Nonverbal Information Recognition and Its Application to Communications

Ryohei Nakatsu
ATR Media Integration & Communications Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
Tel: +81-774-95-1400
nakatsu@mic.atr.co.jp

1. ABSTRACT

The development of sophisticated human interface technologies are highly desired for facilitating human-machine communications. The human interface is defined as technologies for supporting human communications and, therefore, should take the various aspects of communications into considerations. In this paper, various forms of human communications are examined and in particular the nonverbal aspects of communications are shown to be essential. Also it is reported that systematic research to treat *nonverbal information in communications* should be carried out. Several examples of this type of research are described.

1.1 Keywords

communication, nonverbal information, emotion

2. INTRODUCTION

As computers become ever more ubiquitous in society, sophisticated "human interface" between humans and computers need to be developed In recent years, the meaning of the "human interfaces" has been extended from the basic concept of the "man-machine interface." It can now be taken to refer to research into how computers can support person-person communication as well as between humans and computers. For this reason, human interface research must take into account various aspects of human communications. Extensive research has been conducted into the human interface, such as the recognition of speech, hand-written characters , facial expressions and gestures. However, this may lead rise to the following questions. What are the major aspects of human communications? Which of these aspects are connected to the various fields of research? Should research into these aspects of human communication be conducted in an isolated way or integrated with others at a more fundamental stage? This paper examines the issues of human communications and will allude that nonverbal aspects of human communications are essential for the realization of a sophisticated human interface. This paper also reveals the necessity of systematic research on the nonverbal information recognition. Finally, several examples of research into nonverbal information recognition being performed at the ATR Media Integration & Communications Research Laboratories are also presented.

3. NONVERBAL ASPECTS OF HUMAN COMMUNICATIONS

3.1 Human communications

Within narrow definitions, the "human interface" can be taken to mean a "man-machine interface." From this perspective, extensive research has been conducted into how computers can be made easier for people to use. Considerable attention has, therefore, been placed on the methods of communication between humans and computers. However, as computers come to play an increasingly significant role in our lives in a variety of forms, the problem advances beyond how to establish an effective interface between a single computer and a solitary human user. Instead the question arises of how communication among people should develop and how this should be done in an environment characterized by the ubiquitous computer. We have, therefore, viewed the human interface in the following manner.

(1) Researching the human interface means researching human communication activities Communication in this case includes that between humans in addition to that between humans and computers.

(2) The central issue of human interface research is to determine how the computer should support human communication activities

Once the problem is set up in this way, the problem shifts from being how to make it easier for humans to use individual computers to one of being how to make computers support human communication.

Next, from the viewpoint of communication, let's take a look at the way in which the human interface has been researched up to now mainly in the field of engineering Here, it can be seen that engineers have traditionally focused research on such entities as robots and computer agents that integrate functions for communicating with humans. This research has for the most part concentrated on the language component of communication. One example of such research has been speech recognition which aims to extract meaning from human speech, i.e., language-based information. In recent years, however, it has been recognized that the conveyance of emotion and sensitivity, i.e., that of nonverbal communication, also plays an important role in human communication. Technology associated with nonverbal communication has therefore become an important element for future research into the human interface

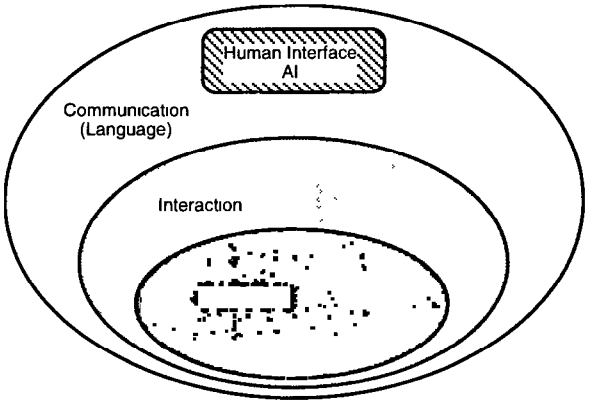


Figure 1 Communication model

3.2 Communication model[1]

Figure 1 schematically represents a human communication model, such that the model itself resembles the structure of the human brain. The top-most layer, which corresponds to the neocortex in the brain, controls language-based communication functions. Researchers investigating the human interface and artificial intelligence (AI) fields have been studying the mechanisms of this layer. As mentioned above, a typical example of this research is speech recognition.

The objective of speech recognition has been to extract just logical information contained in speech. However, logical information is merely one component that makes up speech among a very rich variety of information forms associated with emotions and sensitivity. We can consider these as being generated in the lower layers in the figure, specifically, the interaction layer and reaction layer. The interaction layer here controls actions like the generation of speaking rhythm and the order of taking turns to speak to maintain a coherent communication channel This layer plays an important role in establishing smooth communication between people. Beneath this interaction layer is the reaction layer, which controls basic human actions like the turning of one's head in the direction of an incoming sound and the shutting of one's eyes in response to a bright light stimulus. We can view these as primal functions, that is, as being acquired deep in man's past during a time when he would be classed as being in the animal stages of development.

As such, it can be said that the low-layer functions, which

Verbal/Nonverbal	Type	Contents	Technologies
Verbal Information	Speech	Meaning of utterance	Speech recognition
Nonverbal Information	Body Motion	Facial expression, hand gestures, Body motions, etc	Recognition/generation of facial expression, hang gesutures, and body motions
	Para-language	Emotion, rhythm, timing, etc	Emotion recognition/generation Recognition/generation of rhythms and timing, etc

Table 1 Information used in communication

differ from those that control logical actions and logical information at a higher layer, play an important role in human communication, and that nonverbal information like emotions and sensitivity is generated and understood by these functions in the lower layers. To achieve comprehensive human communication functions that also include the transmission and reception of nonverbal information, therefore, the mechanisms of the above interaction and reaction layers need to be researched and eventually integrated with the functions of the top communication layer.

3.3 Nonverbal information in human communications

Section 3.2 explained the importance of nonverbal information contained in human communication. This section will examine possible research themes and approaches toward the treatment of nonverbal information. Table 1 shows types of representative information used in human communications as well as technologies for handling these forms of information. Here we need to point out the following important issues in the treatment of nonverbal information

(1) For modalities concerning human body, such nonverbal information as facial expressions, hand gestures, and body motions plays a major role in human communications. Therefore, facial expression recognition, hand gesture recognition, and body motion recognition are key research themes. However, for modalities concerning speech, the important aspects for investigation are the recognition of emotions involved in speech, the treatment of rhythms and timing in conversations.

(2) Extensive research has been carried out for many of the research fields described above. In most cases, however, the research has been conducted in isolation from the research into the other forms of human communication. However, human communication is a unified activity where the various kinds of modalities need to be integrated, and as such research needs to be conducted in such a way as to integrate the various kinds of modalities and to develop multimodal recognition technologies.

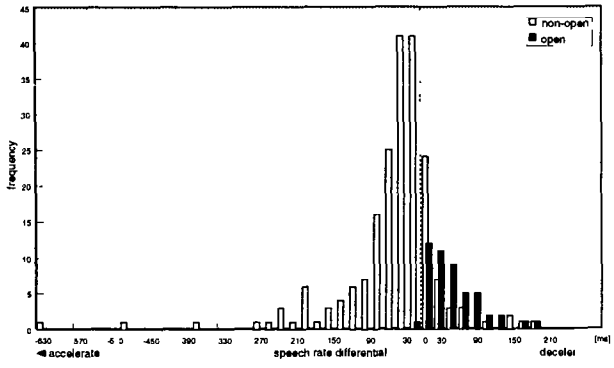
(3) Human do not just passively accept various kinds of nonverbal information, they can also generate and transmit various kinds of nonverbal information. It is, therefore, necessary to study the technologies for generating nonverbal information such as these for emotion generation, facial expression generation and gesture generation. By integrating these nonverbal information generation technologies with nonverbal information recognition technologies, we can develop a multimodal, nonverbal human interface. By further combining this with human interface technologies based on verbal information, it should be possible for computers to communicate with humans in a human-like way.

Based on the above considerations, we have started a project for nonverbal information recognition that investigates these various kinds of modalities. Although it is still at an early stage, we will describe several examples of the research activities being done in our laboratories putting emphasis on speech-related research.

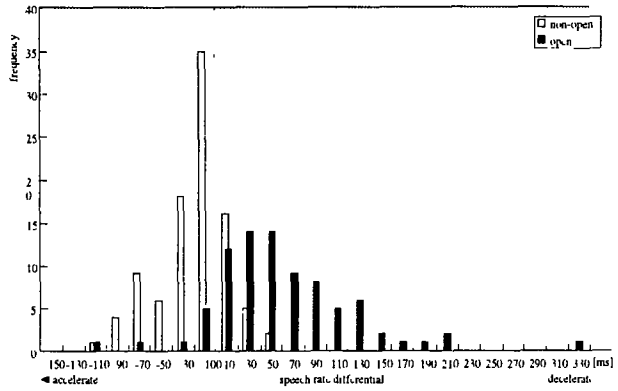
4. RHYTHMS AND TIMING IN HUMAN COMMUNICATIONS

4.1 Analysis of rhythms in conversations[2]

First, we will introduce the research to control speech rhythms in human communications. While it has already been shown that two speakers in engaged conversation will mutually adjust their response time and speech rates, this has only been done qualitatively. Here, we will focus on the "rhythm" in conversation and observe the way in which rhythm is adjusted between two speakers engaged in conversation. The target is a conversation performed over the telephone whereby one speaker informs the other of a route on a map. Here, one unit of speaking remarked by an interval of silence is called a sub-utterance unit (SU), and the average mora duration (AMD) within a single SU is taken to be the evaluation standard when measuring rhythm. Two adjacent SUs are called a non-opening pair (NOP) when made up of elements



(a) single speaker



(b) cross speaker

Figure 2 Speech rate and information openings

like question/answer and request/consent, and an opening pair (OP) when a new topic of conversation begins. The difference in AMDs between two adjacent SUs are measured: a positive result indicates an utterance rhythm that is accelerating, and a negative result an utterance rhythm that is decelerating.

The problem here is to determine how NOPs and OPs accelerate or decelerate for a single speaker and for both speakers in the conversation. The results of analysis are shown in Fig. 2 where it can be seen that. (1) The rhythm of conversation accelerates for NOP and decelerates for OP; and (2) This phenomenon is the same when measurements are made for either the same speaker or between two speakers.

In other words, each speaker in a conversation adjusts his or her speech to the rhythm (speech rate) of the other party's speech. In addition, rhythm speeds up if the topic of conversation remains unchanged, and the rhythm is reset and slows down when the topic changes. These findings indicate that such a mechanism would need to be incorporated in computer agents to make them more human like.

4.2 Generation of natural timing in conversations[3]

We propose a communicative agent model with a "subsumptive structure" which especially aims at the creation of natural timing in conversations. The model consists of multiple layers based on a "subsumption-based architecture" with competence modules. Each competence module independently performs an interaction with the real world and a behavior is expressed in 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 have emerged as a result of the 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 the spreading activation of a "dynamic action selection network". Figure 3 illustrates a communicative agent model consisting of three parts. the environmental context, the intentional context and a set of situated agents. The environmental context places constraints on the actions of the situated agents. The set of situated agents for the behaviors from all of the competence modules is represented in Fig. 3 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.

In the process of conversing, the two contexts and a set of

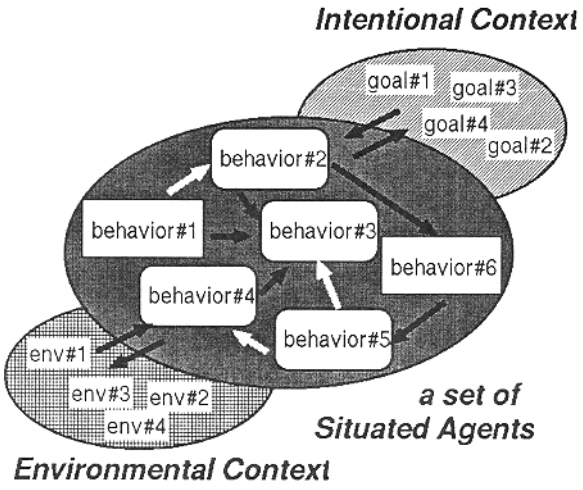


Figure 3 Architecture of communicative agent model

situated agents are influenced by each other. As a result of such interaction, one action is released dynamically using activation/inhibition dynamics which make the accumulated activation energy increase.

We have constructed an interactive agent as a test bed of emergent chatting within the communicative agent model as a result of interactions with a human. The interactive agent is in the shape of an eyeball generated by 3-D computer graphics, which we have called the "Talking Eye"(Fig. 4) 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 speech recognition, prosody of speech and simple motion. Furthermore, it can reproduce about 250 Japanese vocal phrases for chatting by means of speech synthesis. With these capabilities, people can enjoy exchanging simple phrases with it in a more natural way due to the element of natural timing

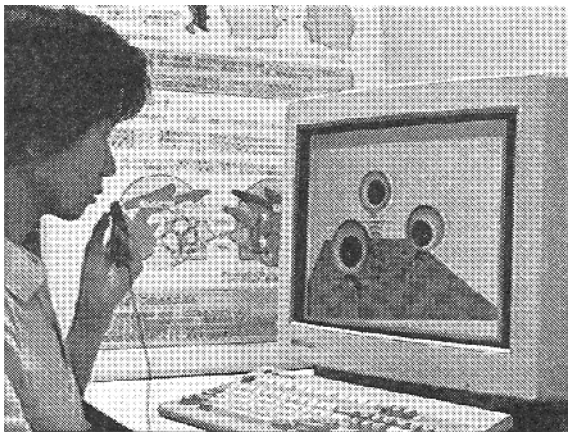


Figure 4 A photograph illustrating chatting with "Talking Eyes"

5. EMOTIONS INVOLVED IN SPEECH

5.1 Recognition of emotions involved in speech[4]

In speech-based communication, emotions play an important role, sometimes playing an even bigger role than logical information included in speech. This phenomenon can be observed in the way that a baby learns to recognize emotional information before understanding semantic information in mother's speech. Adults too recognize emotional information along with semantic information for its speech, and by integrating the two, can come to understand what the other party is trying to say at a deeper level, making for smoother communication.

Based on the above considerations, we investigated the recognition of emotions involved during speech. As a basic architecture for emotion recognition, we adopted a neural network For the kinds of emotions to be recognized, we selected the eight emotional states: neutral, anger, sadness, happiness, fear, surprise, disgust, and teasing

Figure 5 illustrates a block diagram of the process. The

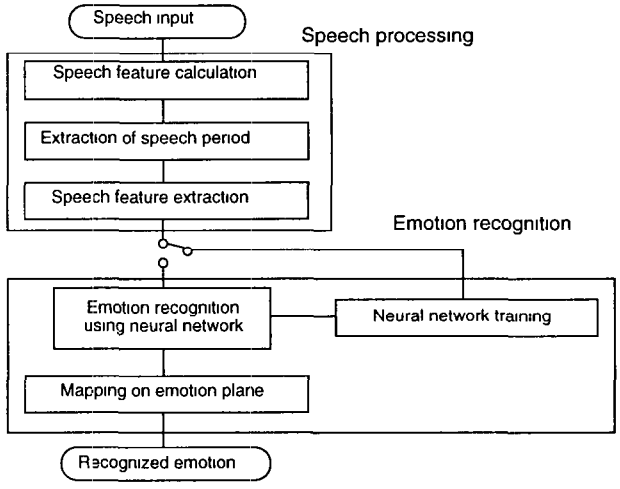


Figure 5 A block diagram of the emotion recognition process

process mainly consists of two parts: speech processing and emotion recognition. In the speech processing part, feature parameters of the input speech are extracted in real time in the feature extraction stage. Then, by observing the speech intensity, i e., volume, the speech is extracted From the extracted speech, feature parameters are extracted and arranged as an output of the feature extraction stage. This output is fed into the emotion recognition part. In the emotion recognition part, a two stage emotion recognition is carried out. In the first stage, a combination of plural neural networks, each of which is designed and trained to recognize a specific emotion included in speech, receive feature parameters and carries out a recognition process. In the second stage, the multiple outputs from the first stage are processed via a specialized logic to obtain emotion recognition results.

For the recognition of emotions, it is necessary to train each of the sub-networks. The following utterances were prepared for the training process:

Words. 100 phoneme-balanced words

Speakers:five male speakers and five female speakers

Emotions.neutral, anger, sadness, happiness, fear, surprise, disgust, and teasing

Utterances:Each speaker uttered 100 words a total of eight times

By carrying out the training process using these utterances, neural networks should be able to carry out speaker-independent, context-independent emotion recognition.

5.2 Application of emotion recognition[4]

As an application of emotion recognition technology, we have created a computer character referred to as MIC that recognizes and responds to emotions. This is a collaborative project with engineers and artists in which the emotion recognition section is developed by engineers and the computer graphics and response animation associated with these computer characters are created by artists. MIC has the following features (1) MIC is capable of recognizing emotion in human speech and responding to it. Thus, MIC can engage in nonverbal communication with humans using emotions. (2) The entire bodily image of these characters is prepared by computer graphics and an emotional reaction is expressed not just by a facial expression but through entire body movements. Moreover, by using images in the background that also involve emotions, the audience is able to easily understand the current emotional state of these characters.

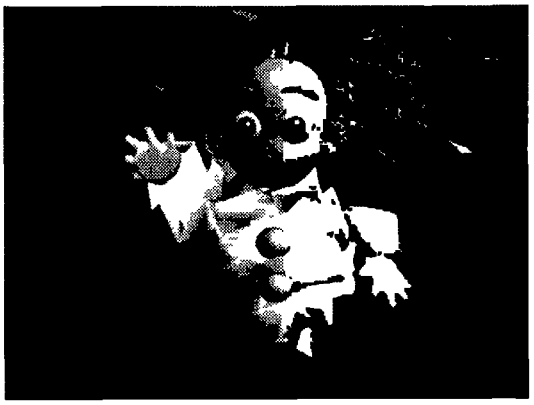


Figure 6 An example of MIC's emotional expression

By utilizing the above capabilities of MIC, people can enjoy conversations based on emotions with MIC A typical MIC response pattern is shown in Fig. 6.

6. FACIAL EXPRESSIONS AND GESTURES

6.1 Recognition of facial expressions and gestures

Facial expression recognition and gesture recognition are the main topics of this conference Our laboratories are also conducting several other research projects [5][6] We have, therefore, omitted the details of these technologies here.

6.2 Application of facial expression recognition[7]

Facial expressions play an important role in natural human communications. We communicate with other people smoothly by recognizing their emotions via facial expressions as well as expressing our own emotions in the same manner. In order to realize an agent which has a human shape in a virtual space, therefore, a technique is required for extracting facial expressions from an image in real time and to reproduce them three-dimensionally on a model face. For this, we studied the real-time recognition of facial expressions and reproduction technologies As an example of applying this technique to the creation of a human-like agent, we examined the reproduction of a three-dimensional face model of a KABUKI actor.

A flowchart for the recognition of facial expressions and the

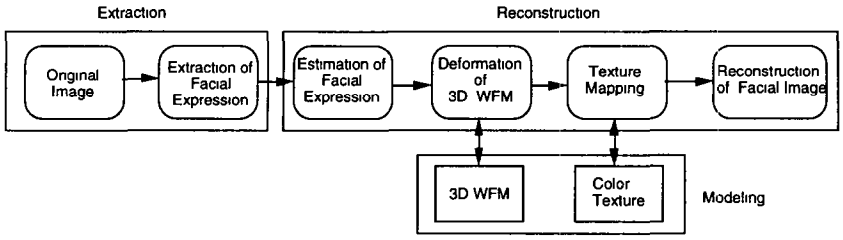


Figure 7 Flowchart for extraction and reconstruction of facial images

creation processing is depicted in Fig. 7. The recognition of facial expressions and the creation processing system consist of three parts: the extraction of expressions, facial reconstruction, and facial modeling. The face model must be created beforehand, a three-dimensional model of the face is created in the form of a wire frame model. With this wire frame model, the facial shape is made in a similar fashion to as if it were assembled from small triangular patches. The color texture of the face is then rendered on these triangle patches In order to extract facial expressions in real time, the subject puts on a helmet and a small video camera attached to the helmet which takes images of the subject's face. If the position or direction of the head changes, the helmet follows these changes to constantly extract facial images. Next, DCT conversion (discrete cosine conversion) is carried out on the images obtained by the camera and changes in facial compositions, such as from the eyes or mouth opening/closing, are extracted. Information on the changes are reflected in the transformation of the three-dimensional model and the extracted facial expressions are reconstructed as facial expressions of a KABUKI actor. An example of a reconstructed KABUKI actor is shown in Fig 8.

Of course, a key point of this system is the technical extraction of facial expressions in real time. At the same time, the transformation technique, which extracts the expressions and reproduces them as a KABUKI actor's facial expressions, is essential. Note that an artist creates the KABUKI actor's face model to add an artistic touch. By adding such an artistic element, anyone can transform himself into a KABUKI actor.

Figure 8 Reconstructed KABUKI actor

7. CONCLUSION

For the realization of smooth communication between human and computers, it is necessary to develop sophisticated human interface In this paper, we have examined the concept of the human interface and shown that it is essential for the development of good human interface to study the nonverbal aspects of communications There are various kinds of research area being concerned into the nonverbal aspect of communication: facial expression recognition, hand gesture recognition, body motion recognition, emotion recognition, treatment of rhythms and timing of speech, and so on. Although extensive research has been carried out in this field, most of it has been conducted in isolation. We have pointed out the necessity of integrated research in these areas. Based on this consideration, we have proposed a project for "nonverbal information recognition." Several examples of research being conducted for this project in ATR Media Integration & Communications Research Laboratories have been introduced

8. REFERENCES

[1] Ryohei Nakatsu, "Image/Speech Processing That Adopts an Ar-

tistic Approach -Toward Integration of Art and Technology," Proc of ICASSP 97, Vol. I/V, pp 207-210 (1997.4).

[2] Hanae Koiso et al., "Information Potentials of Dynamic Speech Rate in Dialogue," Proc. of the 19th Annual Conference of the Cognitive Science Society, pp 394-399 (1997.8)

[3] Noriko Suzuki et al , "Chatting with Interactive Agent," Proc. of Eurospeech 97, pp.2243-2246 (1997.9).

[4] Naoko Tosa and Ryohei Nakatsu, "The Esthetics of Artificial Life," A-Life V Workshop, pp.122-129 (1996.5)

[5] Takahiro Otsuka and Jun Ohya, "Spotting Segments Displaying Facial Expression from Image Sequences Using HMM," Proc. of Int. Conf on Face and Gesture Recognition (1998.4)

[6] Masanori Yamada et al., "A New Robust Real-time Method for Extracting Human Silhouettes from Color Images," Proc. of Int Conf. on Face and Gesture Recognition (1998.4)

[7] Jun Ohya et al , "Virtual Kabuki Theater· Towards the Realization of Human Metamorphosis Systems," Proc of RO-MAN'96, pp 416-421 (1996 11).