ltosa,nakatsul@mic.atr.co.jp

Naoko Tosa & Ryohei Nakatsu (.jp)

ATR Media Integration &
CommunicationsResearch 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 modem 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:
a. anger, sadness, happiness, cheerfulness
b. neutrality, joy, boredom, sadness, anger, fear, indignation
c. anger, fear, sadness, joy, disgust
d. neutral, happiness, sadness, anger, fear, boredom, disgust
e. 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 fon-n 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

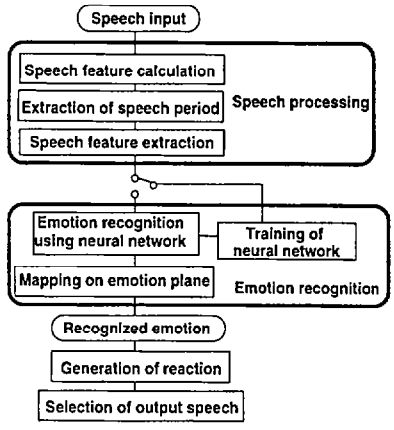

Fig. I Blockdiagram of the processing flow

(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
c. 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, 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 I I 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.

(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..... If 10.
The feature parameters of these ten frames are collected and the output speech features are determined as a 150 (15xlO) dimensional feature vector. This feature vector is expressed as F, = (F,,F2..... Flo)

where F i is a vector of the fifteen feature parameters corresponding to the frame f 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

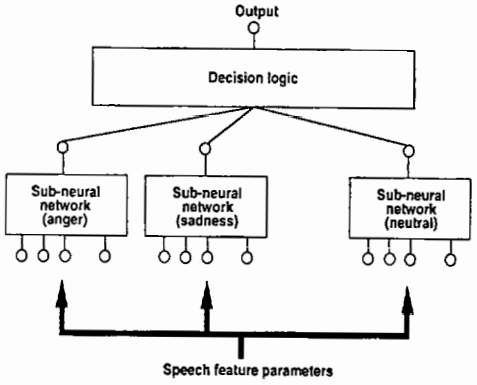

Fig. 2 Configuration of emotion recognition part

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 I 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 speakerindependent and content-independent emotion recognition, the following utterances were prepared for the training process.
Words: I 00 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=(vl, 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=(v I, v2, ..., vg) 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 I and m2 be the first and second maximum values among v I, v2..... v8, and also let (xml, yml), (xm2, ym2) be the emotion positions corresponding to m I and m2, respectively. The final emotion position (x, y) is calculated by

x = c X x , +(1-C)X X .2, y = CXy -1 + (1 -C) X Y .2
(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 Al, 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 al, a2, and a3 be the distances between A and Al, A2, A3.
Then, A is calculated by

A=(A 1 /a 1 + A 2/a 2 +A 3/a 3) /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. 3. Several examples of emotional expressions by MUSE are shown in Fig. 4.

4.6 System configuration

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.

5. Conclusion

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.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 artificial life 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.