Non-verbal Signals for Grounding in Embodied Conversational Agent

by

Yukiko I. Nakano

B.A., Psychology
Tokyo Woman's Christian University, 1988
M.A., Educational Psychology
The University of Tokyo, 1990

Submitted to the program in Media Arts and Sciences, School of Architecture and Planning, in Partial Fulfillment of the Requirements of the Degree of Master of Science in Media Arts and Sciences at the Massachusetts Institute of Technology

September 2002

Signature of Author ______________________
Yukiko I. Nakano
Program in Media Arts and Sciences
August 9, 2002

Certified by ______________________
Justine Cassell
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by ______________________
Andrew B. Lippman
Chairperson
Departmental Committee on Graduate Students

Non-verbal Signals for Grounding in Embodied Conversational Agent

by

Yukiko I. Nakano

B.A., Psychology
Tokyo Woman's Christian University, 1988
M.A., Educational Psychology
The University of Tokyo, 1990

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on August 9, 2002 in Partial Fulfillment of the Requirements of the Degree of Master of Science in Media Arts and Sciences at the Massachusetts Institute of Technology

Abstract

In face-to-face conversation, speakers present non-verbal signals collateral with verbal information. Nodding and gazing at a speaker are known to provide positive feedback from listeners, which contributes to establishing common ground (a process called *grounding*). However, previous theories and computational models of grounding were mainly concerned with verbal grounding acts, and there have not been enough discussion about how nonverbal behaviors are used in the process of grounding.

This thesis first compares face-to-face conversation to conversation without co-presence, revealing how nonverbal behaviors are used in the process of grounding in human communication. Results of the analysis show that, in face-to-face communication, non-verbal behaviors are changing during an utterance and a typical transition pattern of non-verbal behaviors is also different depending on the type of verbal act.

Then, the implementation of grounding functionality onto an Embodied Conversational Agent is presented. The dialogue state updating mechanism in the Dialogue Manager accesses non-verbal information conveyed by a user and judges the groundedness of presented materials based on the results of empirical study,

Thesis Advisor: Justine Cassell
Associate Professor of Media Arts and Sciences

Non-verbal Signals for Grounding in Embodied Conversational Agent

by

Yukiko I. Nakano

The following people served as readers for this thesis:

Reader ______________________________

Deb Roy
Assistant Professor of Media Arts and Sciences
MIT Program in Media Arts and Sciences

Reader ______________________________

Candace L. Sidner
Senior Research Scientist
Mitsubishi Electric Research Laboratories (MERL)

Acknowledgements

I would like to thank all the people who helped me survive in the Media Lab and finish my thesis project. First, I would like to thank my thesis advisor Justine Cassell. She taught me techniques of empirical study and provided guidance in analyzing human behaviors. Insightful and provocative discussions with her allowed me to clarify my thoughts and to find missing pieces in my claim.

I also would like to thank my thesis readers. Candy Sidner has advised me with her profound knowledge in discourse and pragmatics. A fruitful discussion awaited me each time I consulted her. Deb Roy, approaching human-computer interaction in a very different way from our research group, taught me the importance of originality in research.

I really miss all the members of Gesture and Narrative Language group (GNL), and appreciate their valuable help. I could not finish this thesis project without their help. Tim Bickmore gave me insightful comments on my research, and spent much time in discussion with me. Hannes Vilhjálmsson helped me in using animation software and manipulating equipment. He also encouraged me with his constant positive thinking. Kimiko Ryokai made me relax when I had a hard time, and advised me on how to deal with difficult situations. I was touched with her kindness. Cati Vaucelle has been a good friend and we helped each other since the start of our journeys at GNL. Dona Tversky patiently reviewed all my bad English writing. This thesis would not be possible without her extraordinary kindness and patience. Tom Stocky answered my questions about the current implementation of MACK, and Ian Gouldstone designed the wonderful MACK character. Gabe Reinstein and Sha Ma helped me in implementing an eye-tracker and integrating it into MACK. Bob Levine has taken care of producing and submitting the thesis.

I would also like to thank researchers who gave me excellent comments. I would like to express my special thanks to Lisa Harper, Michael Johnston, Barbara Tversky, and Matthew Stone for their stimulating discussions.

Finally, I would like to thank my family for supporting me at home and from a long distance.

Table of Contents

1. Introduction

In conversation, speakers present non-verbal signals collateral with verbal information. Nodding, smiling, and gazing at the speaker have been described as providing positive feedback from the listener ((Clark 1996),(Clark and Schaefer 1989)). These studies claimed that non-verbal signals work with and interact with verbal behaviors in the process of establishing common ground, knowledge or belief that is shared between conversational participants, and they called this process *grounding* ((Clark and Schaefer 1989)). Moreover, a study on eye gaze behavior claimed that there is a very rapid and complex coordination between speech, gaze, and other non-verbal signals ((Argyle and Cook 1976)). However, the previous studies do not answer some basic questions; how do these signals interact and coordinate with each other in grounding? What is the determinant factor that affects the process of grounding in face-to-face communication? Although these are very important issues in establishing a process model of face-to-face communication, they have not been studied yet. This thesis addresses these questions, and approaches them by focusing on the usage of non-verbal behaviors in grounding in face-to-face communication.

1.1. Problem

Figure 1.1-1 shows an example of a face-to-face dialogue. Dotted lines indicate the place of nod, and continuous lines indicate those of gaze at the partner. Lines drawn on the upper side of the words show S's behaviors. Lines drawn at the bottom of the words shows H's. Note that, in this interaction, H never takes a turn, and gives an acknowledgement only once. However, this does not mean that S keeps talking without checking H's understanding, or that H rarely gives feedback to S. After [U2], H nods without saying anything, and keeps nodding until the beginning of the next utterance. At the end of the same utterance, S looks at H, then H gives acknowledgement with a nod. For other parts that don't have any lines, both conversational participants are looking at the map. This example shows that conversational participants change the direction of gaze and use nodding as well as speak utterances, but we don't know how these non-verbal signals contribute to grounding.

[U1] S: And then, you'll go down this little corridor.

[U2] S: It is not very long.

[U3] S: It's about I guess four of five meters.

[U4] H: Um-hum

[U5] S: And there's some vending machines on the left hand side.

[U6] S: And then it just opens up

Figure 1.1-1: Example of a face-to-face conversation

In order to describe the function of these non-verbal signals, this thesis investigates the following sub-questions;

- Is the usage of non-verbal signals different depending on the type of verbal behavior?
- Do non-verbal signals shift during the speech? If so, is there a specific pattern for the transition of non-verbal signals, and are these changes different_depending on the type of verbal behavior?
- Is the usage of non-verbal behavior different depending on communication modality?

1.2. Contributions of the Thesis

For the psychological part of this thesis, I employ an empirical approach to investigate the functions and usage of non-verbal signals in grounding. I collect real human-human dialogues under an experimental setting, and analyze the data to describe how non-verbal behaviors are used in the grounding process. Another aspect of this thesis is to propose a method for implementing grounding function within an interactive agent. I integrate these two aspects of this research by designing the computational mechanism based on the results of empirical study.

1.3. Outline of the Thesis

Chapter 2 gives an overview of some related work on which this thesis is built: common ground and grounding, non-verbal cues in communication, and computer technologies that supports multimodal interaction. Chapter 3 describes the experiment used to collect dialogue data, and the results of the statistical analysis of them. Chapter 4 describes the implementation of ideas for dialogue management to deal with face-to-face grounding within an Embodied Conversational agent. Chapter 5 summarizes the thesis and discusses some future works.

2. Related work

2.1. Theory of grounding

2.1.1. Common ground

When two people are engaged in a conversation as a joint activity, one of the most important purposes of interaction is to establish common ground between them. The technical notion of common ground was introduced by Robert Stalnaker (Stalnaker 1978) based on notions that included common knowledge (Lewis 1969), mutual knowledge or belief (Schiffer 1972), and joint knowledge (McCarthy 1990). Common ground can be defined as the sum of them (Clark 1996).

Ordinarily, people establish a piece of their common ground by pointing to a shared basis for it, which is a joint perceptual experience or a joint action. The joint action may be a physical one or may simply be having a conversation together. Specifically in conversation, common ground is a set of propositions whose truth a person takes for grant as part of the background of the conversation. (Stalnaker 1978) writes;

"Roughly speaking, the presuppositions of a speaker are propositions whose truth he takes for granted as part of the background of the conversation.... Presuppositions are what is taken by the speaker to be the *common ground* of the participants in the conversation, what is treated as their *common knowledge* or *mutual knowledge*" (320, Stalnaker's emphases). .

More generally speaking, common ground can be defined as information based on cultural communities a person belongs to --- such as nationality, occupation, ethnic group, or gender. People take various kinds of information as common ground, but what is important is how common ground is achieved.

2.1.2. Clark's objection to previous discourse theories

Classical theories of discourse in philosophy, artificial intelligence, and psychology presuppose the following three points concerning common ground in discourse.

Common ground: the participants in a discourse presuppose a certain common ground.

Accumulation: in the course of discourse, the conversational participants add shared knowledge to their common ground.
Unilateral action: common ground is added by a speaker uttering the right sentence at the right time.

Clark objected to the third assumption. He claimed that this assumption is not sufficient to handle conversation because these theories are only concerned with a speaker's intention, and assume that what the speaker said is added to the discourse model without any error. The previous theories of discourse were not concerned with dynamics in conversation, and operated on the strong assumption that the hearer understands rationally, and that a speaker's utterance is perfectly understood by the hearer if it is rationally appropriate. As an extension of this discussion, (Walker 1992) proposed IRU (information redundant utterance), which is an utterance that does not add new propositions in the discourse. She claimed that repeating what the speaker said is informationally redundant, but this kind of utterance provides evidence that the mutual understanding is actually achieved.

2.1.3. Grounding

Grounding is a process to make what has been said a part of common ground. (Clark and Schaefer 1989) proposed a model for representing grounding using *contributions*. In their model, a contribution is composed of two main phases.

Presentation Phrase: A presents utterance *u* for B to consider. He does so on the assumption that, if B gives evidence *e* or stronger, he can believe that B understands what A means by *u*.
Acceptance Phase: B accepts utterance *u* by giving evidence *e'*, that he believes he understands what A means by *u*. He does so on the assumption that, once A registers evidence e', he will also believe that B understands.
Through these two phases, people in conversation *contribute* to discourse to reach the *grounding criterion*(Clark and Schaefer 1989);

In addition to these basic processes for grounding, they proposed a notion of "grounding criterion".
Grounding criterion: The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for current purposes.

In achieving common ground, it is often not necessary to assure perfect understanding of each utterance but only understanding "to a criterion sufficient for current purposes" (Clark and Schaefer 1989). Therefore, we can have some sort of common ground without full mutual belief, and also the grounding criterion may change as conversation purposes change.

2.1.4. Evidence of understanding

In a grounding process, the acceptance phase is usually performed by B giving A evidence that he believes he understands what A meant by *u*. (Clark 1996) and (Clark and Schaefer 1989) proposed four main classes of positive evidence of understanding though there are some difference in categories and definition between these proposals.

(1) Continued attention: If the hearer looks away from the speaker, the speaker tries to capture the hearer's gaze, and attention. This is the most basic form of positive evidence.

(2) Assertions of understanding: Asserts understanding using verbal and non-verbal behaviors, such as "uh huh", "I see", or nod or smile.

(3) Presuppositions of understanding: The listener presupposes that s/he has understood the speaker well enough to go on. So, uptake, or initiating relevant next turn is a signal of presupposition of understanding. The clear example of relevant next turn is adjacency pair. When a speaker asks a question, he/she expects that the partner will answer the question.

(4) Displays of understanding: The listener displays part of what s/he construed the speaker to mean. An answer displays in part how the listener construed the speaker's question.

(5) Exemplifications of understanding: The listener exemplifies what s/he has construed the speaker to have meant. Paraphrase or verbatim repetition, grimace, look disappointed, or perform some other iconic gesture can be used for this purpose.

2.1.5. Strength of Evidence Principle

(Clark and Schaefer 1989) proposed, "the participants expect that, if evidence e_0 is needed for accepting presentation u_0, and e_1, for accepting the presentation of e_0, the e_1 will be weaker than e_0." B may accept A's presentation by uttering, "m", but they expect something weaker to be able to accept that "m." The upshot is that every acceptance phase should end in continued attention or initiation of the next turn, the weakest evidence available.

2.1.6. Principle of Least Collaborative Effort

As discussed above, there are different types of evidence, which have different strengths as evidence. When conversational participants choose their behaviors, they consider the cost in collaboration. (Clark and Wilkes-Gibbs 1986) present a Principle of Least Collaborative Effort, which states that "in conversation, the participants try to minimize their collaborative effort - the work both speakers and addressees do from the initiation of each contribution to its completion." Grice expressed this idea in terms of two maxims; Quantity (make your contribution as informative as is required for the current purpose of the exchange, but do not make your contribution more informative than is required), and Manner (Be brief, and avoid unnecessary prolixity).

Note that the principle of least collaborative effort doesn't mean that speakers should always produce proper utterances that includes enough information, but the point is in minimizing the total effort in the collaboration. There are three problems with this principle ((Clark and Wilkes-Gibbs 1986), (Clark and Brennan 1991)).

Time pressure: When the speaker doesn't have enough time to plan and produce an utterance, he/she is more likely to produce improper utterances.

Errors: Speakers often make errors in speech production, and they repair their own speech. Basically, repairing needs effort, but producing a perfect utterance may sometimes need more time and effort.

Ignorance: Speakers sometimes don't know enough about the hearer, and don't know what utterance is appropriate. For example, "Um, the next one's the person ice skating that has two arms?" with try marker[1]. It is often the case that it will take more collaborative effort to design a proper utterance than to design an improper utterance and ask their addressees help.

2.1.7. Levels of conversation

As having a conversation is a joint action, there would be times, when a listener doesn't hear or understand a speaker's presentation entirely. Originally, (Clark and Schaefer 1989) proposed the following four states of understanding;

[1] Sometimes speakers find themselves about to present a name or description they aren't sure is correct or comprehensible. They can present that constituent – often a noun or noun phrase with what Sacks and Schegloff (1979) have called a try marker, a rising intonation by a slight pause, to get their partners to confirm or correct it before completing the presentation (Clark 1996).

Clark & Shaefer's states of understanding
State 0: B didn't notice that A uttered any *u'*.
State 1: B noticed that A uttered some *u'* (but wasn't in state 2).
State 2: B correctly heard *u'* (but wasn't in state 3).
State 3: B understood what A meant by *u'*.

In later work, some modified versions of states definitions were proposed, which are shown in Table 2.1-1. Moreover, based on the definition by (Clark 1994), (Dillenbourg 1996) proposed a relationship between the level of conversation and each participant's action, as shown in Table 2.1-2. This table is helpful to know for what action contributes to what level of conversation. Clark and Shaefer (1989) originally defined the levels from 0 to 3. Brennan and Hulteen (1995) and Clark (1994) added some higher levels in order to describe communication failure in spoken dialogue. (Traum and Dillenbourg 1998) modified their definitions to describe multi-modal communication.

Table 2.1-1: Levels of Conversation

Clark & Shaefer (1989)	Brennan & Hulteen (1995)	Allwood et al (1992)	Clark (1994)	Traum & Dillenbourg (1998)
Psychological model of communication	A model of grounding with a spoken language system	Dimensions in feedback	Problems for maintaining common ground	
State 0: B didn't notice that A uttered any *u'*. **State 1**: B noticed that A uttered some *u'* (but wasn't in state 2).	**State 0: Not attending** The system isn't listening or doesn't notice that the user has spoken. **State 1: Attending** The system has noticed that the user has said something, but it hasn't interpreted the words.	**(i) Contact**: willingness and ability to continue interaction	**Level 1 Vocalization and attention**: is the receiver attending to the speaker and can the producer successfully articulate the message?	**Level 1.Access** : do the collaborators have access to each others communicative actions?
State 2: B correctly hear *u'* (but wasn't in sate 3).	**State 2: Hearing** The system was able to identify some of the user's words, but hasn't parsed the entire utterance.	**(ii) Perception**: willingness and ability to perceive expression and message	**Level 2 Presentation and Identification**: can the message be successfully presented so that the receiver can identify. e.g., the words and structure of a sentence?	**Level 2. Perception**: do the collaborators perceive the communicative actions that are performed?
State 3: B understand what A meant by *u'*.	**State 3: Parsing** The system received what seems to be a well-formed utterance, but hasn't mapped it onto any plausible interpretation. **State 4: Interpreting** The system reached an interpretation, but hasn't mapped the utterance onto an application command.	**(iii) Understanding**: willingness and ability to understand expression and message	**Level 3 Meaning and Understanding**: can the receiver understand what was meant by the message?	**Level 3. Understanding**: do the collaborators understand what is meant by the communicative actions?
	State 5: Intending The system has mapped the user's input onto a command in its application domain, but hasn't acted yet. **State 6: Acting** The system attempts to carry out the command. It is not known yet whether the attempt will have the desired outcome. **State 7: Reporting** The system may or may not has been able to carry out the user's command, and reports any evidence available from the application domain.	**(iv) Attitudinal** reactions: willingness and ability to give attitudinal reactions, such as accept, reject, belief, agreement, etc.	**Level 4 Proposal and Uptake**: will the receiver commit to the proposal made by the producer?	**Level 4. Agreement**: do the collaborators reach agreement about the communicated facts of plans?

Table 2.1-2: Grounding Acts and Conversation Level (Dillenbourg et al., 1996)

Grounding act	From A's viewpoint	From B's viewpoint
Monitoring	*Passive/Inferential (How A reasons about B's knowledge)*	*Pro-active (How B can help A to know about B)*
	Level 1: A infers whether B can access X	Level 1: B tells A about what he can access
	Level 2: A infers whether B has noticed X	Level 2: B tells (or shows) A that B perceived X
	Level 3: A infers whether B understood X	Level 3: B tells A how B understands X
	Level 4: A infers whether B (dis)agrees	Level 4: B tells A that B (dis)agrees about X
Diagnosis	*Active (How A tries to know that B knows X)*	*Reactive (How B participates in A's grounding)*
	Level 1: A joins B to initiate coporesence	Level 1: B joins A
	Level 2: A asks B to acknowledge X	Level 2: B acknowledges X
	Level 3: A asks B a question about X	Level 3: B displays understanding or requests repair of X
	Level 4: A persuades B to agree about X	Level 4: B (dis)agrees on X

2.2. Computational model of grounding

(Traum 1994) proposed the Grounding Acts Model. This model collapses different types of evidence of acceptance mentioned in section 2.1.4, but extends the building blocks of the units of common ground to those that could be realized with a single utterance. Rather than the two phases of presentation and acceptance, the basic building blocks are a set of grounding acts, each of which is identified with a particular utterance unit, and performs a specific function towards the achievement of common ground. In this model, the units of grounded content are called Discourse Units (DU), rather than Contributions. Individual grounding acts could add or change content of the unit. Based on this claim, he proposed a DU state transition diagram, which defines possible sequence of grounding acts to achieve common ground. In Table 2.2-1, S stands for start initial state and F for final state. D stands for dead state, where the conversational material can no longer be grounded. The network is traversed by observing grounding acts as shown in each row in the table.

Table 2.2-1: DU state transition diagram

Next Act	In State						
	S	1	2	3	4	F	D
Initiate^I	1						
Continue^I		1			4		
Continue^R			2	3			
Repair^I		1	1	1	4	1	
Repair^R		3	2	3	3	3	
ReqRepair^I			4	4	4	4	
ReqRepair^R		2	2	2	2	2	
Ack^I				F	1^*	F	
Ack^R		F	F^*			F	
ReqAck^I		1				1	
ReqAck^R				3		3	
Cancel^I		D	D	D	D	D	
Cancel^R			1			D	

(Heeman and Hirst 1995) presented a computational model for grounding a referring expression. They employed a planning paradigm in modeling how conversational participants collaborate in making a referring action successful as well as clarifying a referring expression.

(Paek and Horvitz 1999) claim that the majority of automated dialogue systems as mentioned above focus only on the intention level, but it is necessary for a dialogue system to handle other levels of grounding. They provided infrastructure that recognizes that failures in dialogue can happen at any of the levels proposed representations and control strategies for grounding using Bayesian networks and decision theory. Based on four levels of conversation proposed originally by (Clark and Schaefer 1989), (cf. section 2.1.7), they employed these representations and inference strategies at four levels; Channel level, Signal level, Intention level, and Conversation level.

Although studies in computational linguistics, such as those by (Traum 1994) and (Heeman and Hirst 1995), contributed to establishing a computational model of grounding, they only discussed verbal grounding acts in intention level (level 3). (Peek&Horvitz 1999) proposed a model that can deal with lower level communication failure, which would occur when the system cannot get speech signals from the user. Although their model can deal with a wider range of signals in communication, it was mainly concerned with speech signal, and not concerned with nonverbal signals such as eye gaze and head nod.

2.3. Multimodality in Grounding

Based on the principle of least collaborative effort, conversational participants try to achieve common ground with as little combined effort as needed. However, the effort changes dramatically according to the communication medium. As shown in Table 2.3-1, (Clark and Brennan 1991) proposed eight ways in which a medium may affect the communication between two people. They also proposed various kinds of costs that change depending on the characteristics of the medium (Table 2.3-2). They mention that, in face-to-face conversation, it is easy to nod at interlocutors, and to gaze at interlocutors to show them that they are being attended to, or to monitor their facial expressions. In media without co-presence, gestures cost expensive bandwidth, or are severely limited. This description suggests that the method of displaying positive evidence of grounding is different depending on communication modality.

Note that there is a trade off on the costs of grounding. For example, in a study reported by (Cohen 1984) in which tutors instructed students on assembling a pump, they compared communication by telephone that one by keyboard. In a telephone conversation, producing an utterance and changing speakers does not cost much. On the other hand, in keyboard conversation, the cost for changing a speaker and repair cost are high. Therefore, subjects formulate utterances more carefully in keyboard conversation than in telephone conversation.

Table 2.3-1: Factors for characterizing communication modalities

Modality	Factors
Face-to-face	Copresence, visibility, audibility, cotemporality, simultaneity, sequentiality
Telephone	audibility, cotemporality, simultaneity, sequentiality
Video teleconference	visibility, audibility, cotemporality, simultaneity, sequentiality
Terminal teleconference	cotemporality, sequentiality, viewability
Answering machine	Audibility, reviewability
Electric mail	Reviewability, revisability
Letters	Reviewability, revisability

Table 2.3-2: Costs of grounding

- Formulation costs
- Production costs
- Reception costs
- Understanding costs
- Start-up costs
- Delay costs
- Asynchrony costs
- Speaker change costs
- Display costs
- Fault costs
- Repair costs

(Brennan 2000) provides experimental evidence that reveals how grounding takes place in conversational tasks. She used a computer-based location task, where one party (the director) must describe where on a map the other (the matcher) is to point his cursor. This experiment is broken down along two trials where the director can see where the matcher is vs. where the director cannot, and must rely on verbal descriptions from the matcher. This experimental manipulation changes the strength and type of evidence available for accepting presentations. The results of the experiment revealed the grounding process was shorter when more direct evidence was available.

(Dillenbourg 1996) analyzed grounding across different modes of interaction. They used a virtual environment that the subjects modified by giving on-line commands, such as redirecting the location of the character of the user. In their experiment, the subjects used three modes of communication: dialogue, action command in the virtual environment, and whiteboard drawing. In dialogue, the subjects talked to each other via two commands, "say…" to communicate with anybody in the same room, and "page <Player> …" to communicate with this player wherever he is. Using action commands, they changed the virtual environment, such as the location of the user or other objects. The third mode of communication, whiteboard drawing, was visible in the form of a non-scrollable window that remained the subjects screen until it was deleted. By looking at cross-modal grounding, they found that grounding is often performed across different modes. For example, information presented in dialogue is grounded by an action in the virtual environment. Also, actions in the virtual environment are grounded in the dialogue.

2.4. Non-verbal cues in communication

There are many studies that address conversational functions of nonverbal behaviors in face-to-face communication.

Gesture: (Bavelas, Chovil et al. 1995) proposed a group of gestures that seem to function solely to assist the process of dialogue rather than to convey topical information, and called these kinds of gestures "interactive gestures".

Facial expression: (Chovil and Fridlund 1991) reported that facial expressions can convey discourse information. For example, raising or lowering eyebrows indicates initiation of topics. Back-channel responses[2] are displayed by brow raises, mouth corners turned down, eyes closed, or pressed lips.

Head movement: (Duncan 1974)proposed how speaker and listener nonverbal signals are used in turn taking.. For instance, a speaker within-turn signal that requests feedback from the listener is composed of a set of two cues, (a) the speaker's completion of a grammatical clause, and (b) turning her/his head towards the listener. As a study focusing on non-verbal feedback, (Rosenfeld and Hancks 1980) attempted to subcategorize functions of feedback from the user. They investigated which nonverbal behaviors by listeners were indicative of attention, understanding, and agreement, and how these behaviors were affected by speakers. They asked five independent observers to take a role of a speaker and to rate 250 listener responses. As a result, they found that behaviors of the listener that were associated with judgments of "agreement" were complex verbal listener responses and multiple head nods. The agreeing-type listener response was found to predictably follow the speaker's pointing of her/his head in direction of the listener. In contrast, judgments that the listener was indicating understanding were associated with repeated small head nods by the listener prior to the speech juncture, and did not involve any apparent speaker signals. Thus, signals of understanding, in contrast to agreement, appear to be more subdued in form and more likely to be initiated by the listener than elicited by the speaker. Finally, judgments of listener attention were associated with listener's forward leaning prior to the speaker's juncture, listener's verbal response

2 Back-channel response is a kind of feedback from a listener that contributes to grounding without taking a turn. Clark (1996) uses a term, "background acknowledgement" instead of back channel response. A list of examples of this behavior will be shown in a definition of Acknowledgement in section 3.2.1.

after the juncture, and gesticulation by the speaker after the juncture but prior to resuming speech.

Eye-gaze : (Goodwin 1981) claimed that a speaker should obtain gaze of the listener during the course of a turn, and showed that pausing and restarting of the utterance were used to get the listener's gaze. Moreover, he showed how gestures were used in controlling the focus of listener's eye-gaze, and discussed that, as a strategy for getting the listener's gaze, gestures would be better than an explicit request which would shift focus away from the talk currently progressing (Goodwin 1986).

(Argyle and Cook 1976) discussed that gaze is connected with language as a channel; (a) Speakers look up at grammatical pauses to obtain feedback on how utterances are being received, and to see if others are willing for them to carry on speaking. (b) Listeners look at speakers a lot of the time, in order to study their facial expressions, and their direction of gaze.

(Clark 2001) proposed "Directing-to" and "Placing-for" as techniques for indicating in face-to-face situation. Directing-to is a speaker's signal that directs addressee's attention or gaze to object *o*. Placing-for is a speaker's signal that places object *o* for addressee's attention. Both of these are techniques used to connect a message and the physical world that the message describes, and get the addressee accessible and perceivable to the message.

(Novick, Hansen et al. 1996) investigated gaze patterns between speaker and listener, and examined correlation between eye-gaze pattern and difficulty in conversation. They proposed two patterns of speaker and listener eye-gaze in turn taking, (a) mutual-break pattern: as one conversant completes an utterance, s/he looks toward the other. Gaze is momentarily mutual. (b) mutual-hold pattern: the turn recipient begins speaking without immediately looking away, and break the gaze during the course of the turn. They found that mutual-break gaze pattern was used more frequently when the conversation was proceeding smoothly and mutual-hold was preferred where conversants were having difficulty.

In terms of grounding, (Clark 1996) proposed that eye gaze is the most basic form of positive evidence of understanding, which displays the continued attention to a speaker. A head nod has an equivalent function to verbal acknowledgement, such as "uh huh", "I see", and is used to assert understanding (c.f. 2.1.4). However, he did not provide a profound discussion about how these nonverbal behaviors are used in a grounding process.

As described above, there are many studies that were concerned with communicative functions of nonverbal behaviors. However, there are only a few intuitive descriptions about nonverbal grounding acts, and unfortunately the previous studies didn't provide empirical result that describes how nonverbal behaviors contribute to grounding and how they work with verbal behaviors in a grounding process.

2.5. Computer technology supporting multimodal communication

2.5.1. Video mediated communication

The comparisons between face-to-face communication and video mediated communication (VMC) attempt to identify how effectively VMC mimics face-to-face interaction. In general, it has been found that the closer a communication medium approximates face-to-face interaction, the closer the conversational style is to that produced in face-to-face setting ((Whittaker (to appear)). He also claimed that the most common use of visual channel is to show understanding, such as nodding one's head while another person is speaking. Speakers in face-to-face interaction continually adjusted the content of their utterances based on the addressees' apparent level of understanding.

(Anderson and Casey 1997) compared map task conversations in audio-only, face-to-face and VMC which the subjects can make direct eye contact. Although VMC did not replicate all the benefits of face-to-face communication, the subjects felt able to engage in interaction more freely than in audio-only conditions. For example, the speaker checks whether the listener has understood what s/he said more frequently verbally when they only have an audio link than when visual signals are available. In this respect, VMC seems to have a similar type of benefit to that in face-to-face communication.

By comparing audio and video conference condition, (Daly-Jones, Monk et al. 1998) reported that interpersonal awareness, in terms of an illusion of presence and awareness of the other person's attentional status, was much increased in the video mediated communication than audio condition.

Results of these studies suggest that availability and quality of a visual channel, which convey nonverbal signals, affects the grounding process in VMC as well as human face-to-face conversation. Therefore, in order to use visual channel effectively in VMC, we need to know how nonverbal signals are used in VMC as well as in face-to-face communication, and to compare nonverbal signals in these two modalities. However, as mentioned in the previous section, previous studies have not sufficiently addressed this issue.

2.5.2. Nonverbal behaviors in human-computer multimodal interaction

Embodied interface agent is another computer technology that attempts to mimic face-to-face interaction. (Rickel and Johnson 1999) built a pedagogical agent that is embodied in a 3D virtual environment and demonstrates sequential operations of complex machinery and answers some follow up questions from the student. (Lester, Towns et al. 2000) developed a lifelike pedagogical agent, Cosmo, which can generate deictic behaviors and emotive-kinesthetic behaviors including facial expressions and gestures with arms and hands. This system provides advice to students about Internet packet routing.

Gandalf (Thorisson 1999) is the first good example of embodied interface agent that shows effectiveness of mimicing human non-verbal conversational behaviors. It can generate non-verbal behaviors for turn taking and back-channel responses, such as head nod, gaze towards user and away, and beat gestures. In experimental interaction with human users, it was found that users accepted the agent as more helpful, lifelike, and smooth in its interaction style when it provided the nonverbal conversational behaviors (Cassell and Thorisson 1999).

(Cassell, Bickmore et al. 2000) built Embodied Conversational Agent that exploited several kinds of actual human conversational protocols, such as salutation and farewells, conversational turn taking, and describing objects using hand gestures. They discussed that embodied interface agents can provide a qualitative advantage over non-embodied interfaces if their bodies are used in a way that leverages knowledge of human communicative behavior.

(Traum and Rickel 2002) extended the ideas of embodied conversational agents to multi-party conversation in a virtual world, and discussed how these non-verbal conversational behaviors, which are used for regulating dialogue, are useful in multi-party conversation as well. They mentioned that gaze at a listener at grammatical pauses can work as request-acknowledgement, and gaze at a speaker and nod can work as acknowledge from the listener. However, they did not show any empirical result that supports their proposal.

Previous research in conversational agent showed that embodiment of an agent enhances naturalness in human-computer interactions. Moreover, as people usually use non-verbal signals unconsciously, it is also expected that embodiment facilitates conversational protocols that people use unconsciously. An important point here is that the body of the agent should be used effectively in order to perform communicative behaviors based on human conversational protocol. However, we don't know the protocol for grounding face-to-face.

2.6. Motivation for Empirical Study

As reviewed above, theories of grounding provide the basic idea of grounding, and computational linguists extended and modified the idea to establish computational models. However, their theories and models were mainly concerned with verbal grounding acts. On the other hand, studies in human communication investigated communication functions of nonverbal behaviors. Although they shed light on nonverbal behaviors from different perspectives of communication, there have not been enough discussions about how nonverbal behaviors are used in a process of grounding. In addition, research in VMC and ECAs emphasized the importance of designing computer interfaces based on human conversation protocols, but there is no empirical result and computational model based on which a computer system can be built. Therefore, in this thesis, first, I investigate human verbal and nonverbal behavior to reveal how nonverbal behaviors contribute to grounding in human communication. I will specifically address the following questions:

(1) Is the usage of non-verbal behavior different depending on communication modality?
(2) Is the usage of non-verbal signals different depending on the type of verbal behavior?
(3) Do non-verbal signals shift during the speech? If so, is there a specific pattern for the transition of non-verbal signals, and are these changes different depending on the type of verbal behavior?

If Display Cost is different depending on communication modality as proposed in (Clark and Brennan 1991), it is expected that the way of displaying evidence of understanding is different depending on communication modality. Therefore, question (1) is important in order to reveal the characteristics of face-to-face grounding by comparing it to non-face-to-face grounding.

As described in section 2.4, there are many studies that were concerned with communicative functions of nonverbal behaviors, but the previous studies didn't provide empirical results that describes how nonverbal behaviors contribute to grounding and how they work with verbal behaviors in a grounding process. Question (1) addresses this issue.

In order to implement a Conversational Agent, it is necessary to establish a precise process model of grounding. However, as discussed in section 2.5.2, there is no empirical result and computational model based on which a computer system can be built. Question (3) addresses this issue.

3. Empirical Study

This chapter describes an experiment to collect dialogue data, and reports the results of data analysis.

3.1. Data collection

In this experiment, a student of the MIT Media Laboratory (a direction giver) gives a direction **to** somewhere in the Media Lab to a student outside of the lab (a direction receiver), who does not know about it at all, in different experiment settings. A snapshot of an experiment session is shown in Figure 3.1-1. Camera (A) shows a shared map and movement of subjects' fingers, Camera (B) shows a close up picture of a receiver, Camera (C) shows a close up picture of a giver, and Camera (D) shows an overall picture of the interaction.

Figure 3.1-1: Snapshot of an experiment session

(a) Condition

Two different communication channels that would convey different types of information are controlled in the experiment.

Visual Evidence (VE): Accessibility to a reference to be shared between the conversational participants. This is actually a map of the lab drawn by the direction giver.

Communicative Behavior (CB): Visibility of the partner's face and body.

Two different experimental conditions were set by combining these communication channels above.

(1) Face-to-face condition: Visual Evidence (VE) + Communicative Behavior (CB)

(2) Shared reference condition: VE only

(b) Setting

Face-to-face: Two subjects sit at the table face-to-face. There is a pen and a map drawn by the direction giver on the table. The giver uses them to give a direction.

Shared reference: Two subjects sit at the table face-to-face. There is a pen and a map drawn by the giver on the table. The giver uses them to give a direction. In addition, there is an L-shaped screen between the subjects. Therefore, they cannot see each other's face and body, but they can share the map that the giver drew.

(c) Material

The following four tasks are randomly assigned to each session and condition.

[Direction 1] Give direction from the lobby to the kitchen, then to NeCSys.

[Direction 2] Give direction from the lobby to the Cube, then to the Garden.

[Direction 3] Give direction from the lobby to room 054, then room 335. (Optional)

(d) Design

The following orders of conditions and tasks were randomly assigned to each pair.

	session 1		session 2	
	condition	Task	condition	task
Order 1	face-to-face	[Direction 1]	shared ref.	[Direction 2]
Order 2	face-to-face	[Direction 2]	shared ref.	[Direction 1]
Order 3	shared ref.	[Direction 1]	shared ref.	[Direction 2]
Order 4	[shared ref.	[Direction 2]	shared ref.	[Direction 1]

(e) Procedure and Instructions

(e-1) Draw a map of a route

Before the session, an experimenter asked the giver to draw two maps of what s/he would explain using at least 8 landmarks or signs.

<Instruction>

"First, here is a piece of paper and a pen. I would like you to draw a map from X to Y to Z on this piece of paper. I will give you extra pieces so that you can re-draw the map if you don't like it. But, don't worry. You don't need to draw a perfectly accurate or beautiful map. A rough sketch is fine. The only requirement is to draw at least 8 landmarks or signs in the map. Please do not draw only lines. Draw a map from X to Y to Z, OK? Any questions?"

(e-2) Direction giving task

Each pair of subjects engaged in two conversations in two different experimental settings.

<Instruction>

<Face-to-face> "Hi... (whatever greetings). Okay, here is the task: I'd like you to give (direction giver's name) directions from X to by passing through each of the landmarks on the map. You are welcome to use the map that you drew earlier and use the pen to add more details if you need to. You can take as much time as you need, just make sure that (direction receiver's name) gets to each landmark before you go on to the next leg of the directions (address the receiver). You have to really understand how to get to each landmark before (direction giver's name) goes on to the next step of the directions. So, when you (address the receiver) really understand how to get to a landmark, you move your piece to there. Any questions?"

<Shared reference> "Hi.... Now, why don't you give (direction giver's name) directions from X to Y to Z by passing through each of the landmarks on the map. Again, you are welcome to use the map that you drew earlier and the pen to add more details, but in this case, as you can see, there is a screen in front of you so you shouldn't be able to see much of each other, right? You can take as much time as you need, just make sure that (receiver's name) gets to each landmark before you go on to the next leg of the directions. (address the receiver) You have to really understand how to get to each landmark before (direction giver's name) goes on to the next step of the directions. So, when you (address the receiver) really understand how to get to a landmark, you move your piece to there. Any questions?"

(f) Data storage

Interactions between the subjects were shot from four different angles by using two CCD cameras and two digital video cameras. A map and overall picture of both participants were shot by CCD cameras. A close-up picture of each subject was taken by a digital video camera. These four pictures were combined by a video mixer to display them as one picture that is split into four parts. This picture was video-recorded with a SVHS recorder.

(g) Subjects

Ten students or employees in the MIT Media Lab and ten students outside of the lab were paired. Two were MIT students and eight are students in Boston University. They did not know the floor plans of the Media Lab building.

(h) Data

By running 10 experimental sessions, 10 dialogues in each condition and 20 in total were collected and transcribed.

3.2. Data coding

3.2.1. Coding verbal behaviors

(a) Unit of verbal grounding act

As a unit for a verbal grounding act, we need to define the "utterance unit". Although there is not a consensus as to what defines an utterance unit, most attempts make use of one or more of the following factors;

- Speech by a single speaker, speaking without interruption by speech of the other, constituting a single *Turn*
- Has syntactic and/or semantic completion
- Defines a single speech act
- Is an intonational phrase
- Separated by a pause

While the turn has the great advantage of having easily recognized boundaries, there are some difficulties with treating it as a basic unit of spoken language. Since the turn ends only when another conversant speaks, this may cut off in midstream if the new speaker starts earlier than expected. Likewise, if the new speaker does not come in right away, the first speaker may produce several basic contributions (or units) within the span of a single turn.

(Heeman and Traum 1997) used prosodic feature of speech to split a turn into utterance units. They adopt Pierrehumbert's theory of intonational description for English (Pierrehumbert 1980). According to this view, two levels of phrasing are significant in English intonational structure. Both types of phrases are composed of sequences of high and low tones in the fundamental frequency (F0) contour. An intermediate (or minor) phrase consists of one or more pitch accents plus a phrase. Intonational (or major) phrases consist of one or more intermediate phrases plus a final boundary tone, which may also be high or low, and which occurs at the end of intermediate phrase boundary. In general, major phrase boundaries tend to be associated with longer pauses, greater tonal changes and more final lengthening than minor boundaries.

Another way in which a turn can be segmented is by pauses in the speech stream ((Seligman, Hosaka et al. 1997), (Takagi and Itahashi 1996)). Pause-delimited units are attractive because pauses can be detected automatically, but there are some problems in this approach. For one thing, pauses can occur anywhere in the speaker's turn, even in the middle of a syntactic constituent. There is also often some silence around the point of disfluency during a speech repair.

(Nakatani and Traum 1999) split an utterance into utterance-tokens. The principles for splitting utterances into tokens are based on prosody and grammar, with the intuition that a token should correspond to a single intonational phrase (Pierrehumbert 1980) or perhaps a single grammatical clause (i.e. tensed or untensed unit with predicate argument structure).

As grounding occurs within consecutive utterances by one speaker, it is necessary to define a smaller unit than a *turn*. To tokenize a turn, I employ a method proposed by (Nakatani and Traum 1999), and call the token "utterance unit". The reasons are; as there is a lot of disfluent speech in the data, splitting an utterance at each pause is not an appropriate way of analyzing this data. Second, prosodic feature is helpful to find a token, but this is too subtle to use as the only clue for tokenization. Therefore, I employ a method that combines grammatical information, which is more robust and clear, and prosody to tokenize a turn.

Table 3.2-1: Categories of verbal grounding acts

<Forward looking>

- (a) Statement
    - (a1) Assert (as)
    - (a2) Re-assert (ras)
    - (a3) Other (ost)
- (b) Info-request (ir)
    - (b1) tag (tag)
- (c) Influencing-addressee-future-action (Influence-on-listener) IAF
    - (c1) Request (IAF_r)
    - (c2) Suggest (IAF_s)
    - (c3) Invite (IAF_i)
    - (c4) Other (IAF_o)
- (d) Other (fo)
    - Committing-speaker-future-action (Influence-on-speaker)
    - other

<Backward looking>

- (e) Understanding
    - (e1) Signal-non-understanding (non-u)
    - (e2) Signal-understanding
        - (e21) Acknowledge (ack)
        - (e22) Repeat-rephrase (rep)
        - (e23) Completion (cmpl)
        - (e24) Other (un-o)
- (f) Answer (ans)
- (g) Other (bo)

(b) Categories of UU

The coding scheme used in this data was shown Table 3.2-1. This is defined based on DAMSL coding scheme[3] (Allen and Core 1997). (Core and Allen) reported inter-coder reliability of this scheme. The advantage of using this coding scheme is that the reliability of the scheme has already been known, and the reported reliability would be helpful to estimate the reliability of annotation of this data.

<Forward looking>

(a) Statement: Making claims about the world as in utterances and in answers to questions.

The primary purpose of statements (utterances having a tag in the statement aspect) is to make claims about the world as in utterances such as "It's raining outside" or "I need to get cargo there" (the world includes the speaker) and in answers to questions. As a rule, the content of statements can be evaluated as being true or false. Note that we are only coding utterances that make explicit claims about the world, and not utterances that implicitly claim that something is true. As a intuitive test as to whether an utterance makes an explicit claim, consider whether the utterance could be followed by "That's not true". For example, the utterance "Let's take the train from Dansville" presupposes that there is a train at Dansville, but this utterance is not considered a statement. This suggestion could not be correctly replied with "That's not true" (Allen and Core 1997).

(a1) Assert (as): when the speaker is trying to affect the beliefs of the hearer.

(a2) Reassert (ras): repeating information for emphasis or acknowledgement.

(b) Info-request: Utterances that introduce an obligation to provide an answer

Utterances that introduce an obligation to provide an answer should be marked as Info-request. Note, answers can be nonverbal actions providing information such as displaying a graph. Info-request includes all questions, including yes/no questions such as "Is there an engine at Bath?", "The train arrives at 3 pm right", and even "The train is late" said with the right intonation. The category also includes wh-questions such as "When does the next flight to Paris leave?" as well as actions that are not questions but request information all the same such as "Tell me the time". Requests for other actions that can be used to communicate, such as "Show me where that city is on the map" are also considered Info-Requests. Basically, any utterance that creates an obligation for the hearer to

---

3 http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html

provide information, using any form of communication, is marked as an Info-Request.

(c) Influencing-addressee-future-action (IAF): The primary purpose of this aspect is to directly influence the hearer's future non-communicative actions, as in the case of requests ("Move the train to Dansville" and "Please speak more slowly") and suggestions ("how about going through Corning").

<Backward looking>

(e) Understanding: This aspect concerns the actions that speakers take in order to make sure that they are understanding each other as the conversation proceeds. There are many levels of "understanding", ranging from merely hearing the words to fully identifying the speaker's intention. We group most of these levels together so that if the hearer is said to have understood the speaker, then the hearer knows what the speaker meant by the utterance.

(el) Signal-non-understanding (non-u): Utterances that explicitly indicate a problem in understanding the antecedent.

As an applicability test for Signal-non-understanding, the utterance unit should be able to be roughly paraphrased by an utterance such as "What did you say/mean?". Note that not all clarification questions signal non-understanding. Some of the clarification questions involve acquiring additional information about how or why something was requested or proposed, and do not signal misunderstanding. For example, "Huh?", "What did you say?", "to Dansville?", and "Which train?" are questions that are Signal-non-understanding (SNU).

(e2) Signal-understanding: Utterances that explicitly signal understanding

(e21) Acknowledge (ack): utterances consisting of short phrases such as "okay", "yes", and "uh-huh", that signal that the previous utterance was understood. There are a variety of expressions that are classified into this category, and some typical expressions include:

"I see", "yes", "m", "uh-huh", "yeah", "right", "um-hm", "oh" and "huh".

(e22) Repeat-rephrase (rep): utterances that repeat or paraphrase what was just said in order to signal that the speaker has been understood.

(e23) Completion (cmpl): finishing or adding to the clause that a speaker is in the middle of constructing.

(f) Answer (ans): The Answer aspect is simply a binary dimension where utterances can be marked as complying with an info-request action in the antecedent.

Most questions are answered with one or more declarative sentences although it is possible to answer a question with an imperative as shown in the direction giving example below. Note this imperative-looking answer is also marked as an Assert act as its Forward Function is to provide information rather than to influence *u*'s future action. In fact, answers by definition will always be asserts.

Info-request	utt1: u: How do I get to Corning?
Assert, Answer(utt1)	utt2: s: Go via Bath.

3.2.2. Coding non-verbal behaviors

The coding scheme for non-verbal behaviors was divided into four sub schemes, which are for gaze, head movement, gesture, and map manipulations. The definition of gaze is based on (Exline and Fehr 1982) and categories of head movement were extracted from the body movement scoring system proposed by (Bull 1987). Hand gestures were categorized based on McNeill's hand gestures categorization (McNeill 1992).

(a) Gaze
Gaze: Looking at the partner's eyes, eye region, or face.
Ngaze: Looking away from the partner

(b) Head
HdSh: Head shake. Head turns from side to side in a single continuous movement.
HdNd: Head nod. Head moves up and down in a single continuous movement on a vertical axis, but the eyes do not go above the horizontal axis.

(c) Hand
(c1) Gesture
- Iconic
- Metaphoric
- Deictic
- Beat

(c2) Map manipulation
- pointing with a pen
- pointing with a finger
- gesture on a map with a pen
- gesture on a map with a finger(s)
- drawing with a pen
- piece manipulation (only by a follower)

3.3. Analysis

<Verbal behaviors>

The following categories are used for classifying verbal grounding acts, that is an utterance unit (UU).

<Non-verbal behaviors>

The following four types of non-verbal grounding acts are analyzed as typical behaviors observed in our data[4];

Gaze: gaze at the partner

Map: look at the map

Nod: head nod

Else: look elsewhere

4 In previous studies (c.f. section 2.4), eye gaze and head nod are used in a process of grounding. Therefore, this study focuses on these four kinds of behaviors, and does not analyze hand gestures including map manipulation.

In order to analyze non-verbal grounding acts as dyads, the following nine combinations of non-verbal acts are defined as "non-verbal status (NV status)". [5] For example, *gm* stands for a combination of the speaker's gaze and the listener's looking at the map.

<Forward looking>
(a) Statement
- (a1) Assert (as)
- (a2) Re-assert (ras)
- (a3) Other (ost)

(b) Info-request (ir)
- (b1) tag (tag)

(c) Influencing-addressee-future-action (Influence-on-listener) IAF
- (c1) Request (IAF_r)
- (c2) Suggest (IAF_s)
- (c3) Invite (IAF_i)
- (c4) Other (IAF_o)

(d) Other (fo)
- Committing-speaker-future-action (Influence-on-speaker)
- other

<Backward looking>
(e) Understanding
- (e1) Signal-non-understanding (non-u)
- (e2) Signal-understanding
  - (e21) Acknowledge (ack)
  - (e22) Repeat-rephrase (rep)
  - (e23) Completion (cmpl)
  - (e24) Other (un-o)

(f) Answer (ans)
(g) Other (bo)

Table 3.3-1: Variety of NV status

Combinations of NVs		Listener's behavior			
		Gaze	Map	Nod	Else
Speaker's behavior	Gaze	*gg*	*gm*	*gn*	*ge*
	Map	*mg*	*mm*	*mn*	*me*
	Nod	*ng*	*nm*	*nn*	*ne*
	Else	*eg*	*em*	*en*	*ee*

[5] In this analysis, categories whose case are less than 10 are omitted because a result based on low frequent data is not reliable.

In map task conversation, conversational participants look at the map most of the time (Argyle and Cook 1976). Thus, I can assume that *mm* is a default NV status for both face-to-face and shared reference conditions. However, the participants sometimes change their gaze direction or move their heads. I investigate whether the type of verbal grounding act (UU type) would be a predictive factor that determines NV status change, and whether the usage of verbal and non-verbal grounding acts is different depending on communication mode.

3.4. Results

3.4.1. Basic statistics

In a few interactions in shared reference condition, the camera did not successfully shoot information giver's facial expression. Thus, two dialogues of this condition were eliminated from the data, and eight dialogues are used in the statistical analysis. The results are shown in Table 3.4-1. The mean number of utterances per dialogue is 135.0 in face-to-face, and 143.1 in the shared reference condition. The difference is not statistically significant. The mean length of conversation in face-to-face (3.24 minutes) is longer than that in the shared reference condition (3.78 minutes). The mean length of the utterances is 5.26 words face-to-face, and 4.43 words in the shared reference condition. This difference is statistically significant. In the face-to-face condition, the number of utterances in a dialogue is distributed between 69 and 387. In the shared reference condition, it is distributed between 40 and 409.

(Boyle, Anderson et al. 1994) compared map task dialogues between two conditions: the conversational participants can see each other's face and they cannot see each other's face. They found that conversational participants who could not see each other produced more turns (longer dialogues) than those who could see each other. Although analysis unit in my data is different from theirs, in my data, the number of utterance units per dialogue was not different depending the condition. They also reported that speakers who could not see their partners used fewer word tokens per turn than those who could see each other. I got the same result; the speakers produce fewer words per utterance unit in shared reference condition than face-to-face condition. These results suggest that, in shared reference condition, speakers need to convey information in smaller chunks and speak slower or spend longer pause in their utterance in order to convey information surely.

Table 3.4-1: Basic statistics for each condition

	face-to-face	shared reference	T-test
num. of dialogues	8	8	-
mean length of dialogue (min)	3.24	3.78	t(7)=-1.667 p<.1 (one-tail)
mean num. of utterances (UUs)	135.00	143.13	t(7)=-0.680 (n.s.)
mean length of an utterance (words)	5.26	4.43	t(7)=3.389 p< .01 (one-tail)
minimum num. of utterances per dialogue (UUs)	69	40	-
maximum num. of utterances per dialogue (UUs)	387	409	-

The mean number of the four types of UUs per dialogue is shown in Table 3.4-2, and that of per utterance is shown in Table 3.4-3. The following four types of verbal grounding acts, Acknowledgement, Answer, Information request, and Assertion, are used in statistical analysis because, for other categories defined in section 3.2.1, not enough data was available for statistical analysis. Acknowledgement is used more frequently in shared reference condition than in face-to-face condition. In other UU types, the difference between the face-to-face and shared reference conditions is not statistically significant. Therefore, the distributions of verbal acts are very similar in these two conditions, and this result suggests that nonverbal acts would be more important in characterizing communication in each condition. In the following sections, a more in-depth analysis of the results, including a distribution of NV status changes and transitions of NV status, will be reported. Then, I will discuss how the usage of non-verbal grounding acts is different depending on the communication mode.

Table 3.4-2: Mean num. of different kinds of UUs per dialogue

	face-to-face (n=8)	shared reference (n=8)	paired t-test (n=8)
Acknowledgement	29.8	35.6	n.s.
Answer	10.9	10	n.s.
Info-req	7.3	5.8	n.s.
Assertion	79	86.8	n.s.

Table 3.4-3: Mean num. of different kinds of UUs per utterance

	face-to-face (n=8)	shared reference (n=10)	paired t-test (n=8)
Acknowledgement	0.226	0.259	t(7)=-1.42 p < .1 (one tail)
Answer	0.072	0.043	n.s
Info-req	0.048	0.032	n.s
Assertion	0.591	0.608	n.s

3.4.2. Analysis of the face-to-face condition

In order to address questions proposed in section 2.6, first, I investigate the basic distribution of non-verbal signals and examine whether the distribution is different depending on the type of verbal grounding act. Then, patterns of transition of non-verbal signals in each verbal act are described to figure out how non-verbal signals are used in the process of grounding.

(1) Distribution of NV status with respect to UU type

First, I will investigate what kind of NV status frequently co-occurs with which UU type, and whether frequent NV status varies depending on the UU type. Table 3.4-4 shows the frequency and percentage of NV status out of the total number of NV statuses in four different UU types. Table 3.4-5 and Table 3.4-6 show the results separating the cases according to who is the speaker of the UU; the direction giver or receiver. For example, in Table 3.4-5, *gm* indicates that the direction giver is a speaker and speaks an Acknowledgement while looking at the listener (direction receiver). On the other hand, the listener (direction receiver) looks at the map. The results of Chi-square tests for these tables are all statistically significant.

In addition to Chi-square test, in order to specify which category causes these statistically significant Chi-square values, I calculated a residual for each cell using a method proposed by (Haberman 1973). First, standard residuals are calculated as follows;

$$e_{ij}=(n_{ij}-E_{ij}) / \bullet\ E_{ij} \qquad (1)$$

E_{ij} is an expected frequency, which is calculated as follows;

$$E_{ij} = column\ frequency * row\ frequency / total\ frequency \qquad (2)$$

Estimated variance of e_{ij} is given as follows;

$$V_{ij}=(1-n_{i.}/N)(1-n_{.j}/N) \qquad (3)$$

Based on these variables, adjusted residual for each cell in the crosstab is calculated as follows;

$$d_{ij}=e_{ij}/\sqrt{v_{ij}} \qquad (4)$$

Table 3.4-7, Table 3.4-8, and Table 3.4-9 show adjusted residuals for Table 3.4-4, Table 3.4-5, and Table 3.4-6. Comparing the absolute values of each cell with 1.96, which is the limit of 5% confidence interval in normal distribution, it is found that the residuals are statistically significant in many cells.

Table 3.4-4: Distribution of NV combinations for all cases

All	gg	gm	gn	mg	mm	mn	ng	nm	total
ack	5 (2.1%)	16 (6.8%)	6 (2.5%)	16 (6.8%)	101 (42.6%)	4 (1.7%)	24 (10.1%)	65 (27.4%)	237
ans	27 (29.3%)	9 (9.8%)	5 (5.4%)	10 (10.9%)	29 (31.5%)	2 (2.2%)	6 (6.5%)	3 (3.3%)	92
info-req	16 (20.3%)	29 (36.7%)	1 (1.3%)	14 (17.7%)	18 (22.8%)	1 (1.3%)	0 (0%)	0 (0%)	79
as	02 (13.2%)	207 (26.8%)	71 (9.2%)	68 (8.8%)	227 (29.4%)	87 (11.3%)	6 (0.8%)	3 (0.4%)	771
total	158	282	89	127	413	95	39	72	1276

Chi square test p<.001

Table 3.4-5: Distribution of NV combinations for a Giver's UU

Giver	gg	gm	gn	mg	mm	mn	ng	nm	total
ack	3 (1.8%)	5 (3.1%)	4 (2.5%)	11 (6.7%)	65 (39.9%)	3 (1.8%)	19 (11.7%)	53 (32.5%)	163
ans	11(37.9%)	2 (6.9%)	1 (3,4%)	3 (10.3%)	10 (34.5%)	1 (3.4%)	1 (3.4%)	0 (0%)	29
info-req	9 (17.6%)	21 (41.2%)	1 (2%)	8 (15.7%)	11 (21.6%)	1 (2%)	0 (0%)	0 (0%)	51
as	24 (15.5%)	29 (18.7%)	3 (1.9%)	29 (18.7%)	47 (30.3%)	17 (11%)	5 (3.2%)	1 (0.6%)	155
total	47	57	9	51	133	22	25	54	398

Chi square test p<.001

Table 3.4-6: Distribution of NV combinations for a Receiver's UU

Receiver	gg	gm	gn	mg	mm	mn	ng	nm	total
ack	2 (2.7%)	11 (14.9%)	2 (2.7%)	5 (6.8%)	36 (48.6%)	1 (1.4%)	5 (6.8%)	12 (16.2%)	74
ans	16 (25.4%)	7 (11.1%)	4 (6.3%)	7 (11.1%)	19 (30.2%)	1 (1.6%)	5 (7.9%)	3 (4.8%)	62
info-req	7 (25%)	8 (28.6%)	0 (0%)	6 (21.4%)	7 (25%)	0 (0%)	0 (0%)	0 (0%)	28
as	78 (12.7%)	178 (28.9%)	68 (11%)	39 (6.3%)	80 (29.2%)	70 (11.4%)	1 (0.2%)	2 (0.3%)	616
total	103	204	74	57	242	72	11	17	780

Chi square test p<.001

Table 3.4-7: Adjusted residuals of NV combinations for all cases

All	gg	gm	gn	mg	mm	mn	ng	nm
ack	-5.3208	-6.311	-2.9759	-1.825	3.7373	-3.7418	7.0074	16.1061
ans	5.1287	-2.956	-0.6021	0.3049	-0.1798	-1.9995	2.0046	-1.0278
info-re	2.1928	3.231	-2.0568	2.3813	-1.8794	-2.1602	-1.629	-2.2441
as	1.1352	5.0506	3.8708	-1.671	-2.7588	6.4547	-5.842	-10.049

Table 3.4-8: Adjusted residuals of NV combinations for a Giver's UU

Giver	gg	gm	gn	mg	mm	mn	ng	nm
ack	-2.802	-2.3172	-2.091	-0.188	3.45456	-2.4592	4.1038	8.6996
ans	2.9853	-2.8272	-0.883	1.2127	-0.14802	-2.1825	4.583	1.4657
info-req	1.8814	0.3008	-1.7436	2.9277	-0.69779	-1.7175	-0.6441	-0.8039
as	-0.841	3.4209	2.8906	-2.013	-2.06665	4.013	-5.7238	-6.8699

Table 3.4-9: Adjusted residuals of NV combinations for a Receiver's UU

Receive	gg	gm	gn	mg	mm	mn	ng	nm
ack	-5.132	-5.338	0.2153	-3.0151	2.2755	-2.6808	3.6808	9.1931
ans	4.527	-1.1855	0.4465	-0.4132	0.1264	-0.5089	-0.6531	-2.2159
info-req	1.3836	5.86345	-0.1546	0.65722	-1.9211	-1.1938	-1.9801	-3.0303
as	1.8144	1.99593	-0.3492	2.81039	-1.0453	3.79306	-2.0066	-6.0126

Acknowledgement: In general, the most frequent NV status was *mm* (both participants look at the map) (42.6%), and the next most frequent NV status is *nm* (the speaker nods and the listener looks at the map) (27.4%). In the analysis of residuals, it was found that *nm* occurs more frequently than expected. Therefore, these results indicate that *nm* characterizes typical NV status in Acknowledgement.

An example of a typical interaction is shown in Figure 3.4-1[6]. At [2], a speaker (receiver) was nodding during acknowledging with "Um-hum", and the listener (giver) looks at the map. Also, at [4], the speaker acknowledges, but in this case, both of the participants just look at the task. In comparing Table 3.4-5 with Table 3.4-6, it is revealed that when an Acknowledgement is done by a giver, both participants look at the map (*mm*) 48.6% of the time. In the cases that the receiver acknowledges, *mm* (both look at the map) and *nm* (the receiver nods with Acknowledgement, and the giver looks at the map) are almost equally frequent (39.9% and 32.5% respectively). However, difference depending on the speaker's role in the task was not found in the analysis of residuals in

[6] "G" indicates that the speaker is a direction giver. "R" indicates that the speaker is a direction receiver. Dotted lines indicate the place of nod, and continuous lines indicate those of gaze at the partner. Double lines indicate those of looking away from the partner. Lines on the upper side of the words show G's non-verbal acts. Lines drawn at the bottom of the words shows R's.

[1] G: there are the swinging or revolving doors.

[2] R: Um-hum

[3] G: And elevators are here,

[4] R: Um-hum

Figure 3.4-1: Example of non-verbal acts in Acknowledgement

Table 3.4-8 and Table 3.4-9.

Answer: As shown in Table 3.4-4, generally, *mm* (31.5%) and *gg* (29.3%) are equally frequent in Answer. In analysis of residuals, it was found that *gg* occurs more frequently than expected. Therefore, these results show that *gg* is a typical NV status in Answer. In addition, when a giver answers a receiver's question, *gg* is slightly more frequent than *mm* (37.9% and 34.5% respectively as shown in Table 3.4-5). When the receiver answers the giver's question, *mm* is slightly more frequent than *gg* (*mm* 30.2% and *gg* 25.4% respectively as shown in Table 3.4-6). However, in analysis of residuals, there is no difference depending on the role of the speaker in the task. An ORMAT Table 4.2-1
Table STYLEREF 2 ¥s 4.2 - SEQ Table ¥* ARABC ¥s 2 1 : ApproFigure 3.4-2. Utterance unit [2] and [3] are the Answer from the giver. The giver's answer starts with his looking away from the receiver. After his utterance [2], the conversational participants start looking at each other (*gg*), and keep the mutual gaze until the end of the giver's answer.

[1] R: Will the Garden be obvious?

[2] G: It will, ah, it will be dark.

[3] G: I'm not sure, I think it might be labeled.

[4] R: OK.

Figure 3.4-2: Example of non-verbal acts in Answer

Info-req: In this category, the most frequent NV status is *gm* (36.7%). The distribution is not different depending on who is asking the question. In the analysis of residuals, it was found that *gm* occurs more frequently than expected. Therefore, these results indicate that *gm* is a typical NV status in Info-req. In analyzing difference depending on the role of a speaker, *mg* is more frequent when a giver asks and *gm* is more frequent when a receiver asks a question. An example is shown in Figure 3.4-3. At the middle of utterance [1], the receiver (speaker) starts looking at the giver (*gm*), and quit her gaze at the beginning of the receiver's answer.

[1] R: 100 Memorial Drive is down here, right?

[2] G: Yeah yeah.

Figure 3.4-3: Example of non-verbal acts in Info-req

Assertion: Assertion is the most frequently observed UU type in our data. In general, *mm* (29.4%) and *gm* (26.8%) co-occur with this type of UU most frequently (Table 3.4-4). In the analysis of residuals in Table 3.4-7, it was found that *mn* occurs more frequently than expected. There is no difference depending on the role of the speaker in the task (direction giver or receiver). As shown in Table 3.4-5, when the receiver asserts, *mm* is much more frequent than *gm* (30.3% and 18.7% respectively). However, the difference between *mm* and *gm* is subtle in receiver's assertion (29.2% and 28.9% respectively as shown in Table 3.4-6). An example conversation is shown in Figure 3.4-4. During a sequence of assertions, the giver looks at the receiver at [3] and [5]. At [3], the NV status during G's looking at R is *gm*, then moves to *mm* after the utterance. On the other hand, at [5], the NV status with G's gaze is *gm*, then moves to *gg* after the utterance.

[1] G: And then, you'll go down this little corridor.

[2] G: It is not very long.

[3] G: It's about I guess four of five meters.

[4] R: Um-hum

[5] G: And there's some vending machines on the left hand side,

[6] R: And then it just opens up

Figure 3.4-4: Example of non-verbal acts in Assertion

In summary, *mm* occurs most frequently in the entire conversation. However, by looking at the residuals in Chi-square test, An NV status that characterizes each verbal act was different depending on the type of verbal act. In general, *nm* is more frequent in acknowledgement, *gg* in answer, *gm* in info-req, and *mn* in assertion. Thus, these combinations characterize non-verbal behaviors of each type of UU, and show that although the default NV status is *mm*, there are some NV statuses that co-occur with specific verbal acts.

(2) Transition of NV status

In the previous section, it was found that a dominant NV status is different depending on UU type. However, conversational participants may change their non-verbal behaviors during an utterance. Thus, as the next step, it is necessary to investigate how a dyad's NV status shifts during or after the utterance, and whether the frequent NV status is different depending on the place in an utterance.

Figure 3.4-5, Figure 3.4-6, Figure 3.4-7, and Figure 3.4-8 show frequencies of transitions occurring within an utterance, which is divided into start, middle, and end of the utterance, and during a pause between utterances. A shift that occurs within the first three words of the utterance is classified as "start", and one that occurs within the last three words of an utterance is classified as "end". The rest of the cases are classified as "middle". Because the average length of "middle" is 3.9 words, which is a little bit longer than "start" and "end", the frequency is normalized with the number of words.

Acknowledgement As for within utterance[7], the most frequent transition is from *mm* to *nm*. For between utterances pause, shift from *nm* to *mm* occurs most frequently. This result indicates that a speaker nods during an acknowledgement, but the listener keeps looking at the map without paying attention to the speaker's nod. Therefore, when speakers verbally acknowledge what the listener presented in the previous contribution, in many cases, the listener does not try to perceive the speaker's non-verbal signal (nod) by paying attention to the listener. This suggests that in verbally asserting understanding (e.g. "hum-um" and "OK") as positive evidence in grounding (c.f. section 2.1.4), nod almost always accompanies the verbal behaviors. However the listener of the Acknowledgement does not pay attention to the non-verbal signals.

The next frequent pattern is to shift from *mg* to *ng* during the utterance and *ng* to *mm* after the utterance. In this case, a listener looks at the speaker when the speaker Acknowledges with nodding, and both of the participants return to looking at the map after the utterance.

Answer: the most frequent transition is from *eg* to *gg* at the middle of the utterance. Although the NV status shifts among *gg*, *eg*, and *ge* during the utterance, it rarely shifts to *mm* during an utterance. Most of the shifts to *mm* occur during a pause after the utterance. This means that neither speaker's nor listener's gaze returns to the map once they get mutual gaze in answering a question. Therefore, when answering a question, speakers appear to need the listener to give them mutual gaze as positive evidence of understanding. Although gaze is the weakest evidence in Clark's classification (section 2.1.4), speakers may need this evidence to assure that they have given sufficient information in their answer and the listener successfully perceive the information.

Moreover, in previous research, (Argyle and Cook 1976) discussed that there is aversion of gaze at the beginning of utterances, and there is a lower level of gaze when cognitively difficult topics are discussed. Therefore, the looking away at the beginning of Answer works as a deliberate signal that the speaker is thinking, which is perceived by the listener's gaze, and would be a sort of display that the current speaker understood and accepted the listener's question.

Info-req: shift from *mm* to *gm* is the most frequent at the start, middle, and end of the utterance.

[7] In most cases, an Acknowledgement consists of less than three words. Thus, for this UU type, the analysis only distinguished within utterance from a pause between utterances.

After the utterance, shift from *gm* to *gg* occurs most frequently. This result suggests that speakers need to get mutual gaze right after they finish their question in order to assure that the listener successfully perceives the question, and then a turn is transferred to the listener (the next speaker).

Assertion: A shift from *mm* to *gm* is the most frequent during the utterance. Shift from *gm* to *mm* is the most frequent transition during the pause after the utterance. Thus, this means that a speaker moves her/his gaze from the map to the listener during the assertion, but the listener keeps looking at the map, then, the speaker's gaze moves back to the map during the pause after the utterance. In addition, *mm* to *mn* also occurs frequently after the utterance. These results suggest that speakers need to observe the listener's paying attention to the referent that is referred in the Assertion in order to continue their contribution. Therefore, not only gazing at the speaker, but also paying attention to a referent works as positive evidence of understanding. In grounding Assertion, paying attention to a referent is more important than paying attention to the speaker.

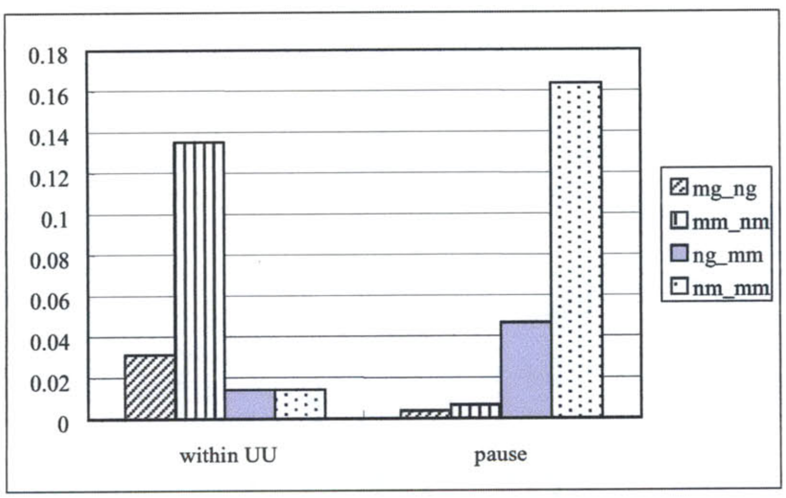

Figure 3.4-5: NV state transition for Acknowledgement

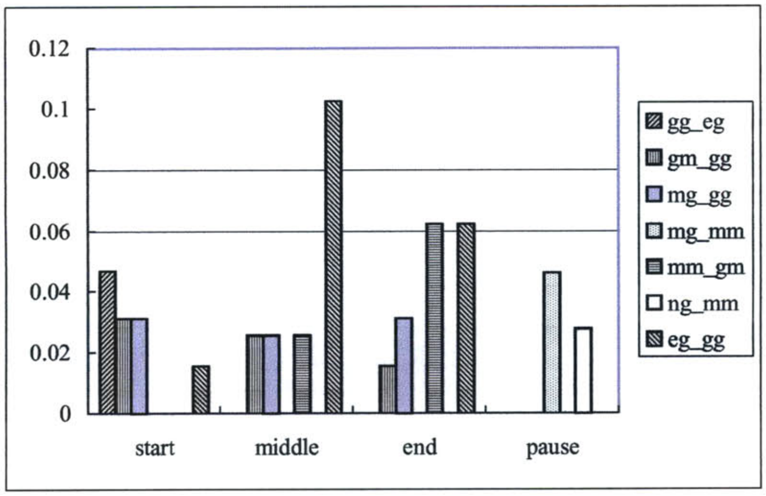

Figure 3.4-6: NV state transition for Answer

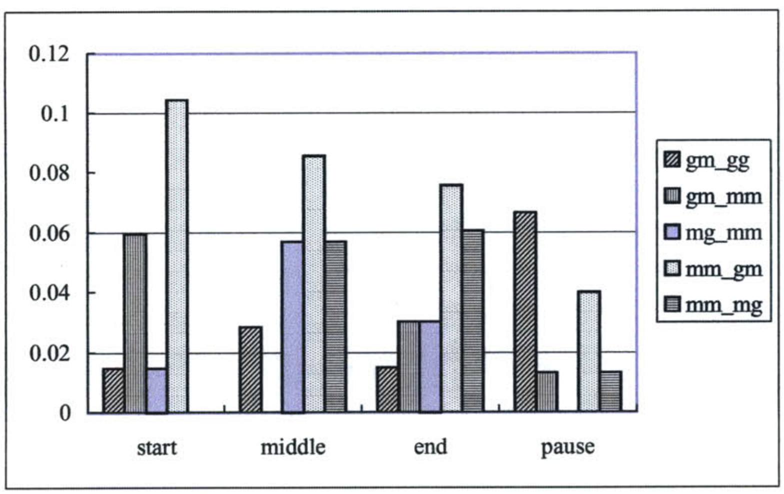

Figure 3.4-7: NV state transition for Information Request

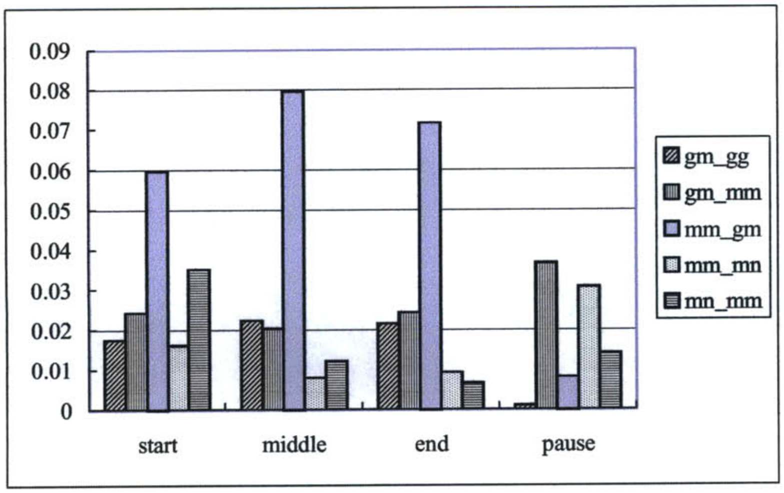

Figure 3.4-8: NV state transition for Assertion

(3) Distribution of behaviors of each conversational participant

Figure 3.4-9, Figure 3.4-10, Figure 3.4-11, and Figure 3.4-12 show the distribution of non-verbal grounding acts for a participant when her/his partner's behavior is given.

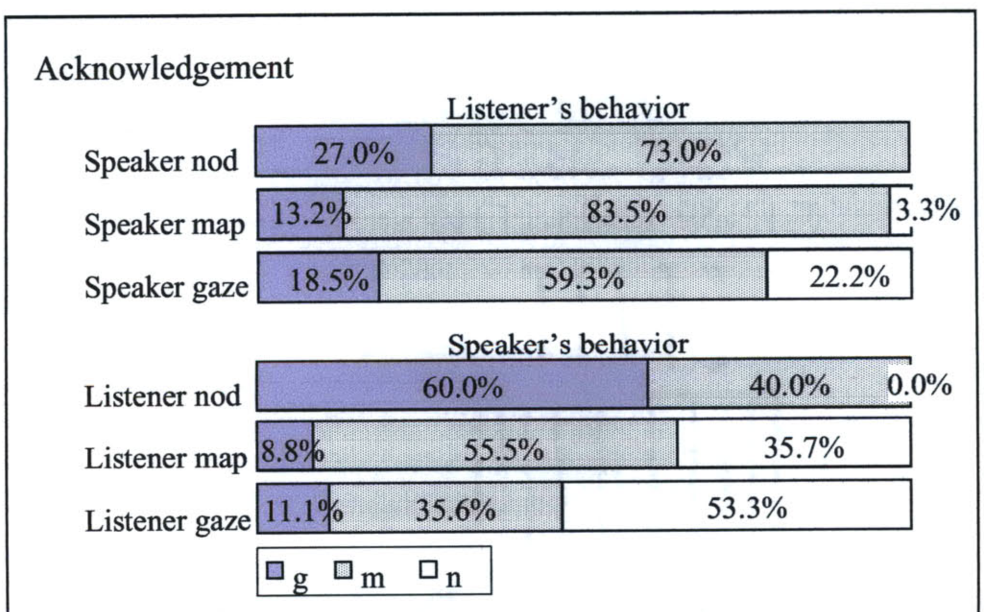

Figure 3.4-9: Distribution of behavior for each participant (Acknowledgement)

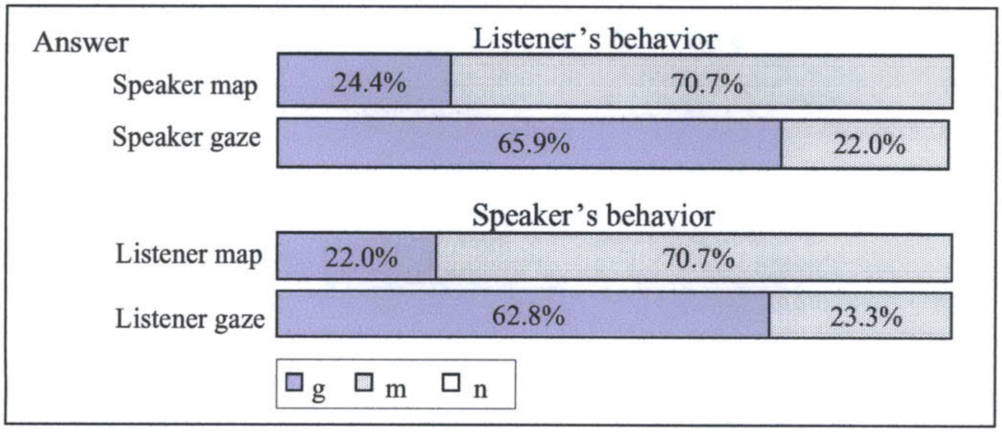

Figure 3.4-10: Distribution of behavior for each participant (Answer)

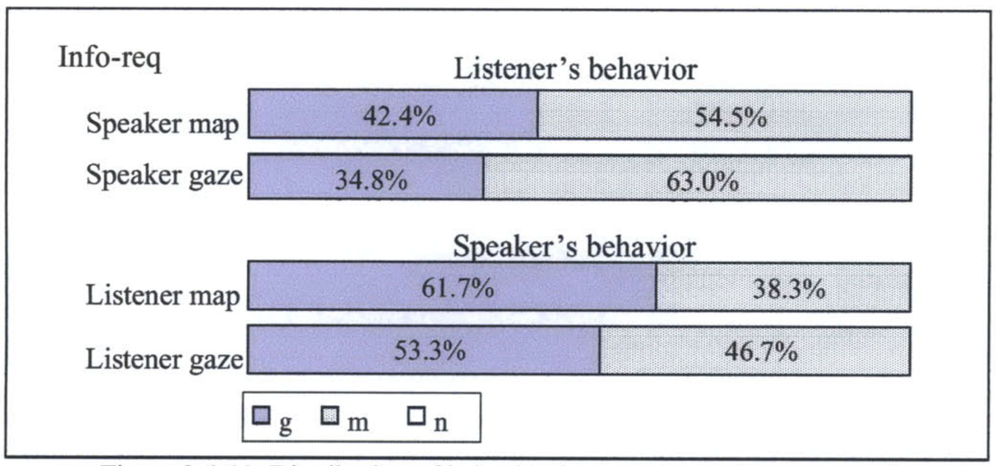

Figure 3.4-11: Distribution of behavior for each participant (Info-req)

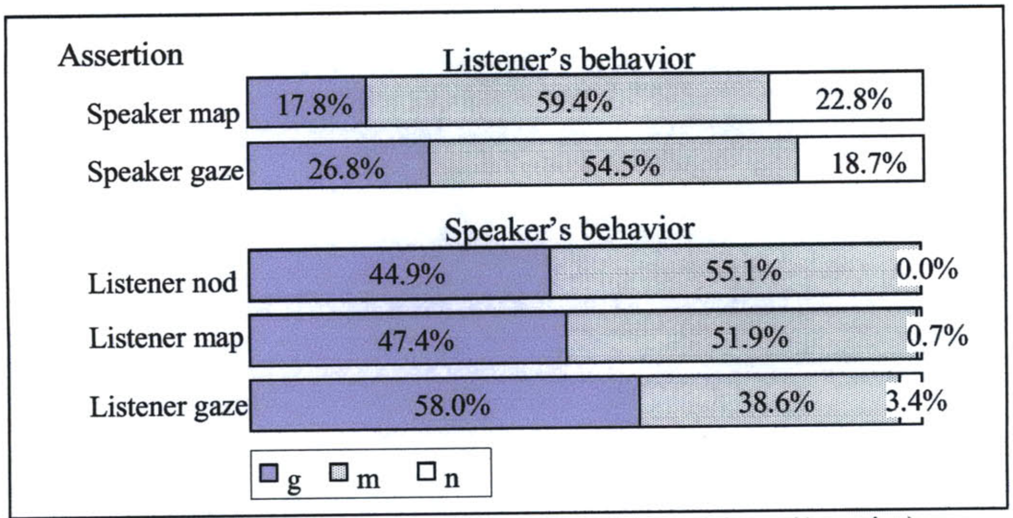

Figure 3.4-12: Distribution of behavior for each participant (Assertion)

Acknowledgement: Regardless of a speaker's behavior, the most frequent listener's behavior is always looking at the map. However, the distribution of a speaker's behavior is different depending on the listener's non-verbal act. A speaker's most frequent non-verbal act is a nod (53.3%) when the listener looks at the speaker. When the listener looks at the map, the most frequent non-verbal act by a speaker is looking at the map (55.5%). Note that a speaker more frequently uses a nod when the listener looks at the speaker (53.3%) than looks at the map (35.7%). This result suggests that the listener's gaze is more likely to co-occur with the speaker's nod.

Answer: When a speaker looks at a listener, the listener looks at the speaker 65.9% of the time. When the speaker looks at the map, the listener looks at the map 70.7% of the time. Interestingly, the same result was found when looking at the speaker's behavior, given a listener's behavior. When the listener looks at the speaker, the speaker looks at the speaker 62.8% of the time. When the listener looks at the map, the speaker looks at the map 70.7% of the time. These results indicate that the speaker's gaze co-occurs with the listener's gaze, and also that the listener's gaze is more likely to co-occur with the speaker's gaze.

Information request: The listener's most frequent behavior is looking at the map no matter where the speaker looks at (When the speaker looks at the map, the listener looks at the map 54.5% of the time. When the speaker looks at the listener, the listener looks at the map 63% of the time.). On the other hand, the most frequent speaker's behavior is a gaze, regardless of the listener's behaviors.

When the listener looks at the speaker, the speaker looks at the listener 53.3% of the time. When the listener looks at the map, the speaker looks at the listener 61.7% of the time.

Assertion: The most frequent listener's behavior is looking at the map regardless of where the speaker looks at. On the other hand, when the listener looks at the speaker, the speaker's most frequent behavior is looking at the listener (58%). When the listener looks at the map, the most frequent speaker's behavior is looking at the map (51.9%). When the listener nods, the speaker looks at the map 55.1% of the time. Note that speaker's gaze more frequently co-occurs with the listener's gaze (58.%), compared to the case that the listener looks at the map (47.4%) and is nodding (44.9%).

In summary, except for answer, the listener's dominant behavior is always looking at the map. In contrast, the distribution of a speaker's non-verbal act is different depending on the type of listener's behavior. This result suggests that the listener's behavior can predict the speaker's behavior to some extent, but the speaker's behavior does not constrain the listener's behavior in most cases. In particular, the listener's gaze more frequently co-occurs with speaker's gaze in Answer and Assertion, and with speaker's nod in Acknowledgement. Thus, it is expected that eye gaze would play important role in grounding in face-to-face communication. In addition, similar result was found regardless of the participation framework. However, in Assertion, the distribution of the speaker's non-verbal acts is different depending on who is asserting. When the giver asserts, if the receiver looks at the giver, the giver looks at the receiver 66.1% of the time. On the other hand, when the receiver asserts, the most frequent giver's behavior is looking at the map (50%).

(4) Summary and discussion for face-to-face condition

In analysis of distribution and the residuals, it was found that salient NV status is different depending on the type of verbal act. In Acknowledgement, *nm* is the salient NV status, *gg* is salient in Answer, *gm* in Info-req, and *mn* in Assertion. Note that all the salient NV statuses revealed by the residual analysis also appear in typical patterns of NV status transition. For example, in Acknowledgement, the salient NV status is *nm*, and this is the second step in the typical transition. Thus, these results consistently show that usage of the nonverbal signals is different depending on the type of verbal act, and, more precisely, depending on the place in the verbal act.

In Answer, keeping mutual gaze during speaker's answering is required as positive evidence of understanding. In Information request, speakers need to get mutual gaze right after the question. In Assertion, the listener's paying attention to the shared referent is observed by the speaker as evidence of accepting the information conveyed with the speech. In these three types of verbal acts, listeners display different kinds of non-verbal signals at different timing and the speaker needs to perceive the signals by looking at the listener.

However, in Acknowledgement, in many cases, speakers and listeners do not look at their partner. The typical NV status that accompanies this verbal act is *nm*, and the typical transition pattern is from *mm* to *nm*, then to *mm*. There are two possible interpretations for this phenomenon. If speakers think that the listener does not perceive the speaker's nod, it is valid to interpret that speakers do not use nod intentionally as a display of evidence of understanding. Therefore, it seems that nod is an introspective behavior, and occurs in a cognitive process of understanding what the partner said. As the second interpretation, it would be possible that listeners perceive the speaker's nod by their peripheral view, and the speaker knows that. If so, speakers use nods intentionally to display their understanding to the listener. The result of analysis in (3), distribution of behaviors of each conversational participant, supports this interpretation. Listener's gaze more frequently co-occurs with speaker's nod in Acknowledgement. This result suggests that the speaker could perceive the listener's gaze and display nods to the listener intentionally. It is impossible to examine these hypotheses with this data because we don't have equipment to observe a subject's view. Looking at the data in shared reference condition could give more information to discuss this issue further.

3.4.3. Analysis of shared reference condition

The same method of analysis was applied to the data in the shared reference condition. In this condition, it is impossible to see the partner's face and body, which are hidden by a board. So, when a participant looks straight toward the board, I categorized this behavior as gaze at the partner though it is actually impossible to look at the partner.

(1) Distribution of non-verbal status with respect to UU type

The distributions of NV statuses are shown in Table 3.4-10, Table 3.4-11, Table 3.4-12. *Mm* is the most frequent status in all UU types. The only exception is that, in Acknowledgement, *nm* is as frequent as *mm* (46.5%). In contrast with the face-to-face condition, the most frequent combination of non-verbal grounding acts is not different depending on the UU type.

Although the most frequent NV status is the same, the results of Chi-square tests for these tables are all statistically significant. In order to reveal the cause of this result, a residual for each cell is calculated and shown in Table 3.4-13, Table 3.4-14, and Table 3.4-15. In general, the most salient NV status is *nm* in acknowledgement, *em* and *en* in answer, and *mn* in assertion. No NV status is statistically significant in info-req. When only looking at a direction giver's utterance, the same result was found. As for the receiver's utterances, the overall frequency of each table was very low, so that I could not get a reliable result for this analysis.

In both face-to-face and shared reference condition, the salient NV status in Acknowledgement and Assertion are *nm*, and *mn* respectively. However, in Answer and Info-req, the salient statuses are different between the two conditions. The interesting point here is that, in Answer and Info-req, the salient NV status includes gaze of either or both of the conversational participants. Therefore, this result suggests that, in grounding these verbal acts in shared reference condition, the conversational participants cannot use the same grounding strategy used in face-to-face condition because they cannot use eye gaze as a nonverbal grounding act.

Table 3.4-10: Distribution of NV combinations for all cases

All	gm	gn	ge	mg	mm	mn	me
ack	1(0.6%)	0	0	1(0.6%)	80(46.5%	4(2.3%)	1(0.6%)
ans	5(10.6%)	0	1(2.1%)	1(2.1%)	17(36.2%	1(2.1%)	1(2.1%)
ir	0	0	0	0	2(66.6%)	0	0
as	22(10.8%	3(1.5%)	1(0.5%)	4(2.0%)	93(45.8%	31(15.3%	1(0.5%)

	ng	nm	nn	eg	em	en	ee
ack	2(1.2%)	80(46.5%	2(1.2%)	0	1(0.6%)	0	0
ans	0	8(17.0%)	1(2.1%)	0	10(21.3%	2(4.3%)	0
ir	0	1(33.3%)	0	0	0	0	0
as	0	9(4.4%)	0	1(0.5%)	35(17.2%	2(1.0%)	1(0.5%)

Chi square test p<.001

Table 3.4-11: Distribution of NV combinations for a Giver's UU

Giver	gm	gn	ge	mg	mm	mn
ack	1(3.2%)	0	0	0	13(41.9%	1(3.2%)
ans	5(13.2%)	0	1(2.6%)	0	14(36.8%	0
ir	0	0	0	0	1(100%)	0
as	22(11.6%	3(1.6%)	1(0.5%)	2(1.1%)	85(45.0%	30

	me	nm	nn	em	en	ee
ack	0	14(45.2%	1(3.2%)	1(3.2%)	0	0
ans	1(2.6%)	5(13.2%)	0	10(26.3%	2(5.3%)	0
ir	0	0	0	0	0	0
as	1(0.5%)	9(4.8%)	0	33(17.5%	2(1.1%)	1(0.5%)

Chi square test p<.001

Table 3.4-12: Distribution of NV combinations for a Receiver's UU

Receiver	mg	mm	mn	me	ng	nm	nn	eg	em
ack	1(0.7%)	67(47.5%	3(2.1%)	1(0.7%)	2(1.4%)	66(46.8%	1(0.7%)	0	0
ans	1(11.1%)	3(33.3%)	1(11.1%)	0	0	3(33.3%)	1(11.1%)	0	0
ir	0	1(50%)	0	0	0	1(50%)	0	0	0
as	2(14.3%)	8(57.1%)	1(7.1%)	0	0	0	0	1(7.1%)	2(14.3%)

Chi square test p<.001

Table 3.4-13: Adjusted residuals of NV combinations for all cases

All	gm	gn	ge	mg	mm	mn	me
ack	-4.11585	-1.43319	-1.16881	-1.1964	0.45603	-3.75132	-0.25275
ans	1.14703	-0.61075	1.75597	0.43799	-0.97406	-1.58427	1.23016
ir	-0.45222	-0.14625	-0.11935	-0.20712	0.61479	-0.51479	-0.14625
as	3.27549	1.812	0.0633	0.92765	0.19148	4.6279	-0.50061

	ng	nm	nn	eg	em	en	ee
ack	1.71924	9.46445	0.92768	-0.8255	-5.60376	-1.65687	-0.8255
ans	-0.49868	-0.91399	1.23016	-0.35262	2.30971	2.48332	-0.35262
ir	-0.11935	0.39058	-0.14625	-0.08436	-0.5848	-0.16895	-0.08436
as	-1.35269	-7.7457	-1.65692	1.04589	3.86889	0.08955	1.04589

Table 3.4-14: Adjusted residuals of NV combinations for a Giver's UU

Giver	gm	gn	ge	mg	mm	mn	me
ack	-1.44955	-0.6424	-0.52349	-0.52349	-0.20269	-1.59842	-0.52349
ans	0.50442	-0.72241	1.41753	-0.5887	-0.91332	-2.4607	1.41753
ir	-0.34883	-0.10846	-0.08839	-0.08839	1.13888	-0.36945	-0.08839
as	0.70634	1.06025	-0.7344	0.86401	0.71677	3.1804	-0.7344

	nm	nn	em	en	ee
ack	6.56464	2.71723	-2.17479	-0.74323	-0.36945
ans	0.50442	-0.41547	1.65752	2.01253	-0.41547
ir	-0.34883	-0.06238	-0.45326	-0.12549	-0.06238
as	-5.15138	-1.64635	0.3323	-1.04267	0.60976

Table 3.4-15: Adjusted residuals of NV combinations for a Receiver's UU

Receiver	mg	mm	mn	me	ng	nm	nn	eg	em
ack	-3.39289	-0.0445	-1.58324	0.42235	0.59911	2.87488	-1.38994	-2.38205	-3.37899
ans	1.75041	-0.88062	1.46174	-0.24015	-0.34066	-0.55192	2.80097	-0.24015	-0.34066
ir	-0.22357	0.06865	-0.25074	-0.11077	-0.15712	0.22563	-0.15712	-0.11077	-0.15712
as	3.02824	0.74791	0.94502	-0.30441	-0.43181	-3.33895	-0.43181	3.30499	4.68819

(2) Transition of NV status

Frequencies of NV status transitions are shown in Figure 3.4-13, Figure 3.4-14, and Figure 3.4-15.

Acknowledgement: The most frequent within utterance transition is from *mm* to *nm*. At a pause between utterances, shift from *nm* to *mm* occurs most frequently. This result is exactly the same as in face-to-face condition.

Answer: During an utterance (start, middle, and end of the utterance), shift from *mm* to *em* occurs most frequently. The number of transition from *em* to *mm* increases toward the end of the utterance. This means that a speaker looks away at an earlier place of an utterance and looks at the map for the rest of the utterance. At a pause between utterances, shift from *nm* to *mm* is the most frequent.

Info-req: Any NV status shift rarely occurs in this UU type. The total number of transition is only two. Thus, in this type of UU, both a speaker and a listener keep looking at the map, and do not change their NV status.

Assertion: At the beginning of the utterance, A shift from *mn* to *mm* is the most frequent. A shift from *mm* to *em* decreases and *em* to *mm* increases during an utterance. A shift from *mm* to *mn* more frequently occurs at the middle of an utterance than at the pause after an utterance. In face-to-face condition, this transition is more frequent at the pause after the utterance. This result indicates that nodding is used in different places depending on the communication modality. In the shared reference condition, nod is more likely to occur during processing the information conveyed by the speaker. In face-to-face condition the listener nods more frequently after processing the utterance.

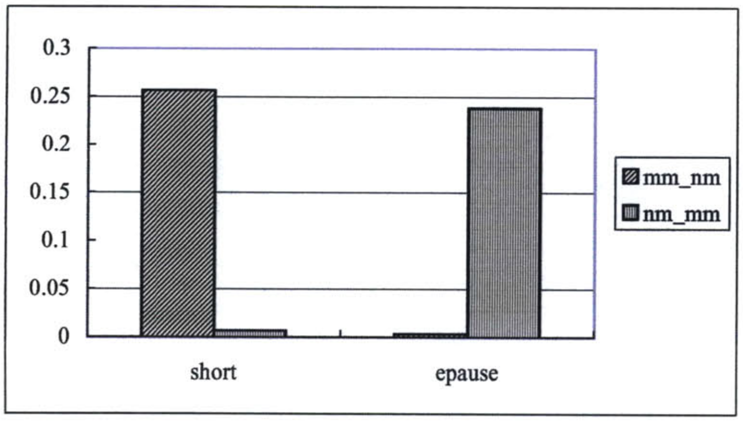

Figure 3.4-13: NV state transition for Acknowledgement

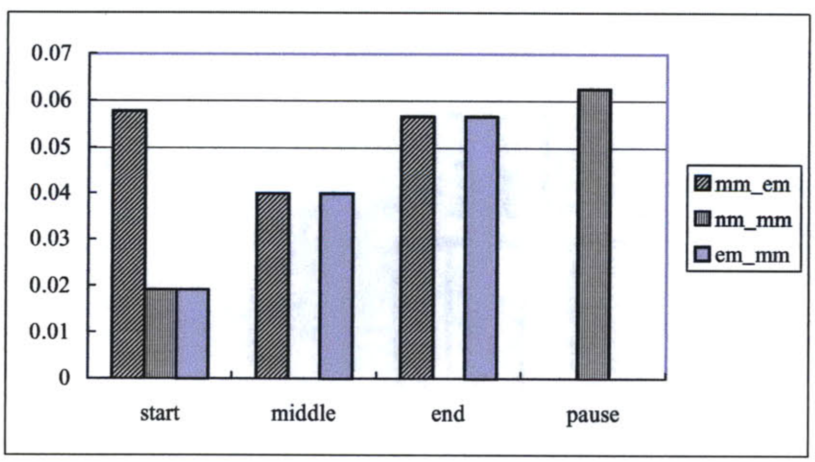

Figure 3.4-14: NV state transition for Answer

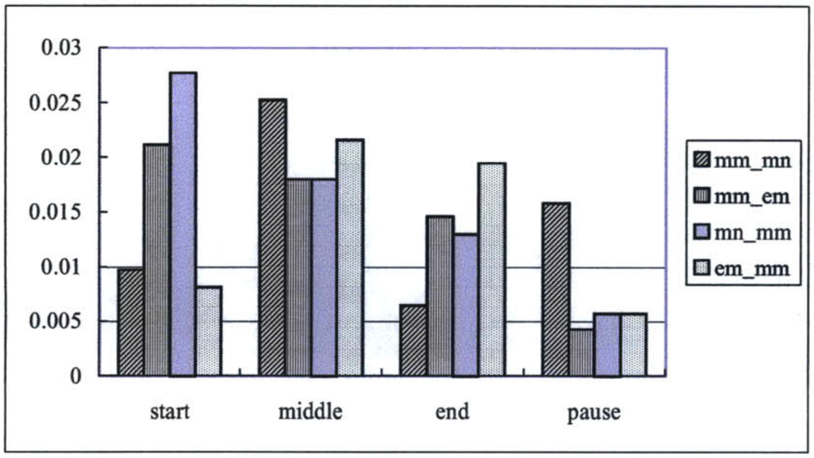

Figure 3.4-15: NV state transition for Assertion

(3) Distribution of behaviors of each conversational participant

Figure 3.4-16, Figure 3.4-17, Figure 3.4-18 show the distribution of non-verbal acts for a conversational participant when her/his partner's behavior is given.

Acknowledgement: Regardless of the speaker's behavior, the most frequent listener's behavior is always looking at the map. On the other hand, while the listener looks at the map most of the time, the speaker looks at the map half of the time (49.4%) and nods the rest of the time (49.4%).

Answer: Regardless of the speaker's behavior, the most frequent listener's behavior is always looking at the map. On the other hand, the speaker's behavior is distributed over different kinds of behaviors though looking at the map is still the most frequent (42.5%).

Assertion: Generally, the most frequent behavior for both a speaker and a listener is looking at the map regardless of the partner's behavior. In addition, there seems to be a correlation between nodding and looking at the map. The listener more frequently uses nodding when the speaker looks at the map (24.0%) than when the speaker looks toward the listener (11.5%). The speaker more frequently looks at the map when the listener is nodding (86.1%) than when the listener is looking at the map (58.5%). This result suggests that the listener more frequently uses nodding when the speaker gives information using the map.

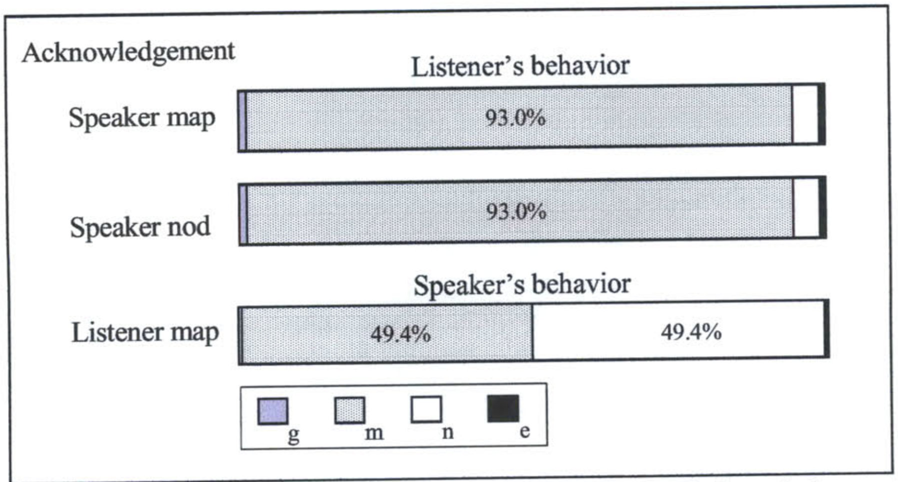

Figure 3.4-16: Distribution of behavior for each participant (Acknowledgement)

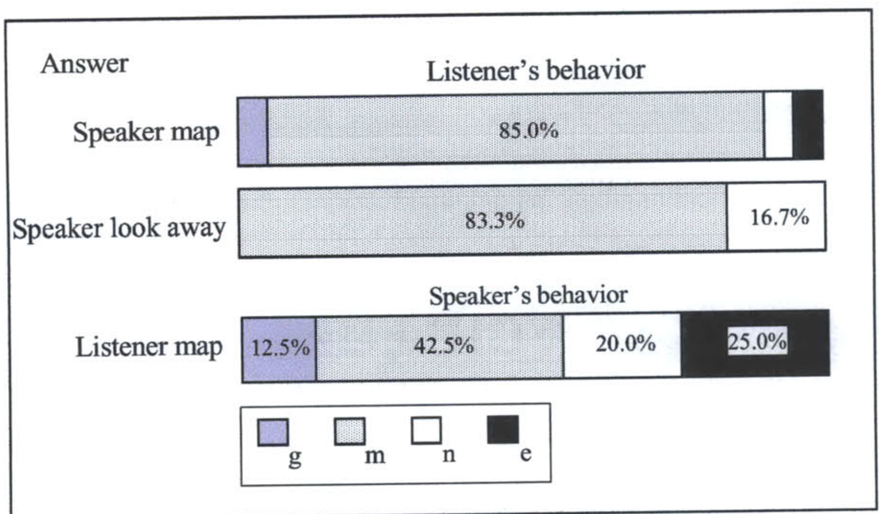

Figure 3.4-17: Distribution of behavior for each participant (Answer)

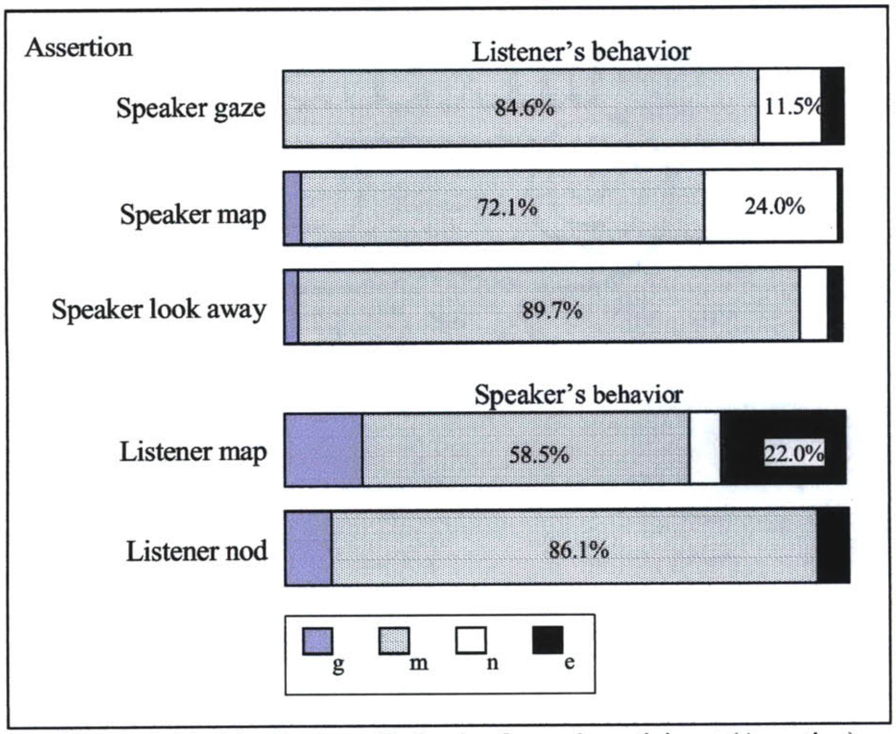

Figure 3.4-18: Distribution of behavior for each participant (Assertion)

(4) Summary and discussion for shared reference condition

In the analysis of distribution and the residuals, it was found that salient NV status is different depending on the type of verbal act. The salient NV status in Acknowledgement is *nm*, *em* and *en* in Answer, and *mn* in Assertion. An interesting result here is that, in Acknowledgement and Assertion, the salient NV statues (*nm*, and *mn* respectively) are the same as those in face-to-face condition, and these NV statuses consist of nodding of either a speaker or listener. In this case, it is impossible for the conversational participants to see the partner's nod, but they still use nods. This result supports a hypothesis that nod is an introspective behavior, and occurs in a cognitive process of understanding.

In the analysis of Distribution of behaviors of each conversational participant, as listeners look at the map most of the time, it was not clear whether the speaker's non-verbal acts are different depending on the listener's behavior as found in face-to-face communication.

3.4.4. Comparison between face-to-face and shared reference condition

In comparing face-to-face to the shared-reference condition, the most remarkable result is that, in shared reference condition, neither speakers nor listeners rarely looking straight although they frequently look straight at the partner in face-to-face condition. This suggests that paying attention to the speaker by a listener and checking the listener's non-verbal behavior by a speaker are very important strategy in grounding in face-to-face communication.

However, in shared reference condition, conversational participants cannot use eye gaze. If they have to pay more display cost as proposed by (Clark and Brennan 1991) than face-to-face condition (c.f. section 2.3), how do the conversants compensate this disadvantage? Do they more rely on verbal communication in shared reference condition to compensate the lack of bodily signal? In comparing frequency of verbal acts in section 3.4.1, it was fond that Acknowledgement occurs more frequently in shared reference condition than in face-to-face condition, and this is the only difference between the conditions (Table 3.4-2, Table 3.4-3). Moreover, in Acknowledgement, the typical NV status and transition pattern are exactly the same in both conditions. This suggests that the lack of co-presence does not affect the grounding process in Acknowledgement. In sum, eye gaze is a very strong device in grounding face-to-face, and communication without co-presence prefers grounding acts which are less influenced by communication modality, such as Acknowledgement.

In addition to eye gaze, previous studies classified head nod as a non-verbal evidence of understanding. However, I found another possibility that nodding is an introspective behavior in a cognitive process of understanding spoken language. Analysis of shared reference condition (section 3.4.3) gives supporting result for this hypothesis. In Acknowledgement and Assertion, nod is used frequently even though it cannot be observed by the partner. On the other hand, results in analysis of face-to-face condition (section 3.4.2) support another hypothesis that nod is used intentionally as a display of positive evidence of understanding. The analysis revealed that speakers more frequently use nod with their verbal Acknowledgement when the listener is looking at the speaker.

Although we cannot observe the subjects view at the experiment, it is possible to discuss some statistics. Comparing the frequency of *nm* in Acknowledgement between two conditions, the difference is not statistically significant (0.219 per utterance unit in face-to-face, 0.261 in shared reference condition). However, *mn* in Assertion is more frequent in face-to-face than in shared reference condition (0.123 in face-to-face, 0.052 in shared reference condition. $t(7) = 5.363$ $p < .01$). In Acknowledgement, speakers use nod even if they know that it is not observed by the listener. On the other hand, listeners use nod less frequently when they know that it cannot be observed by the speaker. These results suggest that function of listeners' nod in Assertion is different from that of speaker's nod in Acknowledgement. The former is a non-verbal display of evidence of understanding, and the latter seems more like an introspective behavior in processing spoken language. Therefore, I can conclude that function of non-verbal act is different depending on what types of verbal act the non-verbal act works with.

Now we go back to the original classification of evidence of understanding in (Clark 1996). This thesis is concerned with these two kinds of evidence.

(1) Continued attention: If the hearer looks away from the speaker, the speaker tries to capture the hearer's gaze, and attention. This is the most basic form of positive evidence.

(2) Assertions of understanding: Asserts understanding using verbal and non-verbal behaviors, such as "uh huh", "I see", or nod or smile.

Verbal and non-verbal behaviors described above were found in direction giving dialogues. In addition, I found that the usage of non-verbal behavior is different depending on not only the modality of communication, but also the type of verbal act with which the non-verbal act co-occurs. In shared reference condition, as the conversational participants cannot use bodily signals, they more frequently use verbal assertions of understanding, such as "Um-hum", or "OK". On the other hand, in face-to-face communication, the conversational participants much more frequently use non-verbal evidence of understanding, such as eye gaze and nod. In Answer, keeping mutual gaze during speaker's answering is required as positive evidence of understanding. In Information request, speakers need to get mutual gaze right after the question. In Assertion, the listener's paying attention to the shared referent is observed by the speaker as evidence of accepting the information conveyed with the speech. In Acknowledgement, speaker's nod is used without being observed by the listener. Note that these results indicate that speakers do not always need the listener's attention. Paying attention to the map can work as positive evidence by co-occurring with Assertion. Moreover, nod is frequently used without paying attention to the partner.

In conclusion, it was found that usage of non-verbal grounding acts is different depending on the communication modality and the type of verbal act. Specifically, in face-to-face communication, non-verbal behaviors are changing during an utterance and a typical transition pattern of non-verbal behaviors is also different depending on the type of verbal act.

4. Implementation

This chapter describes the implementation of grounding functionality in an embodied conversational agent. First, a system architecture of a conversational agent, MACK (Media Lab Autonomous Conversational Kiosk), is described. Second, I propose the dialogue manager and its dialogue state update mechanism that handles grounding processes. In the updating mechanism, results found in the empirical study reported in the last chapter are used in judging whether something is grounded or not. Finally, an example of interaction between a conversational agent and a user is described.

4.1. System

MACK is a conversational kiosk that can answer questions about the MIT Media Lab's various research groups, projects, faculty and academic research staff, and can give directions to demos and research labs using a combination of speech, gesture, and indications on a normal paper map that users place on a table between themselves and MACK (Stocky and Cassell 2002). An example of an interaction with MACK is shown in Figure 4.1-1. After MACK's greeting in [1], the user asks about a group in [2]. In [3], MACK gives a description of the group. In [4], the user asks about a group by pointing at a room on a shared map. In [6], the user asks about a demo, and MACK gives information about the demo in [7]. In [8], the user asks how to get to a room and then MACK gives a direction to the room in [9].

The system architecture is shown in Figure 4.1-2. MACK takes input from a speech recognition system, a pointing sensor, and an Eye-tracking Module (ETM) (the details are described later). We use MIT LCS SpeechBuilder for speech recognition ((Glass and Weinstein 2001)), a Wacom tablet for recognizing where the user points on the map, and IBM Blue Eyes technology (http://www.almaden.ibm.com/cs/blueeyes) for tracking eye movement.

[1] MACK: Hello, my name is Mack. Let me assist you! Let me try to help you make sense of this maze we call <H1> the Media Lab. So, how can <H5> I help you?
[2] USER: Tell me about Gesture and Narrative Language group.
[3] MACK: Oh, GNL ! They're my favorite group. They build machines that can be a part of face-to-face <GB00> conversation and <GB00> story-telling. <GB22> They're like parents to me!
[4] USER: What is here?
[5] MACK: The Lifelong Kindergarten group develops <GB00> technologies and <GB00> activities to help people learn new things i n <H1> new ways. Inspired by the ways children learn in kindergarten, they engage people in <GB05> playful design activities. They're <H1> the ones who spend all day playing with Lego toys!
[6] USER: Can you describe Rea project?
[7] MACK: Rea is kind <GB04> of like my grandma. <GB20> She is an autonomous agent is capable of having a real-time face-to-face <GB20> conversation with a human.
[8] USER: Which way is the Pond?
[9] MACK: To visit The Pond, <GR14> turn right after exiting the elevator on the third level. Follow the hallway to <GR21> the end, go through the double doors, <GR13> turn left and follow the hallway. <GR19> The Pond is on your <GL13> left half way down the hall.

Figure 4.1-1: Example of conversation with MACK

Understanding Module: Outputs from the speech recognition system and the pointing sensor are sent to the Understanding Module, which interprets a user's input by combining the speech and pointing information. The resulting interpretation is sent to the Dialogue Manager.

Eye-tracking Module (ETM): The input from the CCD camera is used for recognizing head nods and eye gaze. Head nods are recognized by using a Hidden Markov Model based network proposed by (Kapoor and Picard 2001). When a head nod is recognized, the result is saved in this module with a time stamp.

For sensing the user's eye gaze information, a neural network learns the distribution of movements of two pupils for the user looking at the agent and looking at the map, and judges which way the user is looking at by using the network. Eye gaze state is checked every 0.1 seconds, and the result (looking at the agent/the map) is maintained with a time stamp.

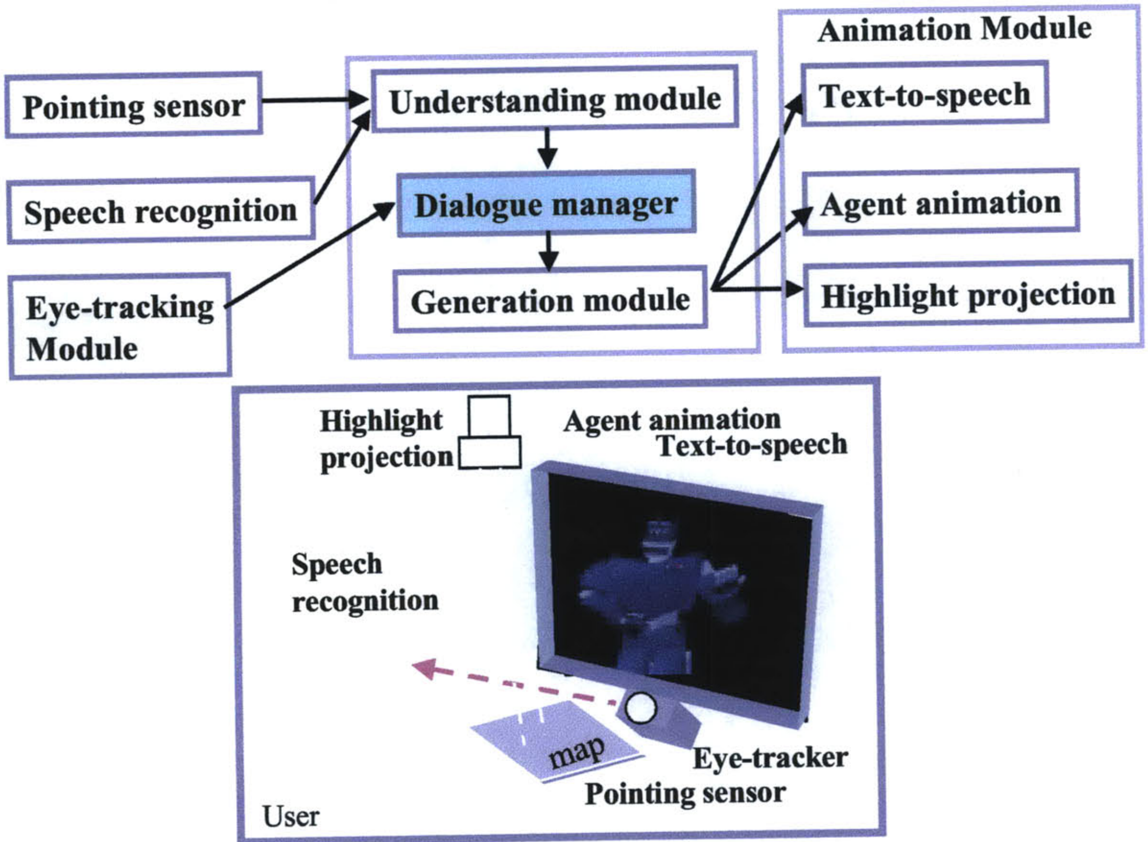

Figure 4.1-2: System architecture

In order to notify the Dialogue Manager of non-verbal information, this module has a function that takes start time and end time as input and returns results of eye gaze state and head nod recognition during the specified time period. When a time stamp of a head nod is within the time period, this function returns "head nod" as the non-verbal state. For eye gaze information, this function computes a major state during the period. For example, suppose that the ETM collects seven data points for a 0.7 seconds period. If more than four of them report that the user is looking at the map, the function returns "looking at the map" as a return value. An advantage_of this method is that it judges more accurately by using more data in a time period. For example, if the accuracy of each judgement is 0.8 (error rate is 20%), theoretically the error rate for a 0.7-second-time-period is about 0.002.

Dialogue Manager (DM): This module updates the dialogue state and decides the next agent action. The Dialogue Manager accesses the ETM to check the user's non-verbal state when it is necessary.

The details will be described in the next section.

Generation Module: Content of the next utterance unit (UU) output by the Dialogue Manager is sent to the Generation Module. This module constructs an utterance along with gestures, facial expressions, and highlights that appear on the map.

Animation Module: Finally, the surface expression of the utterance is sent to a speech synthesizer for speech output (the Microsoft Whistler TTS engine). Gestures and facial expressions are produced by BEAT ((Cassell, Vilhj msson et al. 2001)), which animates MACK's VRML-defined character using a variety of motor skill modules. Highlights on a map are projected through a LCD projector. These three outputs from different channels are produced in a synchronized way.

4.2. Dialogue manager (DM)

The Dialogue Manger updates the state of the dialogue. It keeps track of what is grounded, and what need to be grounded. It also decides the agent's next action.

4.2.1. Dialogue state updating mechanism

The dialogue state is implemented as a blackboard in the Dialogue Manger. Although similar blackboard-based dialogue state representations have been proposed ((Larsson, Bohlin et al. 1999), (Matheson, Poesio et al. 2000)), this thesis is unique in its ability to handle non-verbal information in grounding. The dialogue state consists of Grounded (GND), Un-grounded (UGND), Current Discourse Unit (CDU), and Previous Discourse Unit (PDU).

Grounded (GND): a list of grounded information.

Un-grounded (UGND): a list of ungrounded discourse unit (DU)

Current Discourse Unit (CDU): a set of information about the most recent DU.

Previous Discourse Unit (PDU): A set of information about penultimate DU. This is actually assigned by moving a CDU when a new discourse unit comes in.

A Discourse Unit (DU) includes verbal and non-verbal information that is conveyed with an utterance unit (UU) by either MACK or a user. This consists of the following slots;

(1) ID: identification number **(2) Verbal state:** indicate what is conveyed by verbal acts **Verbal act (VA):** contains of the following information; start and end time of UU, speaker of the UU, UU type, UU description,. **Belief (Bel):** Information conveyed by the utterance. This will be added as a shared belief when the DU is grounded **(3) Nonverbal state:** indicates non-verbal state for within and after the utterance. **userNV_withinUU:** user's NV state within the UU **systemNV_withinUU:** system's NV state within the UU **userNV_pause: user's** NV state for a pause between UUs **systemNV_pause:** system's NV state for a pause between UUs **(4) Score:** indicates how much positive evidence is accumulated for the given discourse unit.

There are two ways of updating the dialogue state, which are shown in Figure 4.2-1. When the Dialogue Manger gets user's UU from the Understanding Module (connected with dotted lines), the dialogue state is updated according to the user's UU. The other case is after the DM decides on the content of the next system's UU to be generated (drawn with solid lines). In this case, the utterance content is used to update the dialogue state.

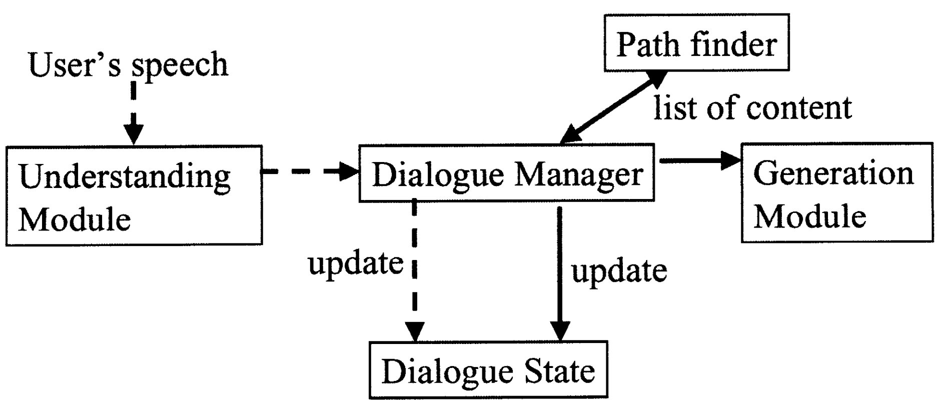

Figure 4.2-1: Flow of updating mechanism

The basic idea for the updating mechanism is as follows; when the mechanism is triggered by a UU, 1.) the mechanism updates the verbal and non-verbal information in a Discourse Unit (PDU and

CDU), 2.) the Dialogue Manager searches and applies an updating rule to add "scores" to a Discourse Unit. The score indicates how much verbal and non-verbal evidence of understanding is accumulated. 3.) If the updated score of a Discourse Unit is higher than a threshold, the Discourse Unit is moved to Grounded field in the dialogue state, and removed from the Un-grounded field.

The details of each step in updating mechanism are as follows. When the DM gets user's speech input or system's utterance content,

(a) Update PDU

(a-1) First, the current CDU is moved to PDU. Then, the Dialogue Manager (DM) gets the user's non-verbal state for a pause between the previous utterance and the new input. This is obtained by accessing the Eye-Tracking Module (ETM) with the end time of the previous utterance and the start time of the new input. The return value from the ETM is saved in a **userNV_pause** in the PDU.

(a-2) Then, the DM searches and applies an update rule for the PDU (described in section 4.2.2).

(a-3) Judges whether the PDU can be moved to GND (described in section 4.2.3).

(b) Update CDU

(b-1) This step creates a new CDU. The DM creates a new DU object, assigns an ID to the DU, and makes it the CDU.

(b-2) Fill the Verbal Act slot with the meaning of the user's input. When the system receives the user's speech input, the output from the Understanding Module (the result of interpretation) is filled in the slot. When the dialogue state is updated after the DM decides the next system's utterance, the content of the output utterance is filled in. The content of information is saved in the Belief slot, which will become a shared belief when the DU is grounded. The belief is computed by extracting a part of the logical form of the Verbal act.

(b-3) This step computes the non-verbal state of the user for the CDU and saves it in a **user_NVstate_withinUU** in the CDU. The DM accesses the ETM with the start time of an utterance and the end time of the utterance, and gets the user's non-verbal state within that time period. In the current implementation, when the updating mechanism is triggered by a user's speech input, the DM can get the end time of the utterance from the Understanding Module, but cannot get its start time. This is because the Understanding Module interprets the user's input after the speech has finished. Therefore, the DM needs to estimate the start time of the utterance. As a temporary solution, the start time is estimated using the number of words in the recognized utterance. On the other hand, when the updating mechanism is triggered by a system's next utterance, the start time is specified when the utterance content is sent to the Generation Module. However, the DM needs to wait for the speech end signal to get the end time of system's utterance. Thus, this step is suspended until the DM gets the speech end event from the Animation Module.

(b-4) This step assigns the non-verbal state of the system (agent) for the CDU and saves it in a **system_NVstate_withinUU** in the CDU. When the updating mechanism is triggered by the system's next utterance, the system's non-verbal state within the utterance is assigned according to the type of the Verbal act (specified in step (b-2)). Appropriate non-verbal states for each type of verbal act are determined based on the results of the empirical study. They are specified as shown in Table 4.2-1

Table 4.2-1: Appropriate NV state change within UU

Verbal acts	Appropriate non-verbal state change within UU
Acknowledgement	Nod
Info-req	look at the map, then looks at the user
Answer	look away, then looks at the user
Assertion	look at the map, then looks at the user

The selected non-verbal state is sent to the Generation Module with the content of the utterance to make the agent perform appropriate non-verbal grounding acts. On the other hand, when the updating mechanism is triggered by the user's utterance, looking at the map is assigned as a default system's non-verbal state. In the current implementation, the Understanding Module cannot interpret the input speech in an incremental way. Therefore, it is impossible to change the agent's non-verbal behavior within the user's on-going speech. However, the agent's non-verbal behavior can be changed according to the expectation of the user's next utterance. For example, when the system asks a question in the current utterance, the next expected input is the user's answer for the question. In this case, the system moves the agent's gaze to the user before the user starts speaking.

(b-5) After the verbal and non-verbal information is filled in the CDU, the DM searches an update rule that can be applied to the CDU, and updates the score of the CDU by applying it (section

4.2.2).

(b-6) Finally, judges whether the CDU can be moved to GND (section 4.2.3).

4.2.2. Searching and applying an updating rule

The dialogue state is changed by applying updating rules defined as follows.

Header: input act type

Condition: condition for applying a rule

Effects: operation to a discourse unit to update the dialogue state

Examples of updating rules are shown in Figure 4.2-2 and Figure 4.2-3.

```
(header: (assertion)
cond: ((Who,MACK)
       (UUType,assertion)
       (CDU_NVstatus_within,gm))
effects: ((addScore (0.324)))
```

Figure 4.2-2: Example of updating rule for assertion

```
(header: (ack)
cond: ((Who,user)
       (UUType,acknowledgement)
       (CDU_NVstatus_within,nm)
       (pdu_id,¬grounded)
       (UUtype_pdu,assertion))
effects: ((push_pdu_to_ground)))
```

Figure 4.2-3: Example of updating rules (acknowledgement)

In updating a DU, first the DM searches an applicable update rule for the DU by checking whether the current dialogue state satisfies the condition of a rule. If the condition is satisfied, the rule is applied to a DU. In Figure 4.2-2, conditions for this rule are: the speaker should be MACK, and UU type should be assertion. These two conditions are checked by looking at Verbal Act in the DU. The third condition is concerned with non-verbal state within DU; the agent looks at the user (g), and the user looks at the map (m). These are checked by looking at the value of **userNV_withinUU** and **systemNV_withinUU** in the DU. If these all conditions are satisfied, 0.324 is added to the score in slot with the meaning ofpdating rule is determined based on the empirical study in Chapter 3.

A score in an updating rule is determined based on the empirical study in Chapter 3. Table 4.2-2 shows proportions of NV status transitions observed in the data. For example, within an UU in Assertion, shift to *gm* occupied 32.4% of all the transitions. However, it only occurs 10.3% of the time during a pause between UUs. These proportions are used as scores which are added in applying an updating rule.

Therefore, NV status changes that are frequently observed in the data have higher scores. Note that this does not specify a transition from one status to another, but, the most frequent transition observed in the data can be reconstructed by traversing the highest score. For example, in a table for Assertion, *gm* has the highest score for within UU (0.324), and *mm* has the highest score for a pause (0.410). If the NV status shifts according to a typical pattern, *gm* to *mm*, the DU gets the highest score.

In addition, a threshold for judging groundedness for each verbal type can be determined by adding the highest score for within UU and that for a pause. As mentioned in section 2.1.3, if the grounding criterion may change as conversation purposes change (Clark and Schaefer 1989), thresholds should be changed depending on the purpose of conversation. Therefore, thresholds used here may not be appropriate for all kinds of conversations, but they are still useful as default thresholds.

Table 4.2-2: Scores for NV status change

Ack	withinUt	pause
gg	0.008	0.034
gm	0.102	0.011
gn	0.008	0.022
mg	0.070	0.056
mm	0.188	0.798
mn	0.016	0.011
ng	0.164	0.022
nm	0.445	0.045
nn	0.000	0.000

Ans	withinUt	pause
gg	0.523	0.050
gm	0.136	0.050
gn	0.023	0.150
mg	0.068	0.050
mm	0.136	0.550
mn	0.000	0.050
ng	0.000	0.050
nm	0.000	0.050
nn	0.000	0.000
eg	0.068	0.000
ge	0.045	0.000

Ir	withinUt	pause
gg	0.160	0.417
gm	0.380	0.250
gn	0.000	0.083
mg	0.160	0.083
mm	0.300	0.083
mn	0.000	0.083
ng	0.000	0.000
nm	0.000	0.000
nn	0.000	0.000

As	withinUt	pause
gg	0.141	0.026
gm	0.324	0.103
gn	0.088	0.154
mg	0.100	0.064
mm	0.267	0.410
mn	0.069	0.237
ng	0.006	0.006
nm	0.004	0.000
nn	0.000	0.000

4.2.3. Judging groundedness

The basic idea for judging groundedness is to calculate the amount of positive evidence and see whether enough verbal and/or non-verbal evidence of understanding is accumulated to constitute the common ground.

```
Add a score specified in the applied rule to the score field in the given DU,
If the score of the given DU is higher than the threshold,
   then
       delete the DU from the UGND list
       put the verbal act in the DU into the Grounded history (gh)
       put the belief in the DU into the Grounded belief (gb)
   else keep the DU in the UGND list
```

The amount of evidence is indicated as the value of score field in a DU. The DM looks at the score value of a given DU, and if the value is higher than the threshold, it judges that this DU is grounded.

4.3. Example

Figure 4.3-1 shows an example interaction between MACK and a user. Dotted lines indicate the place of nods, and solid lines indicate those of gaze at the partner. Lines drawn on the upper side of the words show MACK's non-verbal acts. Lines drawn at the bottom of the words show the user's non-verbal acts. Figure 4.3-2 shows the dialogue state after utterance [3]. At [3], at the beginning of the UU, the NV status is *mm* because both the user and MACK are looking at the map. When MACK starts looking at the user at the middle of the utterance, the NV status becomes *gm*. At the end of [3], the update rule shown in Figure 4.2-2 is applied. As the result of applying the rule, 0.324 point is added to the CDU. If this score is over the threshold, Verbal Act and Belief in this CDU are moved to Grounded history and Grounded belief respectively in GND. Moreover, at [5], the user nods with Acknowledgement, while MACK keeps looking at the map. At this time, a rule for acknowledgement shown in Figure 4.2-3 is applied and the PDU (information presented at [4]) is moved to the grounded filed.

[1] U: How can I get to the Garden?

[2] S: To visit the Garden,

[3] S: make a left,

[4] S: after you exit the elevators on the third floor,

[5] U: Um-hm.

[6] S: Follow the hall way to the end, through the glass doors,

Figure 4.3-1: Example of interaction

Grounded history: [(start=10010,end=10050,speaker=user,VAtype=info-req,
(Info-req(Direction(place the_garden)))]
Grounded belief: [(DU10,SharedKnowledge(Info-req(Direction (place the_garden))))]
UGND: [DiscourseUnit@20]
PDU ID: 20
PDU VA: [start=10100,end=10143,speaker=MACK,VAtype=assertion,
(Inform_PlaceToVisit (place the_garden))]
PDU Bel:[(SharedKnowledge(PlaceToVisit (place the_garden)))
PDU NV: [(systemNV_withinUU,m) , (userNV_withinUU,m) , (systemNV_pause,m),
(userNV_pause,m)]
PDU score: 0.75
CDU ID: DU30
CDU VA: [(start=10150,end=10119,speaker=MACK,VAtype=assertion,
(Inform_Act (act turn_left)))]
CDU Bel: [(SharedKnowledge(Act (act turn_left)))
CDU NV: [(systemNV_withinUU,g) , (userNV_withinUU,m) , (systemNV_pause,?),
(userNV_pause,?)]
CDU score: 0.324

Figure 4.3-2: Dialogue state after [3]

4.4. Discussion for implementation

This chapter described implementation of grounding functionality onto an Embodied Conversational Agent. In order to allow the system to use non-verbal grounding acts in a similar way that human do, frequency of NV status transition observed in real conversation data is exploited as scores to judge groundedness. In addition, a blackboard architecture proposed in previous studies is employed as representation of dialogue state, and it is modified in order to describe non-verbal information in each Discourse Unit (DU). One short point in the current implementation is that the system cannot understand the user's input in an incremental way though human conversants incrementally understand the partner's utterance, and quickly adjust their behavior even in the middle of the utterance. Therefore, in order to approach human-human interaction, it is necessary for the Dialogue Manager to update the dialogue state in a more fine-grained way by accessing non-verbal state more frequently.

5. Conclusion

This chapter gives a summary of this thesis, and discusses some future work.

5.1. Summary

This thesis addressed issues for grounding in face-to-face conversation, especially how non-verbal behaviors, such as eye gaze and head nod, interact with verbal behaviors in a process of grounding. First, in an empirical study, by comparing face-to-face conversation with non-face-to-face communication, it was found that usage of non-verbal behaviors is different depending on the communication modality and the type of verbal act. Specifically, in face-to-face communication, non-verbal behaviors are changing during an utterance and a typical transition pattern of non-verbal behaviors is also different depending on the type of verbal act:

(1) In Answer, keeping mutual gaze during speaker's answering is required as positive evidence of understanding.

(2) In Information request, speakers need to get mutual gaze right after the question.

(3) In Assertion, the listener's paying attention to the shared referent is observed by the speaker as evidence of accepting the information conveyed with the speech.

(4) In Acknowledgement, speaker's nod is used without being observed by the listener.

Then, I implemented grounding functionality onto an Embodied Conversational Agent. The dialogue state updating mechanism in the Dialogue Manager can access information for user's non-verbal signals, and exploits this information in judging whether the speaker's contribution is grounded or not. In order to allow the system to use non-verbal signals in a similar way that human do, frequency of transition of non-verbal behaviors observed in real conversation data is used as scores to judge groundedness.

5.2. Future Work

The most important next step to complete this research is to evaluate the method proposed in this thesis. Even if the proposed method is based on the model of human communication, it is necessary to examine whether the model is also appropriate for human-computer communication. The evaluation would be concerned with the task performance, verbal and non-verbal characteristics of interaction as well as a subjective evaluation using the following criteria: agent's language understanding/use, smoothness of interaction, lifelikeness, social attractiveness, and trustworthiness ((Cassell and Thorisson 1999), (Nass, Isbister et al. 2001)).

As an extension of the empirical study reported in Chapter 3, it would be important to analyze the usage of gestures, and investigate how speech, eye gaze, head nod, and gesture interact each other to achieve common ground. The direction giver frequently uses pointing gestures on the map with looking at the map. On the other hand, the speaker may be more likely to look at the receiver when s/he does a hand gesture in the typical gesture place (McNeill 1992). Analysis of these points will provide more comprehensive model of grounding in face-to-face communication.

Bibliography

Allen, J. and M. Core (1997). Draft of DMSL: Dialogue Act Markup in Several Layers.

Anderson, D. B. and M. A. Casey (1997). "The sound dimension." IEEE Spectrum **34**(3): 46-51.

Argyle, M. and M. Cook (1976). Gaze and Mutual Gaze. Cambridge, Cambridge University Press.

Bavelas, J. B., N. Chovil, et al. (1995). "Gestures Specialized for Dialogue." Personality and Social Psychology Bulletin **21**: 394-405.

Boyle, E., A. Anderson, et al. (1994). "The Effects of Visibility in a Cooperative Problem Solving Task." Language and Speech **37**(1): 1-20.

Brennan, S. (2000). Processes that shape conversation and their implications for computational linguistics. 38th Annual Meeting of the ACL, Hong Kong.

Bull, P. E. (1987). Posture and Gesture, Pergamon Press.

Cassell, J., T. Bickmore, et al. (2000). More Than Just a Pretty Face: Affordances of Embodiment. IUI 2000, New Orleans, Louisiana.

Cassell, J. and K. R. Thorisson (1999). "The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents." Applied Artificial Intelligence **13**: 519-538.

Cassell, J., H. Vilhj • •msson, et al. (2001). BEAT: The Behavior Expression Animation Toolkit. SIGGRAPH 01, Los Angeles, CA, ACM Computer Graphics Press.

Chovil, N. and A. Fridlund (1991). "Why Emotionality Cannot Equal Sociality: Reply to Buck." Journal of Nonverbal Behavior **15**(3): 163-167.

Clark, H. H. (1994). "Managing problems in speaking." SPEECH COMMUNICATION **15**(3-4): 243-250.

Clark, H. H. (1996). Using Language. Cambridge, Cambridge University Press.

Clark, H. H. (2001). Pointing and Placing. Pointing: Where language, culture, and cognition meet. S. K. Mahwah. NJ, Lawrence Erlbaum Associates.

Clark, H. H. and S. E. Brennan (1991). Grounding in communication. Perspectives on socially shared cognition. S. D. Teasley. Washington DC, American Psychological Association: 127-149.

Clark, H. H. and E. F. Schaefer (1989). "Contributing to discourse." Cognitive Science **13**,: 259-294.

Clark, H. H. and D. Wilkes-Gibbs (1986). "Referring as a collaborative process." Cognition **22**: 1-39.

Cohen, P. R. (1984). "The pragmatics of referring, and the modality of communication." Computational Linguistics **10**: 97-146.

Core, M. and J. Allen (1997). Coding Dialogue with the DAMSL Annotation Scheme. AAAI Fall Symposium on Communicative Action in Humans and Machines, Boston, MA.

Daly-Jones, O., A. Monk, et al. (1998). "Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus." International Journal of Human-Computer Studies **49**(1): 21-58.

Dillenbourg, P., D. Traum, et al. (1996). Grounding in Multi-modal Task-Oriented Collaboration. EuroAI&Education Conference.

Duncan, S. (1974). "On the structure of speaker-auditor interaction during speaking turns." Language in Society 3: 161-180.

Exline, R. V. and B. J. Fehr (1982). The assessment of gaze and mutual gaze. Handbook of methods in nonverbal behavior research. P. Ekman. Cambridge, Cambridge University Press: 91-135.

Glass, J. and E. Weinstein (2001). SpeechBuilder: Facilitating Spoken Dialogue System Development. Proc. EuroSpeech, Aarlborg, Denmark.

Goodwin, C. (1981). Achieving Mutual Orientation at Turn Beginning. Conversational Organization: Interaction between speakers and hearers. New York, Academic Press: 55-89.

Goodwin, C. (1986). "Gestures as a resource for the organization of mutual orientation." Semiotica **62**(1/2): 29-49.

Haberman, S. J. (1973). "The analysis of residuals in cross-classified tables." Biometrics **29**: 205-220.

Heeman, P. A. and G. Hirst (1995). "Collaborating on Referring Expressions." Computational Linguistics **21**(3).

Heeman, P. A. and D. Traum (1997). Utterance Units in Spoken Dialogue. Dialogue Processing in Spoken Language Systems. S. LuperFoy. Heidelberg, Springer-Verlag.

Kapoor, A. and R. W. Picard (2001). A Real-Time Head Nod and Shake Detector. Workshop on Perceptive User Interfaces, Orlando FL.

Larsson, S., P. Bohlin, et al. (1999). TrindiKit 1.0 (Manual): http://www.ling.gu.se/projekt/trindi//.

Lester, J. C., S. Towns, et al. (2000). Deictic and Emotive Communication in Animated Pedagogical Agents. Embodied Conversational Agents. E. Churchill. Cambridge, MIT Press: 123-154.

Lewis (1969). Convention: A philosophical study. Cambridge MA, Harvard University Press.

Matheson, C., M. Poesio, et al. (2000). Modelling Grounding and Discourse Obligations Using Update Rules. 1st Annual Meeting of the North American Association for Computational Linguistics (NAACL2000).

McCarthy (1990). Formalization of two puzzles involving knowledge. Formalizing common sense: Papers by John McCarthy. V. Lifschitz. Norwood NJ, Ablex Publishing: 158-166.

McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. Chicago, IL/London, UK, The University of Chicago Press.

Nakatani, C. and D. Traum (1999). Coding discourse structure in dialogue (version 1.0), University of Maryland.

Novick, D. G., B. Hansen, et al. (1996). Coordinating turn-taking with gaze. ICSLP-96, Philadelphia, PA.

Paek, T. and E. Horvitz (1999). Uncertainty, Utility, and Misunderstanding: A Decision-Theoretic Perspective on Grounding in Conversational Systems. Working Papers of the {AAAI} Fall Symposium on Psychological Models of Communication in Collaborative Systems. D. Traum. Menlo Park, California, American Association for Artificial Intelligence: 85-92.

Peek&Horvitz (1999). "Uncertainty, Utility, andMisunderstanding: A Decision-Theoretic Perspective on Grounding in Conversational Systems."

Pierrehumbert, J. B. (1980). The phonology and phonetics of english intonation, Massachusetts Institute of Technology.

Rickel, J. and W. L. Johnson (1999). "Animated Agents for Procedural Traning in Virtual Reality: Perception, Cognition and Motor Control." Applied Artificial Intelligence 13(4-5): 343-382.

Rosenfeld, H. M. and M. Hancks (1980). The Nonverbal Context of Verbal Listener Responses. The Relationship of Verbal and Nonverbal Communication. M. R. Key. New York, Mouton Publishers: 194-206.

Schiffer (1972). Meaning. Oxford University Press.

Seligman, M., J. Hosaka, et al. (1997). Pauseunits and analysis of spontaneous Japanese dialogues: Preliminary studies. Dialogue Processing in Spoken Language Systems. S. LuperFoy. Heidelberg, Springer-Verlag.

Stalnaker (1978). Assertion. Syntax and semantics 9: Pragmatics. P. Cole. New York, Academic Press: 315-332.

Stocky, T. and J. Cassell (2002). Shared Reality: Spatial Intelligence in Intuitive User Interfaces. Intelligent User Interfaces, San Francisco, CA,.

Takagi, K. and S. Itahashi (1996). Segmentation of spoken dialogue by interjection, disfluent utterances and pauses. the 4rd International Conference on Spoken Language Processing

(ICSLP-96), Philadephia.

Thorisson, K. R. (1999). "A Mind Model of Multimodal Communicative Creatures and Humanoids." Applied Artificial Intelligence 13.

Traum, D. and J. Rickel (2002). Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. Autonomous Agents and Multi-Agent Systems.

Traum, D. R. (1994). A Computational Theory of Grounding in Natural Language Conversation. Computer Science. Rochester, NY, University of Rochester.

Traum, D. R. and P. Dillenbourg (1998). Towards a Normative Model of Grounding in Collaboration. ESSLLI-98 workshop on Mutual Knowledge, Common Ground and Public Information.

Walker (1992). Redundancy in collaborative dialogue. Fourteenth International Conference on Computational Linguistics (COLING), Nantes.

Whittaker, S. ((to appear)). Computational communication: a review. The Handbook of Discourse Processes. A. Graesser. Cambridge, MA, MIT Press.