Social Interaction: Multimodal Conversation with Social Agents

Katashi Nagao and Akikazu Takeuchi
Sony Computer Science Laboratory Inc.
3–14–13 Higashi-gotanda, Shinagawa-ku, Tokyo 141, Japan
E-mail: {nagao,takeuchi}@csl.sony.co.jp

Abstract

We present a new approach to human-computer interaction, called *social interaction*. Its main characteristics are summarized by the following three points. First, interactions are realized as multimodal (verbal and nonverbal) conversation using spoken language, facial expressions, and so on. Second, the conversants are a group of humans and *social agents* that are autonomous and social. Autonomy is an important property that allows agents to decide how to act in an ever-changing environment. Socialness is also an important property that allows agents to behave both cooperatively and collaboratively. Generally, conversation is a joint work and ill-structured. Its participants are required to be social as well as autonomous. Third, conversants often encounter communication mismatches (misunderstanding others' intentions and beliefs) and fail to achieve their joint goals. The social agents, therefore, are always concerned with detecting communication mismatches. We realize a social agent that hears human-to-human conversation and informs what is causing the misunderstanding. It can also interact with humans by voice with facial displays and head (and eye) movement.

Introduction

Many artificial intelligence researchers have been seeking to create intelligent autonomous creatures that act like partners rather than tools. They will take the responsibility of doing some social services through interacting with humans. We will depend on the computer that assists us to achieve some tasks and delegate it to the responsibility for working out the details, rather than invoke a series of commands which cause the system to carry out well-defined and predictable operations.

Autonomy is to have or make one's own laws. An autonomous system has the ability to control itself and to make its own decisions. Autonomy is essential to survive in a dynamically changing world such as one we live in. It is the subject of research in many areas including robotics, artificial life, and artificial ecosystems.

However, is autonomy itself sufficient for social services? Although autonomy is vital to survive in the real world, it is only concerned with "self." It is selfish by nature. It seems that it does not work well in human society, since it includes socially constructed artifacts such as laws, customs, culture. Social services provided by computer systems have to incorporate with these artifacts.

Socialness is a higher-level concept defined above the concept of an individual, and is the style of interaction between the individuals in a group. Socialness can be applied to the interaction between humans and computers, and possibly to that between multiple computers. In this paper, we study socialness of conversational interaction between humans and computers. Conversation is no doubt a social activity, especially when more than two participants are involved in it. However, conversation research to date has been biased to problem-solving. Question-answering systems are typical examples. All conversation research based on this view has the following features.

Dialogical: Only two participants, a human (asker) and a computer (answerer), are assumed. Turn-taking is trivial (alternate turns).

Transformational: Computers are regarded as a function that receives an inquiry and produces its answer.

Passive: Computers will not voluntarily speak.

The dialogical and transformational views are well fitted to applications such as natural language interfaces of databases, consulting and guidance systems.

However, our daily conversation is not always functional. One example that is not functional is the co-constructive conversation studied by Chovil (Chovil 1991).

Co-constructive conversation is that a group of individuals, in which, say, people talk about the food they ate in a restraurant a month ago. There are no special roles (like the chair) for the participants to play. They all have the same role. All participants try to

recall the food by relating his or her memory about the food, adding comments, and correcting the other's impression. Turn-taking is controlled by eye contact, facial expression, body gestures, voice tones, and so on. Conversation includes many subconversations, some of them existing in parallel and dividing the group into subgroups. The conversation terminates only when all the participants are satisfied with the conclusion.

Co-constructive conversation closely approximates to our day-to-day conversation. Conversation is a social action. Suchman said that communication is not a symbolic process that happens to go on in real-world settings, but a real-world activity in which we make use of language to delineate the collective relevance of our shared environment (Suchman 1987). To realize a computer that can participate in social conversation such as the co-constructive conversation, described above, is our research goal. To this end, we propose the notion of "Social Interaction" as a new conversation paradigm between humans and computers. In contrast to the problem-solving view, social interaction has the following features.

N-participant conversation: Conversation involves more than two agents that are humans or computers. A computer has to recognize every participant with his/her/its character. There is no fixed roles. Turn-taking is flexible, and highly dependent on the conversational situation.

Social: Every participant is more or less social and follows social rules such as "avoid misunderstandings," "do not speak while other people are speaking," "silence is no good," "contribute whenever possible," etc.

Situated actions: Conversational actions are controlled not only by intelligence and social rules, but also by situations perceived multimodally. Here, we assume that a participant's actions, such as body gestures, eye contact, facial expressions, and coughing, are all included in a situation.

Active: A computer actively joins the conversation, that is, grabs every chance to speak.

We call an autonomous system that can do social interaction with humans a *social agent.* In the following, we study an architecture of a social agent and its behavioral model. This paper is organized as follows. In Section "An Architecture for Social Agents," we present an architecture for a social agent. In Section "Conversation as Situated Action," a situated conversational action based on multimodal cognition is presented. In Section "Conversation as Cooperative Action," we present a model for understanding ill-structured conversation and detecting communication mismatches.

An Architecture for Social Agents

Model

Several agent architectures featuring interaction with a society have been proposed (Cohen & Levesque 1990, Bates, Loyall, & Reilly 1992). Social agents are fully exposed to a real human society, and have to perceive the verbal and nonverbal messages and take actions based on them.

Traditional conversation programs process voice input sequentially, from low-level recognition to high-level semantic analysis. This works well in the domain of transformational question-answering applications. However, conversation such as co-construction requires faster response to other participants' utterances. These reactions are not necessarily deliberate ones executed at a semantic level. Moreover, some reactions may be triggered by nonverbal actions such as eye contact and body gestures. In fact, conversation is supported by multiple coordinated activities at various cognitive levels. This makes communication highly flexible and robust.

Brooks proposed the horizontal decomposition of a mobile robot control system, based on task-achieving behaviors, instead of decomposition based on functional modules (Brooks 1986). His architecure is powerful enough to survive in the real world, as proven by a series of robots he designed. The same argument holds when we design a social agent, since social agents have to be involved in conversations that are *real-world activities going on in real-world settings.* Figure 1 illustrates the horizontal decomposition of a social agent based on task-achieving behaviors. It is important to note that the layers act on sensory data in parallel. There is downward control and upward dataflows.

There has been much debate between those groups that support situated actions and those that support physical symbol systems (Vera & Simon 1993). In our achitecture, these views are placed at opposite ends. Namely, the lower layers rule reactions to multiple sensory input data, while the upper layers administer deliberate actions. The next section explains the lower levels. The section following the next explains the higher levels.

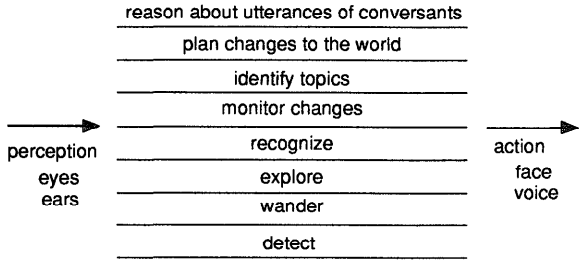

Figure 1: Horizontal decomposition of a social agent

Current Implementation

In the current implementation, a social agent has a face, a voice, eyes, and ears. They are realized by two subsystems, a facial animation subsystem that generates a three-dimensional face capable of various facial displays, and a spoken language subsystem that recognizes and interprets speech, and generates voice outputs. Currently, the animation subsystem is running on SGI 320VGX and the spoken language subsystem on a Sony NEWS workstation. These two subsystems communicate with each other via an Ethernet network.

The face is modeled three-dimensionally. The current face is composed of approximately 500 polygons. The face is rendered using a texture taken from a photograph or a video frame. A facial display is realized by local deformation of the polygons representing the face. We use the numerical equations simulating muscle actions defined by Waters (Waters 1987). Currently, 16 muscles and 10 parameters, controlling mouth opening, jaw rotation, eye movement, eyelid opening, and head orientation are incorporated. These 16 muscles were determined by Waters, considering the correspondence with action units in the Facial Action Coding System (FACS) (Ekman & Friesen 1978). The facial modeling and animation system are based on the work of Takeuchi and Franks (Takeuchi & Franks 1992).

Speaker-independent continuous speech inputs are accepted without special hardware. To obtain a high level of accuracy, context-dependent phonetic hidden Markov models are used to construct phoneme-level hypotheses (Itou, Hayamizu, & Tanaka 1992). The speech recognizer outputs N-best word-level hypotheses. The semantic analyzer deals with ambiguities in syntactic structures and generates a semantic representation of the utterance. We applied a preferential constraint satisfaction technique for disambiguation and semantic analysis (Nagao 1992). The plan recognition module determines the speaker's intention by constructing his belief model and dynamically adjusting and expanding the model as the conversation progresses (Nagao 1993). The response generation module generates a response by using domain knowledge and text templates (typical utterance patterns).

The spoken language subsystem recognizes a number of typical conversational situations that are important in communication. We associate these situations with specific communicative facial displays. The correspondence between conversational situations and facial displays is based on the work of Takeuchi and Nagao (Takeuchi & Nagao 1993). For example, in situations where speech input is not recognized or where it is syntactically invalid, the facial display of "Not confident" is displayed. If the speaker's request is out of the system's knowledge, then the system displays a facial shrug and replies "I cannot manage it without knowing it."

Gaze control is also implemented in the facial animation subsystem using a video camera fixed on top of a computer display. Comparing coming images with the image of the vacant room and segmenting differentiated regions, moving objects are extracted in real-time. Assuming that moving objects are only humans in the room, we can find the 2D position of human participants in the image. Using camera position and direction, the position is translated to 3D orientation, which is applied to eyeball rotation and face rotation when drawing a 3D face.

Figure 2 shows a snapshot of conversation between humans and a social agent.

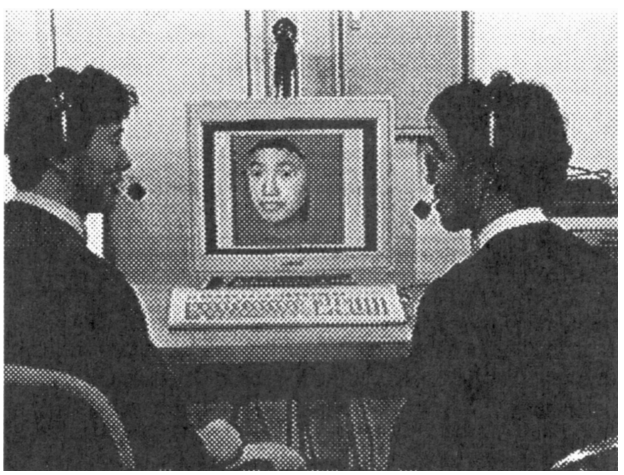

Figure 2: Conversation with a social agent

Conversation as Situated Action

In conversation, people show various complicated behavior. Some are expressed by face, others by body motions. Situated action views this complexity as a result of interaction with an environment including conversants, not as a result of internal complexity. This complexity can be regarded as fruitfulness of communication.

This fruitfulness is related to fruitfulness of sensory data, and is realized by the lower layers in Figure 1. However, when implementing it, there is a severe trade-off between the processing speed and the content of data-processing. The more information, the slower the processing. Since slow reactions are essentially useless in a real-world setting, we have to force ourselves to reduce time-consuming processing in these layers.

In the current implementation, the following processings are considered for image and auditory data in those layers: image scene analysis; face position detection; face identification; facial display recognition; auditory scene understanding; voice position detection; voice identification; speech recognition. These underlined items were already implemented.

The lower four layers of the decomposition in Figure 1, *detect, wander, explore,* and *recognize*, are major

players in a situated conversational action game. *Detect* is to detect an input in any of the sensory channels. An agent may perform quick reactions such as looking in the direction from which the input appeared. Generally, such quick reactions are short-lived. *Wander* is a tendency of distraction. It may distract agent's attention from *detected* input. *Explore* is an opposite action to *wander*, and it looks for something attractive. It may suppress *wander* for a while. *Recognition* is to recognize sensory data that seem to be worth paying attention. It tries to extract its meaning, although its full understanding is left to the higher layers.

From interaction between these layers, various conversational actions emerge. For instance, a composite action of "*detect* then *wander*" appears the sign of ignorance or no interest; "*Explore* then *wander*" appears the sign of refusal; "*Explore* then *recognize*" appears the sign of attendance and interest. Sophisticated social actions like turn-taking are highly depending upon mutual perception of these kinds of signals, so they would be constructed naturally on these situated conversational acitons.

Conversation as Cooperative Action

Language use is an action that influences the human mind. Speech act theory formalized this idea (Searle 1969). Actions are assumed to be performed after planning their effects on the world. So, language use is also assumed to be performed by planning its effect on the mind. The difference between planning for physical action and language use is that planning for language use includes the hearers' beliefs and plans (i.e., plan recognition). Conventional plan recognition models deal only with one hearer's beliefs and plans. Research into two-participant conversation (i.e., dialogue) concentrates on task-oriented, well-structured dialogues that make use of a consistent plan library and a well-ordered turn-taking constraint (Carberry 1990). Group conversation is generally ill-structured. This means that some interruptions by other speakers and *communication mismatches* between conversants occur frequently.

We extended the conventional plan recognition models to deal with communication mismatches by maintaining multiagent's conversational states. Multiagent conversational states consist of agents' beliefs, utterances, illocutionary act types, other communicative signals, turn-taking sequences, and activated plans. Illocutionary act types such as INFORM, REQUEST, etc. are an abstraction of the speaker's intentions in terms of the actions intended by the speaker. Other communicative signals contain facial displays, head and eye movement, and gestures, as described in the previous section. Activated plans are represented as a network that connects preconditions and the effects of plans in an agent's belief space.

Figure 3 illustrates the concept of the multiagent conversational state.

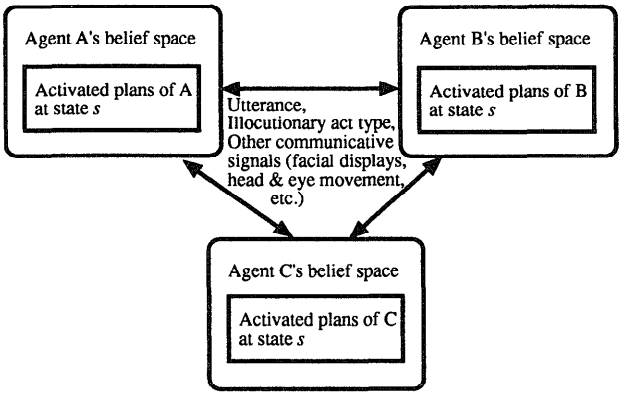

Figure 3: Conceptual model of multiagent conversational state

We will explain our mechanism by using the following discourse fragment. A and B are humans, while C is an agent that overhears their conversation.

A: "Do you know what happened today?"
B: "I don't know exactly, but ..."
C (to B): "I think that he wants to tell you."

C's plan recognition process upon hearing A's utterance "Do you know what happened today?" is traced in Figure 4.

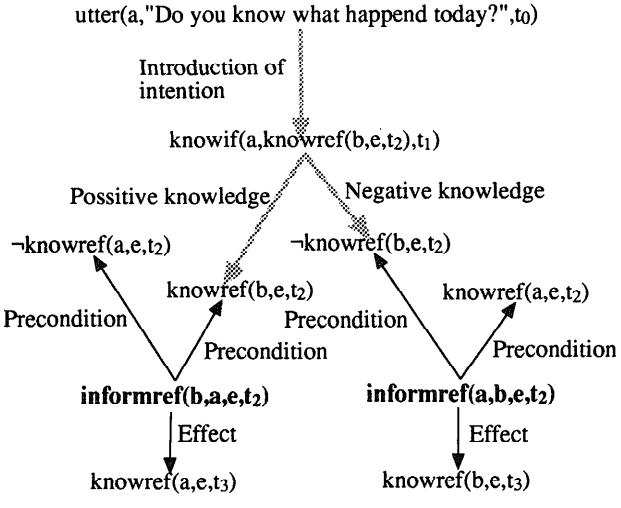

Figure 4: Inferred plan recognition on utterance "Do you know what happened today?"

In this figure, the inference proceeds from the top to the bottom. The directions of the arrows (except the thick ones) indicate logical implication. Thus, the downward arrows correspond to deduction, while the upward ones correspond to abduction. knowif(X,P,T) means that agent X knows that proposition P holds at time T. knowref(X,P,T) means that agent X knows the

contents of proposition P at time T. **informref(X,Y,P,T)** means that X informs other agent Y of the contents of P at time T. **a**, **b**, and **e** denote agents A, B, and the event that 'happened today', respectively.

These plan schemes are assumed to be common to all the conversants, because they are domain-independent and common sense. Domain-dependent plans may differ between conversants, since they may have different experiences on domain-dependent behavior.

Figure 4 shows that, from A's utterance "Do you know what happened today?", the conversants can infer that A's intention is **knowref(a,e,t3)** when A believes **knowref(b,e,t2)** or that A's intention is **knowref(b,e,t3)** when A believes **knowref(a,e,t2)**. After A's utterance, B believes that A does not know **e**. So, he infers that A's intention is **knowref(a,e,t3)**. However, C infers that A knows **e** from another information source.

After the utterance of B, agent C utters utterance "I think that he wants to tell you" because there is a mismatch between beliefs of A and B (in C's belief space), and C infers that it could be an obstacle to the progress of the conversation.

To detect communication mismatches in group conversation, social agents consider assumptions based on other agents' utterance planning. A communication mismatch is judged to have occurred when an agent recognizes the following situations.

1. Illocutionary act mismatch

   An example would be the situation where an agent utters an utterance of illocutionary act type QUESTIONREF (i.e., the agent wants to know about something) and, after that, another agent utters an utterance of type INFORMIF (i.e., yes/no answer). e.g., An agent asked "Do you know what happened today?" with the intention of knowing about the event 'happened today' and the other agent's answer was "Yes, I do."

2. Belief inconsistency

   An agent misunderstands another agent's beliefs. e.g., An agent asked "Do you know what happened today?" with the intention of knowing about the event 'happened today' but the agent misunderstood that the other agent already knew about the event.

3. Plan inconsistency

   An agent misunderstands another agent's intended plans. e.g., An agent asked "Do you know what happened today?" with the intention of knowing about the event 'happened today' and the other agent misunderstood that the agent had a plan to inform about the event.

In general, it is not necessary to exactly determine the conversants' intentions, because it is sufficient to know whether there is a communication mismatch. Decisions on unnecessary occasions should be delayed until required (van Beek & Cohen 1991).

When social agents detect a communication mismatch, they inform the other conversants by saying a few phrases from which they can easily determine their misunderstanding. These utterances are called *minimal utterances*. Minimal utterances of social agents are caused by constraints imposed on their resource-boundedness and socialness. These utterances function to limit the processing for generation, required by resource-bounded agents. They also contribute to avoiding further progress of the conversation without first resolving the misunderstanding. In conversation, timely action is crucial, since delays have some meaning in themselves. We consider the minimal utterance as a situated action. As mentioned before, situated actions in conversation involve multimodality. So, in this case, an agent's response includes facial actions and prosodic actions in voice tones.

Example Conversation

Now, let's look at an example of a conversation between humans and a social agent. The humans (A and B) are talking about cooking, and the social agent (C) overhears their talk. This example is based on Kautz's cooking plan library (Kautz 1990).

A: "I made marinara sauce. What brand of wine do you like?"
B: "Marinara ... Ok. Italian 'Soave' is good."
A (with a perplexed look): "...."
C (to A): "I think that he is thinking of a pasta dish."
A (to B): "Oh. I am making chicken marinara."

Figure 5 shows part of the cooking plan library used to understand the example conversation. In this figure, the upward-pointing thick arrows correspond to is-a (a-kind-of) relationships, while downward-pointing thin arrows indicate has-a (part-of) relationships. We assume that C uses this plan library from the initial state.

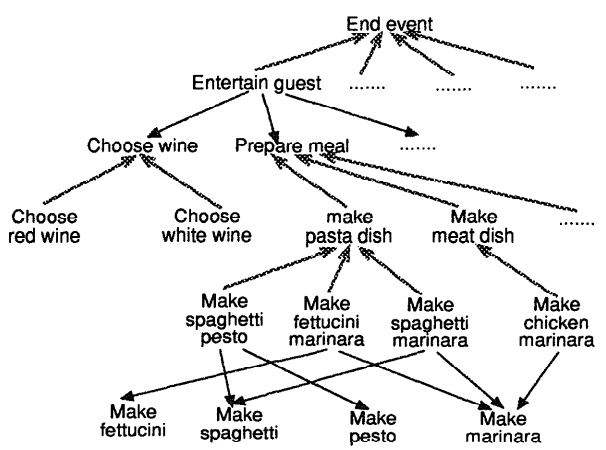

Figure 5: Cooking plan library (Kautz 1990)

After C hears A's first utterance, C infers A's activated plans at that conversational state, as shown in

Figure 6[1].

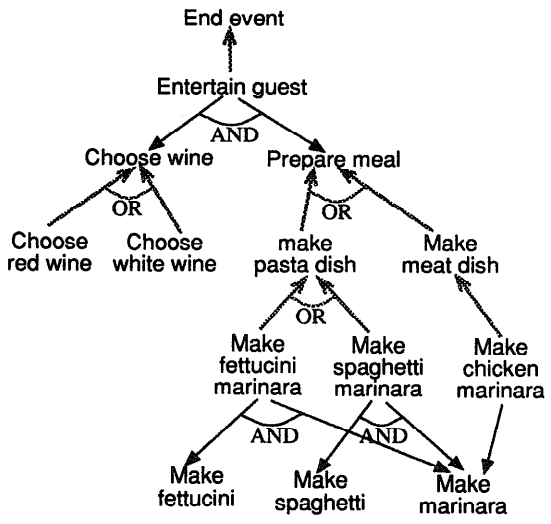

Figure 6: A's activated plan

After hearing B's utterance, C infers B's activated plans (B's recognized plans about A) at that conversational state, as shown in Figure 7. This inference

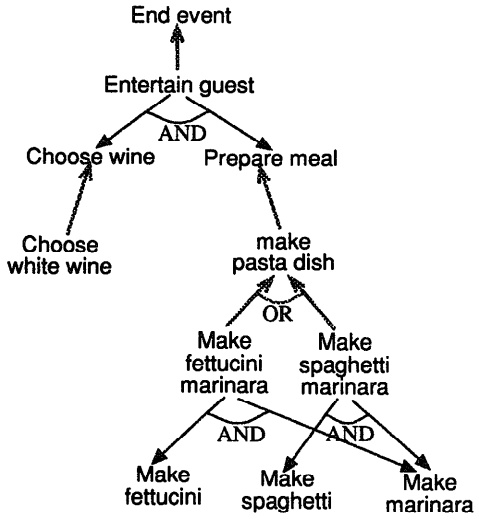

Figure 7: B's recognized plan about A

involves knowledge about the relationships between wines and dishes (i.e., white wines are well-suited to pasta dishes). When C sees A's perplexed facial display, C makes an assumption that A's intention is not to make a pasta dish (but is to make a meat dish) and C detects that there is a communication mismatch (i.e., plan inconsistency) between A and B[2].

Then, C is motivated to inform A (or B) that a misunderstanding has occurred. In this case, C infers that it is better to inform A than to B, since A is now perplexed, and may need help. As a result, C tells A about (C's inferred) B's recognized plan about A.

This example shows that social agents can detect communication mismatches by maintaining multiagent conversational states and can voluntarily take part in conversation to smooth out any misunderstandings when agents detect obstacles to communication. A detailed mechanism of social agents' cooperative (minimal) utterance generation will be presented in a separate publication (Nagao 1994).

Another example of cooperative conversation is that a social agent says some additional information related to the topic in conversation. For example, a guest talks with a desk clerk about a French restaurant, and the clerk tells the best one he knows, then the overhearing agent accesses the agent of that restaurant, and tells the guest that it is fully-booked today.

Concluding Remarks and Further Work

We presented an approach to social interaction, a multimodal conversation with social agents. Socialness is an essential property of intelligent agents as well as autonomy. Social agents consider other agents' (including humans') beliefs and intentions, and behave cooperatively. One example of cooperation is the removal of obstacles to communication caused by misunderstanding between agents. Our model can deal with N-conversant plan recognition and detect communication mismatches between conversants. These mismatches consist of illocutionary act mismatches, belief inconsistency that is a misunderstanding of the beliefs held by other conversants, and plan inconsistency that is a misunderstanding of the intended domain plans of others. An ideal multimodal interaction is modeled by human face-to-face conversation in which speech, facial displays, head and eye movement, etc. are utilized. Our system integrates these modalities for interacting with social agents.

In the future, we plan to simulate human-to-human communication in more complex social environments. We need to design several social relationships between agents and implement social stereotypes (e.g., social standing, reputation, etc.) and personal properties (e.g., disposition, values, etc.). These social/personal properties can dynamically change according to conversational contexts. From these studies, we can propose some design principles for a society of agents.

Of course, future work needs to be done on design and implementation of coordination of multiple com-

[1]For the sake of simplicity, we omit the speech act predicates such as knowif, knowref, informref, etc. described in the previous section.

[2]In the current implementation, the agent cannot recognize/understand human facial displays. So in this case, C detects a communication mismatch because of a break in conversation.

munication modalities. We think that such coordination is an emergent phenomenon from tight interactions with environments (including humans and other agents) by means of situated actions and (more deliberate) cooperative actions. Precise control for multiple coordinated activities, therefore, is not directly implementable. Only constraints or relations among perception, conversational states, and action will be implementable. At the beginning, we are developing a constraint-based computational architecture and applying it to tightly-coupled spoken language comprehension (Nagao, Hasida, & Miyata 1993).

Co-constructive conversation that is less constrained by domains or tasks is one of our future targets to be carried out. We are also interested in developing interactive characters and stories as an application for interactive entertainment. Bates and his colleagues called such interactive systems "believable agents" (Bates *et al.* 1994). We are trying to build a conversational, anthropomorphic computer character that will entertain us with some pleasant stories.

Acknowledgments

The authors would like to thank Mario Tokoro and colleagues at Sony CSL for their encouragement and discussion, anonymous reviewers for their valuable comments on a draft of this paper, and Toru Ohira for his helpful advice on the wording. We also extend our thanks to Satoru Hayamizu, Katunobu Itou, Taketo Naito, and Steve Franks for their contributions to the implementation of the prototype system. Special thanks go to Keith Waters for granting permission to access his original animation system.

References

Bates, J.; Hayes-Roth, B.; Nilsson, N.; and Laurel, B., eds. 1994. *AAAI 1994 Spring Symposium on Believable Agents.* American Association for Artificial Intelligence.

Bates, J.; Loyall, A. R.; and Reilly, W. S. 1992. An architecture for action, emotion, and social behavior. In *Proceedings of the Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agent World (MAAMAW'92).* Institute of Psychology of the Italian National Research Council.

Brooks, R. A. 1986. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation* 2(1):14–23.

Carberry, S. 1990. *Plan Recognition in Natural Language Dialogue.* The MIT Press.

Chovil, N. 1991. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction* 25:163–194.

Cohen, P. R., and Levesque, H. J. 1990. Rational interaction as the basis for communication. In Cohen, P. R.; Morgan, J.; and Pollack, M. E., eds., *Intentions in Communication.* The MIT Press. 221–255.

Ekman, P., and Friesen, W. V. 1978. *Facial Action Coding System.* Palo Alto, California: Consulting Psychologists Press.

Itou, K.; Hayamizu, S.; and Tanaka, H. 1992. Continuous speech recognition by context-dependent phonetic HMM and an efficient algorithm for finding N-best sentence hypotheses. In *Proceedings of ICASSP-92*, I.21–I.24. IEEE.

Kautz, H. 1990. A circumscriptive theory of plan recognition. In Cohen, P. R.; Morgan, J.; and Pollack, M. E., eds., *Intentions in Communication.* The MIT Press. 105–133.

Nagao, K.; Hasida, K.; and Miyata, T. 1993. Understanding spoken natural language with omnidirectional information flow. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1268–1274. Morgan Kaufmann Publishers, Inc.

Nagao, K. 1992. A preferential constraint satisfaction technique for natural language analysis. In *Proceedings of the Tenth European Conference on Artificial Intelligence (ECAI-92)*, 523–527. John Wiley & Sons.

Nagao, K. 1993. Abduction and dynamic preference in plan-based dialogue understanding. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1186–1192. Morgan Kaufmann Publishers, Inc.

Nagao, K. 1994. Minimal utterances of resource-bounded social agents. Technical Report Forthcoming, Sony Computer Science Laboratory Inc., Tokyo, Japan.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press.

Suchman, L. 1987. *Plans and Situated Actions.* Cambridge University Press.

Takeuchi, A., and Franks, S. 1992. A rapid face construction lab. Technical Report SCSL-TR-92-010, Sony Computer Science Laboratory Inc., Tokyo, Japan.

Takeuchi, A., and Nagao, K. 1993. Communicative facial displays as a new conversational modality. In *Proceedings of ACM/IFIP INTERCHI'93: Conference on Human Factors in Computing Systems*, 187–193. ACM Press.

van Beek, P., and Cohen, R. 1991. Resolving plan ambiguity for cooperative response generation. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, 938–944. Morgan Kaufmann Publishers, Inc.

Vera, A. H., and Simon, H. A. 1993. Situated action: A symbolic interpretation. *Cognitive Science* 17(1):7–48.

Waters, K. 1987. A muscle model for animating three-dimensional facial expression. *Computer Graphics* 21(4):17–24.