

>96
● ● ● ● I
S
E
A

SEVENTH INTERNATIONAL SYMPOSIUM ON ELECTRONIC ART

SEPTEMBER 16-20, ROTTERDAM The Netherlands

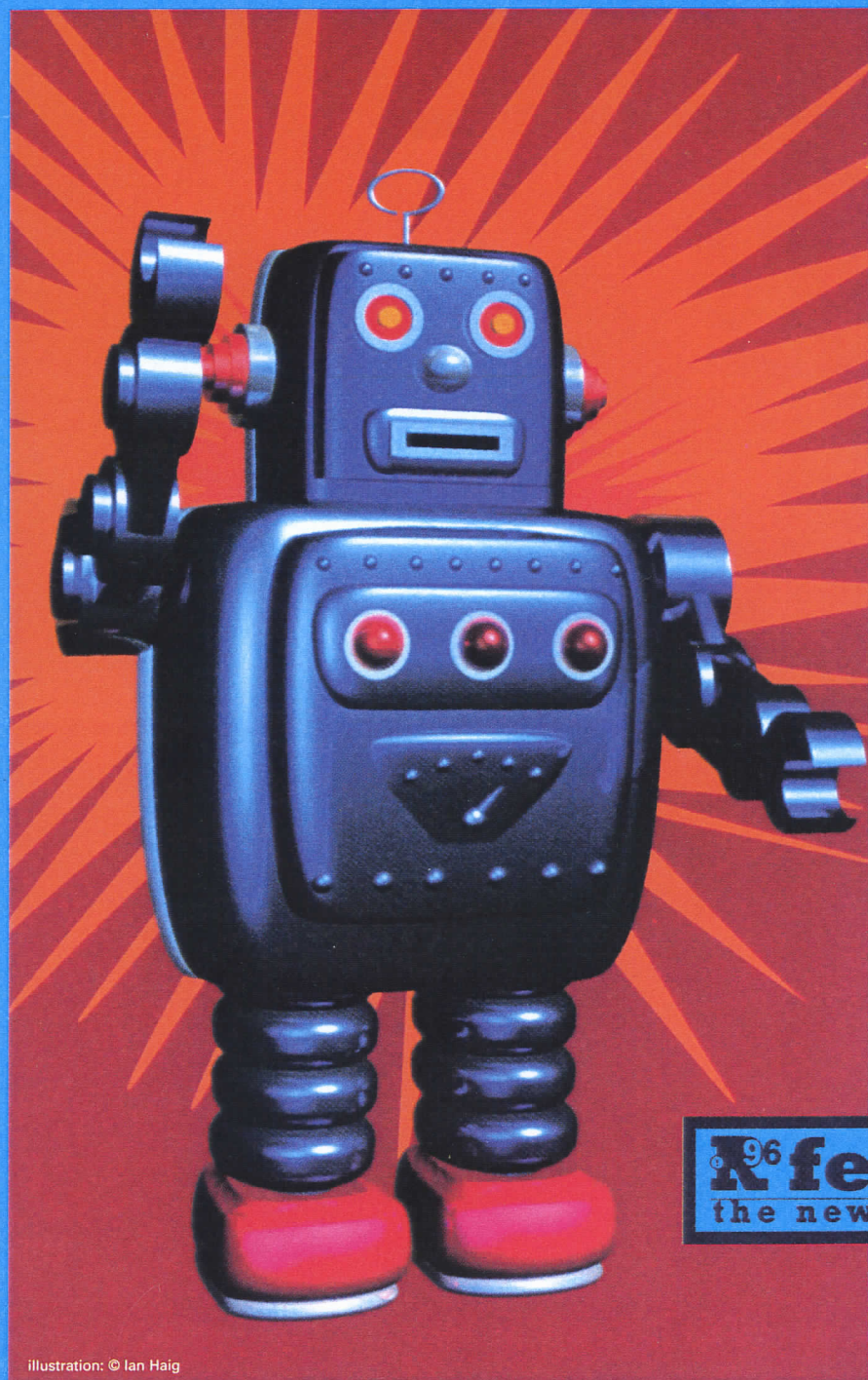
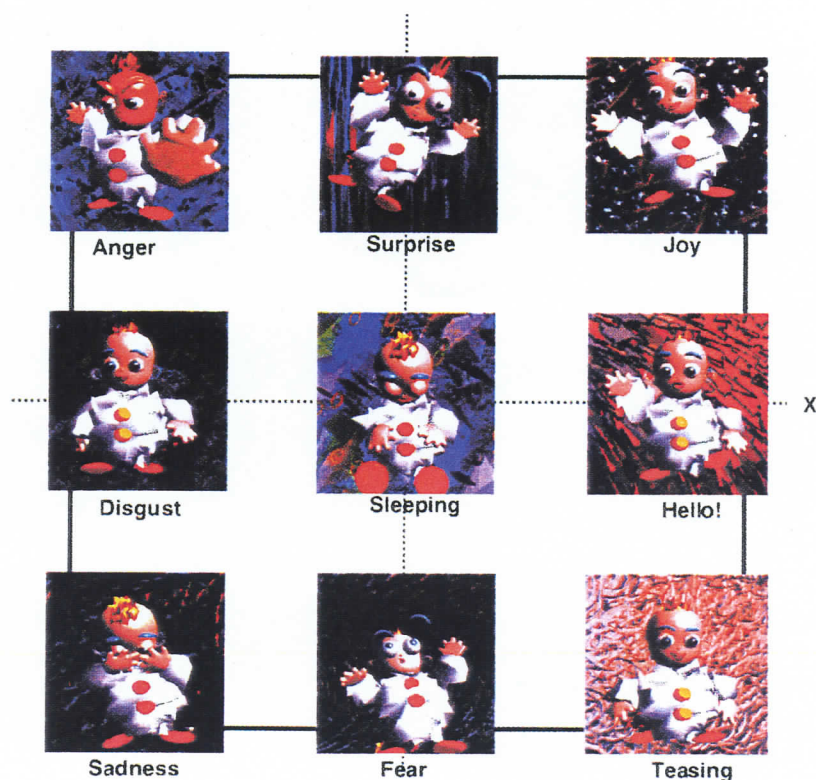


illustration: © Ian Haig

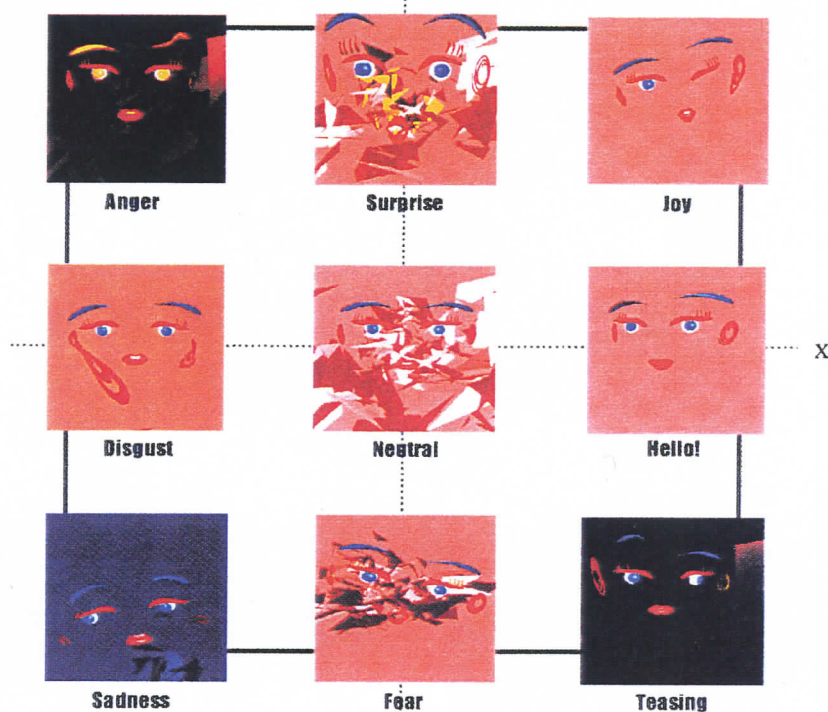
96 festival
the new temptation

PROCEEDINGS

N a o k o T o s a



M I C & M U S E



"MUSE" is a goddess. She is very expressive, has refined manners, is feminine, sensual, and erotic. MUSE's emotions are generated by a musical grammar. For example (joy --- rising musical scale, anger --- vigoroso, sadness --- volante, disgust --- discord, teasing --- scherzando, fear --- pesante) People can communicate with MUSE in an improvisational manner by means of a musical installation.

emerges in which time is suspended between abstraction and reality.

Conclusion

The future is contained in the present . . .
- La Place

Hickory Dickory Dock highlights the constraints that Western temporal perspectives place on the design of interactive multimedia computer programs. By exhibiting the screen designs in the storyboard as finished works of art, the installation critiques the temporal constraints of interactive computing by celebrating the principles it appears to reject.

Temporal orientation is based on our perception of distance and differentiated patterns of activity, both of which are measured in terms of abstract, metric landmarks defined by the clock and the calendar. The computer interface in interactive programs represents a different temporal order in which time must be integrated with actions and events. However, this computing environment differs from the 3-D world of tangible objects because in the computer program, conceptual events take place in a metaphysical space.

In interactive multimedia computing, we can no longer rely on linear temporal structures that limit our perspective to sequential hierarchies and causality. Interactive multimedia computing is a medium that requires new temporal perspectives that transcend the perceptual limitations of the Western temporal order.

References

1. Robert Lawlor, *Voices of the First Day* (Rochester, VT: Inner Traditions, 1991), p. 321.
2. Lawlor, p. 321.
3. Marshall McLuhan and Bruce Powers, *The Global Village* (New York: Oxford University Press, 1989), p. 170.
4. William Friedman, *About Time: Inventing the Fourth Dimension* (Cambridge, MA: MIT Press, 1990), p. 70.
5. Helen Hindle, "Time Estimates as a Function of Distance Traveled and Relative Clarity of a Goal," *Journal of Personality*, 19, 483-501 (1951).
6. Friedman, p. 76.
7. Friedman, pp. 75-77.
8. Paul Davies, *About Time: Einstein's Unfinished Revolution* (New York: Simon and Shuster, 1995), p. 25.
9. Arata Isozaki, MA: *Space-Time in Japan*, exh. cat. (New York: Cooper-Hewitt Museum, 1978), p. 13.
10. Isozaki, p. 16.
11. Quoted in McLuhan and Powers, p. 73.
12. Paul Fraisse, *The Psychology of Time* (New York: Harper and Row, 1963), p. 368.
13. Rudolf Arnheim, *The Dynamics of Architectural Form* (Berkeley: University of California Press, 1977), p. 156.
14. Quoted in Catherine Ellis, *Aboriginal Music* (Queensland, Australia: University of Queensland Press, 1985), p. 109.
15. McLuhan, p. x.
16. Isozaki, p. 14.
17. Ellis, p.

Itosa,nakatsul@mic.atr.co.jp

Naoko Tosa & Ryohei Nakatsu (.jp)

ATR Media Integration &
Communications Research Laboratories

FOR INTERACTIVE FUTURE MOVIE :BODY COMMUNICATION ACTOR, "MIC" & FEELING IMPROVISATION ACTOR, "MUSE"

Abstract

Artist adopt artificial life techniques as a method for artistic expression. Besides this use, it is possible that the medium itself becomes a product concept. Marshall McLuhan's principle that "the medium is the message" does not emphasize sound and image content, but instead draws a link with the technical nature of future media that will break the chains reality has with equivalent symbols. As technical standards rapidly improve, reality as it stands now is becoming alienated from our lives. As we create a virtual life that is nothing short of an artificial life, and communicate with this life itself, we have to ask where our future is leading us.

1. Introduction

In this paper, we address the issues of communication and esthetics of artificial life that possess "human form" in modern society, both from artistic and engineering standpoints. From the standpoint of an image maker, artists seek images that can be touched physically as well as emotionally. This is not, interactive art relying on equipment of the past. Instead, it is interactive art based on communication and on creatures that have a real ability to participate in an interactive process. From an engineering standpoint, researchers have long dreamed of producing human-like robots or computer agents that can communicate with humans in a really human-like way. As it has been recognized that the non-verbal aspect of communications, such as emotion based communications, plays a very important role in our daily life, we have come to the conclusion that if we want to create life-like characters, we have to develop non-verbal communication technologies.

2. Neuro-Baby

Based on the above considerations, one of the authors began a study to create "Neuro-Baby"(NB), a baby-like character that can understand and respond to the emotions of humans. Based on the experiences of developing the early version of NB, we started the development of a revised version, "MIC & MUSE." The basic improvements in "MIC & MUSE" are the following.

2.1 Enriched characteristics and interactions

In the original form, NB had only one visualized figure of a baby. It could recognize emotions of humans and respond to them. Emotion communication, however, is only one aspect of non-verbal communication. In our present study, therefore, we included another kind of non-verbal communication: communication based on music. In addition to "MIC," which is an emotion communication character, we have created "MUSE," which has the capability of musical communication.

2.2 Improvement of non-verbal communication technology

Non-verbal communication technology has been improved to achieve context-independent and speaker-independent emotion recognition. This technology was also applied to the recognition of musical sounds. Details of emotion recognition technology will be stated in Section 4.

3. Design of 11 MIC & MUSE"

3.1 Personality of the Characters

"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. He is the quintessential comic character. "MUSE" is a goddess. She has beautiful western looks, is very expressive, has refined manners, is feminine, sensual, and erotic; these are the attractive features of a modern woman.

3.2 Emotion

How many and what kinds of emotional expressions are to be treated are both interesting and difficult issues. The following are some of examples of emotional expressions treated in several papers:

- anger, sadness, happiness, cheerfulness
- neutrality, joy, boredom, sadness, anger, fear, indignation
- anger, fear, sadness, joy, disgust
- neutral, happiness, sadness, anger, fear, boredom, disgust
- fear, anger, sadness, happiness

In our previous study, we treated four emotional state. Based on the experiences of demonstrating our first version NB to a variety of people and based on the consideration that with an increasing number of emotional states the interaction between NB and humans becomes richer, in this study we have selected seven emotional states.

- (1) MIC recognizes the following seven emotions from intonations in the human voice. An arrow (—>) indicates how to make intonations. The physical form of intonations is called prosody, and how to treat prosody will be stated in Section 4.
- 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, unexpected) —> 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

- (2) MUSE's emotions are generated by a musical grammar (we use moods of the melody and resume of piano)
- Joy —> rising musical scale, elevated, allegro
 - Anger —> vigoroso, 3 times same sound (repetitious)
 - Surprise —> several times same sound (repetitious)
 - Sadness —> falling musical scale, volante
 - Disgust —> dissonant sound, discord
 - Teasing —> scherzando
 - Fear —> pesante

3.3 Communication

In most cases, the content for 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. (1) People use a microphone when communicating with MIC. For example, if one whistles, MIC's feeling will be positive and he responds with excitement. If the speaker's voice is low and strong, MIC's feeling will be bad and he gets angry.

(2) People can communicate with MUSE in an improvisational manner via a musical installation.

4. Processing

In this section, the recognition of emotions included in speech are described. Also, the generation process of Neuro Baby's reactions, which correspond to the emotion received by it, will be explained.

4.1 Feature extraction

(1) Speech feature calculation

Two kinds of features are used in emotion recognition. One is a phonetic feature and the other is a prosodic feature. As the phonetic feature, LPC (linear predictive coding) parameters, which are typical speech feature parameters and often used for speech recognition, are adopted. The prosodic feature, on the other hand, consists of three factors: amplitude structure, temporal structure and pitch structure. For the features expressing amplitude structure and pitch structure, speech power and pitch parameters are used, each of which can be obtained in the process of LPC analysis. Also, a delta LPC parameter that is calculated from LPC parameters and expresses a time variable feature of the speech spectrum are adopted, because this parameter corresponds to temporal structure. Speech feature calculation is carried out in the following way: Analog speech is first transformed into digital speech by passing it through a 6 kHz low-pass filter and then is fed into an A/D converter that has a sampling rate of 11 KHz and an accuracy of 16 bits. The digitized speech is then arranged into a series of frames, each of which is a set of 256 consecutive sampled data points. For each of these frames, LPC analysis is carried out in real time and the following feature parameters are obtained. The sequence of this feature vector is fed into the speech period extraction stage.

(2) Extraction of speech period

In this stage, the period where speech exists is distinguished, and it is extracted based on the information of speech power. The extraction process is as follows. Speech power is compared with a predetermined threshold value PTH; if the input speech power exceeds this threshold value for a few consecutive frames, it is decided that the speech is uttered. After the beginning of the speech period, the input speech power is also compared with the PTH value; if the speech power is continuously below PTH for another few consecutive frames, it is decided that the speech no longer exists. By the above processing, the speech period is extracted from the whole data input.

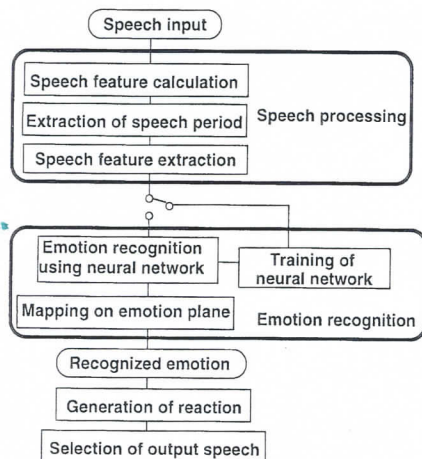


Fig. 1 Blockdiagram of the processing flow

(3) Speech feature extraction

For the extracted speech period, ten frames are extracted, each of which is situated periodically in the whole speech period, keeping the same distance from adjacent frames.

Let these ten frames be expressed as f_1, f_2, \dots, f_{10} .

The feature parameters of these ten frames are collected and the output speech features are determined as a 150 (15x10) dimensional feature vector. This feature vector is expressed as $F = (F_1, F_2, \dots, F_{10})$

where F_i is a vector of the fifteen feature parameters corresponding to the frame f_i . This feature vector F is then used as input to the emotion recognition stage.

4.3 Emotion recognition

As for recognition algorithms, there are two major methods: neural networks and HMMs (Hidden Markov models). Although the HMM approach is main stream in speech recognition, we have adopted the neural network approach here

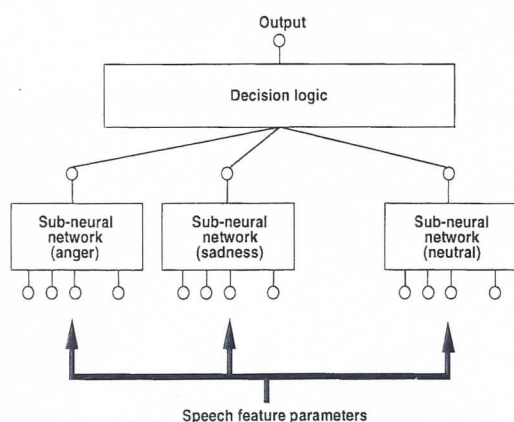


Fig. 2 Configuration of emotion recognition part

because of the following reasons: a. Content independent emotion recognition is our target. Although HMMs are suitable in content recognition, neural networks are considered to be better algorithms. b. HMMs are suitable where the structure of the recognition object is clear to some extent. As phoneme structures are the basis for the content of words or sentences, HMMs are appropriate. In the case of emotion recognition, however, the structure of the emotion feature is not clear. Therefore, a neural network approach is more suitable.

(1) Configuration of the neural network Configuration of the neural network for emotion recognition is shown in Fig.2. This network is a combination of eight sub-networks and the decision logic stage combines the outputs of the eight sub-networks and outputs the final recognition result. Each of these eight sub-networks is tuned to recognize one of seven emotions (anger, sadness, happiness, fear, surprise, disgust, and tease) and neutral emotion. The construction of each sub-network is as follows.

Basically, each sub-network has the same network architecture. It is a three layered neural network with one 150 input nodes corresponding to the dimension of speech features, 20 to 30 intermediate nodes and 1 output node. The reason we have adopted this architecture is based on the consideration that the difficulties of recognizing emotions varies depending on the specific emotion. Thus, it is easier to prepare a specific neural network for each emotion and tune each network depending on the characteristics of each emotion to be recognized. This basic consideration was confirmed by carrying out preliminary recognition experiments. Although negative emotions such as anger or sadness are rather easy to recognize, positive emotions such as happiness are difficult to recognize.

Thus, the detailed architecture of the networks, such as the number of inter-mediate nodes, differs depending on the specific emotion.

As it is necessary to combine the outputs of these eight sub-networks and decide the total output of the emotion recognition stage, a final decision logic is prepared. The details of the decision logic will be described later.

(2) Neural network training

For the recognition of emotions, it is necessary to train each of the sub-networks. As our target is the speaker-independent and content-independent emotion recognition, 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 tease

Utterances: Each speaker uttered 100 words eight times.

In each of the 8 trials, he/she uttered words using different emotional expressions. Thus, a total of 800 utterances for each speaker were obtained as training data. Eight subnetworks were trained using these utterances.

(3) Emotion recognition by a neural network

In the emotion recognition phase, speech feature parameters extracted in the speech processing part are simultaneously fed into the eight sub-networks. Eight values, $V = (v_1, v_2, \dots, v_8)$, are obtained as the result of emotion recognition. To evaluate the performance of emotion recognition, we carried out a small emotion recognition experiment using sub-networks trained by the above process. By the simple decision logic of selecting the sub-network with the highest output value, an emotion recognition of about 60% was obtained.

(4) Mapping on an emotion plane

As described above, the output of the emotion recognition network is a vector $V = (v_1, v_2, \dots, v_8)$ and the final recognition result should be obtained based on V .

To carry out the mapping from V onto E . The simple decision logic shown below is adopted here.

Let m_1 and m_2 be the first and second maximum values among v_1, v_2, \dots, v_8 , and also let (x_{m1}, y_{m1}) , (x_{m2}, y_{m2}) be the emotion positions corresponding to m_1 and m_2 , respectively. The final emotion position (x, y) is calculated by

$$x = c \cdot x_{m1} + (1-c) \cdot x_{m2}, \quad y = c \cdot y_{m1} + (1-c) \cdot y_{m2}$$

(c : constant value).

Through the processes of 4.1 to 4.3, the emotion recognition of MIC is carried out. These recognition processes are mainly designed for emotion recognition, but for the present study is also applied to the musical sound recognition of MUSE.

4.4 Generation of reaction and selection of output speech

(1) The structure of animation

There are four emotional planes, all of which use the same x, y data. a. Plane "a" generates facial animation by choosing the 3 key frames A_1, A_2 and A_3 which are closest to the (x, y) data point. The computation of a weighted mean frame A is done as follows. Let a_1, a_2 , and a_3 be the distances between A and A_1, A_2, A_3 .

Then, A is calculated by

$$A = (A_1 / a_1 + A_2 / a_2 + A_3 / a_3) / (1 / a_1 + 1 / a_2 + 1 / a_3)$$

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.