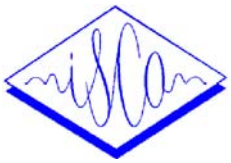

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

Chatting with Interactive Agent

N.Suzuki, S.Inokuchi, K.Ishii and M.Okada
ATR Media Integration & Communications Research Laboratories
Hikaridai, Seika-cho, Kyoto-fu, 619-02 Japan
Tel. +81 774 95 1401, FAX: +81 774 95 1408, E-mail: noriko@mic.atr.co.jp

Abstract

Conventional spoken dialogue systems are based on goal-oriented techniques[8]. The recent expansion of application fields such as cyber space, internet, etc, necessitates the creation of new interaction styles between humans and autonomous agents. Interaction with autonomous agents creates new possibilities for spontaneous conversation in spoken dialogue systems. Within this context, we regard spontaneous, informal chatting behavior as one aspect of spoken dialogue[4][5]. According to this view, an essential property of chatting is the emergence of topics and goals situated within the context of interactions among participants rather than as the result of explicit goals. In this paper, we propose a spoken dialogue system with chatting properties and illustrate sample chatting between a human and a virtual interface agent called Talking Eye using a prototype system.

1 Introduction

"How about it?", "What's that?", "It's your new book.", "Cool!", "That sounds good!"... We spend lots of time on such casual and informal chatting in everyday life. However, do you enjoy chatting like this with your computer now?

Conventional spoken dialogue systems are goal-oriented and pursuing mostly efficiency and accuracy of information transmission. These systems have been designed without considering the exploitation of human-like qualities such as friendliness, spontaneity and flexibility in interactions. Therefore, we have constructed an alternative spoken dialogue system based on a theory of everyday activities[7]. In particular, we focus on chatting as a form of spontaneous dialogue, because the informality and casualness of chatting will enhance the effectiveness of human-computer interaction[4][5]. The essential properties of chatting are the emergence of topics and goals situated within the context of interactions among participants, without a fixed scenario, and the freedom of interpretation of partner's conversational behaviors. Within this context, we study the mechanisms of emergent chatting behaviors between a human and communicative agents on a computer.

From this viewpoint, we propose a communicative agent model aimed at the creation of chatting behavior with humans. The model consists of multiple layers of competence modules based on a subsumption architecture [1].

To demonstrate the effectiveness of our proposed model, we have created an interface agent named "Talking Eye" based on our communicative agent model. Talking Eye has the shape of an eyeball generated by 3-D computer graphics. It can perceive human conversational behaviors by detecting simple phrases[6], prosody of speech and simple motions. Furthermore, it can produce about 250 vocal phrases for chatting using a speech synthesis system[2]. These include, "That's too bad", "How about it?", "That sounds good!", etc. It can also produce about 20 types of eye movements which indicate emotional states such as surprise, agreement, disappointment, etc. Because of these Talking Eye specifications, we can enjoy casual, relaxed chats with Talking Eye in real-time.

First, we will point out some problems that most conventional interactive systems share and we will introduce a novel conception of chatting as emergent phenomena. Second, the communicative agent model is explained. We then give an overview of an interactive agent called Talking Eye that implements the communicative agent model and sample conversation among Talking Eyes and a human. Finally, we discuss the emergence of chatting behavior.

2 Chatting as Emergent Phenomena

Conventional spoken dialogue systems are goal-oriented and emphasize the efficiency and accuracy of information transmission, a typical example being ATIS[8]. These systems force the use of formal, emotion-less interactive style because the systems do not exploit human-like qualities such as friendliness, spontaneity and flexibility in interactions. Therefore, we sometimes feel unnatural and tight for these system.

In contrast, we have considered an alternative spoken dialogue system based on a theory of everyday activities[7]. In particular, we focus on chatting as a form of spontaneous dialogue. The following two points are made concerning the properties of chatting. One is the emergence of topics and the meaning of conversational behavior situated within the context of interactions among participants, without a fixed scenario, rather than having to be assigned an explicit goal

and meaning from the beginning. Another is that a human can correspond adaptively using various ways of interpretation toward a participant's conversational behavior according to the process of chat and emotional state of a human. These properties of chatting produce human-like qualities through which we can perceive another's personality or feeling by the exchange of simple phrases and create a shared sense of empathy.

We think that informality and casualness in chatting will enhance the effectiveness of human-computer interaction[4][5]. Within this context, we try to construct a communication model with the properties of chatting on a computer.

3 Overview of the Model

3.1 Communicative Agent Model

We propose a communicative agent model with a subsumptive structure which specifically aims at the creation of chatting behavior with humans. The model consists of multiple layers based on a subsumption architecture with competence modules[1]. Each competence module independently performs an interaction with the real world and a behavior is expressed toward the real world as a result of activation/inhibition among competence modules. The higher competence module with an intentional behavior subsumes the lower one with a coordinative behavior. The function of the lower competence module is to maintain basic coordination by invoking conversational behaviors constrained by interactions with the real world. The function of the higher competence modules is to coordinate interactions based on a cooperative mechanism for the achievement of goals which is emerged as a result of interactions.

Each competence module consists of a set of behavior modules realized as a situated agent. The conversational behavior is performed by activation/inhibition between a set of situated agents and the real world, as well as, among the situated agents. Each situated agent is activated using spreading activation of a dynamic action selection network[3]. Fig. 1 illustrates the communicative agent model consisting of three parts: the environmental context, the intentional context and a set of situated agents. The environmental context provides constraints on the actions for the situated agents and the intentional context provides constraints on the mental state and motivations for the situated agents. The set of situated agents for the behaviors from all competence modules is represented in Fig. 1 as a single action selection network.

Each situated agent has a condition list, an add list, a delete list and an activation level. The condition list is a list of preconditions which have to be fulfilled before the situated agent can become active. The add list and delete list represent the expected effects of the situated agent's action. In addition, the activation level is the energy value propagated through the network.

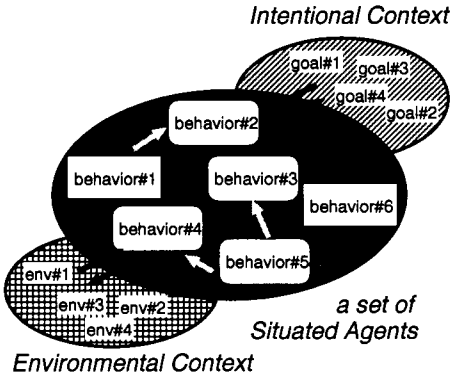

Fig. 1: Architecture of communicative agent model

In the process of conversation, the two contexts and a set of situated agents are influenced by each other. As the result of such an interaction, one action is released dynamically using activation/inhibition dynamics which make activation energy accumulate.

3.2 Example of Behavior Modules

In this paper, we define about 100 behavior modules to perceive the conversational behaviors such as utterance fragment and motion. These modules are classified into the following three types.

Prosodic pattern detection: This focuses on the tail part (last 100 msec) of utterance fragment to perceive the tones of pitch and loudness using f0 and a power pattern. We consider that a human utilizes it adjusting the response to the partner's tone.

Utterance fragment detection: This focuses on the whole part of the utterance fragment to perceive what the partner says. We consider that a human utilizes it to return smart comments.

Motion detection: It focuses on the simple motion to perceive where a human is. We consider that a human utilizes it to decide where to look.

Though these behavior modules consist of a set of simple rules, we do not expect what kind of situated agent is activated in advance, but we do know it after the interaction between a human and the communicative agent.

3.3 Behavior Selection

Details of the mathematical computations used for the spreading activation process are given below.

Perception of events from the real world: After a partner's response, detected fragments and motions become events from the real world which are added to the environmental context.

Support from context: The behavior which suits the current environmental and intentional contexts is

given activation energy from those contexts. When one member existing on the environmental context matches a member on the add list of a situated agent B_k, the environmental context inputs activation energy $e_E(B_k)$ to B_k.

$$e_E(B_k) = \frac{N_{ke}}{N_{ka}} \cdot \delta \quad (1)$$

Moreover, when one member existing on the intentional context matches a member on the add list of a situated agent module B_k, the intentional context inputs activation energy $e_I(B_k)$ to B_k.

$$e_I(B_k) = \frac{N_{ki}}{N_{ka}} \cdot \phi \quad (2)$$

Where, δ and ϕ are the constants for activation energy provided from the two context, N_{ke} and N_{ki} are the number of environmental and intentional context members which match members on the add list of B_k, and N_{ka} is the total length of the add list of B_k.

Activation/inhibition among behaviors: The relationship between one situated agent and others is decided dynamically, as the result of the calculations based on its three internal lists and the current state of the two contexts. If the module B_k has the amount of energy E_{B_k}, direction of activation energy is calculated by three kinds of relationship:

(a) backward: The module B_k is provided with activation energy $e_{(B_{q1} \to B_k)}$ from another module B_{q1} that includes a member on the add list of B_k in the member of the condition list of B_{q1}.

$$e_{(B_{q1} \to B_k)} = \frac{N_{ka}}{N_{q1c}} \cdot \alpha_1 \cdot E_{B_k} \quad (3)$$

Where, N_{ka} is the number of members on the add list of B_k and N_{q1c} is the number of members on the condition list of B_{q1}.

(b) forward: The behavior B_k is provided with activation energy $e_{(B_{q2} \to B_k)}$ from another active module B_{q2} that includes a member on the condition list of B_k in the member of the add list of B_{q2}.

$$e_{(B_{q2} \to B_k)} = \frac{N_{kc}}{N_{q2a}} \cdot \alpha_2 \cdot E_{B_k} \quad (4)$$

Where, N_{kc} is the number of members on the condition list of B_k which matches members in the environmental context, and N_{q2a} is the number of members on the add list of B_{q2}.

(c) take away: The module B_k inhibits activation energy $e_{(B_{q3} \to B_k)}$ from another modules B_{q3} that includes a member on the condition list of B_k in the member of the delete list of B_{q3}.

$$e_{(B_{q3} \to B_k)} = \frac{N_{kc}}{N_{q3d}} \cdot \alpha_3 \cdot E_{B_k} \quad (5)$$

Where, N_{kc} is the number of members on the condition list of B_k, which matches members in the environmental context, and N_{q3d} is the number of members on the delete list of B_{q3}.

The constants, ϕ, δ, α and activation level, also determine the amount of activation for the modules to spread forward, backward, or to be taken away. Consequently, they allow trade-offs between goal-orientedness and data-orientedness (varying parameters ϕ and δ), speed and quality (varying parameter activation level) and variations in sensitivity to goal conflicts (varying parameter α).

Activation of Situated Agent: The amount of activation energy of every situated agent changes as a result of the emergent calculation between a set of situated agent and the contexts as well as among situated agents. Therefore, the active situated agent becomes a candidate to release an action into the real world, if the amount of its activation energy is over a given threshold. The situated agent which has the most activation energy among candidates is released toward the real world as the action. After release, the situated agent module executes the action and changes the environmental context according to the add list and the delete list. Then, the activation energy of the situated agent is removed and threshold is increased.

4 Talking Eye

4.1 Overview of Interactive Agent Talking Eye

We have constructed an interactive agent as one testbed of emergent chatting within the communicative agent model as a result of interactions with a human. The interactive agent has the shape of an eyeball generated by 3-D computer graphics and we have called it "Talking Eye" (Fig. 2). Since a Talking Eye can be represented in such a simple way, we can study a representation of simple modality that the Talking Eye has. At present, it has two modalities, which are utterance generation and motion as actions applied to the real world. It can perceive human conversational behaviors by detecting simple utterance fragments using ATR SPREC[6], prosody of speech and simple motions. Furthermore, it can produce about 250 Japanese vocal phrases for chatting using a speech synthesis system CHATR[2]; these include, "That's too bad", "How about it?", "That sounds good!". It can also produce about 20 types of eye movements which indicate emotional states

such as surprise, agreement, disappointment, etc. With the ATR SPREC system, the Talking Eye can detect one phrase in real-time (less than one second) and at a high phrase detection rate (over 90%). Though the recognition rate is a big problem on most conventional interactive systems, it is not a very big problem for the Talking Eye. Because it interprets its partner's utterance with its own advantage. Therefore, chatting never stop and the Talking Eye can achieve active, non-stop chatting with the communication partner.

Fig. 2: Appearance of chatting with interactive agents "Talking Eyes"

4.2 Sample Conversation

To realize chatting between the Talking Eye and a human, we set up the environmental context, a set of situated agent and the intentional context as follows: the environmental context was set up with some events as a result of perception of conversational behavior; prosodic pattern detection, utterance fragment detection and simple motion detection. Each situated agent was set up with some local constraint of the possibility of action/perception against the real world. The intentional context was set up for mental state and motivation, for example curious, optimistic, mischievous and so on. Table 1 shows a sample conversation among a human (H) and three Talking Eyes (T1–3) using Japanese of Kansai dialect when we provided such an environmental context, a set of situated agents and the intentional context to each Talking Eye.

Although the current implementation uses only simple rules for behaviors and has no model for dialog structure, Talking Eyes exhibit behaviors that appear to be based on the understanding of what a human says. They do not only return a response using the lower level of the communicative agent model, but also comments using the higher one. Consequently, we can confirm to emerge of chatting behavior. Simple rules made complex behavior. Also Fig. 2 shows the appearance of conversation between them.

Table 1: Sample conversation

H:	Anonaa.	(*Well,*)
T1:	Nani nani?	(*Yeah?*)
H:	Kinou no yorunaa.	(*last night..*)
T3:	Hung hung.	(*Hum, hum.*)
H:	Inu ni outen.	(*I met a dog.*)
T1:	Huung.	(*What?*)
T2:	Dongnang?	(*What kind of it?*)
H:	Koinu.	(*It's a puppy dog.*)
T3:	Heee.	(*Really?*)
T1:	Songde?	(*And then?*)
H:	Sorega meqcha chiisakatten.	(*It was small!*)
T1:	Souyaronaa.	(*I guess so.*)
T2:	Sorya, souya!	(*It must have been!*)
H:	Maa eeyang	(*I see.*)

5 Conclusion

In this paper, we proposed a behavioral system to generate spontaneous chatting between a human and computer and introduced an interactive agent called Talking Eye. In further works, we will examine in detail at the chatting behaviors focusing on emergent rhythmic patterns.

Acknowledgment

We would like to thank President Ryohei Nakatsu, Dr. Yasuhiro Katagiri and all the members of dept. 4 for their continuous support in this work.

References

[1] Brooks, R. A.: "A robust layered control systems for a mobile robot", IEEE Journal of Robotics and Automation, Vol. RA-2, No. 2, pp. 14 – 23 (Mar. 1986)

[2] Campbell, N: "CHATR: A high-definition speech re-sequencing system", Acoustical Society of America and Acoustical Society of Japan Third Joint Meeting, pp. 1223 – 1228 (1996)

[3] Maes, P: "How to do the right thing", Connection Science, Vol. 1, No. 3, pp. 291 – 323 (1989)

[4] Okada, et al.: "Incremental Elaborations in Generating and Interpreting Spontaneous Speech", Proc. of ICSLP94, pp. 103 – 106 (1994)

[5] Okada, M.: *Hesitating Computer (in Japanese)*, Info. Frontier Series, Kyoritsu Shuppan (1995)

[6] Shimizu, et al.: "Spontaneous dialogue speech recognition using cross-word context constrained word graphs", ICASSP'96, pp. 145 – 148 (1996)

[7] Suchman, L. A.: "Plans and situated actions", Cambridge University Press (1987)

[8] Zue, et al.: "PEGASUS: a spoken language interface for on-line air travel planning", ARPA Workshop on Human Language Technology, pp. 196 – 201 (1994)