Proceedings of the

International Conference on Multimedia Computing and Systems

June 17 – 23, 1996 **Hiroshima, Japan**

Sponsored by
The IEEE Computer Society Technical Committee on Multimedia Computing

IEEE Computer Society Press
Los Alamitos, California

Washington • Brussels • Tokyo

Life-Like Communication Agent — Emotion Sensing Character "MIC" and Feeling Session Character "MUSE"

Naoko Tosa, Ryohei Nakatsu
ATR Media Integration & Communications Research Laboratories
Seika-cho Soraku-gun Kyoto 619-02 Japan
{tosa,nakatsu}@mic.atr.co.jp

Abstract

Why do people, regardless of age or gender, have an affinity for objects manifested in the human form? From the earthen figures of ancient times to mechanical dolls, teddy bears and robots, is it not true that man has conceived such objects in his imagination, then formed attachments and transferred emotions to them? In this paper, we address the issues of communication and esthetics of artificial life that possess this "human form" in modern society, both from artistic and engineering standpoints. An example is presented in which emotions are interpreted from human voices and emotional responses are triggered within the interactive setting of "MIC & MUSE."

1. Introduction

1.1 From an artistic standpoint

From the standpoint of an image maker, we concentrate on a transformation of emotional reality that keeps pace with the change of times. we seek images that can be touched physically as well as emotionally, this being the real potential that we desire in interactive art.

This is not, interactive art relying on equipment of the past, which simply transforms patterns or uses simple instruments. Instead, it is interactive art based on communication and on creatures that have a real ability to participate in an interactive process. Moreover, we think that by selecting "human" a creature with which we realistically communicate the most we establish a condition that demands a creative character from a creature.

1.2 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. Most research so far has concentrated on the verbal aspect of communication, and communication technologies such as speech recognition or gesture recognition have been studied. As technologies have advanced, however, it has been recognized that the non-verbal aspect of communication, such as emotion based communications, plays a very important role in our daily life. Therefore, we have come to the conclusion that if we want to create life-like characters, we have to develop non-verbal communication technologies. These are expected to give characters the capability of achieving heartful communication with humans by exchanging emotional messages.

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 [1]. 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.

(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)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 " 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

0-8186-7436-9/96 $5.00 © 1996 IEEE
Proceedings of MULTIMEDIA '96

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:

a. anger, sadness, happiness, cheerfulness[1]
b. neutrality, joy, boredom, sadness, anger, fear, indignation[3]
c. anger, fear, sadness, joy, disgust[4]
d. neutral, happiness, sadness, anger, fear, boredom, disgust [5]
e. fear, anger, sadness, happiness[6]

In our previous study, we treated four emotional states[1]. 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.

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, unexpected) ---> 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

(2) MUSE's emotions are generated by a musical grammar (we use moods of the melody and resume of piano)

a. Joy	---> rising musical scale, elevated, allegro
b. Anger	---> vigoroso, 3 times same sound (repetitious)
c. Surprise	---> several times same sound (repetitious)
d. Sadness	---> falling musical scale, volante
e. Disgust	---> dissonant sound, discord
f. Teasing	---> scherzando
g. 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, principles and details for 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 Basic principle

NB has advanced from its original version through several stages[1][2] to "MIC & MUSE." In our present research, we tried to realize higher level processing which achieves more sophisticated interactions between NB and humans. For this purpose, we have considered and emphasized the following issues.

(1) Precise speech processing

What kinds of speech features are to be adopted for the recognition of emotions is another important and difficult issue. One standpoint is that the features to be used for emotion recognition should be totally different from those used for speech recognition, because in the case of emotion recognition, the prosodic features of speech play a more significant role than phonetic features such as speech spectrums. There is another standpoint that the phonetic features are as important as the prosodic features, because prosodic features and phonetic features are tightly combined when uttering speech ,and it is impossible for us to express our emotions by controlling only prosodic features. In our study, therefore a combination of two kinds of features is considered: one is the feature expressing phonetic character-istics of speech and the other is that expressing prosodic characteristics.

(2) Speaker-independent and content-independent emotion recognition

Speaker independence is an important aspect of speech/emotion recognition. From a pragmatic standpoint, 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. Another point is that 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 for the same words or sentences; this is the key to rich and sensitive communications among people. In our study, therefore, by adopting a neural network architecture and by introducing a training stage that use a large number of training utterances, we have developed a speaker-independent and content-independent emotion recognition system.

Figure 1 illustrates a block diagram of the processing flow. The process mainly consists of three parts: speech processing, emotion recognition and generation of reactions.

In the following sections, the detail of each process and the system configuration for carrying out the emotion recognition process is described.

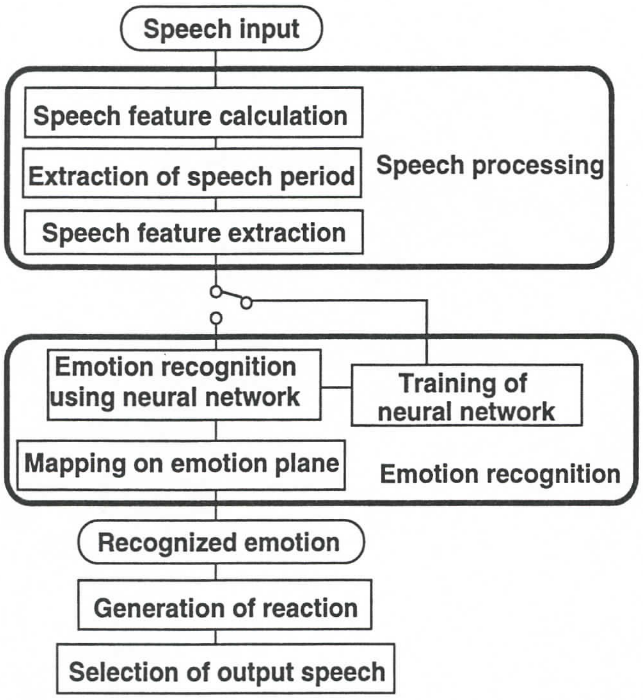
Speech input
Speech feature calculation
Extraction of speech period
Speech feature extraction
Speech processing
Emotion recognition using neural network
Training of neural network
Mapping on emotion plane
Emotion recognition
Recognized emotion
Generation of reaction
Selection of output speech

Fig. 1 Blockdiagram of the prcessing flow

4.2 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 [7], 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.

Speech power: $\mathbf{P}$
Pitch: $\mathbf{p}$

LPC parameters: $\mathbf{c}_1, \mathbf{c}_{2, \ldots}, \mathbf{c}_{12}$

Delta LPC parameter: $\mathbf{d}$

Thus for the t-th frame, the obtained feature parameters can be expressed by $F_t = (P_t, p_t, d_t, c_{1t}, c_{2t}, \ldots, c_{12t})$

The sequence of this feature vector is fed into the speech period extraction stage.

(2) Extraction of speech period

Figure 2 illustrates both the processes of speech period extraction and speech feature extraction. 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.

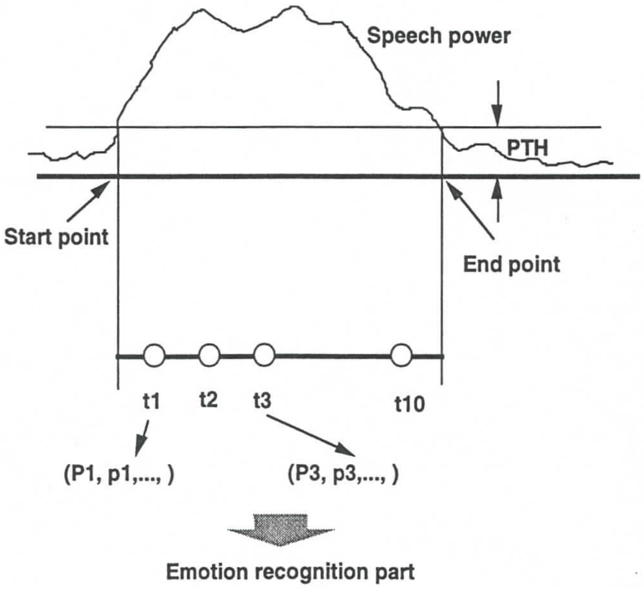

Fig. 2 Speech period exptraction and speech fuature extraction

(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 $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{10}$.

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 $\mathbf{F}_v = (\mathbf{F}_1, \mathbf{F}_{2, \ldots\ldots} \mathbf{F}_{10})$,

where $\mathbf{F}_i$ is a vector of the fifteen feature parameters corresponding to the frame $\mathbf{f}_i$. This feature vector $\mathbf{F}_v$ is then used as input to the emotion recognition stage.

4.3 Emotion recognition

Recognizing emotions is a difficult task. The main reason is that even if human babies obtain the ability of emotion extraction from speech earlier than that of meaning extraction, adults mainly rely on meaning recognition in our daily communication, especially in business commu-nication. This is why speech recognition research has long treated emotions contained in speech as just fluctuations or noise. What makes the situation more complicated is that emotional expressions are intertwined with the meaning of speech, consciously or unconsciously. In the unconscious case, context rather than the emotional feature itself plays a more important role. This means that the intensity of emotional expression varies dramatically depending on the situation. Of course, our final target is to recognize emotions in speech even if emotional expression is unconsciously mixed with the meaning of speech. For the time being, however, this is out of our research target for the above reasons. Therefore, the strategy adopted here is to treat speech intentionally uttered to contain specific emotional expressions ,rather than speech with unconscious emotion expressions. 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 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.3.

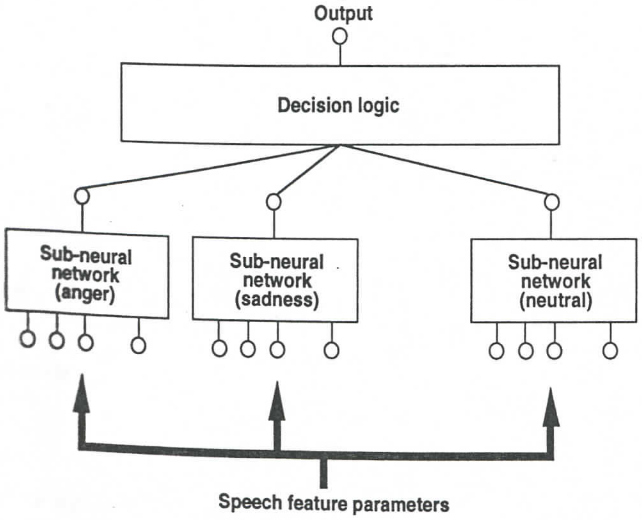
Output
Decision logic
Sub-neural network (anger)
Sub-neural network (sadness)
Sub-neural network (neutral)
Speech feature parameters

Fig. 3 Configuration of emotion recognition part

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 (Fig. 4).

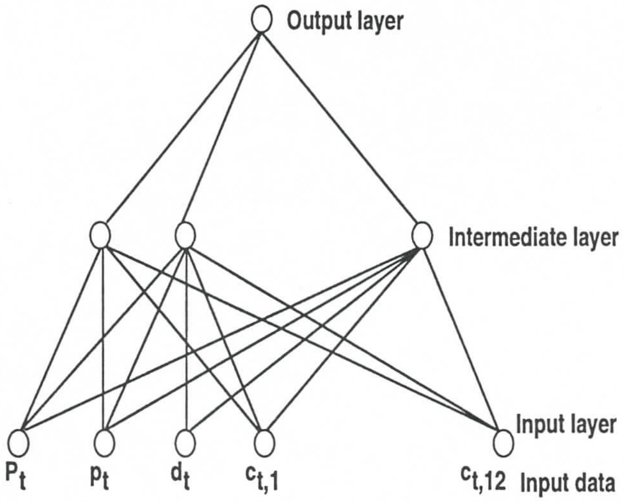

Fig. 4 Configuration of a sub-network

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 experi-ments. 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 sub-networks 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=(v1, v2, ..., v8), 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=(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 our present study we rearranged the positions of the eight emotions on the emotion plane E as shown in Figs. 7 and 8.

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

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

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

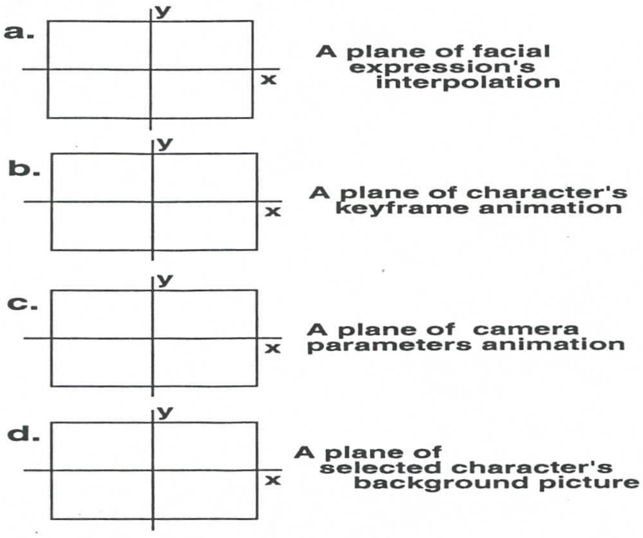
a.
y
x
A plane of facial expression's interpolation
b.
y
x
A plane of character's keyframe animation
c.
y
x
A plane of camera parameters animation
d.
y
x
A plane of selected character's background picture

Fig. 5 The structure of animation.

There are four emotional planes, all of which use the same x,y data (Fig. 5).

a. Plane "a" generates facial animation by choosing the 3 key frames A1, A2 and A3 which are closest to the (x,y) data point. The computation of a weighted mean frame A is done as follows. Let a1, a2, and a3 be the distances between A and A1, A2, A3 as shown in Fig. 6.

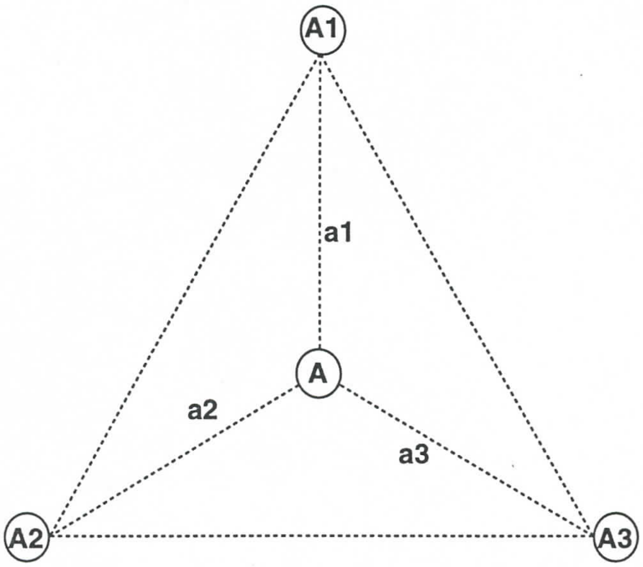

Fig. 6 Computation of a weighted mean value

Then, A is calculated by

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

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.

(2) Selection of output speech

This is a mapping from the (x,y) data points of the emotional plane to 200 sampled speech utterances, and one of the utterances is selected as the output speech. A personal computer is used to play the selected sounds.

4.5 Reaction of the characters.

Reactions of MIC & MUSE were carefully designed and were visualized using computer graphics. Several examples of emotional expressions by MIC are shown in Fig. 7. Several examples of emotional expressions by MUSE are shown if Fig. 8.

4.6 System configuration

Figure 9 illustrates the system configuration along with specific processing assigned to each computer. Two workstations running in parallel to realize real-time interactions are the key to this system.

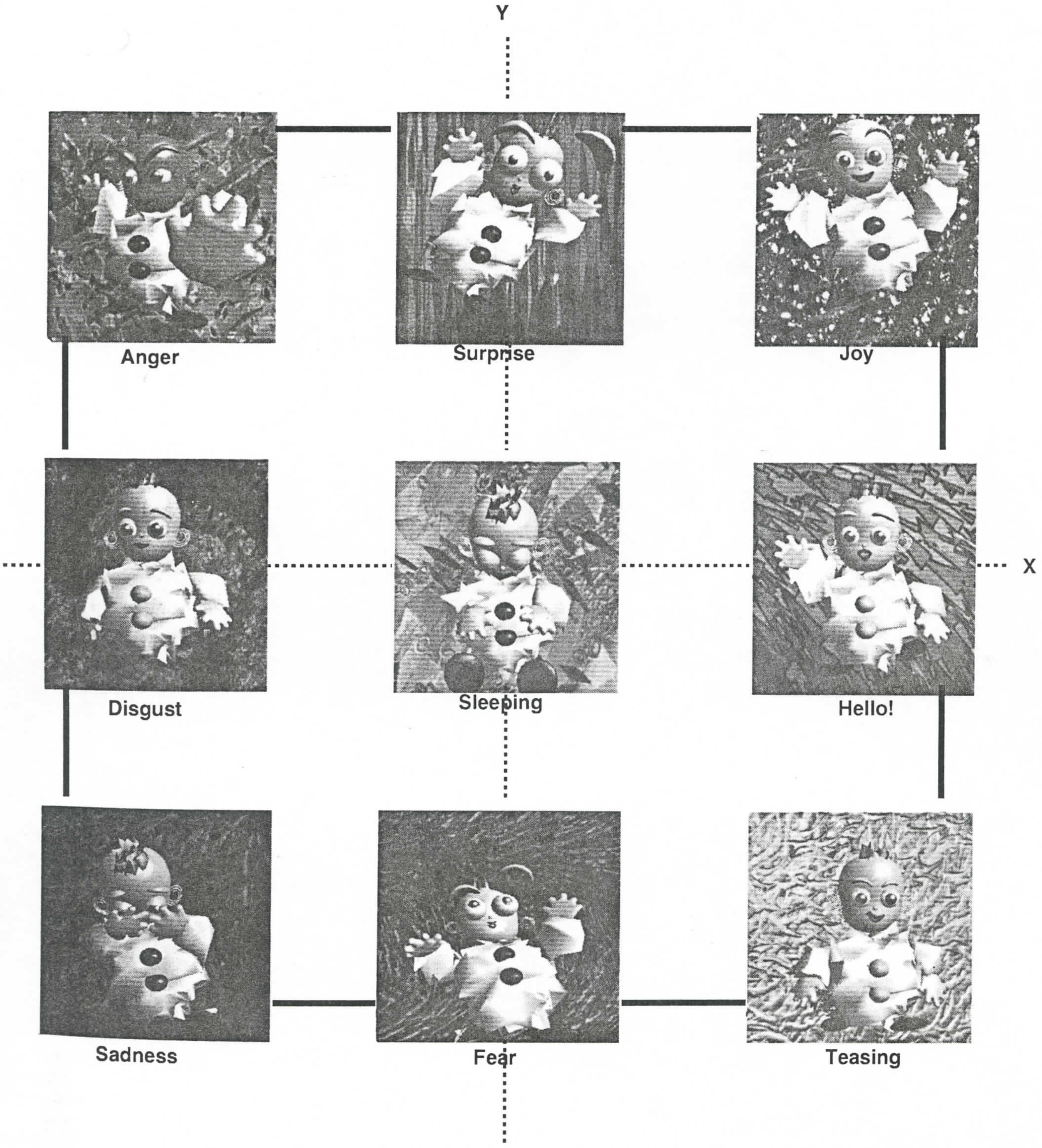

Fig. 7 MIC's emotional expression

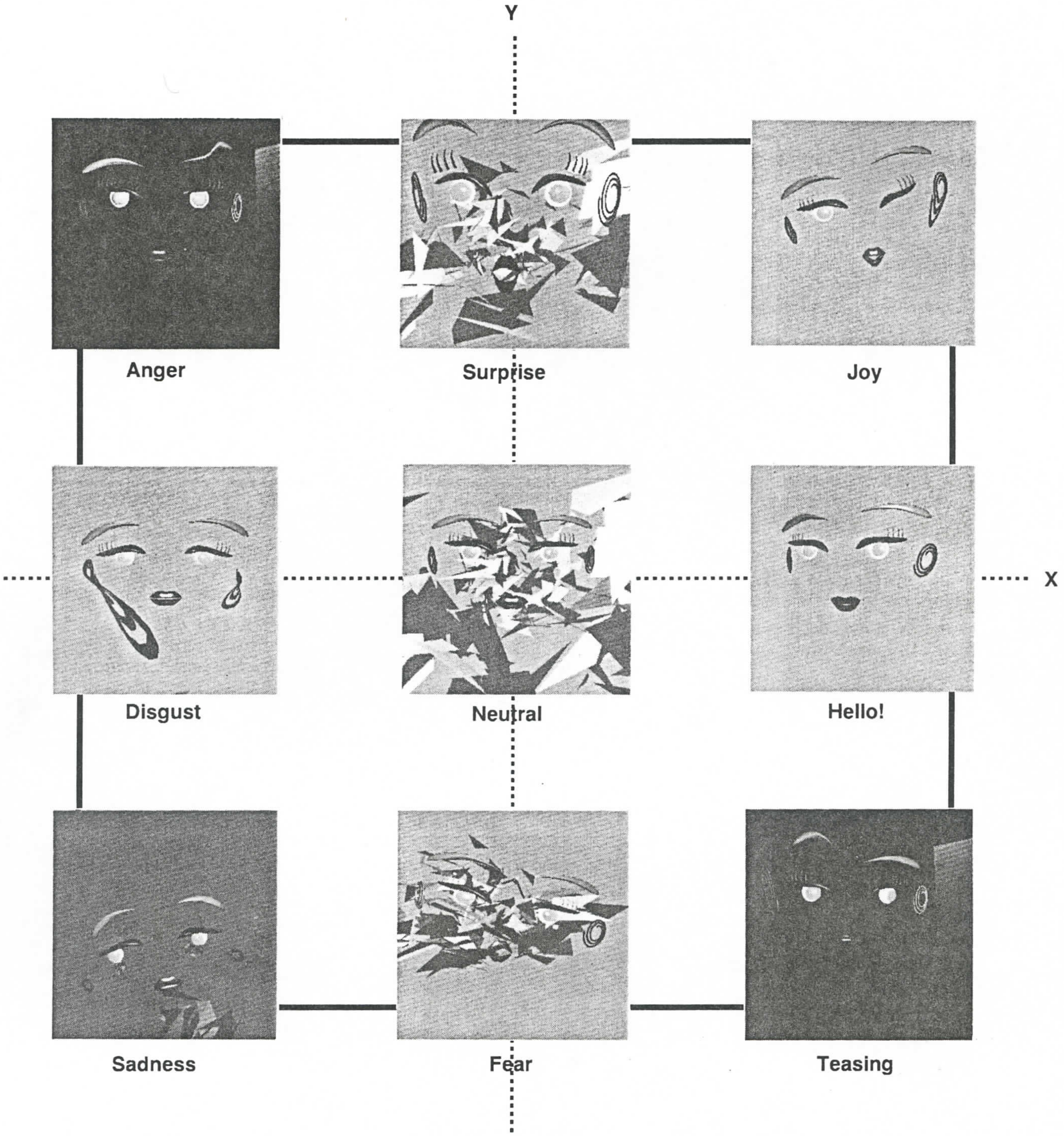

Fig. 8 MUSE's emotional expression

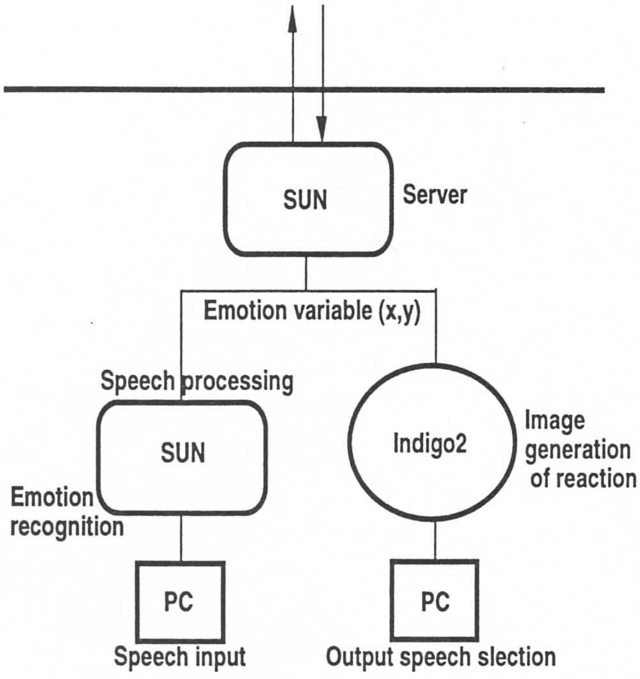

Fig. 9 System Configration

6. Future work

Real human emotions are much more complex and detailed than represented by the simple model introduced here. Therefore, we would like to investigate how to further improve our model. In particular, we would like to extend the model to include the following emotions.
a. Shame (embarrassment);
b. Like (love, dear, intense yearning, desire);
c. Unpleasantness (hate, detestation, dislike, melancholy, pain, dispirit);
d.Offensiveness(impatience, irritation, tension, impression).

Further study to improve emotion recognition is also necessary. A higher emotion recognition rate is expected by preparing more speakers and word/sentence utterances and by designing more sophisticated multiple layered neural networks. On the other hand, it is necessary to develop recognition algorithms dedicated to musical sound recognition of rhythm and melody.

As for the characteristics MIC & MUSE, it is desirable to design a cyberspace where the characters will live and to develop methods that will allow communication between the characters within the cyberspace and interaction with humans.

7. Conclusion

In this paper, a new form of artificial life-like characters, called "MIC& MUSE", are introduced. The basic concept and the details of these life-like characters are discussed both from artistic and engineering standpoints. This research was carried out by a collaboration between an artist and a researcher, where the artist first proposed the basic concept and requested for necessary algorithm and the researcher clarified the specification of the algorithm and realized it on a computer. We think this kind of collaboration is a key to the success of the research.

These life-like characters or "androids" will unravel a new point of view in a new direction which allows the blending of art, computer science, psychology, and philosophy in a kind of novel research on realistic human expression.

References

[1] N. Tosa, et al., "Neuro-Character," AAAI '94 Workshop, AI and A-Life and Entertainment (1994).
[2] N. Tosa, et al., "Network Neuro-Baby with robotics hand," Symbiosis of Human and Artifact, Elsevier Science B.V. (1995).
[3] S. Mozziconacci, "Pitch variations and emotions in speech," ICPhS 95 Vol. 1, p. 178 (1995).
[4] K. R. Scherer, "How emotion is expressed in speech and singing," ICPhS 95, Vol. 3, p. 90 (1995).
[5] G. Klasmeyer and W. F. Sendlmeier, "Objective voice parameters to characterize the emotional content in speech," ICPhS 95, Vol. 1, p. 182 (1995)
[6] S. McGilloway, R. Cowie, and E. D. Cowie, "Prosodic signs of emotion in speech: preliminary results from a new technique for automatic statistical analysis," ICPhS, Vol. 1, p. 250 (1995).
[7] J. D. Markel and A. H. Gray, "Linear prediction of speech," Springer-Verlag (1976).