on this dissertation.

I also take pleasure in thanking Professor Tetsuya Takine, Professor Akihiro Maruta, Professor Seiichi Sampei, Professor Atsuko Miyaji, and Professor Kyo Inoue. Their valuable discussions and favorable comments greatly enhanced the quality of this dissertation.

I would like to express my sincere gratitude to Ms. Aya Inoue, secretary of Babaguchi laboratory, for her enormous support. I also thank Assistant Professor Kazuaki Nakamura for his support for my research activities.

My sincere thanks are due to Professor Yoshinobu Kitamura of Ritsumeikan University, Associate Professor Koji Kozaki, and Assistant Professor Ryu Takeda, who were the teaching staff in Komatani laboratory. They gave me a lot of valuable

suggestions and comments on my research. Of course, I really appreciate the member of Komatani laboratory, students and staff. Especially, I would like to thank Ms. Chizuko Motozono, who was secretary of the laboratory. She supported me so much. I also would like to thank Dr. Yuki Yamagata who always encouraged me.

I would like to express thanks to the member of Sato laboratory; Associate Professor Takuya Matsuzaki, Ms. Rika Shiga, Ms. Kazuko Natsume, and students. They welcomed me when I came back to Sato laboratory and encouraged me greatly. I also would like to express my gratitude to alumni of Sato laboratory. I learned a lot of things from them.

I have really enjoyed working with Dr. Kotaro Funakoshi and Dr. Mikio Nakano at Honda Research Institute Japan Co., Ltd. They gave me invaluable advices and comments on my research. I also express thanks to the members of Honda Research Institute Japan Co., Ltd. for their hospitality during my visit.

I am very grateful to Lecturer Kiyohisa Nishiyama, who is my teacher for technical writing in Nagoya University, for his invaluable comments on this dissertation. I had a great time discussing my research.

I would like to thank my family and grandparents for their love, understanding, constant support, and sincere encouragement. Finally, I express thanks to Yuko for her unconditional support and encouragement in my college life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

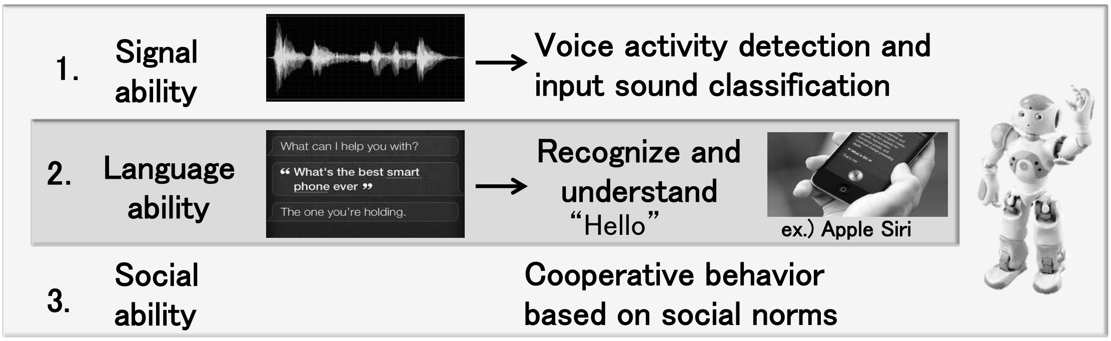
1.1 Humanoid Robot in Public Spaces

1.1.1 Communication Ability of Humanoid Robot

Humanoid robots, which are autonomous robots with a human-shaped body, have potential applications in public spaces, such as hotels, restaurants, exhibition halls, and nursing homes. Many research institutes have tried to develop humanoid robots that can talk like human beings. For example, Honda Motor Co., Ltd. has developed ASIMO [19]; Waseda University has developed SCHEMA [35]; SoftBank Robotics released Pepper [46] in 2015. The motor ability of these humanoid robots, in terms of walking, carrying, and standing, is similar to that of human beings. In contrast, their communication ability, in terms of speaking and hearing, is still far from that of human beings.

The communication ability of human beings can be viewed as a mixture of several abilities. This study classifies these abilities into three categories: a signal ability, language ability, and social ability (see Figure 1.1). Humanoid robots should have these three abilities in order to communicate with people smoothly in public spaces.

On the reception side of communication, signal ability is the ability to detect a sound and determine whether it is a noise or an utterance. The methods of realizing this ability in a robot are input sound classification and voice activity


1. Signal ability
Voice activity detection and input sound classification
2. Language ability
Recognize and understand "Hello"
ex.) Apple Siri
3. Social ability
Cooperative behavior based on social norms

What can I help you with?

“ What's the best smart phone ever ”

The one you're holding.

Figure 1.1: Three categories of communication ability.

detection. They are usually applied before speech recognition as a preprocess.

Moreover on the reception side, language ability is the ability to understand an utterance. This ability can be realized through speech recognition and language understanding. The former converts a detected sound into a language representation; the latter converts the language representation into a semantic representation that a robot (i.e., a spoken dialogue system) can interpret to determine an appropriate action.

On the generation side, these two abilities work in the reverse direction. Language ability is required to generate a language representation (utterance), while signal ability is required to convert the language representation into a sound.

Social ability is the ability to act socially, i.e., according to social norms. To the best of my knowledge, there is no method for realizing this ability in human-robot communication. Spoken dialogue systems on smartphones, such as Apple's Siri, do not need this ability, because they have no bodies. People do not treat smartphones as social beings and do not expect them to act socially in dialogue. In contrast, for spoken dialogue systems installed in humanoid robots, social ability is crucial because people treat them as social beings [23, 39]. Robot's behaviors according to social norms enable the robot to take dialogue situation and users' states[1] into account when interpreting input sound signals and language representation. For example, when a robot determines whether a detected sound is a noise or an utterance (signal ability), the robot can use users' and the robot's states as well

[1] This dissertation uses "user" as a robot user unless there are some particular reasons.

as acoustic information.

1.1.2 Social Ability and Social Norms

Social norms are implicit rules that govern human behaviors in a group [49]. In dialogue, they govern verbal and non-verbal behaviors on the basis of dialogue situations. The followings are typical examples of social norms.

- When a speaker is talking to a listener, the listener should not interrupt the speaker.
- When a speaker is talking to a listener, the speaker should face toward the listener.
- A speaker and a listener should keep a social distance [17], which is the distance that someone implicitly allows others to approach in an interaction.

Social norms depend on not only the dialogue situations, but also human relations. For example,

- when a speaker talks to her boss, she should speak clearly and carefully;
- when a speaker talks to his friends, he may speak frankly;
- when an adult talks to a child, he should use plain expressions.

This study tries to incorporate a social ability to spoken dialogue systems installed in humanoid robots under the assumption that people obey social norms and act *cooperatively* in human-robot dialogue. People have a tendency to treat humanoid robots as social beings; this was demonstrated in a famous psychological experiment conducted by Reeves and Nass [44]. Moreover, this study assumes that people interacting with a robot will not perform uncooperative actions that conflict with social norms. Accordingly, social norms can be used to determine whether a robot should respond to an input sound, i.e., take a turn.

Previous studies on spoken dialogue systems do not focus on the social ability of spoken dialogue systems. This is because the typical target platforms of spoken dialogue systems were mobile phones and smartphones. People do not treat them as social beings because these platforms do not have human-shaped bodies. Thus, in these previous studies, they did not need to consider the social ability for smooth human-machine dialogue. In contrast, several researchers in the field of human-agent interaction (HAI) have paid much attention to social ability [16]. For example, Azaria et al. [2] have built systems that encourage people by using a social channel (human-like utterances and motions).

1.2 Architecture of Spoken Dialogue Systems for Humanoid Robots in Public Spaces

The conventional architecture of spoken dialogue systems consists of five modules: automatic speech recognition, natural language understanding, dialogue management, natural language generation, and text-to-speech synthesis [36, 51]. These systems accept a user utterance (sound) as input and generate a system utterance as output. The first four modules are related to the language ability in Figure 1.1, and the fifth is related to the signal ability.

This architecture works as follows.

1. The automatic speech recognition module converts an input sound into a linguistic representation ("What's your name?").

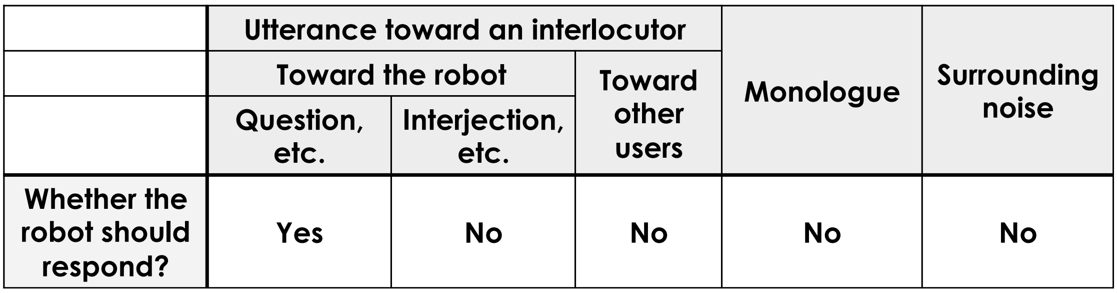
2. The natural language understanding module converts the linguistic representation into a semantic representation ("ASK: NAME").

3. The dialogue management module determines the system's action according to the semantic representation and generates a semantic representation for the response ("OUTPUT (SPEECH: MYNAME)").

4. The natural language generation module converts the obtained semantic representation into a language representation ("My name is Nao.").

5. Finally, the text-to-speech synthesis module synthesizes the sound of the language representation.

Table 1.1: Ideal turn-taking decision for humanoid robots in public spaces.



	Utterance toward an interlocutor			Monologue	Surrounding noise
	Toward the robot		Toward other users		
	Question, etc.	Interjection, etc.			
Whether the robot should respond?	Yes	No	No	No	No

This architecture is designed under the assumption that all input sounds are utterances toward the system. This is because the typical platforms of spoken dialogue systems are mobile phones and smartphones. These devices are used by placing one's mouth next to them, and thus, they clearly receive the sounds of the utterances; surrounding noises can be ignored in this situation.

In contrast, when a spoken dialogue system is installed on a humanoid robot in public spaces, this assumption, i.e., all input sounds are utterances toward the system, is not satisfied. Table 1.1 shows sounds that humanoid robots may detect in public spaces and ideal robot's actions. In public spaces, the number of dialogue participants cannot be controlled; two or more people may join a dialogue with a robot. Therefore, an expectation that all human participants wear microphones is not realistic. The robot has to receive sound through on-board microphones that detect not only utterances but also surrounding noises. This means that the robot has to be able to distinguish utterances from surrounding noises.

In public spaces, moreover, the robot has to recognize the direction of the utterance, i.e., to whom the utterance is directed, when there are two or more human participants. When a user utterance is spoken to the robot and the user expects the robot to respond (for example, questions), the robot should respond to it. When a user utterance is spoken to other users or is a monologue or the user does not expect the robot to respond (for example, interjections), the robot must not respond to it, because such incorrect responses cause the dialogue to break down.

This dissertation presents a novel architecture of spoken dialogue systems for

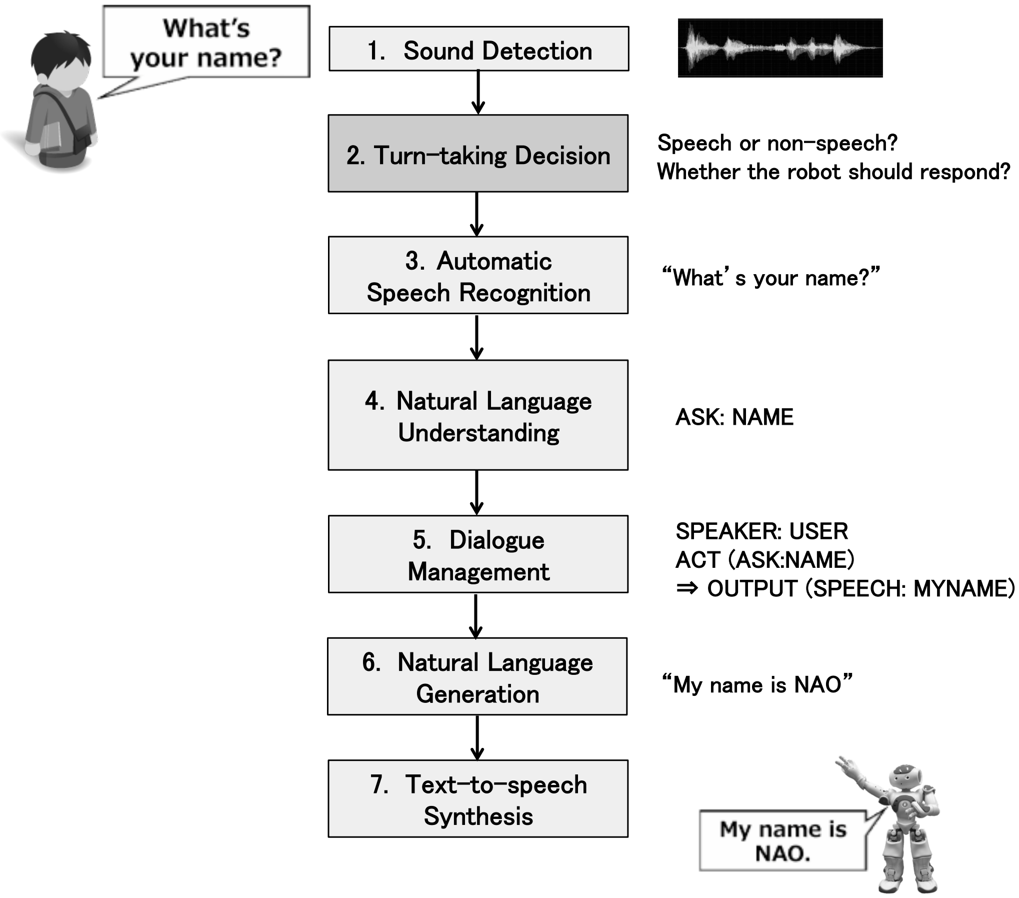


Figure 1.2: Architecture of spoken dialogue systems for humanoid robots.

humanoid robots in public spaces. This architecture (Figure 1.2) includes two new modules: a sound detection module and a turn-taking decision module. The sound detection module discriminates between sound intervals and silent intervals. The turn-taking decision module determines whether the robot should respond to detected sound intervals.

The former is conventionally done as a preprocessing of the speech recognition part and is called *endpointing* [48] or *voice activity detection* [54]. This is because that these conventional studies assumed that sound intervals are speech intervals and can be clearly detected. In contrast, for humanoid robots in public spaces, the sound detection is important because they do not always detect sounds clearly and receive not only utterances but also surrounding noises. Therefore, this study positions the sound detection as a major module.

This study focuses on the development of the turn-taking decision module under the assumption that people obey social norms and act cooperatively with the humanoid robot. To the best of my knowledge, there is no practical method for the turn-taking decision. That is, this study tackles the problem that a humanoid robot in a public space estimates whether it should respond to sound intervals, which are detected by using a conventional method for the sound detection. This module focuses on “taking a turn”, i.e., when to generate a response or an utterance, whereas the term “turn-taking” more generally includes both “taking a turn” and “releasing a turn”. For smooth turn-taking between a humanoid robot and users, the robot should be able to release a turn, as well as take one. However, it seems that “taking a turn” is more important than “releasing a turn” for humanoid robots, because the dialogue will break down if the robot fails to take one, whereas even if the robot fails to release a turn, human participants can save the dialogue.

1.3 Related Work

Table 1.2 is a classification of related work. Studies on spoken dialogue systems can be classified on two axes: the number of participants and the dialogue environment. “Two-party” indicates a one-to-one conversation, i.e., a human talks with a system, while “multi-party” indicates a conversation among groups of three or more participants. “Silent environment” indicates that the input sounds are only utterances, i.e., do not include noises, while “public space” indicates that both utterances and noises are included in the input sounds.

Many conventional studies on spoken dialogue systems fall into the class of two-party dialogue in a silent environment (the upper-left area in Table 1.2). This is because the typical target platforms of spoken dialogue systems are mobile phones and smartphones. Numerous studies on interactive voice response (IVR) systems were actively conducted in the early 2000’s [10]. These studies have dealt with utterances toward the robot (system) only in Table 1.1.

Recent research on spoken dialogue systems has explored two directions. Some

Table 1.2: Classification of related work.

		Number of participants	
		Two-party	Multi-party
	Target platform	Mobile phone, car navigation system	Humanoid Robot
Dialogue environment	Silent environment (w/o noise)	Many conventional studies	Talking with multiple users
	Public space	Dealing with various sounds	**This study**

studies have aimed at the development of multi-party spoken dialogue systems (the upper-right area in Table 1.2), while others have worked on the development of spoken dialogue systems which can be used in public spaces (the lower-left area in Table 1.2). The former's typical platform is a humanoid robot. These studies have dealt with whether utterances are toward the robot or other users in Table 1.1. The latter's is a car navigation system. These studies have dealt with whether input sounds are utterances toward the robot or surrounding noises in Table 1.1.

This study explores the area of multi-party dialogue in public spaces (the lower-right area in Table 1.2). This study deal with all input sounds in Table 1.1. This is a more realistic setting for spoken dialogue systems installed in humanoid robots in public spaces. The following subsections discuss the conventional techniques of turn-taking, talking with multiple users, and dealing with various sounds.

1.3.1 Turn-Taking

Turn-taking is the fundamental way that humans interactively organize dialogue by guessing the timing of taking and releasing their turns to speak or respond. Humans unconsciously take turns depending on the interlocutors' status in terms of their utterances, motions, and facial expressions. Sacks et al. [47] were the first researchers to propose a model of turn-taking, i.e., linguistic rules, in human-human dialogue. Other studies analyzed human-human dialogue and revealed that non-verbal behaviors are cues for turn-taking [13, 26].

These findings have been introduced into human-machine interaction. Raux and Eskenazi [43] proposed a method for controlling the timing of turn-taking over the single modality of speech. Skantze et al. [52] monitored users' attention by tracking their head movements for smooth turn-taking. They recently proposed a method for generating turn-taking cues at an appropriate timing in a face-to-face setting [53]. Vertegaal et al. [57] constructed a conversational system that uses eye movements to determine to whom the user is speaking.

Smooth turn-taking is very important for spoken dialogue systems for humanoid robots. Humans talk naturally to robots that have a human-like appearance. Therefore, the robots need to be able to estimate the timing of taking or releasing a turn from the dialogue situation. Hatice et al. [31] developed a humanoid robot Kaspar and used probabilistic models for natural turn-taking in drumming interaction games. Chao et al. [9] showed that a humanoid robot's gaze can control the utterance timing of human listeners.

1.3.2 Talking with Multiple Users

Some studies have aimed to develop spoken dialogue systems for humanoid robots that can engage in speech interactions with multiple users. Bohus and Horvits [3] devised a method for managing engagement in multi-party dialogue and implemented it in a spoken dialogue system for a robot. Engagement is the process by which two or more users join, maintain, and leave their conversation [50]. They estimated dialogue participants and speakers from multi-modal information. They installed their system in a humanoid robot named Nao [45], which works as a navigator in front of elevators in a building [4]. Keizer et al. [25] explored an approach involving automatic learning of dialogue strategies for multi-party human-robot dialogue and demonstrated it using an example of a robot bartender.

The problem of addressee identification in multi-party dialogue situations is closely related to this study, in which the addressee is identified by using acoustic and visual information. Nakano et al. [40] showed that their agent can identify addressees by using face tracking and prosodic analysis of user utterances in human-agent interactions with high accuracy. By using this method, the agent can respond to an input sound when it identifies the addressee as itself only. However, this method is limited to identifying an addressee of an utterance under the assumption that every input sound is an utterance toward one of the interlocutors. The experiments of these studies excluded noises, despite that they used the on-board microphones to obtain sounds; this is an unrealistic setting.

1.3.3 Dealing with Various Sounds

As described in Section 1.2, spoken dialogue systems used in public spaces receive surrounding noises as well as utterances. Furthermore, a user may talk to himself, i.e., conduct a monologue. Incorrect responses to noises and monologues can cause dialogue breakdowns.

Surrounding noises occur at arbitrary timings in public spaces. Noises in public spaces include, for example, ringtones, users' footsteps, and even motor noises of the robot. Many researchers have proposed methods for making noise-robust spoken dialogue systems. These methods can be categorized as noise robust speech recognition or input sound classification.

As for the studies on noise robust speech recognition, Gales et al. [15] proposed hidden Markov model (HMM) composition by using parallel model combination to optimize a speech recognition system to surrounding noises. Boll [5] proposed a noise suppression algorithm for reducing the spectral effects of acoustically added noise in speech. Recent studies have proposed methods of blind dereverberations that can be used for recovering unknown sources from observations without the aid of information about the sources [58].

As for the studies on input sound classification, Lee et al. [34] collected sounds data in public spaces and divided these sounds into five categories (adult, child, laugh, cough, and noise). They built five GMMs by using acoustic features, and distinguished these categories. Brueckmann et al. [8] proposed a method to adaptively reduce noise. These methods need to build sound classification models based on acoustic features beforehand. Moreover, Huang et al. [20] proposed a method for voice activity detection by using lip movements obtained from a vision system. This method uses the shape of the region between the lips to determine whether the mouth is open or closed. Recently, a voice activity detection method that integrates audio and visual information has been proposed [62]. By using these methods, the systems can ignore noises and respond only to utterances.

Monologues are acoustically similar to utterances toward someone. In order to discriminate between these utterances, non-acoustic features are needed in addition to acoustic features. Komatani et al. [29] proposed a method to distinguish monologues from utterances toward a spoken dialogue system with information peculiar to spoken dialogue systems, such as the duration of an utterance and elapsed time since the end of the system's utterance.

All of these methods introduced in this subsection fall into the binary classification category. Therefore, they cannot be applied to multi-class classification for sounds in public spaces. For example, a method for distinguishing between utterances and noises cannot correctly distinguish an announcement in a station and utterances directed toward the robot because these sounds are acoustically equivalent.

1.4 Utilizing Social Norms in Human-Robot Dialogue

This study utilizes social norms for multi-party human-robot dialogue in public spaces, by implementing a turn-taking decision module. This means that a robot infers the dialogue situation from several observations and selects an appropriate action (whether to respond or not) according to social norms. This study calls it *situation understanding of turn-taking.*

The observations can be divided into two types: observations about the robot and observations about users. The former is the robot states, such as its postures, motion, and utterance; the latter is the users' states. This dissertation presents two methods for situation understanding based on each observation.

First, this dissertation focuses on the observations about the robot. This dissertation presents a method to predict when a user is likely to begin speaking to a robot based on the robot's states, such as its posture, motion, and utterances. If the robot can predict whether the user is likely to begin speaking, it can reject sounds as noise when the user is not likely to begin speaking.

Second, this dissertation focuses on the observations about users. This dissertation presents a method to determine whether the robot should respond to its input sounds on the basis of the users' states. In such environments, the robot would receive various sounds from its surroundings including noises and utterances. This method gives each input sound interval one of two labels: *ought-to-respond* and *ought-not-to-respond*. This method is useful for preventing the robot from making incorrect or unnecessary responses in a multi-party situation.

1.5 Outline of this Dissertation

The rest of this dissertation is organized as follows. Chapter 2, 3, and 4 describe my study, which was addressed in Osaka University, on human-robot dialogue. Chapter 2 describes a model for predicting when a user is likely to begin speaking to a humanoid robot [A-3, B-5][2]. The problem is cast as a binary machine-learning task with the input consisting of features describing the robot's state, such as its postures, motions, and utterances. Chapter 3 shows that the prediction model described in Chapter 2 can be used for spoken dialogue systems of humanoid robots [A-2, B-2, B-3, B-4]. This chapter discusses two issues that arise when applying this model to the spoken dialogue system and describes the actual construction of the system. Chapter 4 discusses how a robot determines whether it should respond to its input sounds in a realistic situation, i.e., in a multi-party human-robot dialogue in a public space. Here, typical user behaviors in human-robot dialogue are exploited for making the estimation [A-1, B-1]. Finally, Chapter 5 concludes this dissertation and points to directions that future studies can take.

[2] These labels correspond with the publication list.

Chapter 2

Predicting When a Human is Likely to Begin Speaking to a Humanoid Robot

2.1 Introduction

This chapter describes a method for predicting *ease-of-speaking*, i.e., when a user is likely to begin speaking to a robot. It focuses on observations about a robot (described in Section 1.4) and uses the following social norm for building the method: a human speaker considers her interlocutor's situation when she determines whether to begin speaking. For example, if her interlocutor is speaking and/or looks busy, she refrains from beginning speaking.

Incorrect or unnecessary responses are a critical problem in speech interaction with humanoid robots. It is important for communication robots to prevent such responses because automatic speech recognition by using the robot's head-mounted microphones is more difficult than that of using microphones in headsets [32]. The robot needs to ignore unnecessary sounds, such as environmental noises, a user's unintentional mutters, and laughter. In order to ignore unnecessary sounds, many researchers have focused on sounds which are detected by systems [28, 61, 34], because acoustic information can be extracted from detected sounds. However, these methods cannot distinguish between human voices and sounds similar to human voices, such as voices and singing on the radio and TV, because these sounds are acoustically equivalent.

We adopt a novel approach to ignoring unnecessary sounds: we focus on the robot's state rather than the users' state. This idea is inspired by the fact that human speakers usually take the state of their addressee into consideration and chooses when to begin speaking to the addressee. We assume that this convention can be applied to human-robot interaction and have built a model that uses the robot's state to predict when the user is likely to begin speaking. If the robot can predict the user's timing, it can know the timing when a cooperative user is likely to begin speaking. Conversely, audio received when the user is unlikely to begin speaking can be discarded as non-speech with high probability. The proposed model can be used in conjunction with existing approaches focusing on input sounds to help distinguish user utterances. Furthermore, the robot can produce situations in which the user is not likely to begin speaking and thus can control when the user begins speaking.

The approach of the method is twofold:

1. The problem is formulated within a machine-learning framework.

2. Training data for the machine learning is collected from an experiment involving multiple participants.

First, we define the robot states, i.e., the observation about the robot, to be used as input features for the machine learning. These states are defined at any timing by using the robot's behaviors, such as its posture, motion, utterances, etc. We then predict ease of speaking, i.e., whether the user is likely to begin speaking to the robot. That is, we cast the problem as a binary machine-learning task, where the input consists of features describing the robot's state, and the output indicates whether the user is likely to begin speaking or not likely to begin speaking.

Second, we collect training data for the machine learning by using multiple participants. Different people may have different ideas about when to begin speaking, so we conducted an experiment in which we asked several participants to label some real-time behaviors. While they were watching a sequence of behaviors of a humanoid robot, they indicated in real time whether they would be likely to begin speaking. We used the parts to which the participants gave the same labels in the training to predict *ease of speaking* that most people feel. In essence, this enabled us to extract the parts in which multiple participants are likely to begin speaking, thereby eliminating the effect of individuality.

2.2 Definition and Formalization

2.2.1 Definition

The proposed method predicts whether a user is likely to begin speaking to a robot. The prediction target corresponds to the transition relevance place (TRP) in human-human dialogue. The TRP was first advocated by Sacks et al. [47] and is a well-known notion in the human conversation analysis community. It indicates places where the addressee is likely to begin speaking, i.e., when she will take a turn.

This chapter assumes three conditions:

1. The content that the user is trying to convey to the robot is not urgent.

2. The user regards the robot as a social being.

3. The time when most cooperative users are and are not likely to begin speaking exists.

Here, this chapter assumes that the content the user is trying to convey is not urgent, e.g., asking the robot to turn on the air conditioner. The content of an utterance is crucial to determining whether a user is actually likely to begin speaking to the robot. Very urgent content, e.g., asking the robot to call an ambulance, would be dealt with regardless of the robot's state. Section 3.3.2 will discuss how the proposed method handles the variation caused by degree of urgency.

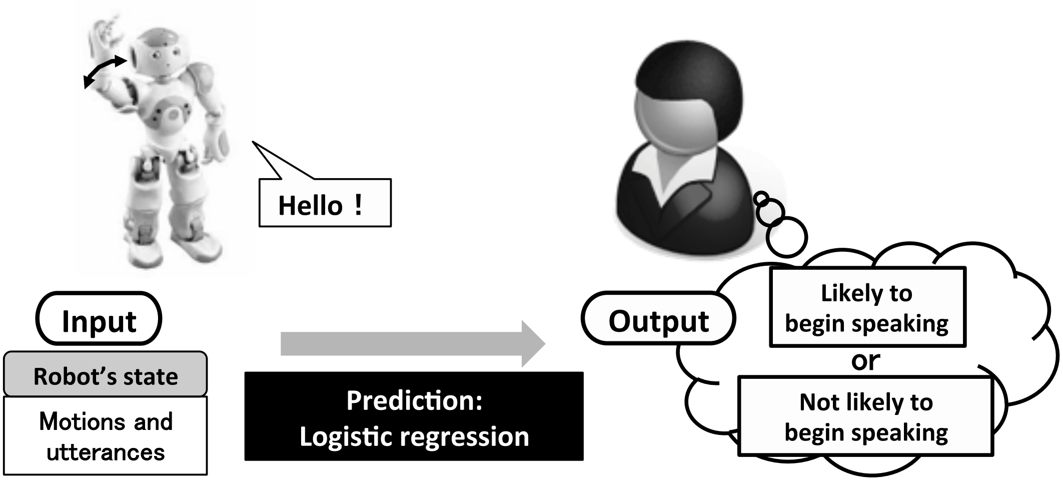


Figure 2.1: Overview of proposed method for predicting when a user is likely to begin speaking.

In addition, this study assumes that the user treats the robot as not a machine but as a social being; that is, we assume that the user feels a kind of anthropomorphism. It is well known that robots such as Geminoid [21] and Repliee R1 [38], which closely resemble humans, make users feel anthropomorphism [44, 39, 23]. If a robot moves only when instructed, people begin speaking to it without really considering its state. Here, we use a humanoid robot that speaks and moves like a human, thereby satisfying this assumption.

Finally, this study assumes that the time when most cooperative users are and are not likely to begin speaking exists. In general, ease of speaking depends on individuality. However, we believe that the timing when most people feel common ease of speaking regardless of individuality exists. This study aims to build the model that can predict the timing.

2.2.2 Formalization

An overview of the proposed method is shown in Figure 2.1. This method predicts *ease of speaking*, i.e., the output indicates whether the user is likely to begin speaking or not likely to begin speaking. We cast the problem as a binary machine-learning task with the input consisting of features describing the robot's state, such as its postures, motions, and utterances. Whether the user is likely to begin speaking to the robot is estimated every 0.1 seconds.

The prediction model is defined as:

$$y^{(t)} = f_p(\mathbf{x}^{(t)}), \tag{2.1}$$

where $\mathbf{x}^{(t)}$ is the input feature vector obtained at time t, and the function f_p is a predictor taking the feature vector as the input and outputting whether the user is likely to begin speaking to the robot. We used logistic regression as the predictor f_p. This is because that we can easily understand which input features contribute to predict ease of speaking and how likely users are to begin speaking by observing coefficients and the probability value of the function f_p. The output $y^{(t)}$ is a binary value:

$$y^{(t)} = \begin{cases} 0 & (\textit{not likely to begin speaking}) \\ 1 & (\textit{likely to begin speaking}). \end{cases} \tag{2.2}$$

In the training phase, '1' means likely to begin speaking, and '0' means not likely to begin speaking.

The input features consist of the nine features listed in Table 2.1. x_1 through x_3 represent the previous robot's utterance, x_4 through x_7 represent the robot's motion, and x_8 and x_9 represent the robot's head/eye direction. Section 2.3.3 describes the input features in detail.

We obtained these features every 0.1 seconds; therefore, the prediction is performed every 0.1 seconds. Although our prediction target, when a user is likely to begin speaking, depends on context in dialogue, we did not use contents of robot's utterance as features. This is because that using the contents as features may make the model depend a domain and need to retrain it when using in another domain. We use "speech interval", "utterance pattern", and "prosody" as features of the prediction model instead of the contents. These features implicitly represent dialogue context and are independent in domains.

We use non-verbal information as well as verbal information as features. Because robot behaviors can be controlled by the system developer, one of the ambitious applications of the prediction model is to control the users' behaviors in such a way that they will ignore noises. Moreover, it is difficult to predict whether the user is likely to begin speaking by using only verbal information. When the robot is not talking, but is moving with its back to the user, for example, the user is not likely to begin speaking. Thus, we predict the situation by using the robot's motion and head/eye direction.

Table 2.1: Input features obtained from robot behaviors.

$\mathbf{x}$	Input features	Details
x_1	Speech interval [seconds]	Elapsed time from end of previous robot's utterance
x_2	Utterance pattern (0 or 1)	Whether previous robot's utterance is interrogative
x_3	Prosody (0 or 1)	Whether previous robot's utterance ends with rising intonation
x_4	Motion (head) [degrees]	Angle difference from previous frame
x_5	Motion (left arm) [degrees]	Angle difference from previous frame
x_6	Motion (right arm) [degrees]	Angle difference from previous frame
x_7	Motion (legs) [degrees]	Sum of angle differences of legs from previous frame
x_8	Head/eye direction (horizontal) [degrees]	Angle position from front
x_9	Head/eye direction (vertical) [degrees]	Angle position from front

2.3 Building the Prediction Model

We faced two issues in building the model for predicting ease of speaking. First, we needed to prepare training data. We had to collect data from people that indicated when they are likely to begin speaking in a dialogue because there are no other studies on ease of speaking. Next, we needed to design features for the machine learning by using the robot's states.

The above considerations led us to proceed with this study as follows.

1. We made the robot's behavior sequences including its various motions and utterances of the robot.

2. We asked several participants to watch the sequences and to assign labels indicating whether they are likely to begin speaking to the robot at any time they felt appropriate during the sequence.

3. We built a model to predict ease of speaking by using the labels as supervisory signals and the robot's states obtained from the sequences as features.

2.3.1 Making Robot Behavior Sequences

To determine whether a user is likely to begin speaking to the robot, we took the following factors into consideration: the robot's posture, motion, and utterances; specifically, whether the robot turns towards the user, moves, or utters something. We only used factors that the robot can automatically obtain because these are what would be used as input features in the online interactions.

Whether a user is likely to begin speaking cannot be determined by just one factor because an actual robot's behaviors contain these factors in a *continuous* and *compound* manner. By *continuous*, we mean that directly after the robot exhibits one behavior, it exhibits another behavior. If the robot exhibits two or more behaviors continuously, a user needs to consider the relationship between them when deciding whether to speak. For example, a user may be likely to begin speaking when the robot turns towards her after speaking in another direction. By *compound*, we mean that the robot exhibits behaviors containing many factors at the same time. When the robot is exhibiting such behaviors, whether the user is likely to begin speaking depends on the specific combination of factors. For example, a user is likely to begin speaking when the robot says nothing. However, the user is not likely to begin speaking when the robot bows and makes no utterance.

Thus, we made two sequences that contain the various factors in a continuous and compound manner. The content of the sequences was a self-introduction by Nao [45], which is a humanoid robot made by SoftBank Robotics. We used Voice-Text [11] as the text-to-speech (TTS) engine. We called these two sequences X and Y. Sequence X was 150.0 seconds long, and sequence Y was 259.3 seconds long. We made sequence X first because it was simpler and contained mostly speech information. Sequence Y was longer and contained a more varied combination of factors than sequence X. Figures 2.2 and 2.3 show parts of sequences X and Y. Sequences X and Y include 30 and 39 utterances altogether. We used these sequences for making training and evaluation data sets after making manual annotations on them.

Part of sequence X is depicted in Figure 2.2. In (i), the robot waves its hand, turns right, and says "Hello". In particular, the robot moves its head, then its right arm, and then its right hand and left hand. Since we assume that the user is in front of the robot, the robot is not facing the user at the time. In (ii), the robot faces forward — that is, in the direction of the user — and does not move or speak. In (iii), the robot speaks for about one second and looks around. Here, (i) and (iii) exhibit compound factors, and (i) and (ii) exhibit continuous factors.

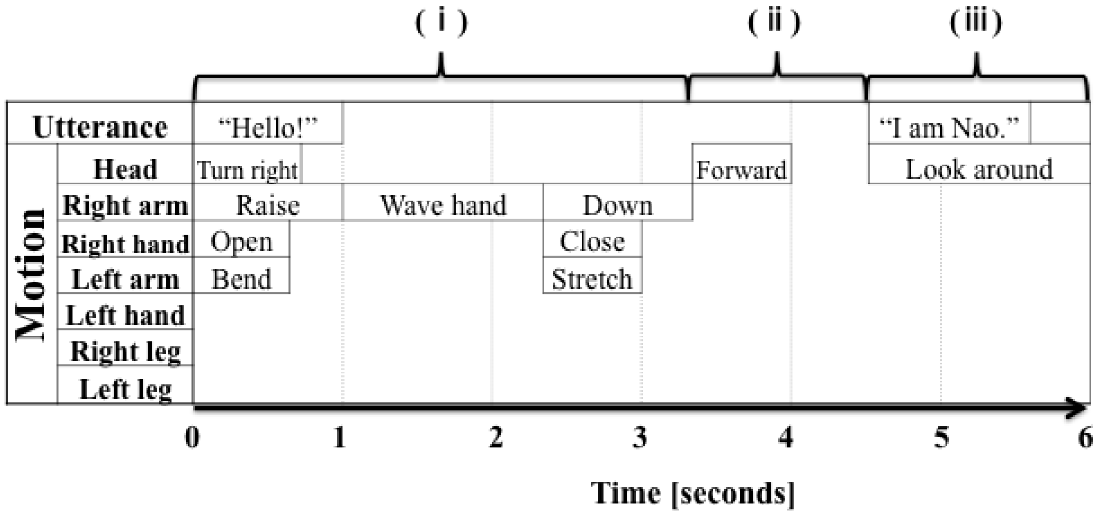


Figure 2.2: Part of sequence X.

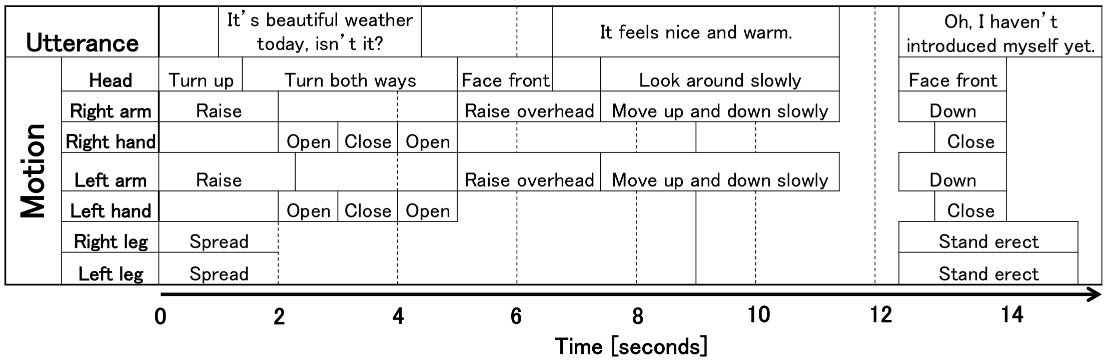


Figure 2.3: Part of sequence Y.

2.3.2 Data Collection

We collected a data set and had participants give it labels indicating whether they would be likely to begin speaking to the robot. The target data was the two robot behavior sequences described in the previous section. Because the prediction model should not be dependent on a specific user, we specified that

- multiple participants must be used, and
- each participant must annotate the data multiple times.

We asked three participants to make annotations as they watched a sequence of behaviors of a humanoid robot. The participants were students in our laboratory. The data collection procedure was as follows:

1. Participants were instructed on the experimental procedure and usage of a graphical user interface (GUI) for annotation and then allowed to practice with the GUI for a while.
2. Participants sat down in front of the robot.
3. Participants watched the sequence several times to prevent them from being surprised when watching the sequence for the first time.
4. Participants made annotations as to whether they would begin speaking by using the GUI. They watched the same sequence three times.

For simplicity, we conducted the data collection one by one. In general, there may be several users in public spaces. If several users are participating in the same conversation, whether a user is likely to begin speaking will vary depending on his position. For example, suppose a robot turns towards user A and stops silently, and user B is located to the left of the robot. In such a case, we expect that user A will begin speaking and user B will not because the robot is facing user A.

The GUI rigged to a computer display showed and recorded whether the participants indicated they were likely to begin speaking: “likely to begin speaking” was indicated by their keeping the mouse button pushed down and “not likely to

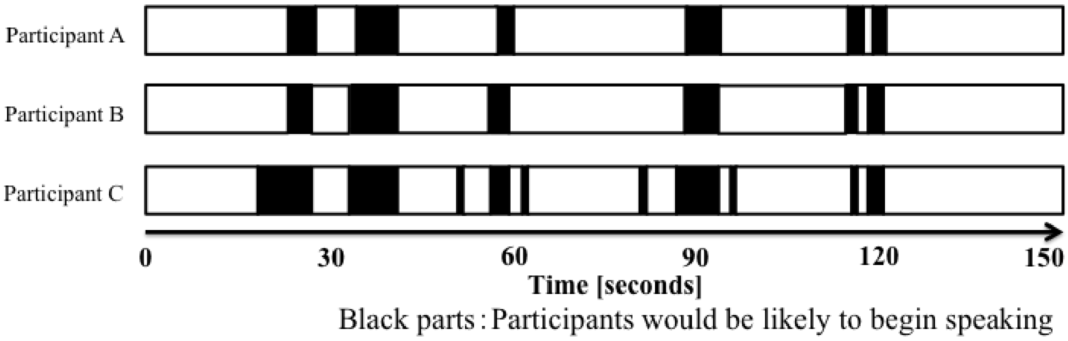


Figure 2.4: When the three participants would be likely to begin speaking for data X.

begin speaking" was indicated when they stopped pushing it. The robot's behavior sequences do not change regardless of whether the participants push down the mouse button. In general, their decisions would be affected by the preceding robot behaviors. However, we asked the participants to watch the entire sequence of the robot's behaviors to collect annotation data on the same conditions.

Written instructions were provided before the experiment to avoid any inaccuracies stemming from unreliable oral instructions. Participants could ask questions if necessary. Participants were also given the following instruction:

> Please indicate when you can say "Hey" to the robot to ask it to speak a little bit more loudly.

The labels given by the three participants are shown in Figures 2.4 (data X) and 2.5 (data Y). These figures show when each participant pushed the mouse button; that is, when they would be likely to begin speaking. The black parts denote when the participants would be likely to begin speaking, and the white parts denote when they would not be likely to begin speaking. The horizontal axis is the time of the robot behaviors. We analyze these data in subsection 2.4.1 to make training and evaluation data.

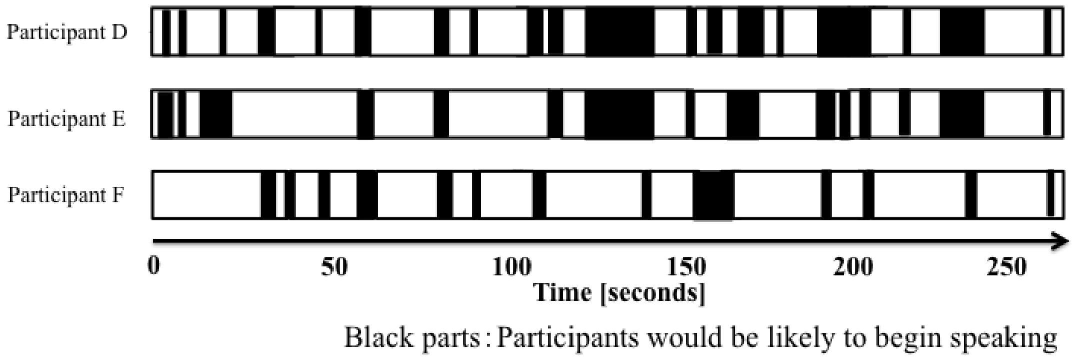


Figure 2.5: When the three participants would be likely to begin speaking for data Y.

2.3.3 Input Features

The proposed method uses the robot's state as the input feature to predict ease of speaking. Because the robot's state is controllable by the system developer, the proposed method has the potential of being used to control user behaviors as well as make the robot ignore noises. These states are defined at any time by using the robot's behaviors, such as its posture, motion, and utterances. In particular, we use the nine features listed in Table 2.1. These features were calculated by using the robot's behavior sequences every 0.1 seconds, as shown in subsection 2.3.1.

Speech interval (x_1) Feature x_1 represents the elapsed time from the end of the robot's previous utterance. An example is shown in Figure 2.6, where the horizontal axis is time, and the robot utterances are indicated by the gray blocks. The length of each gray block indicates the elapsed time of each utterance. A space between a gray block and the next bar represents a speech interval.

Generally, when a speaker talks continuously, a listener is not likely to begin speaking. In particular, as shown in Figure 2.6, the shorter the pause is, the more difficult it is for the listener to begin speaking to the speaker. On the other hand, the listener is likely to begin speaking when the pause is long. Thus, speech intervals are useful for predicting ease of speaking.

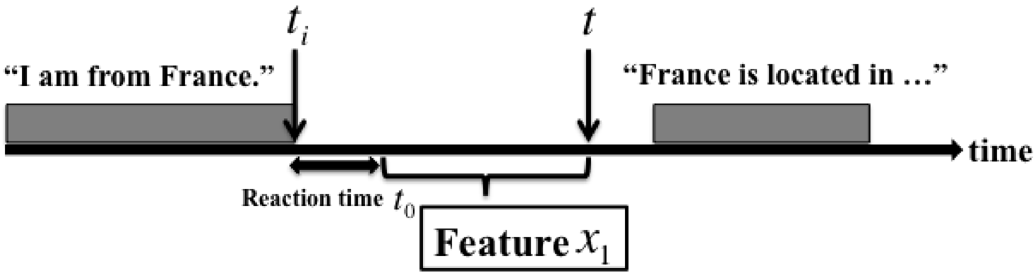


Figure 2.6: Feature x_1: Speech interval.

Feature x_1 is defined as

$$x_1 = \begin{cases} t - (t_i + t_0) & (x_1 > t_0) \\ 0 & (otherwise). \end{cases} \tag{2.3}$$

Here, t_i is the end time of the previous utterance i of the robot, and t is the current time. t_0 is a constant offset reflecting the time until the user perceives the end of a robot utterance. We set t_0 to 1.1 after a preliminary experiment. This feature was calculated for every time t of both robot's behaviors.

Utterance pattern (x_2) Humans can smoothly take turns in human-human dialogue because they predict the end of an interlocutor's utterance by using utterance patterns. For example, "question" and "request" evoke taking turns; i.e., the interlocutor will finish speaking soon.

Feature x_2 represents whether a previous robot utterance is a question. In particular, when the previous robot utterance is interrogative, this feature remains '1' until the next robot utterance starts. In this study, we manually assigned '1' or '0' to Feature x_2 every 0.1 seconds.

Prosody (x_3) Humans also uses prosodic information to predict the end of utterances. For example, people can predict the end of an utterance when the interlocutor says "How can I help you?" or "You are a student, right?". Duncan et al. [13] have shown that high rising intonation is correlated with the occurrence of turn-taking attempts.

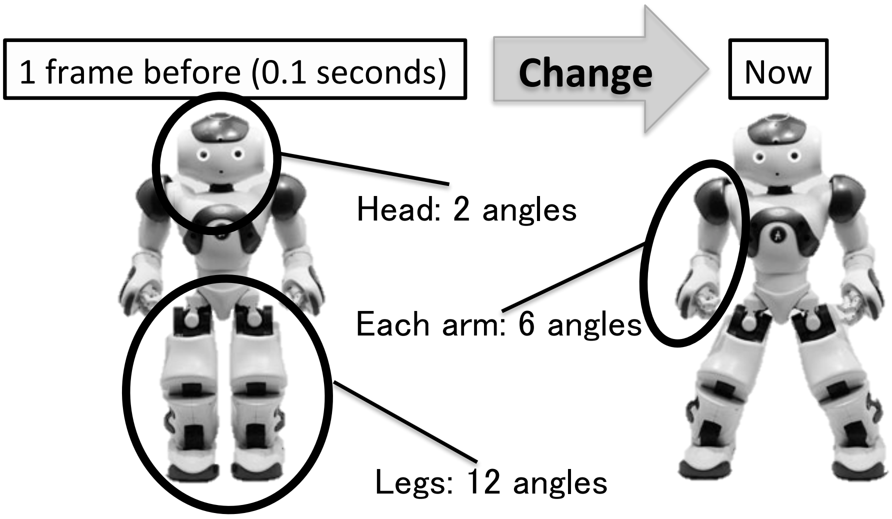


Figure 2.7: Features x_4 through x_7: Motion.

As well as for Feature x_3, when the previous robot utterance ends with a rising intonation, this feature remains '1' until the next robot utterance starts.

Motion (x_4, x_5, x_6, and x_7) When the robot is moving, a user is not likely to begin speaking. Features x_4 through x_7 represent the robot's motion and are defined by changes in the joint angles of the robot. The robot we used, Nao, has 26 joint angles, as shown in Figure 2.7, and their angle positions can be obtained via its Application Programming Interface (API). Here, "change" is defined as the absolute difference between the angle position of the current frame and that of the previous frame (i.e., 0.1 seconds before). We summed the differences for each part (head, left arm, right arm, and legs) to roughly represent the robot's motion and ignored small noises from the position sensors. We separately used each difference as a feature.

Head/eye direction (x_8 and x_9) When the robot looks towards the user, she is likely to begin speaking. On the other hand, when it looks in another direction, she is not likely to begin speaking.

Features x_8 and x_9 represent the robot's head/eye direction and are defined by using the angle positions of the robot's neck, which can also be obtained via

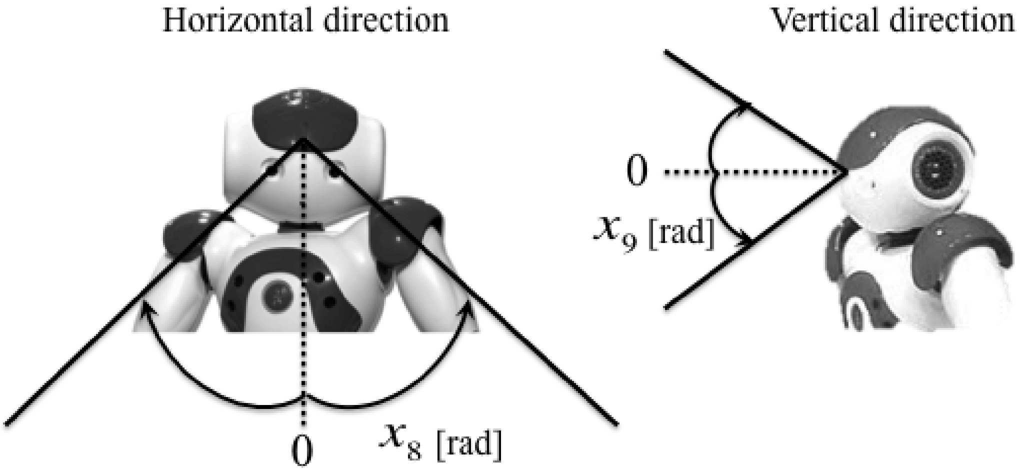


Figure 2.8: Features x_8 and x_9: Robot's head/eye direction.

its API. These features are defined as the absolute angle difference (in radians) between the user position and the robot's head/eye direction, since both indicate whether the robot has turned towards the user or not. We assume that the user is sitting down in front of the robot, so these features are simply angle positions from the front. The robot's head/eye direction is shown in Figure 2.8. The left and right figures are for features x_8 and x_9, respectively.

2.4 Experiment

2.4.1 Making Training Data

To make training data, we analyzed data X and Y collected in subsection 2.3.2. In particular, we investigated the differences among the participants on the basis of data X and Y. We tried to determine whether there were common parts at which the three participants were likely to begin or not begin speaking. This was done to check whether the collected data could be used as training data for the machine learning.

Using the common parts for training data enables us to make the prediction model by a small amount of training data. As described in subsection 2.2.1, we aim to build the model that can predict the timing when most people feel common ease of speaking regardless of individuality. It is supposed to ask many participants to make annotations with robot's behavior sequences including various motions and utterances and then use all the parts for training data, however this requires high cost for these data collection. Moreover, while this approach eliminates parts peculiar to individuals, it is difficult for the prediction model to handle the variation for any individual in public spaces. Therefore, we took an approach that the common parts were used as training data of the prediction model.

We obtained three data sets from each participant (each participant watched the same sequences three times). We used the second data set from each because we felt that the participants may not have been skilled enough to use the GUI during the first trial. We also thought that the participants might have forgotten to give labels for the first data set, although we tried to prevent them from being surprised. As for the third data set, some participants seemed tired by this time. Consequently, we used the two sequences from the second trial and denoted them as data X and Y. Each data set contained the annotations of all three participants.

We tried to determine whether the three participants gave the same labels to the same behaviors. For example, in data X (Figure 2.4), all three participants would be likely to begin speaking at 25, 40, 60, 90, 115, and 120 seconds during the robot behaviors. There were also several common parts where they were not likely to begin speaking. The results show that, in general, there are several common parts mixed in with a few that the participants did not agree on. We used the common parts as the training and evaluation data.

Here, we discuss the details of the collected data. Table 2.2 shows that the three participants gave the same labels for 135.0 of 150.0 seconds in data X and 143.0 of 259.3 seconds in data Y. The details of the parts given the same labels are shown in Table 2.3. In data X, they indicated they would be likely to begin speaking for 14.2 seconds and were not likely to begin speaking for 120.8 seconds during the 135.0 seconds. In data Y, they were likely to begin speaking for 16.1 seconds and not likely to begin speaking for 126.9 seconds during the 143.0 seconds. We assigned weights in accordance with the ratio of the two parts during the training

Table 2.2: Duration in which the three participants gave the same labels (in seconds).

	Data X	Data Y
Same labels were given	135.0	143.0
Different labels were given	15.0	116.3
Total	150.0	259.3

Table 2.3: Details of common labels (in seconds).

	Data X	Data Y
Likely to begin speaking	14.2	16.1
Not likely to begin speaking	120.8	126.9
Total	135.0	143.0

and evaluation phase.

2.4.2 Experimental Evaluation by Cross-Validation

First, we evaluated the performance of our model by cross-validation; i.e., within each data set. We used logistic regression as the machine-learning method as already described in subsection 2.2.2. The training and evaluation data sets are the common parts to which the three participants gave the same labels (Table 2.3). Because the data was discretized every 0.1 seconds, the data items numbered 1,350 for data X and 1,430 for data Y. We used the commonly agreed upon labels as the teaching signals of the target variable; that is, we assigned '1' when the agreed upon label was "likely to begin speaking" and '0' otherwise. The explanatory variables were the nine input features described in the previous section. We gave weights to the cases of likely to begin speaking in accordance with the numbers of the two labels (8.5 and 7.9) for data X and Y. The performance was measured in terms of prediction accuracy, i.e., the ratio of the number of correctly predicted labels to the number of labels agreed upon by everyone. A stratified ten-fold cross-validation was then performed.

The prediction accuracies of the proposed method were 87.4% and 92.1% for

Table 2.4: Prediction accuracy and subsets of features (%).

	Data X	Data Y
All features (proposed)	87.4	92.1
Utterance only	85.9	91.2
Motion only	63.2	76.0
Head/eye direction only	63.4	71.5

data X and Y, respectively (Table 2.4). We also showed how much the individual features contributed to the performance. We classified the nine features into three groups: "*Utterance*", "*Motion*", and "*Head/eye direction*". "*Utterance*" includes x_1 (speech interval), x_2 (utterance pattern), and x_3 (prosody). "*Motion*" includes x_4 (head motion), x_5 (left arm motion), x_6 (right arm motion), and x_7 (legs motion). "*Head/eye direction*" includes x_8 (horizontal head/eye direction) and x_9 (vertical head/eye direction). Their prediction accuracies are also listed in Table 2.4. The prediction accuracies of "*Utterance*" were 85.9% and 91.2 % for data X and Y, respectively. "*Utterance*" was the most effective for the prediction. "*Motion*" and "*Head/eye direction*" were less effective when used by themselves, but were helpful when used together with other features.

2.4.3 Experimental Evaluation on Open Data

We performed an additional experiment in which we completely separated the evaluation data set from the training data set to show that our model is effective on data sets other than the original training data set. We trained the model using data sets X and Y and then evaluated these models on data X. An evaluation using ten-fold cross-validation was also performed with data X. The case of using data Y was an open test. The results are summarized in Table 2.5. The prediction accuracies of the open test and the cross-validation were 88.5% and 87.4%, respectively.

The accuracy of the open test, 88.5%, was only 1.1% higher than that of the cross-validation. Since it was almost equivalent to that of the cross-validation, the model trained with a specific data set is also effective on another data set. This result demonstrates that our model does not depend on a specific training data

Table 2.5: Comparison of models with different training data (open vs. cross-validation).

Training data	Relationship with test data	Prediction accuracy for data X
Data Y	Open	88.5 % (1,195/1,350)
Data X	Cross-validation	87.4 % (1,180/1,350)

set.

2.5 Summary of this Chapter

In human-human dialogue, people usually take the state of their addressee into consideration when taking a turn; that is, they do not begin speaking at a random time. Assuming that humans begin speaking to humanoid robots in a similar manner, we constructed a model for predicting when a user is likely to begin speaking to a robot. In particular, we predicted this time by using a machine-learning method that takes a humanoid robot's behaviors (its posture, motion, and utterances) as input features.

We evaluated the prediction model by ten-fold cross-validation and evaluated the prediction accuracies in each test set: 92.1% and 87.4%. After that, we evaluated the prediction model on completely different data sets for the training and open test. The prediction accuracy in this case was 88.5%, which is almost equivalent to that of cross-validation (87.4%). This result demonstrated that this model does not depend on a specific training data set. Furthermore, it showed that the robot's motion, posture, and utterances are useful features for predicting ease of speaking.

Chapter 3

Building a Spoken Dialogue System for Humanoid Robots with a Prediction Model based on Social Norms

3.1 Introduction

In this chapter, we discuss that how the prediction model described in Chapter 2 can be used in actual spoken dialogues. In the research reported in Chapter 2, we showed the proposed method can predict whether the user is likely to begin speaking. However, before incorporating the proposed method in a spoken dialogue system for a humanoid robot, we need to discuss two issues of the prediction model.

First, in Chapter 2, we collected data from three students in our laboratory and used it as data for training the prediction model. In contrast, people who talk to a spoken dialogue system in public spaces would be members of the general public who vary widely in age and are of different sexes. Therefore, we need to determine whether the prediction model built in Chapter 2 was influenced by the specific participants (students in my laboratory).

Second, in public spaces, the robot has to talk with not only users who have been already known by it but also unknown users. The robot cannot also know what the user will speak in advance. Therefore, the robot handles variations in these. Ease-of-speaking depends on their character and on what they want to say. For example, some people are talkative, while others are taciturn. Talkative people may quickly respond to the robot when it finishes talking. In contrast, taciturn people may not quickly respond to it. Furthermore, when a person wants to talk about something urgent (for example, he wants to go to the bathroom), he may talk to the interlocutor immediately. In contrast, when he wants to adjust an air conditioner's temperature a little, he does not need to talk immediately. The prediction model needs to be able to handle variations such as these when it is used in public spaces.

In this chapter, we show three primary contributions to utilizing the prediction model based on social norms for humanoid robots in public spaces. First, we demonstrate that the method described in Chapter 2 does not depend on the specific participants. We collected data from 25 participants recruited from the general public and used it to retrain the prediction model. We determined its performance by conducting cross-validation and open tests. Second, we demonstrate that the prediction model can handle two variations that arise when the participants give labels. Essentially, when people are likely to begin speaking depends on two things: who the individual is, and which instructions are given prior to the data collection. Finally, we incorporate our method in an actual spoken dialogue system for a humanoid robot and show an example that the prediction model enables the robot to ignore irrelevant sounds.

3.2 Evaluating a Model that Predicts When People Will Speak to a Humanoid Robot

3.2.1 Data Collection using Participants from the General Public

This study aims at building a prediction model that does not depend on specific participants. To this end, we recruited 25 human participants from the general public and had them label the robot's behaviors sequences (X and Y) as in the experiment described in the previous chapter.

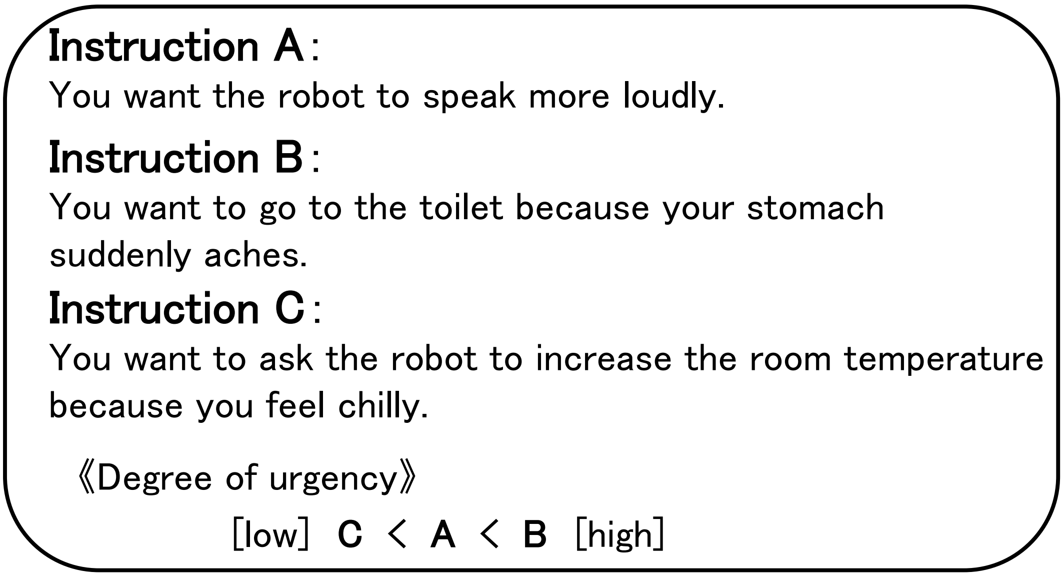
Table 3.1 compares the two data collections. The participants (13 males and 12 females) were from the general public, and their ages ranged from 20 to 50 years. The participants were equally distributed by age group. The average age of the participants was 37.9 years.

Participants were given three separate instructions related to the situation in each data collection, including the level of urgency (Figure 3.1). These instructions were not to limit the utterance content but to inform the participants of the degree of urgency. We specified the degree of urgency so that all participants would experience the same situation.

The data collection procedure is shown in Figure 3.2. Before the actual experiment, participants watched the sequences and practiced with the GUI so that they would be accustomed to the task. Data collections 1 and 2 differed only in the sequences. The 25 participants annotated each sequence three times with instruction A. We had them perform the annotation three times to habituate them to each sequence. Each data set was then used as training and test data. We gave different instructions (as shown in Figure 3.1) during data collection 3 from that used during data collections 1 and 2, to investigate the influence of different instructions. In particular, the participants labeled sequence Y by following instructions B and C. At the end the data collection process, the participants were asked to fill out a questionnaire with responses on a seven-point Likert scale, part of which is shown in Figure 3.3. The participants rated how much their feelings

Table 3.1: Differences in the data collections.

Data collections	Section 3.2.1	Section 2.3.2
Participants	The general public	Students in my laboratory
Number of participants	25 people	3 students
Sex	13 males and 12 females	All three males
Age	20-50 years old	About 20 years old



Instruction A:
You want the robot to speak more loudly.

Instruction B:
You want to go to the toilet because your stomach suddenly aches.

Instruction C:
You want to ask the robot to increase the room temperature because you feel chilly.

《Degree of urgency》

[low] **C** < **A** < **B** [high]

Figure 3.1: Part of instructions used in data collections.

changed when instructions B or C were given. These were rated relative to the case when instruction A was given.

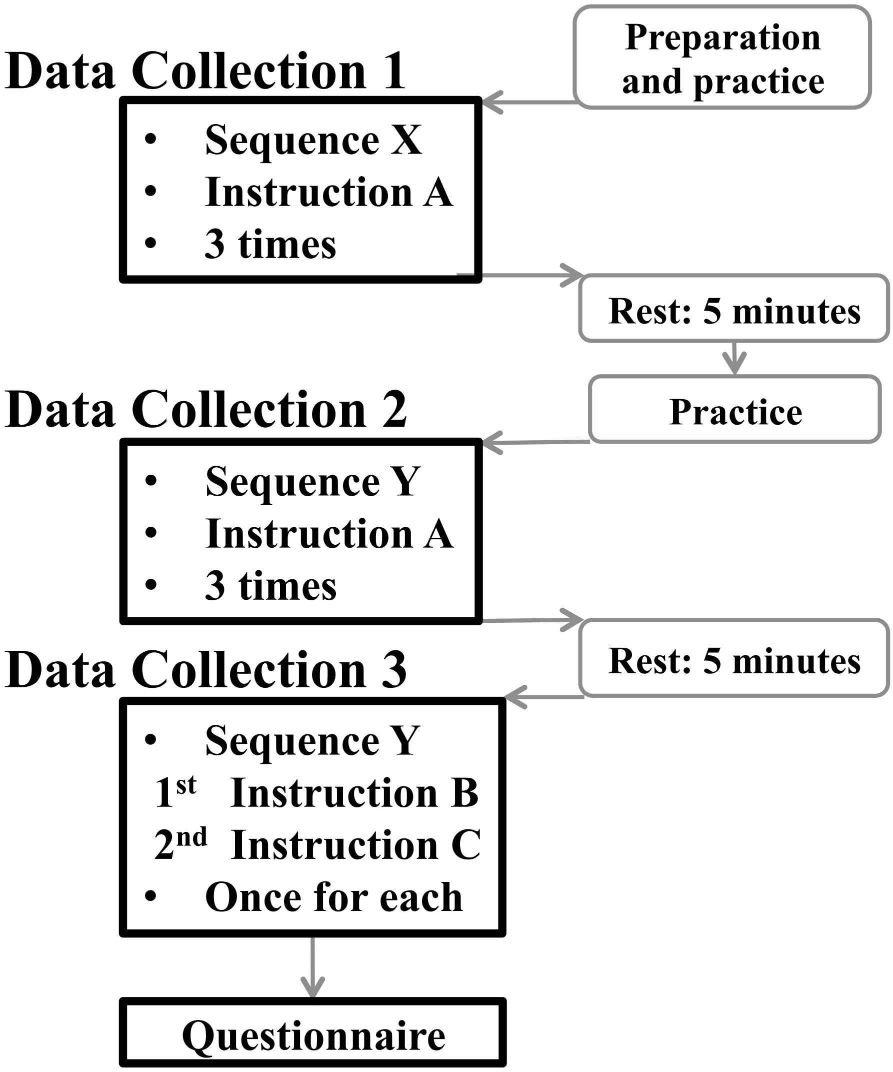
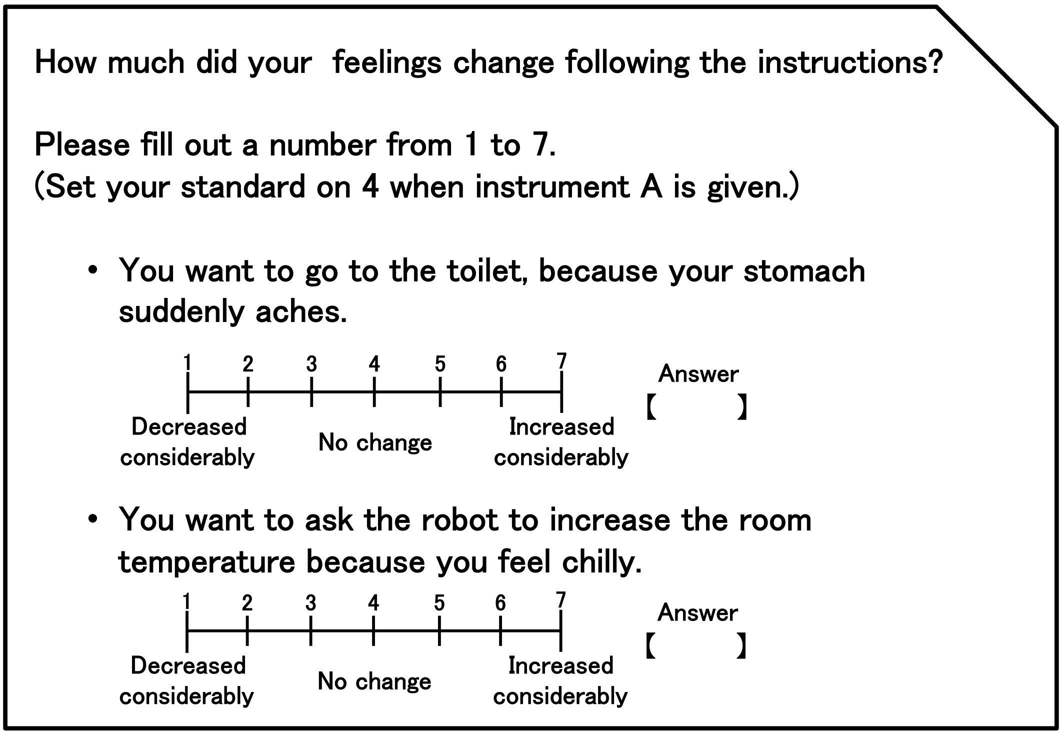


Figure 3.2: Data collection procedure.



How much did your feelings change following the instructions?

Please fill out a number from 1 to 7.
(Set your standard on 4 when instrument A is given.)

- You want to go to the toilet, because your stomach suddenly aches.

1 2 3 4 5 6 7

Decreased considerably — No change — Increased considerably

Answer 【 】

- You want to ask the robot to increase the room temperature because you feel chilly.

1 2 3 4 5 6 7

Decreased considerably — No change — Increased considerably

Answer 【 】

Figure 3.3: Part of the questionnaire.

Table 3.2: Training data and details of common labels (in frames).

	Data Y'	Data Y
Data collections	Section 3.2.1	Section 2.3.2
Sequence	Y	Y
Used parts	21 or more participants	All 3 participants
Likely to begin speaking	259	161
Not likely to begin speaking	1,123	1,269
Total	1,382	1,430

3.2.2 Making Training Data

We prepared two training data sets: data Y and data Y'. Data Y was the same as in Section 2.3.2. Data Y' was obtained from data collection 2 described above. We used the second out of the three trials for the same reason as before Section 2.4.1. The data are summarized in Table 3.2.

For data Y', we used the parts to which 21 or more participants gave the same labels, indicating whether they would be likely to begin speaking. The number 21 is 7/8th of 25. This number was derived from the probability when the labels by the three participants agree by chance is $(1/2)^3$, which corresponds to the condition in the previous data collection (i.e., data Y). Table 3.2 also shows the numbers of frames to which 21 of the participants gave the same labels: 259 are labeled likely to begin speaking, and 1,123 are labeled not likely to begin speaking. We gave weights to the cases of "likely to begin speaking" based on the ratio of the two labels (4.34).

3.2.3 Evaluation of Dependency on Specific Participants

We tried to determine whether the proposed method depended on the participants whose data were used in the previous data collection. We did so by comparing the new prediction model based on data Y' with the model built from the previous data (the one based on data Y).

The performance was measured in terms of the macro-average F1, which is the arithmetic average of the two F-values for the two labels, "Likely to begin speaking" and "Not likely to begin speaking". The macro-average F1 is a popular measure used in the text categorization community [60]. The F-value is the harmonic mean of the precision and recall rates. The numbers of the two labels were not equal, and we used this measure to equivalently evaluate the prediction performance of the two labels. In addition, we set a majority baseline to verify that a simple method did not perform well. We used the average of two F-values when all labels were '1' and '0' as the baseline method. Its macro-average F1 was 47.2.

We calculated the macro-average F1 for the following three conditions.

(1) Ten-fold cross-validation

(2-1) Open test for all data

(2-2) Open test for the parts to which t or more participants gave the same labels

We used the data for sequence X consisting of 1,500 frames as the evaluation target for Conditions (2-1) and (2-2). These data were completely separate from the training data set (sequence Y); this means the test was an open one.

For Condition (2-1), we used all the data, which means that all frames, including the uncommon parts, were used in the evaluation. The performances for Condition (2-1) are the averages of the macro-average F1 values for all participants. For Condition (2-2), we used the parts to which most of the participants gave the same labels. In particular, we set this condition as 18 or more participants giving the same labels for which the macro-average F1 was the highest. Eighteen or more participants gave the same labels to 1,362 frames; the numbers of frames at which they were likely and not likely to begin speaking were 135 and 1,227, respectively.

Table 3.3 compares the performances obtained under these conditions. The difference in performance between data Y' and data Y was at most 1.6 points under all three conditions. This difference was statistically insignificant ($P < 0.05$). This result shows that the new model based on data Y' performed almost equivalently to the model based on data Y. Namely, despite the difference in the participants' backgrounds (students in my laboratory vs. the general public) and the different numbers of participants, these models performed almost the same. This demonstrates that the proposed method does not depend on specific participants whose data were used in the previous data collection.

In contrast, the performance of (2-1) was lower than that of (2-2). For training data of the prediction models, we used the parts to which almost participants commonly gave the same labels to eliminate the effect of individuality. That is, it is difficult for the models to predict the parts to which participants gave different labels. Thus, the proposed model needs to be able to handle individual differences, which we will investigate in the next section.

Table 3.3: Performance obtained with the two training data sets (macro-average F1).

Training data (Sequence Y)	(1) 10-fold cross-validation	Open test for Sequence X		
		(2-1) All parts		(2-2) Common parts
		25 recruits from the general public	3 students	
Data Y’	90.2	69.8 ± 8.4	69.0 ± 2.5	84.9
Data Y	88.6	69.6 ± 9.1	69.4 ± 2.5	84.3
Baseline	47.2			

3.3 Handling Variations Caused by User-Annotated Labels

3.3.1 Handling Individuality

Next, we determined whether the variation caused by individuality, i.e., differences between interlocutors, can be represented by changing the threshold of the logistic regression. If this were possible, it would enable us to automatically set the threshold according to individuality. Moreover, this approach does not need to retrain the prediction model. If the robot predicts whether a specific user is likely to begin speaking, we should use the user's annotation for training data. In contrast, we want to use the prediction model in public spaces, and so building the model for each user is impractical.

In particular, we needed to verify that we set the optimal threshold for each individual and that the performance improved. We compared two cases. One is when the fixed threshold of the logistic regression is used for all participants. We examined the effect of setting the optimized threshold (0.5) in the training data and setting the threshold (0.4) of the best performance for the test data. The other case is when the threshold is changed for each individual. Here, we used all the data collected from the 25 participants in data collection 1 as the test data. After this verification, we discuss how the optimized threshold is set.

We verify that we set the optimal threshold for each individual and that the performance improved. Table 3.4 shows the average and standard deviation of the macro-average F1 of each trial of the 25 participants. The case when the threshold is fixed is the same as the case in Section 3.2.3 and corresponds to Condition (2-1) in Table 3.3. The optimized thresholds were set per participant to those when the macro-average F1 was the highest by varying the threshold from 0.1 to 0.9 in increments of 0.1. Figure 3.4 shows the performance improvement from the fixed threshold (0.5) to the optimized threshold for each individual. The performance improved 72 times out of the total 75 trials. The maximum performance improvement was 18.0 points in a trial by one participant. Table 3.4 shows that,

Table 3.4: Effect on performance when thresholds were changed.

	Macro-average F1		
Trials	First	Second	Third
Optimized	76.1 ± 6.1	74.3 ± 8.2	74.7 ± 7.7
Fixed (0.5)	70.0 ± 7.2	69.8 ± 8.4	69.1 ± 7.8
Fixed (0.4)	73.1 ± 8.8	72.8 ± 9.4	72.4 ± 9.3

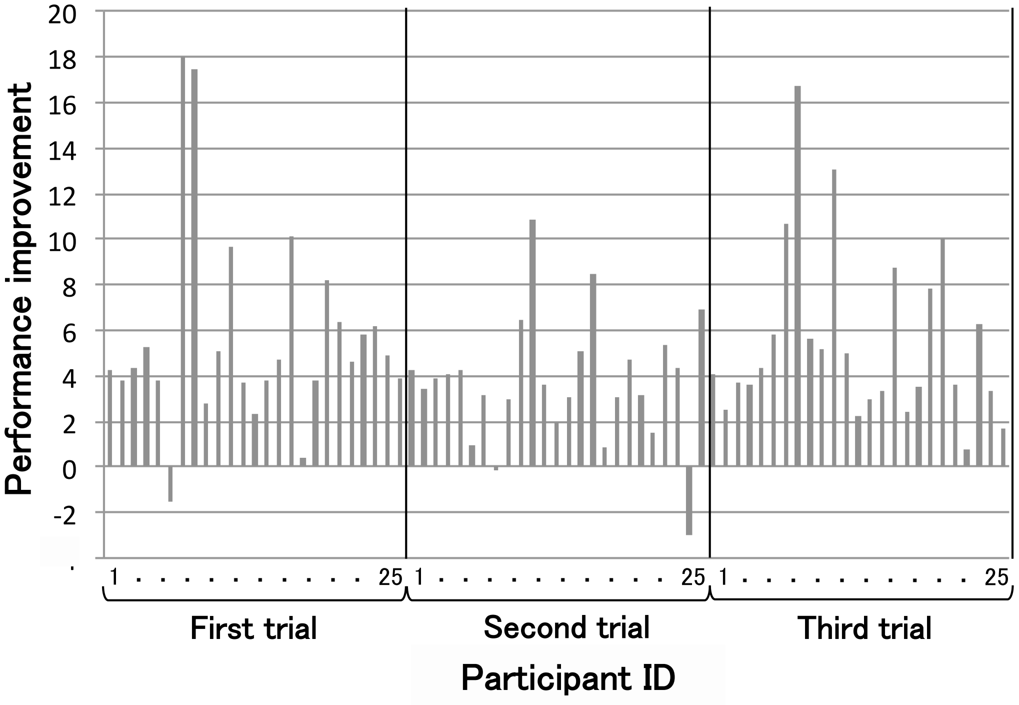


Figure 3.4: Performance improvement in going from a fixed threshold (0.5) to a changing threshold.

overall, the performance when the threshold was changed was higher than when the threshold was fixed (0.4 or 0.5). This demonstrates that changing the threshold for each individual results in better performance.

Next, we discuss how the optimized threshold is set. In particular, we analyzed the results in more detail by focusing on the relationship between the optimized thresholds and the number of frames in which the user is likely to begin speaking. We did so because we thought the optimized threshold might correspond to how much the individual participants felt likely to begin speaking. As an example, the optimized threshold for one participant was 0.3 and the number of the frames was

300, while another optimized threshold was 0.8 and its number was 83. In other words, when the number of the frames was higher, the threshold was lower, and vice versa. The relationship is shown in Figure 3.5. The correlation coefficient was -0.70, and its regression equation was $y = -447x + 400$, where x is the optimized threshold and y is the number of frames at which the user is likely to begin speaking. This result suggests that if we can set the threshold according to this personal attribute, i.e., how likely participants are to begin speaking, the performance of the proposed model will improve.

Furthermore, we consider whether this threshold can be automatically changed according to individuals online. To use the relationship in actual human-robot dialogue, we need to know how likely the individual is to speak and use it as her attribute. In particular, we need to prepare the number of the frames that can be obtained by having the users the labels indicating whether they are likely to begin speaking to a sequence prepared in advance. The author believes that the number of frames can be estimated from the information obtained during a conversation, such as the utterance frequency of the user, because it corresponds to how likely the user is likely to begin speaking. That is, the threshold is set to 0.5 in the first place as the default when a robot talks to a stranger. After that, the robot estimates the optimal threshold from the utterance frequency during a certain interval. Finally, the robot set the optimized threshold.

3.3.2 Handling the Influence of Instructions

To verify that the model can handle the variation caused by the influence of the instructions given to the participants, we investigated the relationship between the instructions given before the data was collected and the variations caused by users labeling the data differently. As described in Section 3.2.1, the instruction corresponded to the degree of urgency. Therefore, if we estimate the degree of urgency for each individual, we may be able to adjust the threshold of the logistic regression with the proposed model online.

We investigated the relationship between how much each participant felt it was

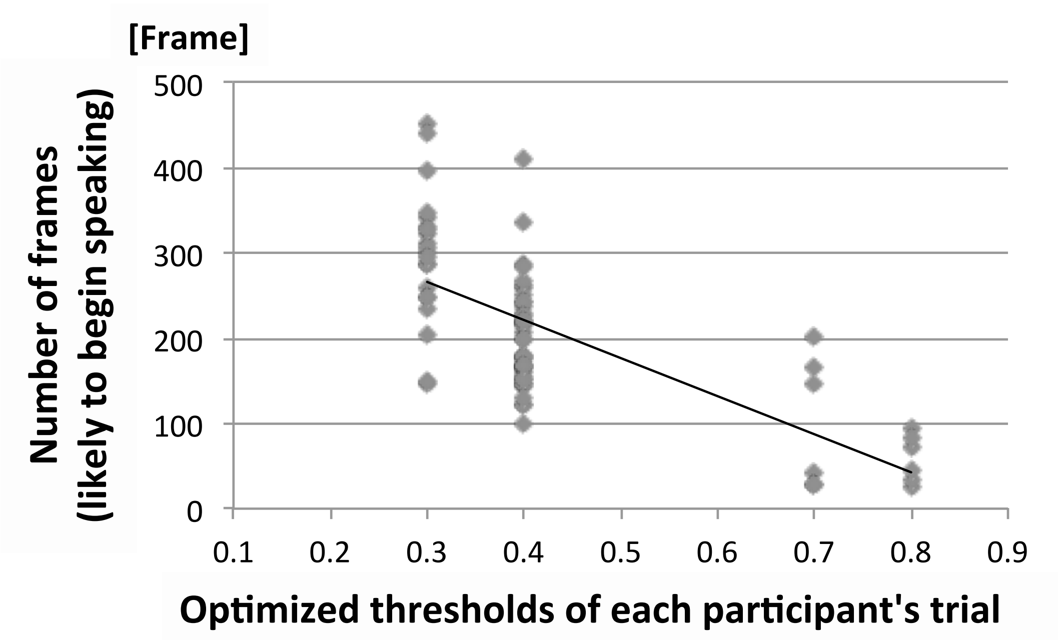


Figure 3.5: Optimized thresholds and number of frames.

possible to begin speaking and the participant's ratings on how likely they felt to begin speaking in comparison with the case of instruction A. Note that how much each participant felt it was possible to begin speaking means the difference in the number of frames in which the user is likely to speak. The relationship is shown in Figure 3.6. The correlation coefficient was 0.83, and its regression equation was $y = 204x - 834$, where x is each participant's rating and y is the fluctuation in the number of the frames compared with the case when instruction A was given. This result shows that these measures are correlated. That is, when participants feel more urgency, the number of the frames increases, and vice versa.

This tendency is helpful for adjusting the threshold during a conversation if we also obtain the user's degree of urgency in addition to the number of the frames explained in Section 3.3.1. The user's degree of urgency corresponds to the numerical ratings of the participants (Figure 3.6). Komatani et al. [30] showed that the user's degree of urgency could be estimated from the information obtained during a conversation, such as the frequency of barge-ins (a user begins speaking when a robot is speaking). This method enables us to estimate the user's degree of urgency and its results can then be used to change the threshold.

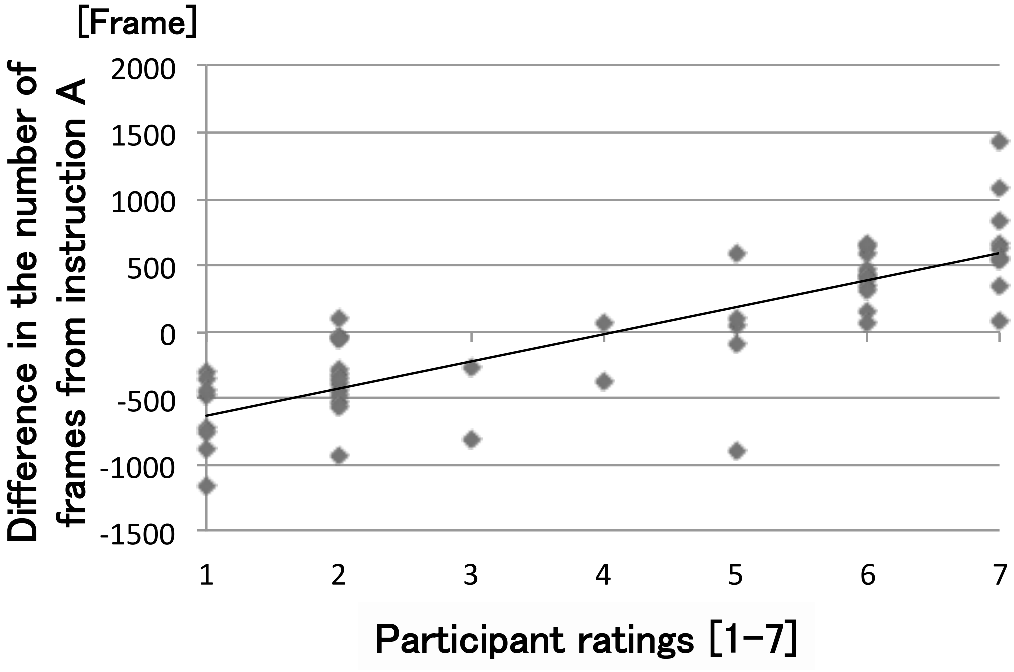


Figure 3.6: Relationship between ratings and how much each human participant felt it was possible to begin speaking.

3.4 Incorporating the Prediction Model in a Spoken Dialogue System for a Humanoid Robot

The previous section showed how humanoid robots can predict when a user is likely to begin speaking and handle two variations in dialogue. In this section, we introduce an example of a spoken dialogue system for a humanoid robot in a public space and how the prediction model is incorporated in the spoken dialogue system.

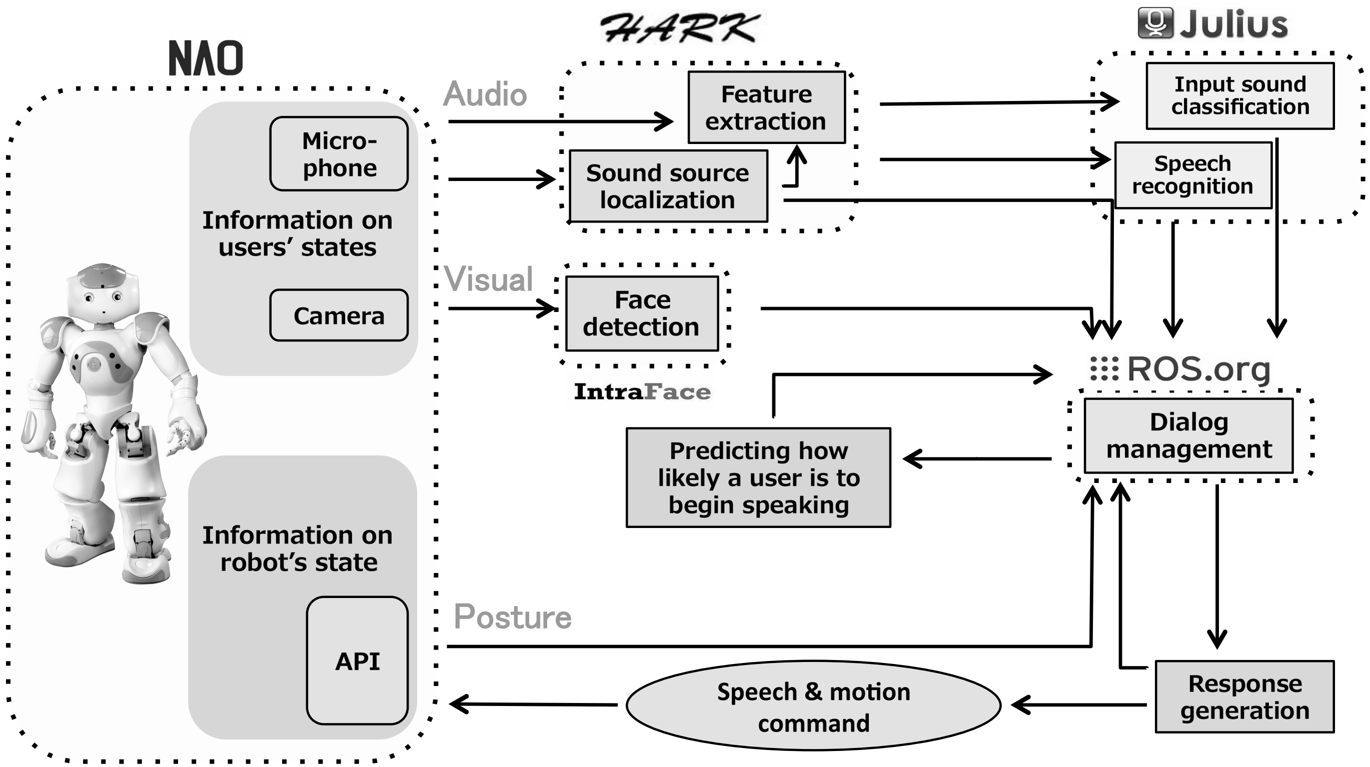


Figure 3.7: Overview of the spoken dialogue system for the humanoid robot.

3.4.1 Overview of Spoken Dialogue System

Figure 3.7 is an overview of the spoken dialogue system for a humanoid robot. We used the humanoid robot Nao [45], the speech recognition decoder software Julius [33], and the text-to-speech engine VoiceText [11].

We used multi-modal information (audio and visual information, as well as the robot's posture) as the system's input. Conventional spoken dialogue systems take only audio information as input. In contrast, in public spaces, the interpretation of sounds obtained from the robot's head-mounted microphones is much more difficult than the interpretation of sounds obtained a headset or directional microphone [32]. Therefore, to enable the robot to ignore certain sounds, we used not only audio information but also visual and posture information as input. For noise-robust human-robot dialogue, it is essential that the robot interpret this information comprehensively.

Audio information was obtained through the robot's four head-mounted microphones. This information was used for sound source localization, sound source separation, speech recognition, and for distinguishing human voices from surrounding noises. To distinguish these sounds, we made Gaussian mixture models (GMM) using the Hidden Markov Model Toolkit [63] (HTK) on the basis of acoustic features.

Visual information was obtained from the robot's head-mounted camera. From it, we obtained the interlocutor's position by using face detection and lip movement analysis using Intraface [12]. Several other studies have used visual information to detect a user's facial and lip movements [6]. By using this method, it is possible to recognize whether a user's mouth is moving, regardless of the acoustic environment. Thus, a sound can be regarded as noise if the user's mouth is not moving at the time.

Posture information was obtained via the application programming interface (API) and the dialogue manager. The posture information was used for calculating the input features of the prediction model based on social norms.

All the above information was collected by the I-O manager, which sent it to the language understanding and response generation module. The language understanding and response generation module sent speech and motion commands to the robot's API, NAOqi. As the robot's response, a voice sound file, loaded in Nao in advance, was played from the robot's speaker. The robot moved by sending commands from the language understanding and response generation module to the API.

The dialogue system ran the main processes, including speech recognition and sound source localization and separation, on an external computer because the robot's CPU did not have enough computational power to perform computationally these intensive processes.

Processing delay is often a problem when transferring large amounts of data from a robot to an external computer, especially if a multi-modal function is used. For example, if there is a delay in the input of the visual information at one time t, this input may be transferred at the same timing as the next audio signal at another time $t+1$. To prevent this from happening, we used the Robot Operating System (ROS) [42] as an I-O control. ROS has a function that manages each input and output of modules for each time series. We avoided the processing delay by managing the information needed to generate the response of the robot via ROS in advance. Furthermore, it is easy for ROS to add a new module to the dialogue system because ROS manages each module in parallel.

3.4.2 Discussion

We introduce a dialogue situation where the spoken dialogue system is useful. In particular, the situation is when the robot detects an unnecessary sound. Figure 3.8 shows the robot introducing my laboratory to a user. Figure 3.9 shows successful and failed examples of dialogues when a warning sounded from an ambulance nearby. In this case, the ambulance driver was giving an audible warning to pedestrians.

In the failed dialogue, the robot responded incorrectly to the warning from the ambulance. In the case of using an input sound classifier by using audio

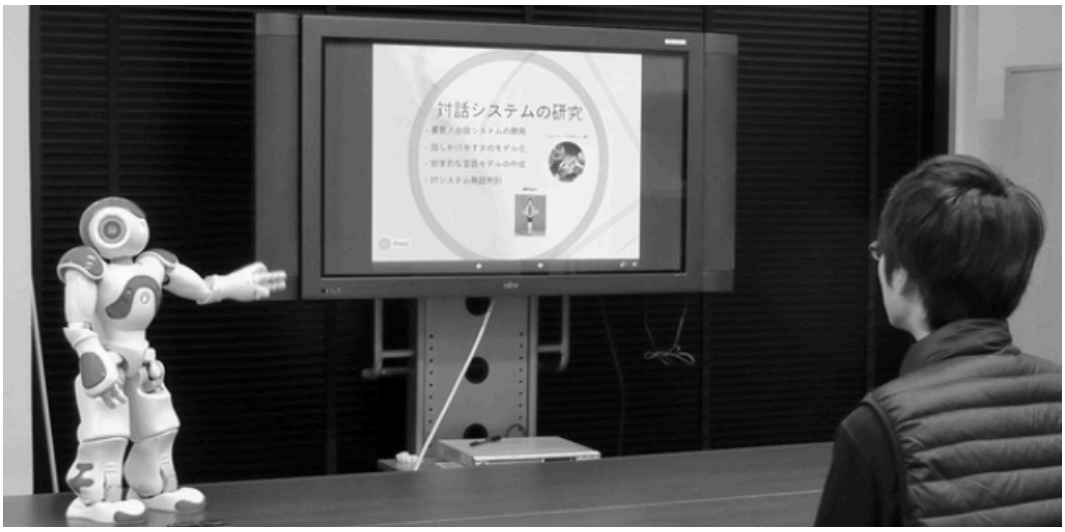

対話システムの研究
・複数人会話システムの開発
・話しかけやすさのモデル化
・効率的な言語モデルの作成
・対システム発話判別

Figure 3.8: The robot is introducing my laboratory to a user while the user is looking at screen.

information only, the output may be classified as a "human voice" because part of this warning is actually a recording of a human voice.

On the other hand, by using the proposed method, the robot knew when the user would likely to begin speaking. For example, when it turns on the TV and talks to the user, as shown in Figure 3.8, the user is not likely to begin speaking. Therefore, the robot estimated that the audio information detected when the user was unlikely to begin speaking as probably not speech, and hence it ignored this noise and waited for the next sound.

Utterance		Robot’s gaze
Robot:	What do you want to know?	Toward user
User:	I want to know something about the research field.	
R:	This laboratory's research field is spoken language systems and …	Toward slide
Noise:	Wee Woo Wee Woo…	
	Turn to the left. Please be careful.	

【Failed Example】	
R:	I’m Nao. I'm from France… （Robot responds incorrectly to warning from ambulance）

【Successful Example】		
R:	（Robot rejects noise and then waits for next sound）	Toward user
U:	I want to know about the boss.	
R:	Our boss is very kind…	Toward user

Figure 3.9: Successful and failed human-robot dialogues when an ambulance passed by.

3.5 Summary of this Chapter

This chapter showed that the method described in Chapter 2 is useful in actual human-robot dialogues. We described three primary contributions for incorporating the prediction model, i.e., the model for predicting whether a user is likely to begin speaking to a robot, in humanoid robots in public spaces.

First, we demonstrated that the prediction model described in Chapter 2 does not depend on specific participants. We rebuilt the prediction model with data collected from 25 participants recruited from the general public. We compared its prediction performance with that of the prediction model built (in Chapter 2) using data collected from participants who were students in my laboratory. Experimental results showed that these models performed almost the same despite the differences in the participants' background and the number of participants.

Second, we demonstrated that the prediction model can handle two variations, (1) individuality and (2) influence of instructions, that appear when participants give labels. Regarding (1), we found that the threshold of the logistic regression could be changed by investigating the relationship between the thresholds and the number of the frames. Regarding (2), we investigated the relationship between the number of the frames and the influence of the instructions given before the data collection, i.e., the user's degree of urgency. We showed that the user's degree of urgency could be used to predict whether the user is likely to begin speaking from the relationship between the participant ratings and the fluctuation in the number of the frames.

Finally, we introduced a spoken dialogue system for a humanoid robot in a public space and how the prediction model is incorporated in it. We described a dialogue situation where the proposed method enabled the robot to ignore unimportant sounds.

Chapter 4

Estimating Response Obligation in Multi-Party Human-Robot Dialogue

4.1 Introduction

This chapter discusses how a robot can determine whether it should respond to input sounds in a realistic situation, i.e., a multi-party human robot dialogue in public spaces. This chapter focuses on observations about users, while Chapter 2 focused on observations about the robots.

Communication robots are expected to be capable of interacting with people in public spaces, such as a reception desk of a restaurant or a hotel. In such environments, the robot receives various sounds, such as noises from its surrounding and users' voices, and furthermore, it needs to be able to interact with multiple people at the same time. Moreover, it is often suddenly talked to. If it incorrectly determines that it should respond to these sounds, it may, for example, erroneously speak to the surrounding noises or ignore a user talking to it.

In this chapter, we present a machine-learning based method to estimate a *response obligation*, i.e., whether an input sound should be responded to by a robot or not. The input sounds are all sounds that occur while the robot is interacting

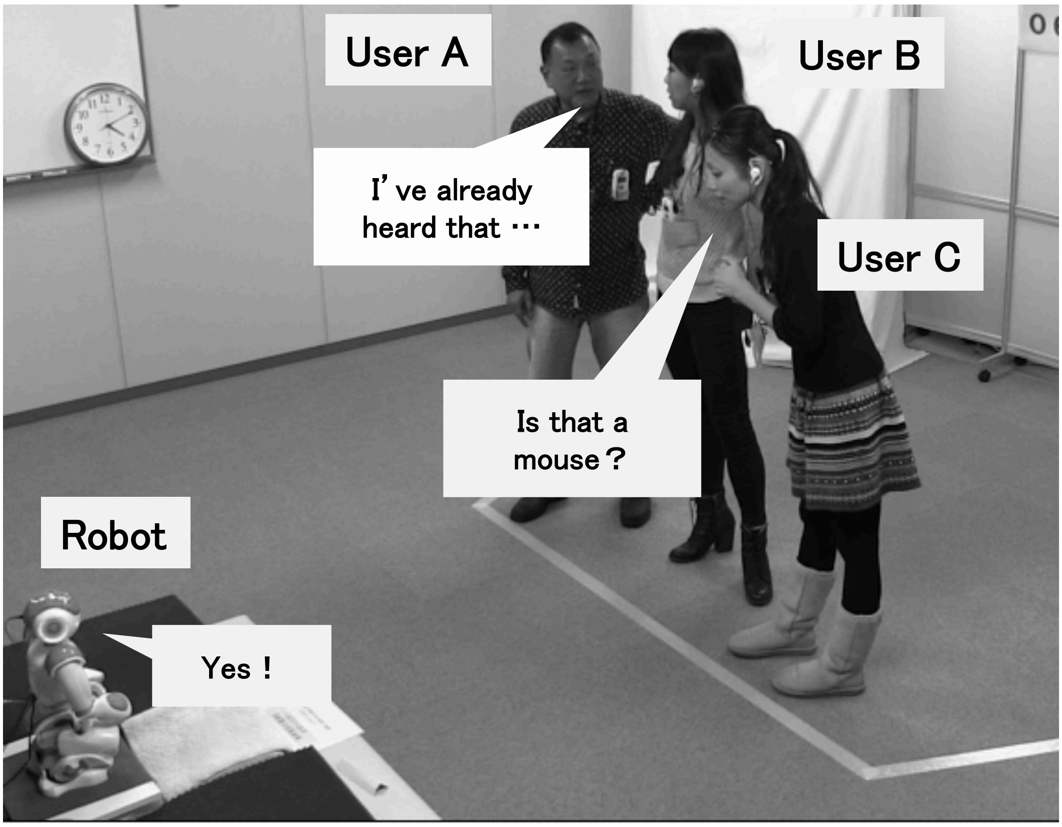


Figure 4.1: Interaction among three people and a robot (condition of data collection).

with users. We give each sound segment one of two labels: *ought-to-respond* and *ought-not-to-respond*. The former is given to those to which the robot should respond. The latter is given to user monologues, user utterances towards other users, surrounding noises (e.g., footsteps and motor noises), and user utterances to which the robot should not respond, such as interjections.

This method is useful for preventing the robot from making incorrect or unnecessary responses in a multi-party situation. For example, let us assume a situation in which three users and a robot interact, as shown in Figure 4.1. Here, user C is speaking to the robot, and user A is talking to user B. If the robot correctly estimates a sound segment corresponding to the utterance of user A as *ought-not-to-respond*, the robot can reject it and continue talking with user C. If the robot estimates a sound segment corresponding to an utterance of user C as *ought-to-respond*, the robot can correctly respond to the utterance.

Response obligation is basically the same concept as discourse obligation [55].

However, response obligation includes the case of multi-party dialogue; discourse obligation has only been discussed in the context of two-party dialogue. Furthermore, we take into consideration other input sounds, not just human speech. Furthermore, Jahansson et al. [22] dealt with the same problem setting, i.e., estimating response obligation in multi-party dialogue. They suggested a method to estimate response obligation by using the state of play and players' gazes while playing cards. They used typical features of playing cards and their input sounds were recorded by microphones in the users' headsets; we do not use such information because we want to build a robot that can talk with users in a real environment. That is, we pursue a method that can be used in wider dialogue domains.

To estimate the response obligation, we mainly use acoustic and visual information: an input sound classification result and a user's whole body motion both during and after the sound segment. The input sound classification result is useful for distinguishing speech and non-speech sounds. The whole body motion enables us to make use of differences in user behaviors when they speak to the robot and those when they speak to other users.

The novelty of this study is two-fold. First, we address a more realistic problem setting than those assumed in previous studies. The previous studies assume that the input sounds are utterances only that are towards one of interlocutors, while we deal with not only utterances but also noises. Therefore, our problem setting is closer to actual interactions in public spaces. We demonstrate that our model can determine whether the robot should respond to input sounds and that it outperforms a conventional model based on a previous study [40]. Second, we introduce new features that improve the estimation accuracy of the response obligation in a more realistic situation. We investigate the performances of models using various combinations of features and find that the new features are helpful for estimating the response obligation.

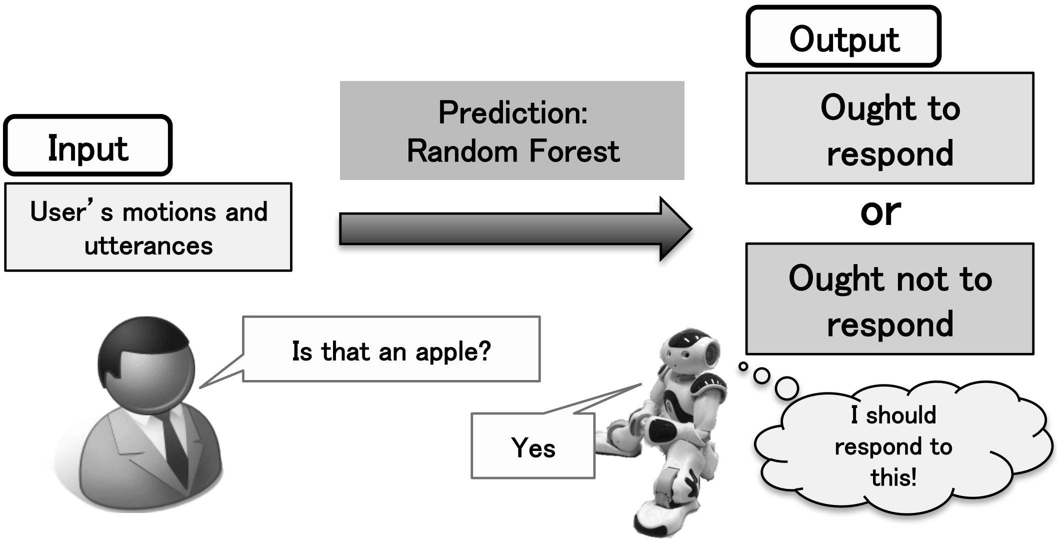

Input
User's motions and utterances
Prediction: Random Forest
Output
Ought to respond
or
Ought not to respond
Is that an apple?
Yes
I should respond to this!

Figure 4.2: Overview of estimating response obligation.

4.2 Estimating Response Obligation

4.2.1 Overview of the Proposed Method

An overview of the method is shown in Figure 4.2. When a user asks a robot, e.g., "Is that an apple?", it detects an input sound segment and estimates the response obligation to it, i.e., classifies the sound as *ought-to-respond* or *ought-not-to-respond*. This is a binary classification task using various input features obtained from acoustic and visual information during and after the segment. Although the response obligation might be gradual in nature, we treat it as a binary notion. If it is classified as *ought-to-respond*, the robot responds to the speaker. Otherwise, it ignores the sound. We do not handle situations when several users simultaneously talk to the robot because they do not occur frequently in current human-robot interaction.

4.2.2 Features for Handling Various Sounds in Public Spaces

Two issues affect the response obligation estimate.

1. The robot has to classify whether the detected sound is a voice.
2. If the sound is a voice, it has to determine whether the addressee is the robot

and whether the robot should respond to it.

The second issue is strongly related to a previous study on addressee identification in multi-party dialogue [40], which deals with identifying which interlocutor is the addressee for each utterance. The response obligation is easier to estimate in this study than in the previous one because the robot only has to determine whether the addressee is itself. On the other hand, our study does not assume that the addressee of an utterance always exists; i.e., our study includes monologues as an estimation target. In that sense, our study is more difficult than the previous one because the robot needs to be able to determine when the utterance is a monologue.

To handle various sounds in public spaces, we use the following feature sets.

(a) User's face direction during the sound segment,

(b) Prosodic information,

(c) Duration of the detected sound segment,

(d) Input sound classification result,

(e) User's motion during the sound segment,

(f) User's motion and face direction after the sound segment,

(g) Robot's dialogue act type just before the sound segment.

Table 4.1: List of feature sets.

		Feature sets	Features	Units and types	Number of features
Previously used features	(a)	The user’s face direction during the sound segment	Average face rotation angle during the sound segment	[degrees]	9
			Average angular velocity, maximum angular velocity	[degrees/frame]	
	(b)	Prosodic information (voice probability, F0, loudness)	Average of each information, average of inter-frame differences, maximum inter-frame difference (only), difference between the intrasegmental average and the average for all sound segments	Voice probability [%] F0 [Hz] Loudness [phons]	10
	(c)	Duration of the sound segment		[seconds]	1
Newly introduced features	（d）	Input sound classification results	Classification result [speech/non-speech]	labels	2
			Relative likelihood	probability （0.00-1.00）	
	(e)	The user’s motion during the sound segment	Average velocity (head, spine, right elbow, and, left elbow), Total change, Maximum velocity (head)	[meters/frame]	18
	(f)	The user’s motion and face direction after the sound segment	Velocity (head, spine)	[meters/frame]	9
			Average angular velocity (yaw, pitch, roll)	[degrees/frame]	
	(g)	The robot’s dialogue act type just before the sound segment	25 dialogue act types	Dialogue act labels	1

Feature sets (a) through (c) were used in the previous study on addressee identification [40]. We use these feature sets for estimating response obligation. In contrast, we do not use the other features used in the previous study [40] because the following assumption was satisfied to use them: the system has already recognized a speaker. For example, they used the ratio of the time that the speaker keeps looking at the system to the utterance time as a feature for identifying the addressee. We cannot use such features because we do not limit the detected sounds to utterances. Thus, we only use the features that do not need to limit the detected sounds to utterances.

Feature sets (d) through (g) are new ones for estimating the response obligation. Section 4.3.3 and Table 4.1 describe these features. The next subsection explains the concepts of this new information.

4.2.3 New Feature Groups Used in This Study

We use four new feature sets to estimate the response obligation.

(d) Input sound classification result

Our target sounds include non-speech sounds, such as surrounding noises and the robot's motor noise. Therefore, to estimate the response obligation, the robot needs to determine whether a detected sound is a user utterance. To distinguish speech and non-speech sounds, we used the results of a previous study on input sound classification [34] into two classes of Gaussian Mixture Models. We use the classification result and the relative likelihood as features.

(e) User's motion during the sound segment

The user's whole body motion during the sound segment enables us to capture differences in user behaviors between when they speak to the robot and when they speak to another user. For example, when speaking to the robot, a user tends to stand still. In contrast, when speaking to other users, she tends to move her head a little and slowly sways because she is relaxed. We thus use these tendencies as features.

(f) User's motion and face direction after the sound segment

We use the user's whole body motion after the sound segment to exploit typical user behaviors in human-robot interaction. In general, when a user asks the robot a question, a few second delay occurs before it responds to her. During this time, she tends to stand still while facing it, because she expects the robot to respond. On the other hand, when she speaks in monologue or talks to other users, she usually sways slightly because she is relaxed. Turnhout et al. [56] analyzed dialogues among two people and a system and empirically demonstrated that a longer silence can be observed after system-directed utterances than after utterances directed to the other user. Taking this tendency into account, the case will likely be *ought-to-respond* if a user stops moving after her utterance, and *ought-not-to-respond* otherwise.

(g) Robot's dialogue act type just before the sound segment

When a robot greets cooperative users, it is typical for users to respond instantly. On the other hand, when it asks a question, the users tend to insert interjectional utterances or consult with each other. Thus, the dialogue act type of the robot's preceding utterance is expected to be useful for estimating the response obligation.

We do not use verbal information except the dialogue act type of the preceding robot utterance. Katzenmaier et al. [24] identified the addressee in human-robot dialogue by using automatic speech recognition and face tracking. Zuo et al. [64] suggested a method to detect robot-directed utterances for understanding object manipulation requests in a real environment. They used speech recognition and the consistency of the dialogue situation to detect these utterances. We believe that it is desirable to estimate the response obligation without using verbal information such as speech recognition results. This is because that speech recognition results vary from one domain to another, and avoiding speech recognition errors is difficult for a robot, especially when it stands apart from users who are freely talking with each other, as shown in Figure 4.1. That is, even if it is useful in this study, it might not always be useful in another domain. Instead, we use the dialogue act types of the robot utterances.

4.2.4 Formulation

Response obligation is estimated for each input sound segment k, whose start and end times are denoted as $s^{(k)}$ and $e^{(k)}$, respectively. The estimation can be represented as $y^{(k)} = f(\mathbf{x}^{(k)})$, where f_e denotes the estimator and $y^{(k)}$ is a binary value:

$$y^{(k)} = \begin{cases} 1 & (\textit{ought-to-respond}) \\ 0 & (\textit{ought-not-to-respond}). \end{cases}$$

$\mathbf{x}^{(k)} \equiv (x_1^{(k)}, ..., x_N^{(k)})$ is an input feature vector having N dimensions. We calculated feature sets (a), (b), (c), (d), and (e) from interval $(s^{(k)}, e^{(k)})$, as listed in Table 4.1. For (f), we obtained the feature set from interval $(s^{(k)}, s^{(k)} + \alpha)$. Here, the constant α denotes the duration in which to collect the user's motion and face direction after the sound segment. For (g), we used the dialogue act type of the robot utterance starting just before $s^{(k)}$. We used Random Forests as the estimator f_e because it performed best among several methods, including logistic regression, support vector machine, and decision tree, in our preliminary experiment. Moreover, we believe that decision tree models are more effective than log-linear models because our feature sets are strongly correlated. For example, when a user looks toward a robot (user's face direction), his voice power (prosodic information) is important for estimating response obligation, and vice versa. That is, we chose Random Forests, which are one of decision tree models and are considered combinations of input features. We trained the estimator with the data described in the next section.

4.3 Experiments

4.3.1 Dialogue Corpus for Training and Test Data

We used the multi-party dialogue corpus collected in a Wizard-of-Oz manner [27] as the training and test data. This corpus consists of 60 human-robot multi-party interactions, where each of 30 trios (90 participants in total) engaged in two interactions with a humanoid robot, Nao. In each interaction, up to three participants played a quiz game with the robot, as shown in Figure 4.1. Each interaction lasted about 25 minutes. The participants could enter or leave the place at any time; there were times when one to three participants took part in the game. The robot was controlled by a human operator in another room, who judged whether it should respond to input sounds. Participants were instructed that they should speak in Japanese or English, while it spoke in English.

We used the 12 sessions that had already been fully tagged from all the sessions as the training and test data. The total duration of the 12 sessions was about 320 minutes. The 12 sessions of the dialogue corpus consist of three kinds of data (each kind of data was saved in a file.):

1. Kinect data: Videos (color and depth) recorded by a Microsoft Kinect [37] camera set behind the robot. We used Kinect for Windows v1.

2. Dialogue sound data: Sounds recorded as WAV format files by a microphone set behind the robot.

3. Annotation data: Speakers, speech segments, addressees, gazes, participation statuses, dialogue act types, and transcriptions were manually annotated with ELAN [59].

The dialogue act types of the annotation data were annotated with the 25 kinds of dialogue act types proposed by Kennington et al. [27].

Table 4.2: Data size and breakdown by target labels.

Sound segments	Response obligation label	Number of segments	Total
Speech	*ought-to-respond*	871	871
	ought-not-to-respond	2,421	3,135
Non-speech	*ought-not-to-respond*	714	

4.3.2 Annotating Labels for Sound Segments

Table 4.2 is a breakdown by target label in our data. Sound segments consist of speech and non-speech segments. We gave two labels, i.e., *ought-to-respond* and *ought-not-to-respond*, to both segments. Speech segments were manually labeled, from which various features were extracted. This corresponds to an ideal condition in which speech segments are successfully detected. We labeled the speech segments with dialogue act type tags and addressee tags. *Greeting*, *Answer*, and *Time-Management* are examples of dialogue act types. An addressee tag indicates a person to whom a user or robot speaks. When the robot said "Hello" to user A, this segment was labeled with *Greeting* as the dialogue act type and with "To_A" as the addressee.

We gave the *ought-to-respond* label to the speech segments to which the robot actually responded. As mentioned before, the robot was controlled by an operator. In other words, the robot's responses to speech segments meant that the operator judged the segment as *ought-to-respond.* Taking this point into account, we set the *ought-to-respond* label based on the robot's actual responses. The sound segments labeled with *ought-to-respond* did not include non-speech segments because the robot should not respond to non-speech sounds such as surrounding noises. Figure 4.3 illustrates a speech segment labeled *ought-to-respond.* We gave the *ought-to-respond* label to a user utterance directed to the robot that the robot replied to within ten seconds in which no other utterances of the same user were directed to the robot. We treated the robot utterances as replies only when they had a specific reply-related dialogue act type such as *Answer* to exclude vague cases where it was hard to determine whether the robot actually responded or not.

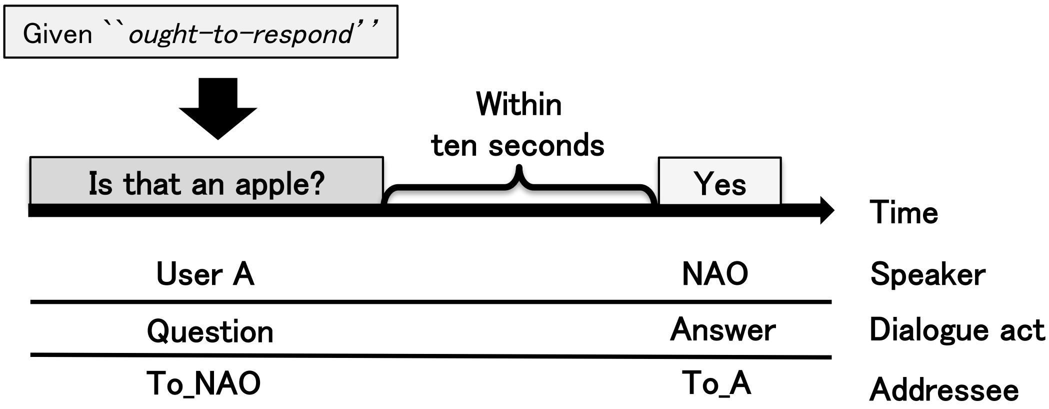


Figure 4.3: Example of annotation of speech segment of *ought-to-respond*.

We gave the *ought-not-to-respond* label to the following sound segments:

1. speech segments that should not be responded to by the robot, and

2. non-speech segments.

First, we gave *ought-not-to-respond* labels to user speech segments that do not necessarily require the robot to make a response, such as *Time-Management* and *Monologue*. Next, we gave *ought-not-to-respond* labels to non-speech segments corresponding to surrounding noises (footsteps, hand clapping, etc.) and the robot's motor noise. The non-speech sounds had not been annotated, and thus, we automatically extracted the segments where the energy was larger than a certain threshold by using the Julius adintool software[1].

Table 4.2 lists the data sizes of the labels. The number of speech segments for *ought-to-respond* is 871, and the number of speech and non-speech segments for *ought-not-to-respond* is 3,135.

4.3.3 Input Features for the Response Obligation Estimation

This section explains the feature vector $\mathbf{x}^{(k)}$. Note that section 4.2.3 describes the concepts of the new features. We used 50 features grouped into seven sets, as

[1] http://julius.sourceforge.jp/juliusbook/en/adintool.html

listed in Figure 4.1. Feature sets (a) through (c) were used in the previous study on addressee identification [40]. Feature sets (d) through (g) are the new features.

(a) User's face direction during the sound segment (nine features): This feature set consists of the following features calculated per input sound segment: the average of each face rotation Euler angle (yaw, pitch, roll) during the sound segment, the average angular velocity of each face rotation angle, and the maximum angular velocity of each face rotation angle. This results in nine features.

We can obtain the face direction information of the user from the Kinect data by using Kinect SDK every 30 msec (33.3 frames per second). The face direction information for each frame can be obtained as Euler angle components (yaw, pitch, roll). Each value of this feature set was calculated by using the face direction information. The average of each face rotation Euler angle was calculated by averaging out each face rotation Euler angle obtained in each frame during the sound segment. Furthermore, we calculated the absolute difference in the angle from the previous frame and used the average of the difference and the maximum difference during the sound segment as features.

We can use Kinect SDK to capture direction information for up to two users' faces (and skeleton information as well) in the order of their distance from the Kinect. We determined the target user whose face and body are to be used as features on the basis of the manual annotations of the speech segments, that is, the speaker who talked. This corresponds to a situation where we can correctly obtain the localization results of the sound segments. We did not use automatic sound source localization methods. On the other hand, state-of-the-art sound source localization technology can identify the speaker [1]. For the non-speech segments, we used the face and body information of the user who was the closest to Kinect.

(b) Prosodic information (ten features): We obtained three pieces of prosodic information from the dialogue sound data for every 10-msec frame by using openSMILE[14]:

1. voice probability (the ratio of the harmonic components to the total power)

2. F0

3. loudness

This feature set consists of the following features calculated per input sound segment: the average of each piece of prosodic information over the sound segment, the difference between the intrasegmental average, and the average for all sound segments of the three pieces of information, the averages of inter-frame differences for the three pieces, and the maximum inter-frame difference in only loudness, resulting in ten features.

The three pieces of information were calculated for each sound segment (4,006 segments in total), as shown in Table 4.2, and averaged. The averages of the inter-frame differences were gotten by calculating the absolute difference in each piece of information from the previous frame. The maximum inter-frame difference in loudness is the largest inter-frame difference in the sound segment.

(c) Duration of sound segment (one feature): This feature set represents the duration of the sound segment. We used the speech intervals and non-speech intervals extracted in Section 4.3.2 as speech segments and non-speech segments, respectively.

(d) Input sound classification result (two features): This feature set consists of results of input sound classifications using GMMs and their relative likelihood. The Gaussian Mixture Model has been a major method for these verification [34] because of its text-independency and powerful classification performance. For the classifier, we built two GMM classes: speech and non-speech. As the training data for GMMs, we used sound segments extracted from ten sessions. The speech GMM was trained with the speech segments that were manually annotated in the ten sessions included in the dialogue corpus. The non-speech GMM was trained with non-speech segments that were automatically extracted from the sessions by using adintool, as mentioned before. We regarded the sound segments detected outside the annotated speech segments as non-speech segments. The total durations of these training data for the “speech” and “non-speech” GMMs were about 7,320 and 671 seconds, respectively.

We built GMMs using HTK [63]. We set the number of mixtures to 16 on the basis of the results of a preliminary experiment. GMMs were trained with 26-dimensional features, i.e., 12-dimensional Mel-Frequency Cepstrum Coefficients (MFCC) vector, 12-dimensional Δ MFCC vector, power, and Δ power. We used the foreground noise detection function of the Julius speech recognizer as an input sound classifier.

(e) User’s motion during the sound segment (18 features): This feature set consists of the following features calculated per input sound segment: the average velocity of each coordinate value in 3D for four joints (the head, the spine, the right elbow, and the left elbow), the average velocity of each coordinate value of the whole upper body, and the maximum velocity of each coordinate value of the head. This results in 18 features.

As well as face direction information, we obtained skeletal information of the user from the Kinect data by using Kinect SDK every 30 msec as well as face direction information. For each frame, 3D coordinate values of joints were obtained via Kinect SDK. The origin of the coordinate system was the Kinect. These features were calculated by using this skeletal information.

The average velocity of each coordinate value for four joints can be gotten by calculating the velocity of each coordinate per frame and then taking the average in the sound segment. The average velocity of the whole upper body was calculated as follows. First, we calculated the velocities of each coordinate value for six joints (the head, the spine, the center of the hips, the center of the shoulders, the right shoulder, and the left shoulder) per frame. Second, we summed these velocities per coordinate value. Finally, we took their average in the sound segment. The maximum velocity of each coordinate value of the head is the largest velocity of each coordinate value calculated per frame.

(f) User's motion and face direction after the sound segment (nine features): This feature set consists of the average velocity of each coordinate value for two joints (the head and the spine) and the average angular velocity of each face rotation angle (yaw, pitch, roll), resulting in nine features.

We obtained the skeletal and face direction information during t seconds after the sound segment and calculated these features. The calculation procedure is the same as those during the sound segment. We set t to 2.0 because we found the shortest speech interval between the end of a user utterance and the start of a robot utterance was about two seconds.

(g) Robot's dialogue act type just before the sound segment (one feature): This feature set consists of the robot's dialogue act type just before the sound segment only. One of 25 dialogue act types [27] was used as this feature for the input sound segment.

4.3.4 Evaluation

We investigated the performance of our model in estimating the response obligation. In particular, we compared the results of three models:

1. Proposed model

   This model was trained by using all feature sets listed in Table 4.1, i.e., feature sets (a) through (g).

2. Baseline model

   This model was trained by using the previously used features, i.e., (a) through (c), corresponding to the features used for the conventional addressee identification [40].

3. Only GMM model

   This model was trained by using only feature set (d). This corresponds to estimating the response obligation by classifying as speech or non-speech.

We used Weka [18] ver. 3.7.5 machine-learning software to train and evaluate the models. We chose Random Forests [7] as described in subsection 4.2.4. We set the number of trees to 18 because it performed best with 18 trees in our preliminary experiment.

The number of *ought-to-respond* segments was less than the sample number of *ought-not-to-respond* ones, as shown in Table 4.2. Therefore, we gave the ratio of the two labels (3.60) as a weight for the cases of *ought-to-respond* in the training phase to take this deviation into account.

We evaluated the three models by 10-fold cross-validation. The performance was measured in terms of precision, recall, F-measure, and the arithmetic mean between the F-measures of *ought-to-respond* and *ought-not-to-respond.*

Table 4.3 compares the performances of the three models. The average F-measures of the three models (proposed model, baseline model, and only GMM model) were 0.823, 0.772, and 0.767, respectively. The average F-measure of the only-GMM model was lowest of the three models because this model could not distinguish between utterances toward other users and the robot. The average F-measure of our model was the highest of the three models. This difference was found to be statistically significant by z-test ($p = .0017 < .01$). Furthermore, the F-measure of *ought-not-to-respond* of our model was also the highest of the three models. We believe that the performance of *ought-not-to-respond* is more important than that of *ought-to-respond.* For example, if the robot incorrectly responds to surrounding noises and user's monologues, it might confuse the users and cause the subsequent interaction to fail. On the other hand, even if the robot incorrectly does not respond to a user utterance to which it should respond, it can go on talking as long as he/she repeats the utterance.

Table 4.3: Performance comparison.

	Ought-to-respond			*Ought-not-to-respond*			
	Precision	Recall	F-measure	Precision	Recall	F-measure	Arithmetic mean of F-measure
Proposed model	**0.884**	0.745	**0.809**	0.780	**0.902**	**0.837**	**0.823**
Baseline model	0.839	0.677	0.750	0.730	0.870	0.794	0.772
Only GMM	0.723	**0.876**	0.792	**0.843**	0.664	0.743	0.767

Next, we investigated the results in more detail. In particular, we divided the sound segments labeled *ought-not-to-respond* into four categories: robot-directed utterances, user-directed utterances, monologues, and non-speech sounds. Sound segments labeled *ought-to-respond* are all robot-directed utterances. We calculated the accuracies, i.e., recall rates, of the estimation results for the five categories (Table 4.4). The accuracies of each category correspond to those in Table 4.3[2].

We found that our model outperformed the baseline model in all categories. The difference in the robot-directed utterances was the largest of all categories. This is because the new feature sets were the most effective when estimating robot-directed utterances. On the other hand, the performances of both models for non-speech sounds were comparable. This result showed that non-speech sounds were only estimated by the previously used features, i.e., duration of the sound segment and prosodic information.

Moreover, we compared the accuracies of our model with those of the only GMM model. Here, our model outperformed the only GMM model in the four categories of *ought-not-to-respond*. Furthermore, in each category, we found that while the performances of the only GMM model for robot-directed utterances of *ought-to-respond* and non-speech sounds were relatively high, those for the others were low. These results showed that the only GMM model could determine whether the sound is speech only, and almost all speech sounds were thus classified as *ought-to-respond*. We also found that the accuracy of our model for non-speech sounds was 0.02 higher than that of the only GMM model.

4.3.5 Identifying Effective Features

We investigated how much the performance varied when one of the feature sets was removed. If the performance deteriorates when we remove a feature set, the feature set is effective for estimating the response obligation. In this analysis, we changed only feature sets and evaluated with the same data and evaluation method as in the previous section.

[2] Although the recall rates should agree, differences arise from the implementation of the 10-fold cross-validation in Weka.

Table 4.5 shows the effect of removing each feature set. Removal of any feature set decreased the average F-measure compared with the performance of the proposed model as shown in Table 4.3. Thus, all the feature sets were effective for estimating the response obligation. Additionally, when we removed (d), the performance showed the largest decline. We can easily guess that (d) is especially useful for estimating the response obligation because the only GMM model could even classify non-speech sounds as *ought-not-to-respond*. The second largest drop in performance was when we removed (f); i.e., the average F-measure was 0.028 lower than that of the proposed method. The decline was worse than in the cases of removing (a) and (e). We found that the user's motion after the sound segment was more effective than the user's motion and face direction during the sound segment.

Table 4.4: Accuracies per category of input sound (Recall rates).

	Ought-to-respond	*Ought-not-to-respond*				
	Robot-directed	Robot-directed	User-directed	Monologue	Non-speech	All
Data size of each kind of input sound	871	813	735	833	714	4,006
Proposed model	0.735	**0.809**	**0.884**	**0.920**	**0.993**	**0.862**
Baseline model	0.684	0.766	0.863	0.897	0.987	0.833
Only GMM	**0.820**	0.703	0.557	0.782	0.972	0.762

Table 4.5: Performance deterioration in terms of F-measure when individual feature sets are removed from the estimation.

		Response obligation			Performance
	Removed feature set	*Ought-to*	*Ought-not-to*	Average F-measure	decline
(a)	User's face direction during the sound segment	0.800	0.827	0.814	−0.009
(b)	Prosodic information	0.779	0.820	0.802	−0.021
(c)	Duration of the sound segment	0.791	0.806	0.799	−0.024
(d)	Input sound classification result	0.768	0.808	0.788	−0.035
(e)	User's motion during the sound segment	0.794	0.825	0.810	−0.013
(f)	User's motion and face direction after the sound segment	0.779	0.812	0.795	−0.028
(g)	Robot's dialogue act type just before the sound segment	0.795	0.826	0.811	−0.012

4.4 Summary of this Chapter

When a robot interacts with two or more users in public spaces, it needs to appropriately estimate a response obligation, i.e., whether the robot should respond to input sounds. This chapter presented a machine-learning based method to estimate the response obligation for dialogue situational understanding. This method focuses on the observation about users. We cast the problem as a binary classification task using features based on this observation. We used not only acoustic information but also the users' motions during sound segments and the results of the input sound classification. We additionally used users' motions and head directions after sound segments.

The evaluation showed that our model can accurately determine whether the robot should respond to input sounds. The arithmetic mean of the F-measures for *ought-to-respond* and *ought-not-to-respond* was 0.823, which was 0.051 higher than the baseline corresponding to Nakano et al. [40] with statistical significance. We examined the accuracies of the estimations for the five categories of input sounds and showed that the proposed model can more accurately estimate the response obligation in user-directed utterances and monologues. Furthermore, we investigated how much the performance varied when each of the seven feature sets was removed. As a result, we found that the input sound classification result and the users' whole body motions were helpful for estimating the response obligation.

Chapter 5

Conclusions

5.1 Summary

This dissertation presented a novel architecture of spoken dialogue systems for humanoid robots in public spaces. Two new modules were introduced to the conventional architecture of spoken dialogue systems: sound detection and turn-taking decision. These modules contribute to resolving the following two issues affecting humanoid robots in public spaces. First, a humanoid robot in a public space needs to be able to talk with multiple users. The robot has to classify utterances into two categories: utterances toward it and utterances toward other users. Second, the robot needs to be able to deal with various sounds. It receives not only utterances from its interlocutors but also surrounding noises and even user monologues. The robot has to distinguish utterances toward it from other input sounds.

This study focused on the implementation of the turn-taking decision module using social norms. In particular, this dissertation presented two methods of situation understanding for the turn-taking decision module and installed them in a humanoid robot, Nao. These methods enable the robot to estimate the current dialogue situation and behave according to social norms.

Chapter 2 described a model utilizing a social norm wherein a human speaker considers his interlocutor's situation when he takes a turn. In particular, this

Table 5.1: Summary of proposed methods and applicable conditions.

Proposed methods	Chap. 2 and 3 Predicting ease of speaking	Chap. 4 Estimating response obligation
States used as features	Humanoid robot's states	Users' states
Performance (F-measure)	About 75% if it was adapted to variations	About 82%
Difficult case	Timings depending on a specific user	When a robot detects overlapping sound events
Applicable condition	All features can be obtained appropriately.	

chapter presented a method to predict when a user is likely to begin speaking to a humanoid robot. The method is a machine-learning based method that uses the robot's behaviors, such as its posture, motion, and utterances, as input features. Experimental results showed that this method can accurately predict when the user is likely to begin speaking and that the features about the robot's behaviors are useful for the prediction.

Chapter 3 showed that the method described in Chapter 2 is useful in actual human-robot dialogues. In Chapter 2, we collected data from three students in my laboratory and used it as training data of the prediction model; as such, the model might have been influenced by the specific participants. In this chapter, the prediction model was rebuilt by using newly collected data from 25 participants. Experimental results showed that the method described in Chapter 2 does not depend on specific participants. Moreover, this chapter showed that individual variations caused by the label annotation can be handled by adjusting the threshold of the logistic regression. The threshold could be optimized by the correlation between the number of frames annotated as positive (how many frames in which an individual participant feels likely to speak) and his rating of the dialogue situation. Finally, we introduced a spoken dialogue system for a humanoid robot in a public space. We described an example that it is useful for ignoring unnecessary sounds.

Chapter 4 presented a method for deciding whether to respond in a multi-party human-robot dialogue. This method is a machine-learning based one that estimates the response obligation. It uses not only acoustic information, but also the users' motions during sound segments and the results of the input sound classification. In addition, it uses the users' motions and head directions after sound segments. The experimental results showed that the proposed model can determine whether the robot should respond to input sounds more accurately than a conventional method of addressee identification. The results also revealed that two of these features, i.e., the sound classification results and users' whole body motions, were especially helpful for estimating the response obligation.

Table 5.1 summarizes the two proposed methods (predicting ease of speaking and estimating response obligation) for situation understanding and applicable conditions. Each method used information which obtained from robot's states and users' states. The performance of predicting ease of speaking was about 75% when applied to individuality, i.e., 25% of the predictions were incorrect. This is because that the prediction model could not predict for timings depending on a specific user. As described in Chapter 2, we eliminated parts peculiar to individuals the from collected data. To predict these timings appropriately, training data should be made by the specific user's annotation, but this is not a proper way for humanoid robots in public spaces. Moreover, the performance of estimating response obligation was about 82%. This estimator could not predict correctly when a robot detected overlapped sound events, such as when a surrounding noise overlapped with a user utterance. To solve this problem, we can separate overlapped sounds by sound source separation technology [41]. To utilize the proposed methods in actual human-robot dialogue, all input features have to be obtained appropriately. For example, if the robot fails to obtain users' states because of an overlap of users and too noisy environment, it cannot estimate response obligation. That is, the experiments in this study were conducted under ideal environment, i.e., the robot can obtain information appropriately.

We describe three technical contributions in this study to the development of spoken dialogue systems for humanoid robots in public spaces as follows:

(1) New architecture of spoken dialogue systems for humanoid robots in public

spaces was demonstrated. The two modules newly introduced in this study are essential processes for humanoid robots in public spaces and can be utilized for interpretation of input sounds.

(2) Social norms in human-human dialogue were applied to human-robot dialogue. Both presented methods in Chapter 2 and 4 were used social norms for input features of machine-learning based models. It is essential for humanoid robots in public spaces to select an appropriate action according to social norms. The new findings and ideas in this study can contribute to develop innovative methods for human-robot dialogue.

(3) The robot states were used for situation understanding in human-robot interaction. Chapter 2 confirmed that the robot states were useful for predicting ease of speaking under the assumption that users act cooperatively in human-robot dialogue. Using the robot states as well as users' states enables the robot to estimate a dialogue situation.

5.2 Future Direction

First of all, all the experiments in this dissertation were performed offline as described in the previous section. Therefore, the following works should be conducted preferentially.

1. Incorporating the estimator into the spoken dialogue system in Figure 3.7
2. Investigating how valid the proposed methods are for actual human-robot dialogue in public spaces

Moreover, there are some directions in which to expand the study of my dissertation.

The first direction is integration of the proposed methods. This dissertation described the developments of two methods for situation understanding in human-robot dialogue: one for predicting when a user is likely to begin speaking to a robot from the observations about the robot (described in Chapter 2) and one for determining whether the robot should respond to input sounds from the observations about users (described in Chapter 4). Although these methods are not integrated, they can be integrated to enable a more accurate estimation of whether the robot should respond to input sounds because the two methods use different observations (the observations about the robot vs. the observations about users). Therefore, the method for integrating these methods should be investigated.

The second direction is to prove the generality of the proposed methods. In Chapter 2 and 3, we used two robot's behavior sequences when conducting data collection to make training data. To show that the prediction model can be used in variety of dialogue situations, how various behaviors are needed to include in these sequences should be investigated. Moreover, we did not use a time sequence model that can train context, such as hidden markov models and recurrent neural network, because we used the common parts which participants gave same labels only as training data, i.e., this is not time series data. In contrast, if we sufficiently collected training data including various dialogue situations, we would have used all parts which participants gave labels and trained the prediction model by the time sequence model. Furthermore, this dissertation only showed their effectiveness in experiments using Nao. To verify the generality of the methods, experiments with other humanoid robots that have different body parts should be conducted. For example, while Nao has human-shaped body parts, such as a head, legs, and arms, Pepper [46] has wheels instead of human-shaped legs. Finally, experiments with humanoid robots that have good communication abilities, such as quick reactions and accurate detection of input sounds, should be conducted. In Chapter 4, we used as input features user behaviors specific to when they are talking to the humanoid robot used in the experiment; for example, a user stands still when he is talking to the robot although he moves his head and slowly sways when he talks to the other user. These user behaviors may depend on the robot's communication ability; if the robot can talk more smoothly, users may talk to it in the same manner as they do to other users.

The third direction is to control dialogue situations by generating behaviors according to social norms. This study used social norms for determining whether a robot should respond to input sound intervals. This is a passive aspect of utilizing social norms. On the other hand, social norms can be actively used in dialogue situations. For example, when it is very noisy in a dialogue, it is difficult for the robot to recognize utterances accurately. In this situation, the robot could produce situations in which a user is not likely to begin speaking, such as by putting its finger on its lips and saying “wait a moment”, because the robot does not want users to talk to it.

References

[1] Sylvain Argentieri, Patrick Danes, and Philippe Soueres. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112, 2015.

[2] Amos Azaria, Ya'akov Gal, Sarit Kraus, and Claudia V. Goldman. Strategic advice provision in repeated human-agent interactions. *Autonomous Agents and Multi-Agent Systems*, 30(1):4–29, 2016.

[3] Dan Bohus and Eric Horvitz. Models for multiparty engagement in open-world dialog. In *Proceedings of The Annual SIGdial Meeting on Discourse and Dialogue*, pages 225–234, 2009.

[4] Dan Bohus and Eric Horvitz. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of The ACM International Conference on Multimodal Interaction*, pages 2–9, 2014.

[5] Susanne Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.

[6] Christoph Bregler and Yochai Konig. Eigenlips for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 669–672, 1994.

[7] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] Robert Brueckmann, Andrea Scheidig, and Horst-Michael Gross. Adaptive noise reduction and voice activity detection for improved verbal human-robot

interaction using binaural data. In *Proceedings of The International Conference on Robotics and Automation*, pages 1782–1787, 2007.

[9] Crystal Chao and Andrea L Thomaz. Turn-taking based on information flow for fluent human-robot interaction. *AI Magazine: Special Issue on Dialog with Robots*, 32(4):53–63, 2011.

[10] Ross Corkrey and Lynne Parkinson. Interactive voice response: review of studies 1989–2000. *Behavior Research Methods, Instruments, & Computers*, 34(3):342–353, 2002.

[11] Hoya Corp. Voicetext. `http://voicetext.jp/`. Last accessed on 2016/12/12.

[12] Fernando De la Torre, Wen-Sheng Chu, Xuehan Xiong, Francisco Vicente, Xiaoyu Ding, and Jeffrey Cohn. Intraface. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 1–8, 2015.

[13] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.

[14] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the ACM international conference on Multimedia*, pages 1459–1462, 2010.

[15] Mark JF Gales and Steve Young. An improved approach to the hidden markov model decomposition of speech and noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 233–236, 1992.

[16] Michael A. Goodrich and Alan C. Schultz. Human robot interaction: A survey. *Foundations and Trends in HumanComputer Interaction*, 1(3):203–275, 2008.

[17] Edward Twitchell Hall. *The hidden dimension*. Doubleday & Co, 1966.

[18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11:10–18, 2009.

[19] Honda Motor Co., Ltd. ASIMO. `http://world.honda.com/ASIMO/`. Last accessed on 2016/12/12.

[20] Yung-Jieh Huang, Hiroshi Dohi, and Ishizuka Mitsuru. Man-machine interaction using a vision system with dual viewing angles. *IEICE TRANSACTIONS on Information and Systems*, E80-D(11):1074–1083, 1997.

[21] Hiroshi Ishiguro and Syuichi Nishio. Building artificial humans to understand humans. *The Japanese Society for Artificial Organs*, 10(3):133–142, 2007.

[22] Martin Johansson and Gabriel Skantze. Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. In *Proceedings of The Annual SIGdial Meeting on Discourse and Dialogue*, pages 305–314, 2015.

[23] Takayuki Kanda, Hiroshi Ishiguro, Michita Imai, and Tetsuo Ono. Development and evaluation of interactive humanoid robots. In *Proceedings of the IEEE*, volume 92, pages 1839–1850, 2004.

[24] Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of The ACM International Conference on Multimodal Interaction*, pages 144–151, 2004.

[25] Simon Keizer, Mary Ellen Foster, Oliver Joseph Lemon, Andre Gaschler, and Manuel Giuliani. Training and evaluation of an MDP model for social multi-user human-robot interaction. In *Proceedings of The Annual SIGdial Meeting on Discourse and Dialogue*, pages 223–232, 2013.

[26] Adam Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 26(1):22–63, 1967.

[27] Casey Kennington, Kotaro Funakoshi, Yuki Takahashi, and Mikio Nakano. Simple timing for probabilistic multiparty dialogue management. In *Proceedings of the Workshop at the Conference of Human-Robot Interaction*, 2014.

[28] Wooil Kim and Hanseok Ko. Noise variance estimation for Kalman filtering of noisy speech. *IEICE Transactions on Information and Systems*, E84-D(1):155–160, 2001.

[29] Kazunori Komatani, Akira Hirano, and Mikio Nakano. Detecting system-directed utterances using dialogue-level features. In *Proceedings of The Annual Conference on the International Speech Communication Association*, pages 230–233, 2012.

[30] Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. User modeling in spoken dialogue systems to generate flexible guidance. *User Modeling and User-Adapted Interaction*, 15(1):169–183, 2005.

[31] Hatice Kose-Bagci, Kerstin Dautenhahn, and Chrystopher L. Nehaniv. Emergent dynamics of turn-taking interaction in drumming games with a humanoid robot. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, pages 346–353, 2008.

[32] Ivana Kruijff-Korbayov, Heriberto Cuayhuitl, Bernd Kiefer, Marc Schrder, Piero Cosi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser, Hichem Sahli, Georgios Athanasopoulos, Weiyi Wang, Valentin Enescu, and Werner Verhelst. Spoken language processing in a conversational system for child-robot interaction. In *Proceedings of the Workshop on Child Computer Interaction*, pages 132–134, 2012.

[33] Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Julius an open source realtime large vocabulary recognition engine. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1691–1694, 2001.

[34] Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura, Hiroshi Saruwatari, and Kiyohiro Shikano. Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. In *Proceedings of The Annual Conference on the International Speech Communication Association*, pages 173–176, 2004.

[35] Yoichi Matsuyama, Kosuke Hosoya, Hikaru Taniyama, Hiroki Tsuboi, Shinya Fujie, and Tetsunori Kobayashi. SCHEMA: Multi-party interaction-oriented humanoid robot. In *Proceedings of the ACM SIGGRAPH Conference and Exhibition in ASIA*, pages 82–82. ACM, 2009.

[36] Michael McTear. *Spoken dialogue technology: toward the conversational user interface*. Springer Science & Business Media, 2004.

[37] Microsoft. Kinect for xbox 360. `http://www.xbox.com/en-US/xbox-360/accessories/kinect`. Last accessed on 2016/12/21.

[38] Takeshi Minato, Michihiro Shimada, Hiroshi Ishiguro, and Shoji Itakura. Development of an android robot for studying human-robot interaction. In *Proceedings of the International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 424–434, 2004.

[39] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. The uncanny valley. *The Robotics Automation Magazine*, 19(2):98–100, 2012.

[40] Yukiko Nakano, Naoya Baba, Hung Hsuan Huang, and Yuki Hayashi. Implementation and evaluation of a multimodal addressee identification mechanism for multiparty conversation systems. In *Proceedings of The ACM International Conference on Multimodal Interaction*, pages 35–42, 2013.

[41] Michael S Pedersen, Jan Larsen, Ulrik Kjems, and Lucas C Parra. A survey of convolutive blind source separation methods. *Multichannel Speech Processing Handbook*, pages 1065–1084, 2007.

[42] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: an open-source Robot Operating System. In *Proceedings of The International Conference on Robotics and Automation*, volume 3, page 5, 2009.

[43] Antoine Raux and Maxine Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 629–637. Association for Computational Linguistics, 2009.

[44] Byron Reeves and Clifford Nass. The media equation: How people treat computers, televisions, and new media as real people and places. *Cambridge University Press*, 1996.

[45] SoftBank Robotics. Nao. `https://www.ald.softbankrobotics.com/en/cool-robots/nao`. Last accessed on 2016/12/12.

[46] SoftBank Robotics. Pepper. `https://www.ald.softbankrobotics.com/en/cool-robots/pepper`. Last accessed on 2016/12/12.

[47] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

[48] Mohammad Hasan Savoji. A robust algorithm for accurate endpointing of speech signals. *Speech communication*, 8(1):45–60, 1989.

[49] Muzafer Sherif. *The psychology of social norms.* Harper, 1936.

[50] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. Where to look: a study of human-robot engagement. In *Proceedings of the international conference on Intelligent user interfaces*, pages 78–84, 2004.

[51] Gabriel Skantze. *Error Handling in Spoken Dialogue Systems.* PhD thesis, KTH Royal Institute of Technology, 2007.

[52] Gabriel Skantze and Joakim Gustafson. Attention and interaction control in a human-human-computer dialogue setting. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pages 310–313, 2009.

[53] Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. Turn-taking, feedback and joint attention in situated humanrobot interaction. *Speech Communication*, 65:50–66, 2014.

[54] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE signal processing letters*, 6(1):1–3, 1999.

[55] David R. Traum and James F. Allen. Discourse obligations in dialogue processing. In *Proceedings of The Annual Meeting of the Association for Computational Linguistics*, pages 1–8, 1994.

[56] Koen van Turnhout, Jacques Terken, Ilse Bakx, and Berry Eggen. Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 175–182, 2005.

[57] Roel Vertegaal, Robert Slagter, Gerrit C. van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the Annual SIGCHI Conference on Human Factors in Computing Systems*, pages 301–308, 2001.

[58] Longbiao Wang, Kyohei Odani, and Atsuhiko Kai. Dereverberation and denoising based on generalized spectral subtraction by multi-channel lms algorithm using a small-scale microphone array. *EURASIP Journal on Advances in Signal Processing*, 2012(1):12, 2012.

[59] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of The Language Resources and Evaluation Conference*, page 5.

[60] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1:69–90, 1999.

[61] Sukhyun Yoon and Chang D. Yoo. Speech enhancement based on speech/noise-dominant decision. *IEICE Transactions on Information and Systems*, E85-D(4):744–750, 2002.

[62] Takami Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno. An improvement in audio-visual voice activity detection for automatic speech recognition. In *Proceedings of the International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 51–61. Springer-Verlag, 2010.

[63] Steve Young. *The HTK Hidden Markov Model Toolkit: Design and philosophy.* University of Cambridge, Department of Engineering, 1993.

[64] Xiang Zuo, Naoto Iwahashi, Ryo Taguchi, Shigeki Matsuda, Komei Sugiura, Kotaro Funakoshi, Mikio Nakano, and Natsuki Oka. Robot-directed speech detection using multimodal semantic confidence based on speech, image, and motion. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 2458–2461, 2010.

Publication List

A. Journal Papers

A-1 Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, and Kazunori Komatani. Estimating Response Obligation by Focusing on User States in Multi-Party Human-Robot Dialogues. Journal of the Japanese Society for Artificial Intelligence, Vol. 31, No.3, pages 1-10, 2016 (in Japanese).

A-2 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Evaluating Model that Predicts When People will Speak to Humanoid Robot and Handling Valiations by Individuality and Instruction. Journal of the Japanese Society for Artificial Intelligence, Vol. 29, No.1, pages 32-40, 2014 (in Japanese).

A-3 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Building Model to Predict How Likely User is to Talk to Humanoid Robot. Journal of the Japanese Society for Artificial Intelligence, Vol. 28, No.3, pages 255-260, 2013 (in Japanese).

B. International Conferences (refereed)

B-1 Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, and Kazunori Komatani. Estimating Response Obligation in Multi-Party Human Robot Dialogues. In Proceedings of The IEEE-RAS International Conference on Humanoid Robots, Seoul, Korea, Nov. 3-5, pages 166-172, 2015.

B-2 Takaaki Sugiyama. Position Paper. In Proceedings of The Young Researchers' Round table on Spoken Dialogue System, Prague, Czech Republic, Aug. 31 - Sep. 1, pages 75-76, 2015.

B-3 Takaaki Sugiyama. Realizing Robust Human-Robot Interaction under Real Environments with Noises. In Proceedings of The International Conference on Multimodal Interaction, Istanbul, Turkey, Nov. 12-16, pages 354-358, 2014.

B-4 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Evaluating Model that Predicts When People will Speak to Humanoid Robot and Handling Variations of Individuals and Instructions. In Proceedings of The International Workshop on Spoken Dialog Systems, Napa, United States, Jan. 18-20, pages 62-72, 2014.

B-5 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Predicting When People will Speak to a Humanoid Robot. In Proceedings of The International Workshop on Spoken Dialog Systems, Paris, France, Nov. 28-30, pages 187-198, 2012.

C. Domestic Symposium

All the following papers are written in Japanese.

C-1 Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, and Kazunori Komatani. Estimating Response Obligation by User Behaviors for Multi-Party Dialogue Robot. Special Interest Group on Spoken Language Understanding and Dialogue Processing, B5-02, 2015.

C-2 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Predicting How Likely User is to Talk to Humanoid Robot. Human-Agent Interaction Symposium, P6, 2013.

D. Domestic Annual Conference

All the following papers are written in Japanese.

D-1 Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, and Kazunori Komatani. Estimating Whether Robot should Respond to Input Sounds in Multi-Party Dialogue. The 29th Annual Conference of the Japanese Society for Artificial Intelligence, 2L3-1, 2015.

D-2 Takaaki Sugiyama, Ryu Takeda, and Kazunori Komatani. Estimating Evaluation Values for Ambiguous Labels Included in Chat-oriented Dialogue. The 21th Annual Meeting of Natural Language Processing, B1-5, 2015.

D-3 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Noise-Robust Speech Interaction based on Online Prediction of How Likely the User is to Talk to Humanoid Robot. The 28th Annual Conference of the Japanese Society for Artificial Intelligence, 2M5-OS-20b-3, 2014.

D-4 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Online Prediction of How Likely User is to Talk to Humanoid Robot. The 76th National Convention of Information Processing Society of Japan, 1ZB-3, 2014.

D-5 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Empirical Evaluation of Model that Predicts When People will Speak to Humanoid Robot. The 27th Annual Conference of the Japanese Society for Artificial Intelligence, 1K3-OS-17a-6, 2013.

D-6 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Classifying Input Sounds Including Overlapped Speech in Spoken Dialogues with Humanoid Robot. The 75th National Convention of Information Processing Society of Japan, 4T-3, 2013.

D-7 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Building Model to Predict How Likely User is to Talk to Humanoid Robot. The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 1O2-OS-18-7, 2012.

D-8 Takaaki Sugiyama, Kazunori Komatani, and Satoshi Sato. Modeling Ease of Talking to Humanoid Robot. The 74th National Convention of Information Processing Society of Japan, 5Q-3, 2012.