Multimodal Agent Interface based on Dynamical Dialogue Model

–MAICO:Multimodal Agent Interface for COmmunication–

Toshiro Mukai, Susumu Seki, Masayuki Nakazawa, Keiko Watanuki, Hideo Miyoshi
Multimodal Functions Sharp Laboratory, Real World Computing Partnership
in System Technology Development Center, SHARP Corporation
1–9–2 Nakase
Mihama-ku, Chiba-shi,Chiba 261–8520 Japan
+81–43–299–8712
{mukai, seki, nakazawa, watanuki, miyoshi}@iml.mkhar.sharp.co.jp

ABSTRACT

In this paper, we describe a multimodal interface prototype system based on Dynamical Dialogue Model. This system not only integrates information of speech and gestures, but also controls the response timing in order to realize a smooth interaction between user and computer. Our approach consists of human-human dialogue analysis, and computational modeling of dialogue.

KEYWORDS: Multimodal Interface, Dynamics, Dynamical Dialogue Model, Nonverbal Information, Back Channels, Response Timing

Introduction

It is well known that people smoothly communicate with each other by exchanging various information, especially nonverbal information such as speech intonation, rhythm, timing and gesture[3],[4],[6]. We believe that introducing these kinds of information exchanges makes human interface more natural and attractive. The goal of our research is to develop a multimodal human interface through which human can communicate with computer to realize a natural communication environment.

To achieve the goal, it is necessary to analyze how humans exchange mutual information [3],[4], and to construct a dialogue model suitable for exchanging multimodal information [1],[2],[5]. In this paper, we focus on back channels or listener's feedback (*aizuchi* in Japanese) as a smoother of dialogue. We describe a model for generating *aizuchi* based on DDM (Dynamical Dialogue Model)[1] and a prototype dialogue system MAICO (Multimodal Agent Interface for COmmunication; Figure 1).

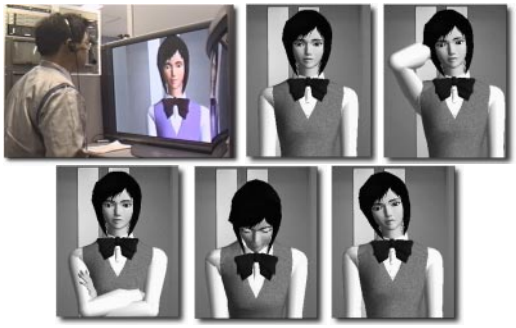


Figure 1: Dialogue scene and MAICO's actions (normal, at a loss, confusing, bow, question)

Dynamical Dialogue Model

In order to realize a smooth dialogue between human and computer, we believe that computer should have an ability not only to passively respond to the user's input but to make active interaction based on dialogue contexts. While nonverbal information plays a very important role in dialogue, for traditional dialogue models based on verbal information, it is difficult to handle nonverbal information effectively. Nonverbal information is a better represented in the form of time sequential patterns which are not discrete symbols.

Under this consideration, we have proposed DDM[1],[2] for handling nonverbal information. It is based on the dynamical system in physics, in which current status decides how those status will change in the future, and is described in differential equations. The model described with the equation (1),(2) enables the computer to insert *aizuchi* of its own rhythm as well as to go along with user's utterances.

$$\frac{dx}{dt} = q \tag{1}$$

$$\frac{dq}{dt} = f_{int}(x) + f_{ext}(y) \tag{2}$$

$$f_{int}(x) = -\omega^2 x \tag{3}$$

$$f_{ext}(y) = \begin{cases} b(y-a)^2 + c & \text{if } y \leq 0 \\ -b(y+a)^2 - c & \text{otherwise} \end{cases} \tag{4}$$

UIST '99. Asheville, NC

where x, q and y represent a desire level to insert *aizuchi*, a velocity of its change and the difference of enthusiasm for the dialogue between user and computer, respectively. q is considered as the context or the trend of the dialogue on the computer side. f_{int} and f_{ext} mean a desire function of generating *aizuchi* and user influence function. ω is frequency, a, b and c are positive constants which characterize the computer's personality.

The model can naturally handle continuity of nonverbal information, and can have the dialogue context as internal variables. We believe these features of our model enable the computer to respond autonomously and flexibly based on the context of nonverbal information represented by time-sequential patterns of multiple channels. Although a computer should generate *aizuchi* in response to user's utterance, passive *aizuchi* is not enough to facilitate user's speech.

Implementation of MAICO

We constructed a testbed of this DDM for evaluation. The system (see Figure 2) consists of Image Recognizer, Speech Feature Extractor, Keyword Spotter, Natural Language Processing, DDM, CG Agent Generator, and Speech Generator. In the testbed, we have implemented *aizuchi* system in response to user's voice power and some keywords which generates *aizuchi* with understanding its meaning. The system controls facial expressions and speech of the agent according to the DDM output. When the velocity of desire change to insert *aizuchi* is below a certain threshold, the system outputs only nod. When above the threshold, it outputs nod with utterances such as ”yes” and ”sure”. As the agent's voice, we used some pre-recorded intersections such as ”*hai*” and ”*ee*” (”yes” and ”sure” in English) uttered by a person. In addition to those words, the agent gives responses to some specified phrases by using keyword spotting technique. The timing of the response is decided by DDM.

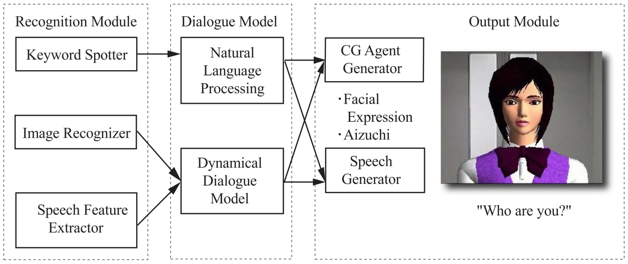
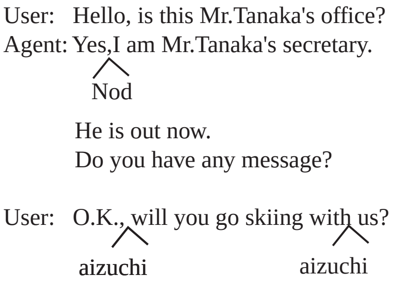


Figure 2: System Configuration

Examples of interaction

We applied our dialogue model to Auto-Answering Visual Telephone. The user talks to an agent some messages. The system asks some questions to complete the message, for example asking for information of who, when, where and what. If any information isn't filled out because of lack of information or recognition failure, the system keeps asking the user for more information.


User: Hello, is this Mr.Tanaka's office?
Agent: Yes,I am Mr.Tanaka's secretary.

He is out now.
Do you have any message?

User: O.K., will you go skiing with us?

Figure 3: Example of Dialogue

Summary

We have focused on back channels or listener's feedback (*aizuchi* in Japanese), and developed the prototype dialogue system based on the DDM. The DDM can explicitly treat continuous time, and can naturally deal with time-sequential patterns. We showed that the model was very promising in both simulation and real-time system.

In the future, we will extend the model by applying it to various tasks. It is also necessary to determine parameters automatically from dialogue database.

REFERENCES

1. S.Seki, et al:Human Interface with Rhythm based on Dynamical Dialogue Interface, IPSJ SIG-HS, 98-HI-80, pp.39-40, 1998(in Japanese).
2. K.Sakamoto et al: Multimodal Interface Prototype Based on Rhythm and Timing of Interaction”, Proc. RWC'97, pp.31-36, 1997.
3. K.Watanuki, et al: Analysis of Multi-Modal Interaction Data in Human Communication, Proc. ASJ Spring Conference, 1-7-20, pp.39-40, 1994(in Japanese).
4. K.Watanuki, et al:Study of Listener's Back Channels, ASJ Spring Conference, 3-6-15, pp111-112,1998(in Japanese).
5. K.Sakamoto, et al: A Response Model for a CG Character Based on Timing of Interactions in a Multimodal Human Interface, Proc.IUT'97,pp.257-260,1997.
6. N.Ward: In Japanese a Low Pitch Means 'back-Channel Feedback Please', SIG-SLP-11-2,1996.
7. R.A.Bolt : The Integrated Multi-Modal Interface, The Transactions of the Institute of Electronics Information and Communication Engineers, Vol.J-70-D,No.11,1987.