ORIGINAL ARTICLE

Naoko Tosa

Artistic communication for A-life and robotics

Received: August 3, 1998 / Accepted: September 3, 1998

Abstract In research concerning communication between machine and man and involving artificial life and robotics today, the realization of sensitive communication produced by introducing an artistic interface is receiving considerable attention. In this paper, we discuss artistic human-like communication research, including the artificial life and robotics research which would become necessary with an increasing need for man–machine communication. I have consulted an artist involved with this theme, a critic in media art, a scientist, and a biologist. Their research will be introduced from their own view points, and will cover artificial intelligence work capable of communicating with humans.

Introduction

Humans cannot live alone. A human is basically a creature who longs for communication with other humans and things. By showing affection to someone or something, teasing someone, or personalizing things, a human is spiritually satisfied. In the future, humans will tend to love or hate the computers that will fill our lives. In order to co-exist with computers, we are forced to change our lifestyles. In our life today it is nearly impossible to avoid communication with computers. That is why sensitive communication with computers becomes important. Although computers are essentially unfriendly, they can be made friendly by the skillful design of computer software and hardware. In this paper, we introduce an AI computer system featuring an interactive theater in which a person can create an impromptu poem and participate in a play while communicating with AI characters capable of sensing human emotion.

Interactive poem

We propose a new type of speech-based interaction system called the "Interactive poem" (Fig. 1). A human and a computer agent create a poetic world by exchanging poetic phrases, thus realizing emotion-based communication between computers and humans. As a first step toward emotion-based communication between computer agents and humans, we have developed several computer agents such as "Neuro baby"[1] and "MIC and MUSE".[2] These are computer characters that are capable of recognizing several emotions in speech and reacting to them by changing their facial expression and body motions. These agents have been very successful and have been demonstrated at various exhibitions. As a next step toward the realization of feeling-based communications between computer agents and humans, we selected "poem" as a means of communication.

There are several reasons for this approach. The main reason is that in a poem not only the meanings of words or phrases, but also the rhythms and moods created by their sequence, play an essential role. Therefore, the poem is intended to transmit feeling information such as mood and sensitivity rather than logical information. The second reason is that poems were originally expressed by oral reading rather than in writing. This means that a poem is suitable for interaction between computers and humans. Recently, researchers have shown increased interest in the realization of feeling-based interactions and communications between computers and humans.[3–5] However, only a few have worked on voice communication, despite the fact that voice

N. Tosa (✉)
ATR Media Integration and Communications Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
e-mail: tosa@mic.atr.co.jp

This work was presented, in part, at the Third International Symposium on Artificial Life and Robotics, Oita, Japan, January 19–21, 1998

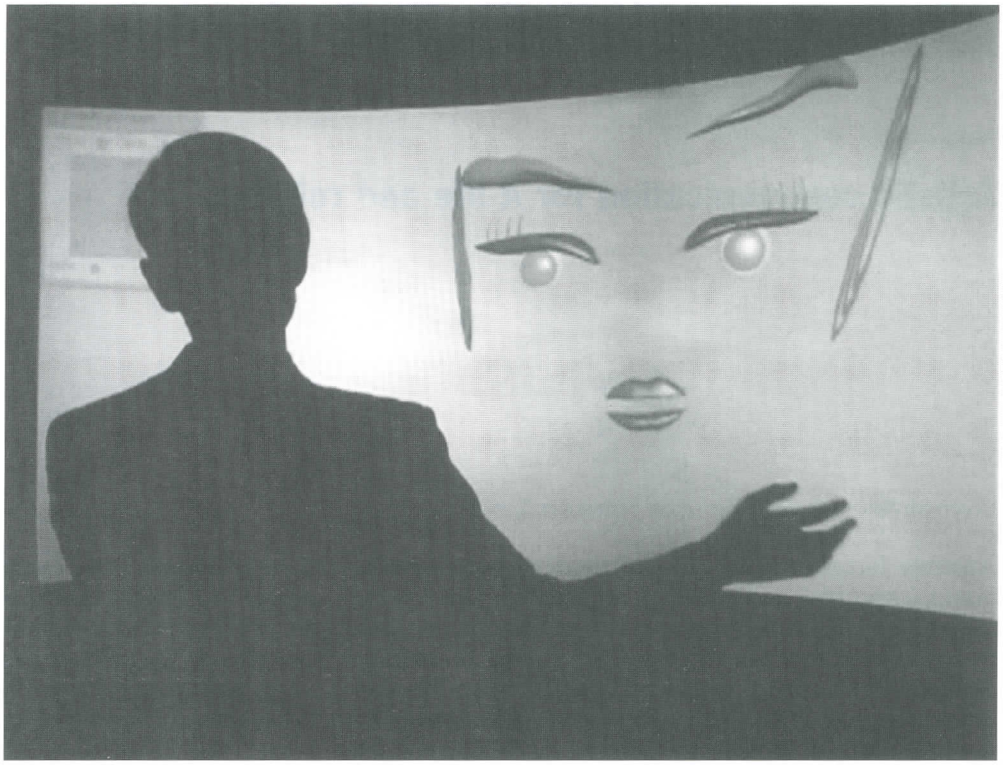


Fig. 1 Interactive poem

is an essential means of feeling-communications. This is the third reason for our interest in developing communication based on an uttered poem. This paper next explains the basic principles of the interactive poem system we have developed based on the above concept. The software and hardware configurations are then described in detail. Finally, a typical installation of the interactive poem system is introduced. The interactive poem is a new type of poem that is created by a participant and a computer agent collaborating in a poetic world full of inspiration, emotion, and sensitivity.

Concept of an interactive poem

My interest is in how to generate feeling in communication between people and intelligent characters. Also, I am interested in creating an intelligent character's consciousness. The computer-generated poet "Muse" can make a poem with you interactively and in real time. Interactive poetry has its roots in old Japanese culture as a type of poem called a "renga." The Renga is a kind of haiku. The haiku was created in the Edo era "beginning in the 1600s" and is a typical expression of Japanese sensibility. The renga is a combination of short poems generated by several people. For example, one person makes the first short poem, and another person makes the second short poem.

A computer agent called MUSE, which has been carefully designed with a face suitable for expressing the emotions of a poetic world, appears on the screen. It will utter a short poetic phrase to the participant. Hearing it allows the participant to enter the world of the poem and at the same time feel an impulse to respond by uttering one of the optional phrases or by creating their own poetic phrase. Exchanging poetic phrases through this interactive process allows the participant and MUSE to become collaborative poets generating a new poem and a new poetic world.

Software configuration

The system used to create the interactive poem consists of four main units: system control, speech recognition, computer graphics generation, and speech output (Fig. 2A).

The system control unit manages the behavior of the whole system by utilizing the interactive poem database. In this system, the most important issue is constructing the interactive poem, so we must first explain how the interactive poem database is constructed. A conventional poem is considered to be a sequence of poetic phrases. In other words, the basic construction of a conventional poem can be expressed by a simple state-transition network where each phrase corresponds to a given state, and for each state there is only one successive state (Fig. 2B).

The basic form of the interactive poem is expressed by this simple transition network, but it differs from a conventional poem in that phrases uttered by the computer agent and phrases uttered by the participant appear in turn. This

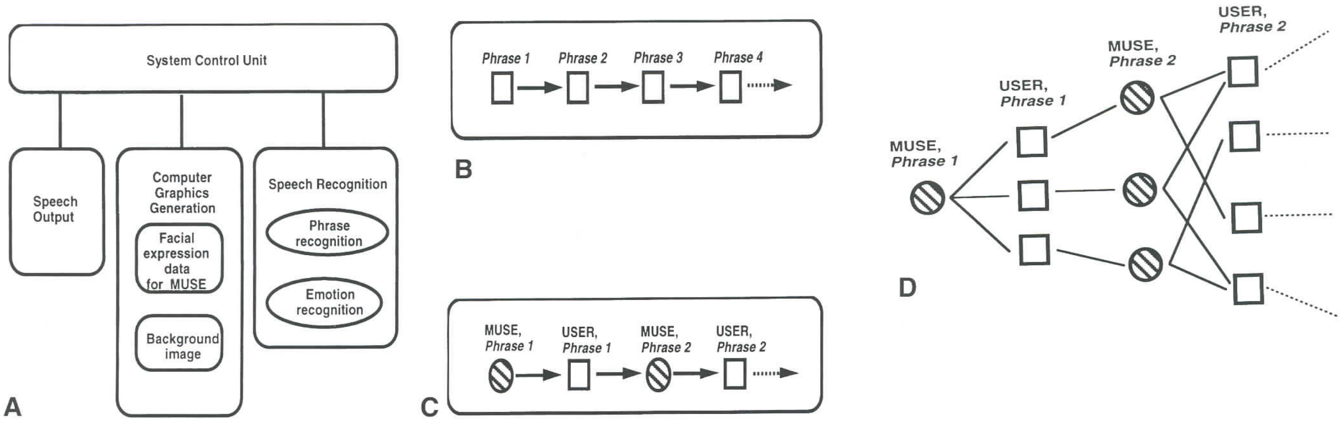


Fig. 2 **A** Block diagram of the interactive poem. **B** Conventional poem. **C** Construction of the interactive poem. **D** Modified method of constructing an interactive poem

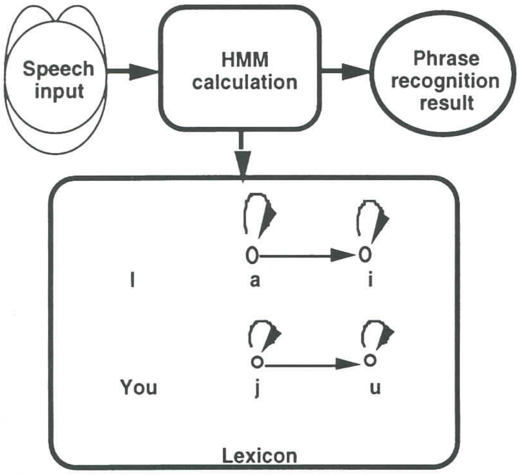


Fig. 3 Phrase recognition

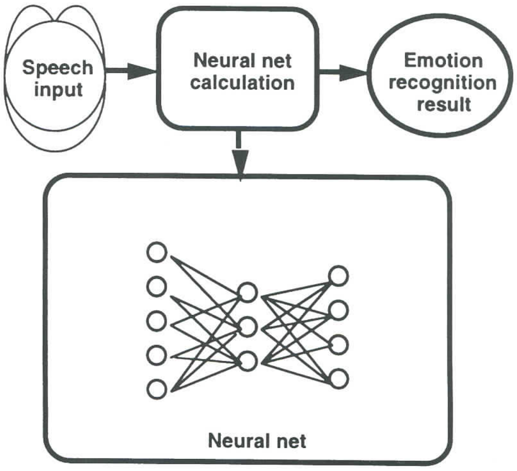


Fig. 4 Emotion recognition

corresponds to a simple interaction where the computer agent and the participant alternately read a predetermined sequence of poetic phrases (Fig. 2C).

To introduce improvisational interaction into our system, we modified this simple transition so that multiple phrases are connected to each phrase of the computer agent (Fig. 2D). These phrases are carefully created and chosen by taking into account how well their rhythms are formed and the meaning of each phrase. this transition network is stored in the interactive poem database and used to control the whole process.

The speech recognition unit has two different speech recognition functions: phrase recognition and emotion recognition. To recognize each phrase uttered by a participant, we have adopted a HMM (hidden Markov model)-based speaker-independent speech recognition technology. Each phrase to be uttered is represented in the form of a phoneme sequence and stored in the lexicon (Fig. 3). To simultaneously detect the emotional state of a participant, an emotion recognition function is introduced. A neural network architecture has been adopted as the basic architecture for emotion recognition. This neural network is trained with the utterances of many speakers to express the eight emotional states of joy, happiness, anger, fear, teasing, disgust, disappointment, and neutrality. As such, speaker-independent and content-independent emotion recognition is realized (Fig. 4).

The reaction of the computer agent to the utterances of the participant is expressed through speech and by images. In the speech output unit, speech data for each phrase to be uttered by the computer agent is digitally stored and generated when necessary.

The computer graphics generation unit controls the image reactions of the computer agent. Image reaction consists of two kinds of images: facial expressions for the computer agent MUSE and various scenes. The facial expressions of MUSE express its reactions to the emotional state of the participant. These images are represented by keyframe animations, each of which corresponds to the eight emotions (Fig. 5). To express the atmosphere of the interactive poem, several kinds of scenes are digitally stored. Each scene image corresponds to a group of states in the transition network, and each correspondence is carefully determined in advance.

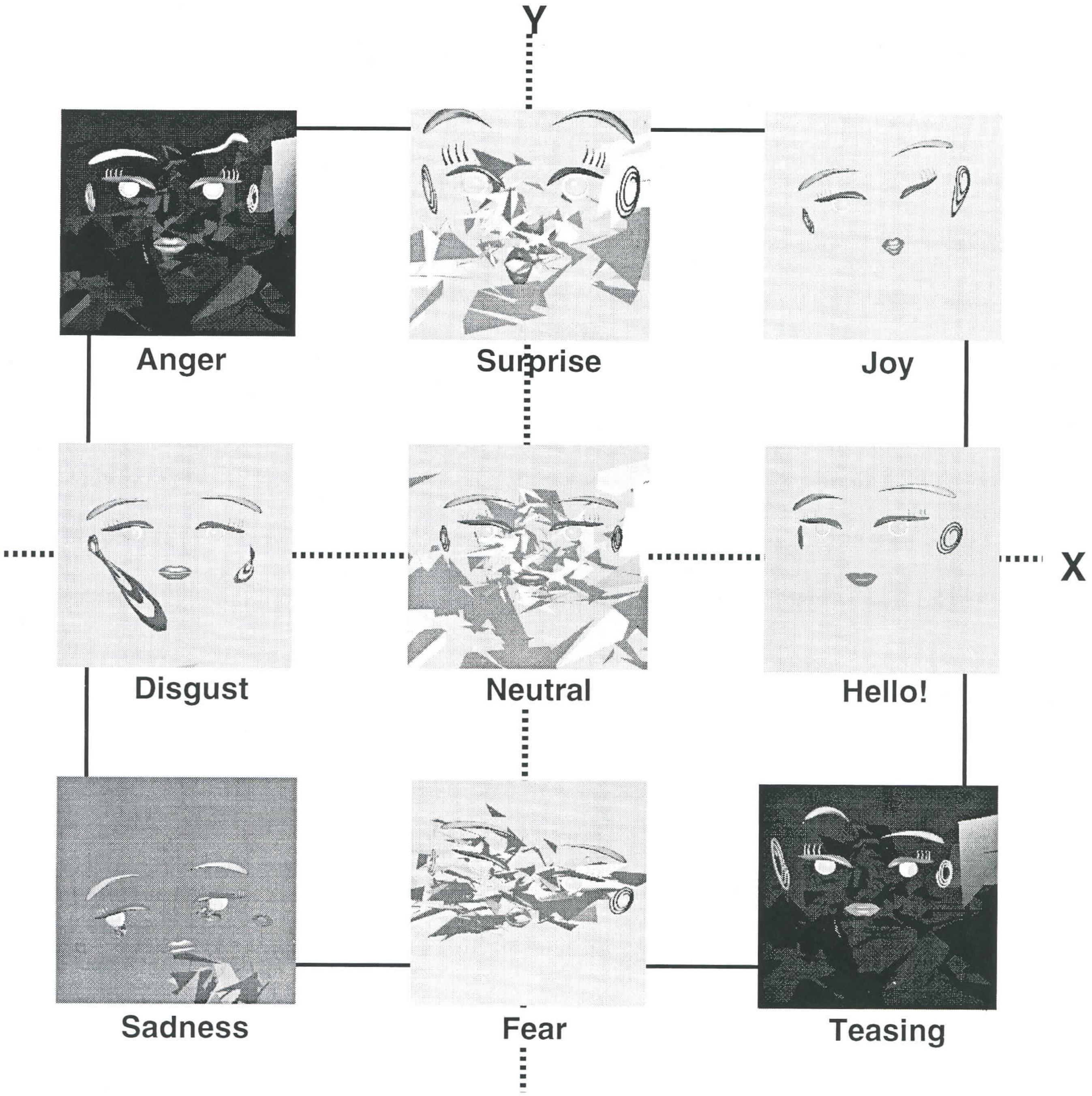


Fig. 5 MUSE's emotional expressions

Hardware configuration

The hardware configuration mainly consists of several workstations and a PC: a workstation for computer graphics generation, a workstation for both system control and phrase recognition, a workstation for emotion recognition, and a PC for speech output. For the participant's convenience, optional phrases that may be uttered following an utterance by MUSE appear on the display. The participant can choose one of these phrases based on their feelings and sensitivity, or they can create their own poetic phrase. Regardless, the emotion recognition function can produce a result. In addition, the phrase recognition function selects the pre-existing phrase that most closely resembles the uttered phrase. Therefore, the participant feels as if the interactive poem process continues in a natural way (Fig. 6).

Interactions

The interaction mechanism operates as described below.

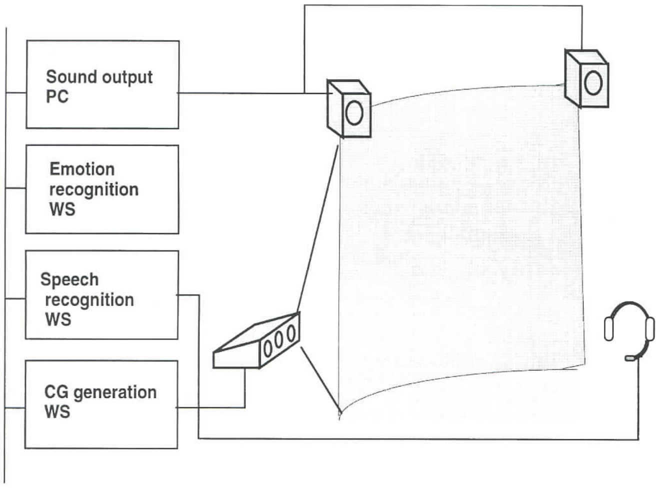


Fig. 6 Interactive poem hardware configuration

1. When MUSE utters a phrase, the recognition process is activated. The participant then utters a phrase, and this is recognized by the phrase recognition function, which uses the lexicon subset corresponding to the next set of phrases in the transition network. At the same time, the emotion contained in the utterance is recognized by the emotion recognition function.
2. Based on information pertaining to the recognition and transition networks, the system's reaction is decided. The facial expression of MUSE changes according to the results of emotion recognition, and the phrase MUSE utters is based on the results of phrase recognition and the transition network. The background scene changes as the transitions continue.
3. In the above manner, poetic phrases between MUSE and the participant are consecutively produced.

Interactive cinema with emotion recognition

As a media artist, I have always been fascinated by the idea of entering the world of movies I created myself. In interactive art, one can interact with movies in virtual reality. My vision was to create a work in which I can talk to characters of my own creation as if they were alive, and feel the excitement of dramatically changing a virtual reality world. One could say that this experience resembles the world of dreams. In a dream, we autonomously communicate with the characters and objects we encounter, and the world is a series of fragments of images with or without context. Characters in a dream fascinate us with their mysterious behavior. What happens in a dream resembles being emotionally involved with movies in a movie theater. The difference is that we are more autonomous in dreams. We have begun research on this electronic daydreaming as the next generation of cinema. We now introduce the design of attractive characters that are capable of recognizing emotion from the tone of a human voice.

Interactive character's design

"MIC" is a male child character. He has a cuteness that makes humans feel they want to speak to him. He is playful and cheeky, but does not have a spiteful nature. He is the quintessential comic character.

Emotion

MIC recognizes the following emotions from intonations in the human voice. The text after the asterisk indicates how the user should make the intonations. The physical form of intonations is called prosody, and how to treat prosody is discussed below.

- Joy (happiness, satisfaction, enjoyment, comfort, smile) * exciting, vigorous, voice rises at the end of a sentence.
- Anger (rage, resentment, displeasure) * voice falls at the end of a sentence.
- Surprise (astonishment, shock, confusion, amazement, unexpectedness) * screaming, excited voice.
- Sadness (sadness, tearful, sorrow, loneliness, emptiness) * weak, faint, empty voice.
- Disgust * sullen, aversive, repulsive voice.
- Teasing * light, insincere voice.
- Fear * frightened, sharp, shrill voice.

Communication

In most cases, the content of a media transmission conceals the actual functions of the medium. The content is impersonating a message, but the real message is a structural change that takes place in the deep recesses of human relations. We aim for this kind of deep communication. People use a microphone when communicating with MIC. For example, if the participant whistles, MIS's feeling is positive and he responds with excitement. If the speaker's voice is low and strong, MIC's feeling is bad and he gets angry.

Processing

This section describes the principles and operational details for the recognition of emotions included in speech. It also explains the generation process of neuro baby's reactions, which correspond to the emotions it receives.

Basic principle

Neuro baby (NB) has advanced from its original version through several stages [1,6] to MIC and MUSE. In our present research, we tried to realize higher-level processing that can achieve more sophisticated interactions between NB and humans. For this purpose, we have emphasized the following issues in our work.

Treatment of various emotional expressions

How many and what kinds of emotional expressions to be adopted are both interesting and difficult issues. In our previous study, we investigated four emotional states.[1] Based on our experiences of demonstrating the first version of NB to various people, and in the belief that an increased number of emotional states would make the interaction between NB and humans richer, this study encompasses the seven emotional states described in the next section.

Speaker-independent and content-independent emotion recognition

Speaker-independence is an important aspect of speech/emotion recognition. From a pragmatic viewpoint, a speaker-dependent emotion recognition system requires a tiresome learning stage each time a new speaker wants to use the system; therefore, it is not easy to use. Moreover, humans can understand the emotions included in speech as well as the meaning conveyed by speech, even for arbitrary speakers. Content-independence is also indispensable for emotion recognition. In daily communication, various kinds of emotions are conveyed by the same words or sentences; mastering such nuances is the key to rich and sensitive communications among people. Therefore, by adopting a neural network architecture and by introducing a training stage that uses a large number of training utterances, we have developed a speaker-independent and content-independent emotion recognition system.

Block diagram of the processing

Figure 7 is a block diagram of the processing flow. The process mainly consists of three parts: speech processing, emotion recognition, and the generation of reactions. In the speech processing part, feature parameters of input speech are extracted in real time in the feature extraction stage. Then by observing the speech power, the period where speech exists is extracted. For 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, where 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 in speech, receives feature parameters and carries out a recognition process. In the second stage, the multiple output of the first stage is processed through a specialized logic, and the emotion recognition results are expressed as points on a two-dimensional space on which eight emotions, including a neutral state, are displayed according to our criteria listed above. The position of the result on the emotion plane and its movement determine the reaction of neuro baby, including its facial expression and action. These facial expressions and actions were previously created by an intuitive design process developed by one of the authors. These reactions are visualized with computer graphics along with appropriate speech output.

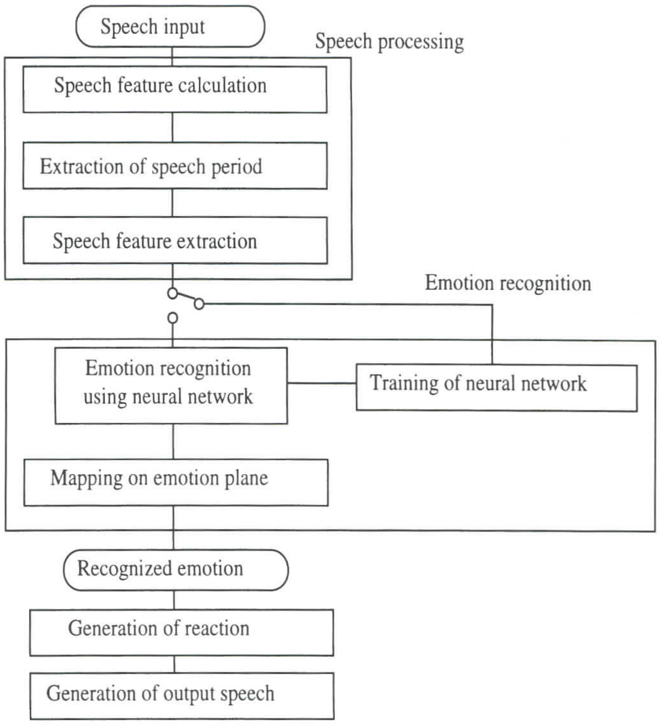

Speech input
Speech processing
Speech feature calculation
Extraction of speech period
Speech feature extraction
Emotion recognition
Emotion recognition using neural network
Training of neural network
Mapping on emotion plane
Recognized emotion
Generation of reaction
Generation of output speech

Fig. 7 Block diagram of the processing flow

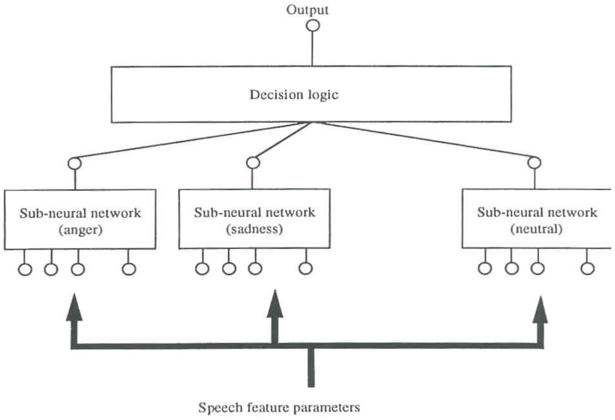

Output
Decision logic
Sub-neural network (anger)
Sub-neural network (sadness)
Sub-neural network (neutral)
Speech feature parameters

Fig. 8 Configuration of the emotion recognition part

Configuration of the neural network

The neural network for emotion recognition is a combination of eight subnetworks (Fig. 8). The decision logic stage combines the outputs of these subnetworks and outputs the final recognition result. Each subnetwork is tuned to recognize one of seven emotions (anger, sadness, happiness, fear, surprise, disgust, and teasing) emotional neutrality. Each subnetwork has basically the same network architecture (Fig. 9). It is a three-layered neural network with 150 input nodes corresponding to the dimension of speech features 20–30 intermediate nodes and one output node.

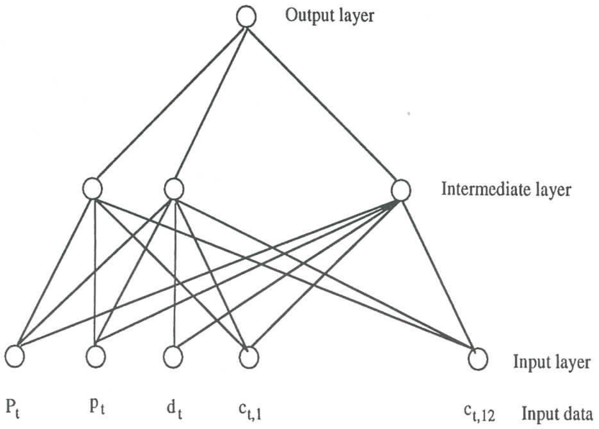


Fig. 9 Configuration of a subnetwork

Neural network training

To recognize emotions, it is necessary to train each of the subnetworks. Since our target is speaker-independent and content-independent emotion recognition, the following utterances were prepared for training.

- *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 eight times.

In each of the eight trials, the speaker uttered the words using different emotional expressions. Thus, a total of 800 utterances for each speaker was obtained as training data.

Using these utterances, we carried out various preliminary training tests. It turned out that preparing two kinds of networks for each emotion, one for male speakers and the other for female speakers, is better than preparing only one network to handle both male and female utterances. In other words, the emotional expressions of males and females are somewhat different and cannot be handled together. The reason for this is not clear and will require further research.

Emotion recognition by a neural network

In the emotion recognition phase, speech feature parameters extracted in speech processing are simultaneously fed into the eight subnetworks, which are trained as described above. Eight values, $V = (v1, v2, \ldots, v8)$, are obtained as the result of emotion recognition.

Mapping on an emotion place

As described above, the output of the emotion recognition network is a vector $V = (v1, v2, \ldots, v8)$ and the final recognition result should be obtained based on V. In our previous study, we expressed the final emotion state by a point on a two-dimensional plane. Based on the experiences of previous research, in the present study the positions of the eight emotions have been rearranged on emotion place E (Fig. 10).

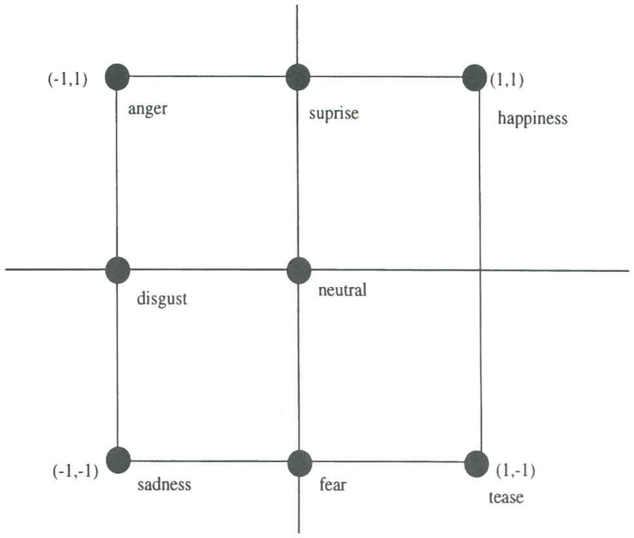


Fig. 10 Emotional plane

It is necessary, therefore, to carry out the mapping from V onto E. Let $m1$ and $m2$ be the first and second maximum values among $v1, v2, \ldots, v8$), and also let $(xm1, ym1)$, $(xm2, ym2)$ be the emotion positions corresponding to $m1$ and $m2$, respectively. The final emotion position (x, y) is calculated by

$$x = c^*\mathrm{xm}1 + (1 - c)^*xm2, \qquad y = c^*ym1 + (1 - c)^*ym2$$

where c has a constant value.

Generation of reaction and selection of output speech

The structure of animation

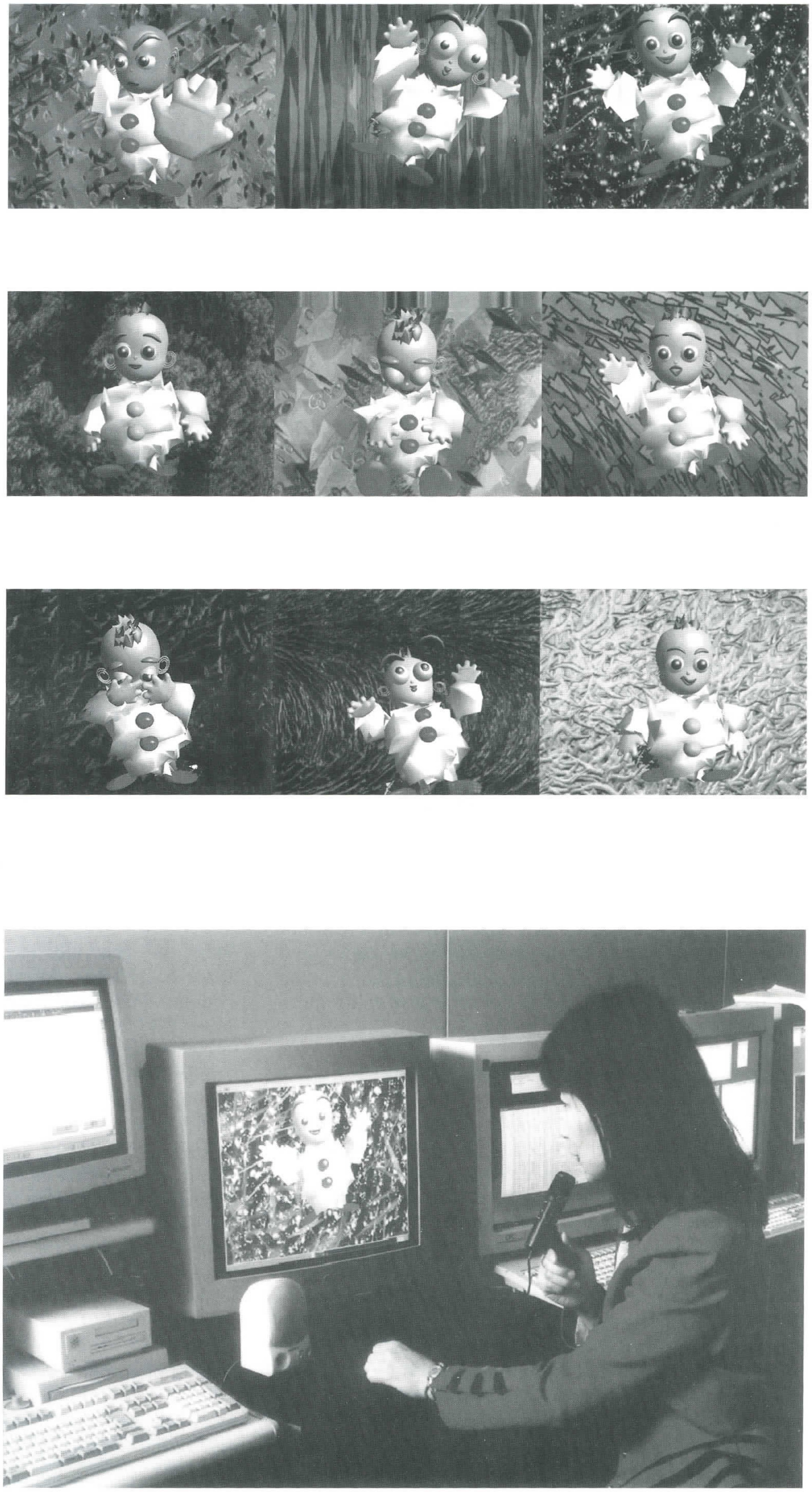
There are four emotion planes, all of which use the same x, y data (Fig. 10).

- Plane "a" generates facial animation by choosing the three key frames closest to the (x, y) data point. The computation of a weighted mean frame is done as follows. Let "A" be the area formed by the three key frames, and $a1$, $a2$, and $a3$ the areas shown in Fig. 6. Accordingly, the three weights for each key frame are $A/a1$, $A/a2$, and $A/a3$, respectively. The interpolated frame is then calculated by a weighted average of the three key frames.
- Plane "b" generates an animation of the character's body by mapping each (x, y) data point on the plane to a body key frame.
- Plane "c" is a mapping of each (x, y) data point to camera parameters such as zoom, tilt, and pan.
- Plane "d" is a mapping of each (x, y) data point to background tiles.

Selection of output speech

From a mapping from the (x, y) data points of the emotion plane to 200 sampled speech utterances, one of the utter-

Fig. 11 Example of the interaction between MUSE and a participant



ances is selected as the output speech. A personal computer is used to play the selected sounds.

Reaction of the characters

The reactions of MIC could be carefully designed and were visualized with computer graphics. Several examples of the emotional expressions of MIC are shown in Fig. 11.

Conclusion

This paper describes two new areas of applied research in artificial life: the "interactive poem" created by the interaction between a human and an anthropomorphic computer-generated poet named "MUSE", and, as part of next-generation interactive movies, the design of characters that can sense human emotions based on computer graphics. Research in these areas is now expanding into a movie description system consisting of an interaction manager, scene manager, script manager, and other functions that enable character scenes and stories to change dynamically. Here, while the properties of anthropomorphic artificial-life characters are often set by internal design, they may also turn "good" or "bad" according to environmental settings and the scheme adopted for communicating with humans. The internal design of characters depends heavily on artificial-life technology, and the ability to communicate requires that an environment be established in which emotions can be introduced naturally. The ultimate objective of combining art with artificial-life research is to achieve works that have an impact on viewers in a new and powerful manner. In this regard, the time is coming when artists and researchers must develop a sharp sense of the best techniques for efficient artistic expression, or in other words, the means of achieving a balance between technology and creativity.

References

1. Tosa N et al. (1994) Neuro-character. AAAI'94 Workshop, AI and A-Life and Entertainment
2. Tosa N, Nakatsu R (1996) Life-like communication agent – emotion sensing character "MIC" and feeling session character "MUSE". Proceedings of the international conference on multimedia computing and systems, pp 12–19
3. Maes P, Darrell T, Blumberg B, Pentland A (1995) The ALIVE system: full-body interaction with autonomous agents. Proceedings of the computer animation'95 conference
4. Perlin K (1995) Real-time responsive animation with personality. IEEE Trans Visualization Comput Graphics 1:5–15
5. Bates J, Loyall B, Reilly S (1992) An architecture for action, emotion, and social behavior. Proceedings of the fourth European workshop on modeling autonomous agents in a multi-agent world
6. Tosa N et al. (1995) Network Neuro-Baby with robotics hand. Symbiosis of human and artifact. Elsevier, Amsterdam