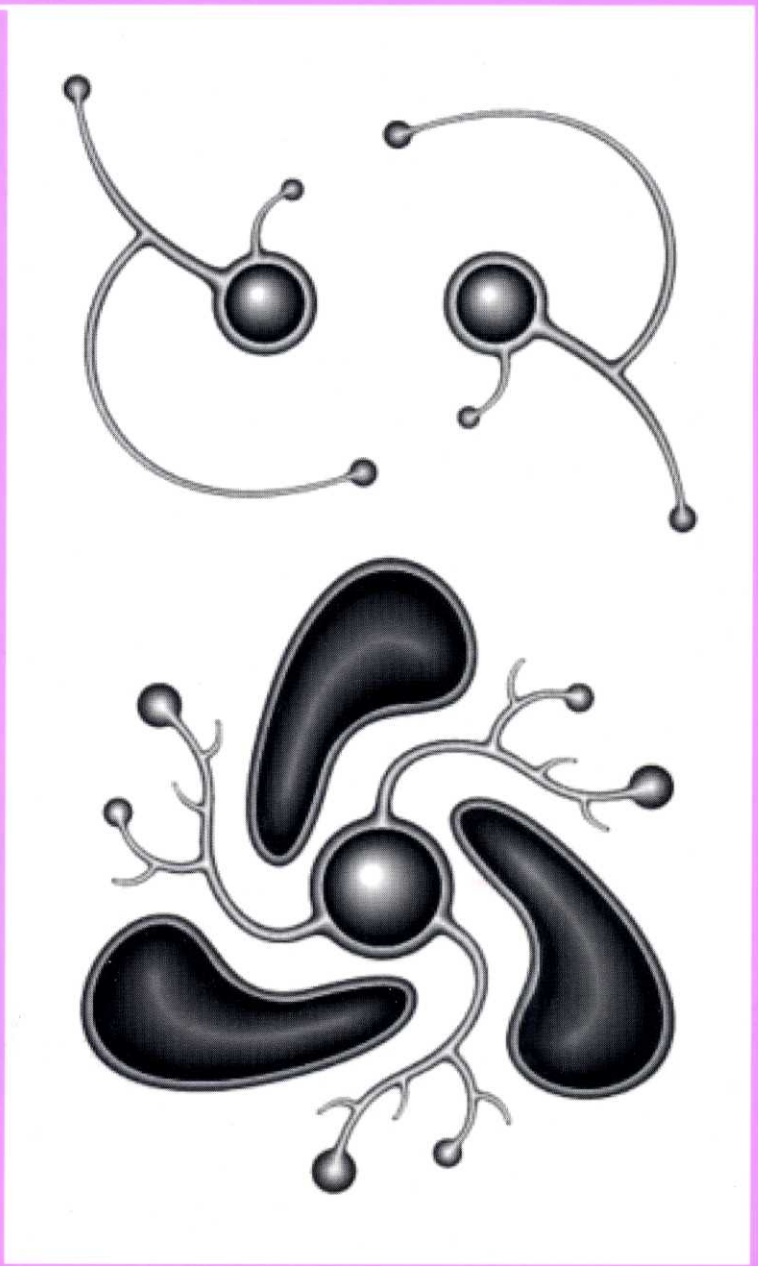
Volume 9 Number 1

ISSN 1462-6268

DIGITAL CREATIVITY

Special Issue

Computers & Post-Biological Art



SWETS & ZEITLINGER PUBLISHERS

Digital Creativity
1998, Vol. 9, No. 1 pp. 53–61

1462-6268/98/0901-0053$12.00

Artistic Communication for A-Life and Robotics

Naoko Tosa & Ryohei Nakatsu

ATR Media Integration & Communications Research Laboratories, Japan

tosa@mic.atr.co.jp

Abstract

A computer-generated poet, MUSE, conveys short poetic words and emotions to a person using the "Renga" format. "Renga" is generated by multiple people as a combination of short Japanese poems such as the "Waka" or "Haiku" which were created in the ancient era and have been used as a medium to express Japanese spiritual emotions. By hearing these words, the person is able to enter the world of that poem and, at the same time, he or she is able to speak to MUSE with poetic words. Through this process of exchanging poetic words, the interactive poem allows the user and computer to work together to build the world of an improvised poem filled with inspiration, feeling, and 'emotion'.

Keywords: *emotion recognition; cross-cultural understanding; spiritual interactive art; computer poetry; sensitive communication*

Introduction

Humans cannot live alone. A human is basically a creature who longs for communication with humans and things. By showing affection to someone or something, teasing someone, or personalising 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, however, 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 skilful design of computer software and hardware. In this paper, we introduce an AI computer system featuring an interactive theatre 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 *Interactive Poem* (Fig. 1). A human and a computer agent create a poetic world by exchanging poetic phrases, thus realising emotion-based communications between computers and humans. As a first step toward emotion-based communications between computer agents and humans, we have developed several computer agents such as Neuro Baby (Tosa, 1994), MIC and MUSE (Tosa & Nakatsu, 1996). These are computer characters that are capable of recognising several emotions in speech and reacting to them by changing their facial expressions and body motions. Fortunately, these agents have been very successful and have been demonstrated at various exhibitions. As a next step toward the realisation of feeling-based communications between computer agents and humans, we selected the 'poem' as a means of communication. There are several reasons for this approach. The main

Digital Creativity, Vol. 9, No. 1

reason is that in a poem not only the meaning of words or phrases but also the rhythms and moods created by their sequence plays 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 realisation of feeling-based interactions and communications between computers and humans (Maes et al, 1995; Perlin, 1995; Bates et al, 1992). However, only a few have worked on voice communications, despite the fact that voice is an essential means of feeling-communications. This is the third reason for our interest in developing communications 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 configuration and hardware configuration are then described in detail. Finally, a typical installation of the *Interactive Poem* system is introduced.

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. A computer agent called MUSE, who has been carefully designed with a face suitable for expressing the emotions of a poetic world, appears on the screen. She will utter a short poetic phrase to the participant. Hearing it allows him/her 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 his/her own poetic phrase. Exchanging poetic phrases through this interactive process allows the participant and MUSE to become collaborative poets who generate a new poem and a new poetic world.

The Concept of Interactive Poem

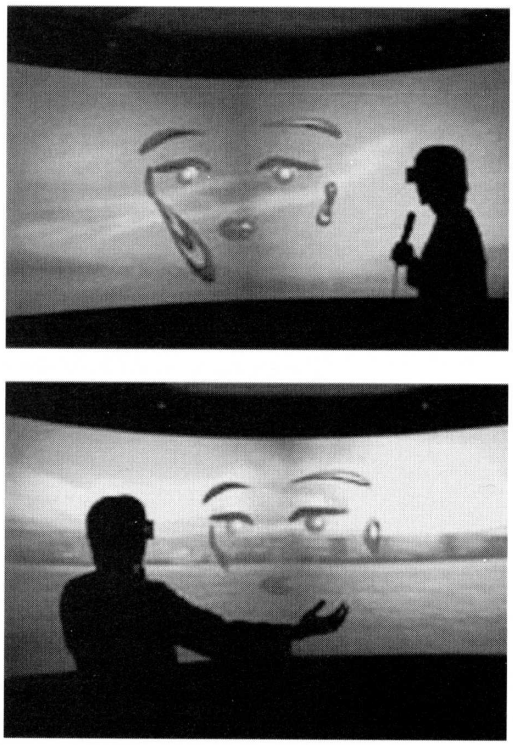


Figure 1

Interactive Poem

My interest is in how to generate feeling in communication between people and intelligent characters. Also, I'm 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 "renga". Renga is a kind of haiku. Haiku was created in the Edo era (beginning in the 1600s) and is a typical expression of Japanese sensibility. Renga is a combination of short poems and generated by multiple people. For example, one person makes the first short poem, and another person makes the second short poem.

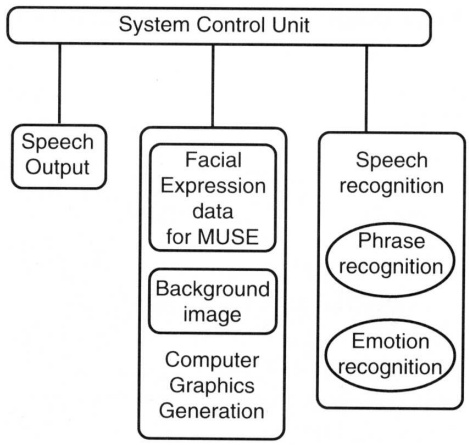


Figure 2 Block diagram of the Interactive Poem

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. 2). The system control unit manages behaviour of the whole system by utilising 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 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. 3a). 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 a participant appear in turn. This corresponds to a simple interaction where the computer agent and the participant alternately read a predetermined sequence of poetic phrases (Fig. 3-B).

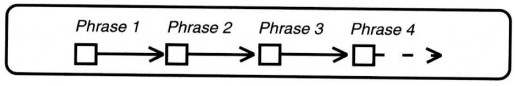


Fig. 3a Construction of a conventional poem

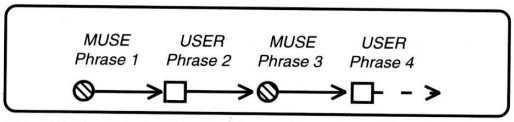


Fig.3b Construction of the Interactive Poem (a)

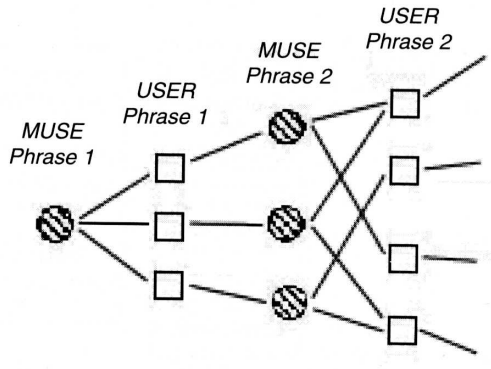


Fig. 3c Construction of the Interactive Poem(b)

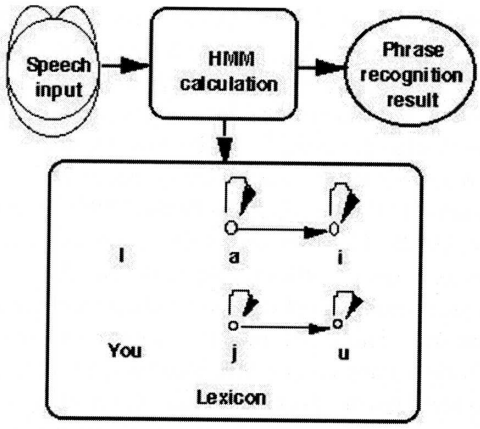


Figure 4 Phrase recognition

Reaction of the computer agent to utterances of a participant is expressed through her 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 reaction 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 her reactions to the emotional state of the participant. These images are represented by keyframe animations, each of which corresponds to the eight emotions (Fig. 6). 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.

Digital Creativity, Vol. 9, No. 1

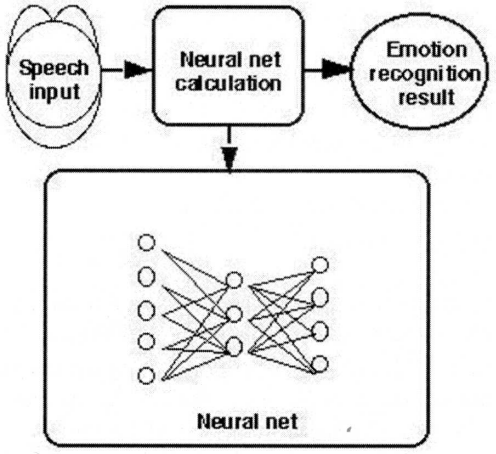


Figure 5 Emotion recognition

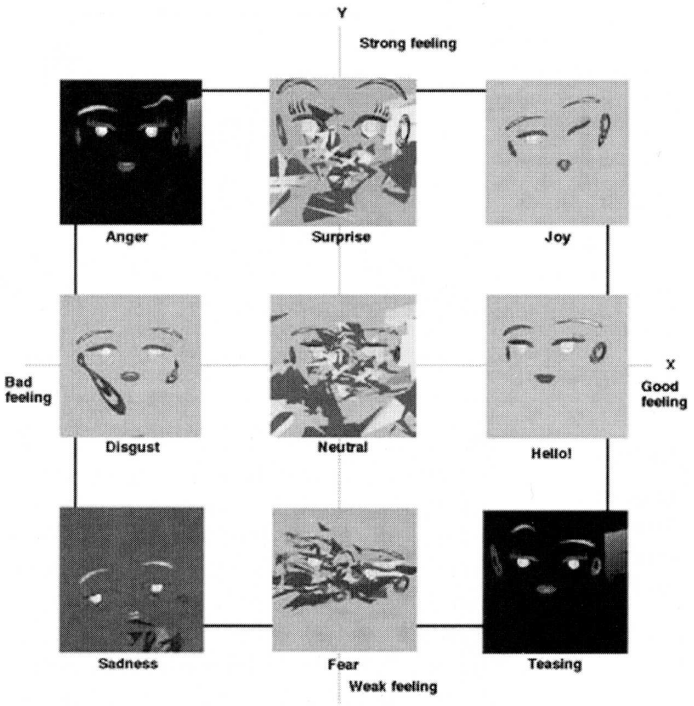


Figure 6

MUSE's emotional expression

Hardware Configuration

This 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 of 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. 7).

Interactions

The interaction mechanism operates as follows.
(1) When MUSE utters a phrase, the recognition process is activated. A participant then utters a phrase and it is recognised 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 recognised by the emotion recognition function.
(2) Based on information pertaining to recognition and the transition network, 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've 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

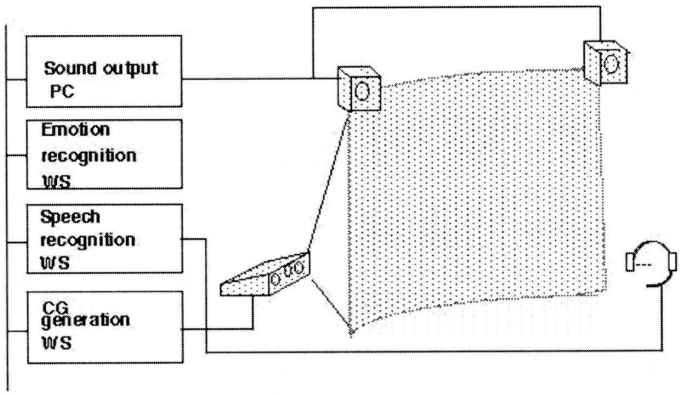


Figure 7

Interactive Poem — hardware configuration

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 behaviour. What happens in a dream resembles being emotionally involved with movies in a movie theatre. The difference is that we are more autonomous in dreams. We have begun research on this electronic daydreaming as the next generation of cinema. In this paper, we introduce the design of attractive characters that are capable of recognising emotion from the tone of a human voice.

Interactive Character 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 doesn't have a spiteful nature. For example, he is the quintessential comic character.

Emotion

MIC recognises the following emotions from intonations in the human voice. An asterisk indicates how the user should make intonations. The physical form of intonations is called prosody.

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

Communication

In most cases, the content of a media transmission conceals the actual functions of the medium. This 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, MIC'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 (Tosa, 1994; 1995) to MIC & MUSE. In our present research, we tried to realise higher-level processing that can achieve more sophisticated interactions between NB and humans. For this purpose, we have emphasised the following issues in our work.

(1) 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

Digital Creativity, Vol. 9, No. 1

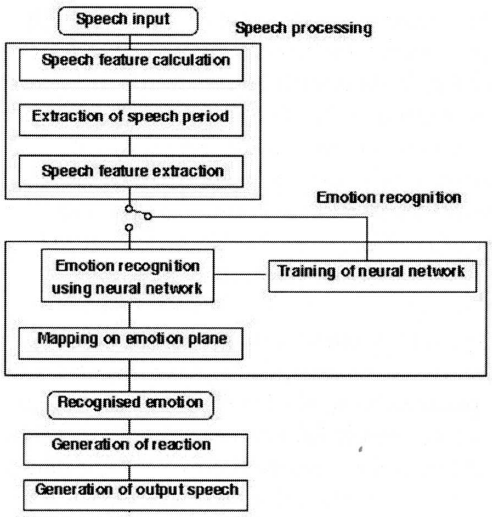


Figure 8 Block diagram of the processing flow

investigated four emotional states [1]. Based on our experiences of demonstrating the first version NB to a variety of people and on 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 later.

(2) 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. Also, content independence is indispensable for emotion recognition. In daily communication, various kinds of emotions are conveyed by the same words or sentences; mastering such nuance 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

Fig. 8 is a block diagram of the processing flow. The process mainly consists of three parts: speech processing, emotion recognition and 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. 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, 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

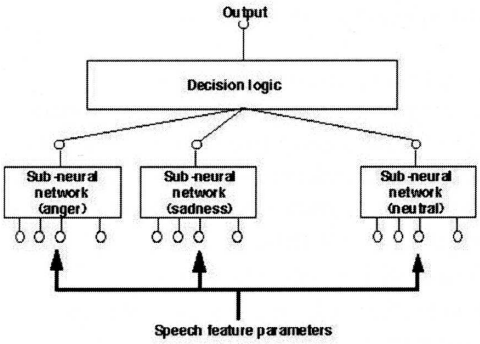


Figure 9 Configuration of emotion recognition part

recognise 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 specialised logic, and the emotion recognition results are expressed as points on a two-dimensional space on which eight emotions including neutral

state are displayed according to our criteria listed earlier. The result's position on the emotion plane and its movement determine the reaction of Neuro Baby, including its facial expressions and actions. These facial expressions and actions were previously created by an intuitive design process developed by one of the authors. These reactions are visualised with computer graphics along with appropriate speech output.

Configuration of the neural network

The neural network for emotion recognition is a combination of eight sub-networks (Fig.9). The decision logic stage combines the outputs of these sub-networks and outputs the final recognition result. Each sub-network is tuned to recognise one of seven emotions (anger, sadness, happiness, fear, surprise, disgust, and teasing) and neutral emotion. Basically, each sub-network has the same network architecture (Fig. 10). It is a three-layered neural network with 150 input nodes corresponding to the dimension of speech features, 20 to 30 intermediate nodes and one output node.

Neural network training

To recognise emotions, it is necessary to train each of the sub-networks. 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, he/she uttered words using different emotional expressions. Thus, a total of 800 utterances for each speaker were 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 between 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 sub-networks and trained as described above. Eight values,

$$V = (v1, v2, \ldots, v8),$$

are obtained as the result of emotion recognition.

Mapping on an emotion plane

As described above, the output of the emotion recognition network is a vector V=(v1, v2, ..., 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 plane E (Fig. 11). 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, ..., 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

$$x = c*xm1 + (1-c)*xm2$$
$$y = c*ym1 + (1-c) * ym2$$

(c: constant value).

Generation of reaction and selection of output speech

The structure of animation

There are four emotion planes, all of which use

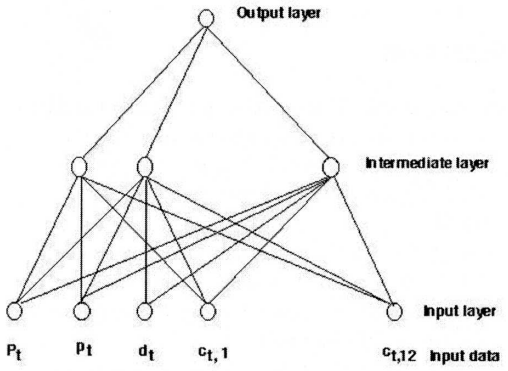


Figure 10 Configuration of a sub-network

the same x, y data (Fig. 11).

a. 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.
b. 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.
c. Plane *c* is a mapping of each (x, y) data point to camera parameters such as zoom, tilt, and pan.
d. 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 utterances 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 visualised with computer graphics. Several examples of emotional expressions by MIC are shown in Fig. 12.

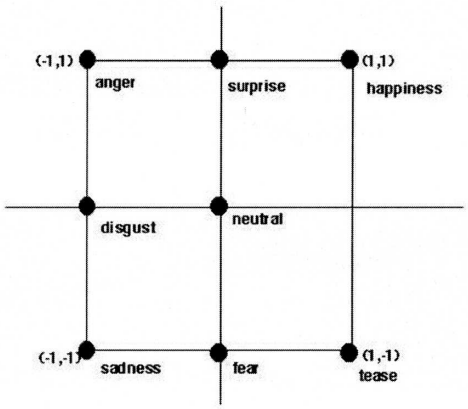


Figure 11 Emotional plane

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

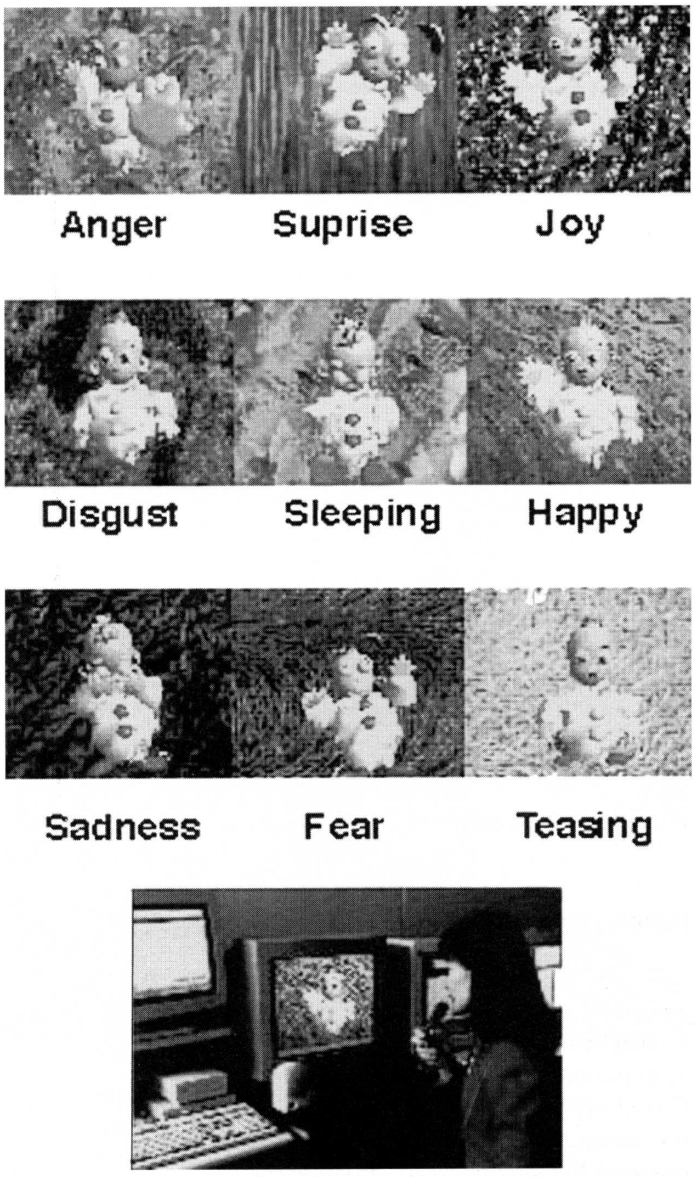


Figure 12

Example of interaction between MUSE and participant

research is to achieve works that impact 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 bridge between technology and creativity.

References

Bates, J., Loyall, B. and Reilly, S. (1992) An architecture for action, emotion, and social behaviour. *Proceedings of the Fourth European Workshop on Modelling Autonomous Agents in a Multi-Agent World.*

Maes, P., Darrell, T., Blumberg, B. and Pentland, A. (1995) The ALIVE system: Full-body interaction with autonomous agents. *Proc. of the Computer Animation '95 Conference.*

Perlin, K. (1995) Real-time responsive animation with personality. *IEEE Transactions on Visualization and Computer Graphics*, **1**(1), pp.5–15.

Tosa, N. et al. (1994) Neuro-Character. *AAAI'94* Workshop, AI and A-Life and Entertainment.

Tosa, N. et al. (1995) Network Neuro-Baby with robotics hand. *Symbiosis of Human and Artifact.* Elsevier Science B.V.

Tosa, N. and 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.

Naoko Tosa is Artistic Director & Researcher in the Interactive Movie Project, ATR Advanced Telecommunications Research Laboratories, Associate Professor at Kobe University and Lecturer at the Department of Media Arts & Science, Musashino Art University