芸術としてのソフトウエアロボット

Life-like,Autonomous Character "MIC" & Feeling Session Character "MUSE"

Naoko Tosa
ATR Media Integration & Communications Research Laboratories
Seika-cho Soraku-gun Kyoto, Japan
Phone: +81 774 95 1427
tosa@media.kyoto-u.ac.jp
http://www.mic.atr.co.jp/~tosa/

Ryohei Nakatsu
ATR Media Integration & Communications Research Laboratories
Seika-cho Soraku-gun Kyoto, Japan
Phone: +81 774 95 1400
nakatsu@mic.atr.co.jp
http://www.mic.atr.co.jp/~nakatsu/

概要

　人間が、老若男女問わず「人型」の物を好むのはなぜだろうか？昔から、土偶、からくり人形、ぬいぐるみ、ロボットにいたるまで、人は自分と同じ形を作り、愛着を覚え、感情移入しているではないか。 本論では、現代社会におけるこの「人型」ラベルを持つ人工生命の美学とコミュニケーションにアートとエンジニアリングの立場から論点をあてる。自己の分身であり自分に最も近い他者をキーワードに、話し掛ける人の声から音声認識により感情抽出をし、インタラクティブに答えるヒューマノイド-エージェント「 MIC 」と「MUSE」との新しいコミュニケーションの形態を紹介する。

Table of Contents

1　はじめに

1.1　アーティストの立場から

　私が映像に求めた現実感とは、映像に触れるという行為と、自分が創造するイメージが頭脳を持ち能動的に自己生成し、自己判断でき対応するシステムを求めた。コンピュータベースのインタラクティブ映像にその可能性を感じた。しかし、従来のパターン化した装置の様なインタラクティブ映像ではなく、インタラクティブの機能をリアルな対応のできる生物と、そのコミュニケーションに置き換えて考える。そして私達にとって身近で一番コミュニケーションをする生物”人間”を選び、心理状態や、感情表現、性格付けや知能、行為とコミュニケーションをテクノロジーを用いてどこまでリアルに個性を持って表現できるのかという興味のもとに研究を始めた。

<-- Table of Contents

2　Neuro-Baby

　筆者の一人が始めたニューロベイビーという、人工知能の赤ちゃんの顔が、人の声の抑揚から感情的リアクションをするというインタラクティブアートの作品を基に、"MIC" & "MUSE"という音声や音楽から感情認識して、リアクションを行うというキャラクターを新しいノンバーバルコミュニケーションの研究として、開発した。

<-- Table of Contents

3　"MIC" & "MUSE"のデザイン

3.1 性格設計

MIC（ミック）は、ニューロベイビーを基に、さらに知識を習得して我々現代人と共に、現在進行形で成長していく人間の子供をモデルにしたパーソナリティーを持つ仮想の男の子のキャラクターである。容貌は人が話しかけたくなるようなかわいらしさを持ち、お茶目で生意気であるが、憎めない性格をしている。いわゆる漫画の主人公的キャラクターである。　"MUSE"は、女神のキャラクターである。容貌は西洋的な女性の美しさを持ち、表現豊かで優雅にふるまい且つ、賢く強いといった現代的な女性美の象徴である。

３.２感情

MICは、下記のような８つの感情を＊に指示している声の抑揚から認識できる。 例えば、だれも話しかけないと居眠りをし、だれかが話しかけると、機嫌が良いときは、「こんにちは」、悪い時は、「バイバイ」と返事をする。低い声でばかにすると怒り、からかうと逆立ちをする。口笛を吹いてあげると、エキサイティングしてジャンプをし、人間が不機嫌な顔をして咳払いをすると、悲しくなって手で顔を覆い後ろをむいてしまう。たまに愛想をつかし人間に愚痴をこぼす。

```
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
h. neutral------*normal voice
```

MUSE（ミューズ）は、人間と音楽のメロディとリズムでコミュニケーションをとる。ミックとは言葉で記号的な感情表現のコミュニケーションを行うが、ミューズとは感情表現する言葉に満たない、もっと微妙な気分のコミュニケーションができる。彼女はピアノで我々に話しかけるので、人間は下記の感情を表わすミュージカルグラマー（楽典）を用いて彼女と即興的なフィーリングセッションを楽しむ。

```
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
```

[<-- Table of Contents]

4　処理内容

本章では、音声に含まれる感情の認識の基本方針と処理の内容について述べる。また、認識結果の感情に対応したMIC&MUSEの反応パターンの生成法についても述べる。

4.1基本方針

MIC&MUSEは、その原形であるニューロベビーから何段階かの変遷を経て現在の形に成長した[1][2]。現在のMIC&MUSEでは、人間とエージェント間のより洗練されたインタラクションを実現するために必要な技術を開発することを狙っている。このために、以下に述べる項目について検討した。

(1) リアルタイム処理

人間とエージェント間のリアルタイムのインタラクションを実現するためには、音声特徴抽出、感情認識、応答生成の各部分をリアルタイムで動作させる必要がある。本システムでは、音声処理およびシステム構成の部分の構成をリアルタイム向けにすることにより、リアルタイム処理を実現している。

(2) 種々の感情の取り扱い

何種類のかつどのような感情を認識対象とするかは、重要なかつ困難な問題である。これまでに、感情を扱った種々の研究がなされているが、そこで扱われている感情の例を以下に示す。

a. 怒り、悲しみ、幸せ、喜び[1]
b. 普通、喜び、退屈、悲しみ、怒り、恐れ、からかい[3]
c. 怒り、恐れ、悲しみ、喜び、軽蔑[4]
d. 普通、幸せ、悲しみ、怒り、恐れ、退屈、軽蔑[5]
e. 恐れ、怒り、悲しみ、幸せ[6]

我々の以前の研究では、 a に示してある４つの感情を扱った[1]。最初のバージョンのニューロベビーを各種の機会に展示した経験、および感情の種類が増えることが人間とエージェントのインタラクションの内容の豊かさにつながるという考察の上に立って、本研究では、２．２で述べた７つの感情＋普通、計８つの感情を扱うこととした。

(3) 精密な音声処理

感情認識のためにどのような音声特徴量を用いるかは重要な問題点である。１つの考え方として、感情認識に用いる音声特徴量は音声認識に用いる音声特徴量とまったく異なるようにすべきだというものがある。これは、音声認識においては、周波数スペクトラムなどの音韻特徴が重要な役割をしているのに対し、感情認識においては、音声の韻律特徴が主要な役割をしているという一般的な考え方に基づいている。しかしながら、また別の見方がある。すなわち、音声を発声する場合は、音韻特徴と韻律特徴は分かち難く結び付いており、韻律特徴の制御のみによって感情を表現することは困難であるため、韻律特徴と同時に音韻特徴も用いるべきであるという考え方である。本研究では後者の立場に立ち、２つの特徴パラメータ、すなわち音韻特徴を表現するパラメータと韻律特徴を表現するパラメータを両方用いることにした。

(4) 不特定話者コンテキスト独立型感情認識

音声認識や感情認識において、不特定話者を扱えるというのは重要な機能である。実用的な観点からすると、話者が変わる度に大変な学習処理を行う必要があるというのは望ましくない。別の観点からすると、人間は不特定話者の声に含まれる意味内容と感情を同時に認識出来るという事実がある。また、感情認識においてはコンテキスト独立性は重要な機能である。われわれの日常の会話では、同じ単語や文章に異なった感情を乗せることがひんぱんに生じる。本研究では、

認識のためのアーキテクチャとしてニューラルネットを用いると共に、大量の学習サンプルを用いた学習処理を行うことにより、不特定話者に対応できかつコンテキスト独立型の感情認識を実現することをめざした。

図１は処理の流れのブロック図である。全体の処理は３つの部分、すなわち、音声特徴抽出部、感情認識部、および反応生成部から構成されている。音声特徴抽出部では、まず入力音声から音声特徴パラメータがリアルタイムで抽出される。次に、音声パワーを用いて音声区間の抽出が行われる。抽出された音声区間を用いて、入力音声のそれぞれに応じた音声特徴量 が決定される。この音声特徴量は、感情認識部への入力として用いられる。感情認識部では、２段階の感情認識が行われる。第１段階ではニューラルネットが用いられる。第１段階は、８つの感情のそれぞれを認識出来るよう学習された８つのニューラルネットが並列におかれた構成になっており、音声特徴抽出部の出力が８つのニューラルネットに同時に入力される。次に、第２段階として論理部があり、８つのニューラルネットの出力を論理処理して、２次元の感情平面 への写像を行う。２次元の感情平面上には８つの感情をあらかじめ適切に配置しておく。２次元の感情平面 上の認識感情の位置およびその動きに応じて、MIC&MUSEの反応、すなわち顔の表情と体の動きが生成される。これらの顔の表情および体の動きは著者の一人であるアーティストの直感と感性と経験によって注意深く事前に定めておく。反応パターンはコンピュータグラフィクスで表示されると共に適切な音声もしくは音楽が出力される。

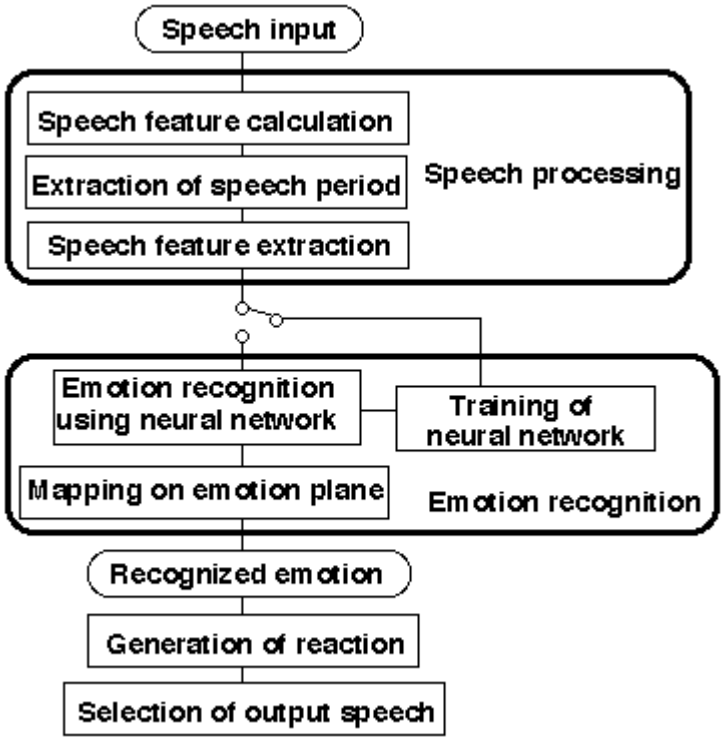

図１

以下に、それぞれの処理の詳細、および感情認識を行うためのシステム構成を述べる。

４．２　音声特徴抽出

(1) フレーム毎の特徴パラメータ抽出

感情認識のために２種類の特徴パラメータが用いられる。１つは音韻特徴を表わすパラメータであり、もう１つは韻律特徴を表わすパラメータである。音韻特徴パラメータとしてはLPCパラ

メータが用いられる[7]。LPCパラメータは、音声特徴を表現するための代表的なパラメータであり、音声認識においてよく用いられる。これに対し、韻律特徴としては３種類のパラメータが用いられる。すなわち、エネルギー、音韻特徴の時間変化、およびピッチである。エネルギーおよびピッチに対応するパラメータとしては、LPCパラメータから求められる音声パワーおよびピッチ情報を用いる。また、LPCパラメータの時間変化を表わすパラメータを時間変化パラメータとして用いる。

入力音声の音声特徴パラメータの計算は次の手順で行う。アナログ音声は、まず6 kHz の低域通過フィルタを通した後、11 KHz、16 bitでサンプリングされ、ディジタル音声にy変換される。ディジタル音声は、２５６個のサンプリング点を含むフレームの連続として表現され、各フレーム毎にLPC分析が行われ、以下の特徴パラメータが求められる。

```
音声パワー: P
ピッチ: p
LPC パラメータ: c1, c2, ..., c12
時間変化パラメータ: d
```

tフレームに関する音声特徴パラメータは以下で表現される。

```
Ft = ( Pt, pt, dt, c1t, c2t, ..., c12t)
```

このパラメータの時系列が音声区間抽出部へ送られる。

(2) 音声区間抽出

図２に音声区間抽出と音声特徴抽出の処理を示してある。音声区間抽出では、音声パワー情報を用いて以下のようにして音声区間を抽出している。音声パワーはあらかじめ定めたいき値Pthと比較される。Pthより音声パワーが大きい区間が連続してある程度続くと音声が存在すると判断される。またその後で音声パワーがいき値より小さい区間がある程度続くと音声が終了したと判断される。背景雑音が音声抽出に大きな影響を与えるため、システムが動作開始した段階で背景雑音レベルを測定し、これに基づいてPthを定めている。

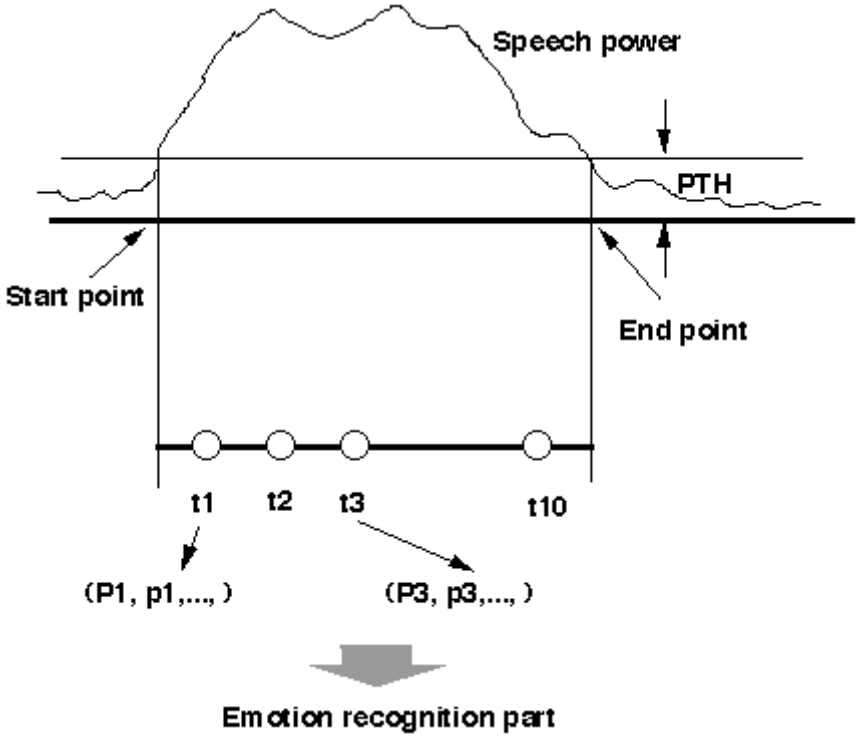

図２

(3) 音声特徴抽出

抽出された音声区間の全体から、等間隔になるように配置された１０フレームを取り出す。この抽出された１０フレームを、 f1, f2, ..., f10とする。これらの１０フレームの特徴パラメータをまとめることにより、音声の特徴量は１５０次元（１５×１０）の特徴ベクトル、

FV = (F1, F2, ..., F10)

として表現される。ここで、Fi は、iフレームの特徴パラメータである。 FVは、感情認識部への入力として用いられる。

４．３　感情認識

感情認識は困難な研究対象である。その主な原因として、人間は赤ん坊の時には感情の認識に基づいて主として行動しているが（赤ん坊は意味内容の理解を始める前に感情の認識ができるといわれている）、大人の場合は主として音声の中に含まれる意味内容の理解に基づいてコミュニケーションを行っているという点にある。このために、音声認識研究においては、長い間音声に含まれる感情の情報を単なる変動もしくは雑音として扱ってきた。さらに問題を複雑にしているのは、人間が発声する音声において意識的な場合も無意識的な場合も含めて、意味内容と感情が分かち難く結び付いている点である。特に、コンテキストは極めて重要な役割をしており、コンテキストが我々が無意識に表出し知覚する感情のレベルを左右していることが多い。このことは言い替えれば、音声に含まれる感情表現の量が状況に極めて依存していることを示している。もちろん、最終的には無意識の感情表現と意味内容表現がミックスされている場合にも感情の認識ができることが望ましいが、当面は無意識的な感情表現を認識することは上記の理由によって困難であると考えられる。したがって、ここでは無意識的な感情表現がされた音声を扱うののではなく、意識的に感情表現を伴って発声された音声を取り扱いの対象とする。

次に、認識のアルゴリズムであるが、現在音声認識で主として用いられているのはニューラルネットワークとHMM（隠れマルコフモデル）の２種類である。現在の音声認識における主流のアルゴリズムはHMMであるが、ここでは以下の理由によってニューラルネットを用いることとした。 ａ．我々の目的はコンテキスト独立型感情認識である。HMMは、意味内容の認識には適しているが、コンテキスト独立型の認識にはニューラルネットが適していると考えられる。 ｂ．HMMは、認識対象の構造がある程度明確になっている場合に適している。意味内容の認識の場合は、音韻連鎖という音声構造が事前にわかっているためにHMMが適当である。しかしながら、感情認識の場合はまだ感情の構造が明らかでないため、ニューラルネットを用いる方が適当と考えられる。

(1) ニューラルネットの構造

感情認識のためのニューラルネットの構造を図３に示す。このネットワークは８つのサブネットワークの集合とそれらのサブネットワークの出力を統合する論理部から構成されている。８つの各々のサブネットワークは８つの感情（怒り、悲しみ、喜び、恐れ、驚き、愛想をつかす、からかい、および普通 ）のそれぞれにあらかじめチューンしてある。

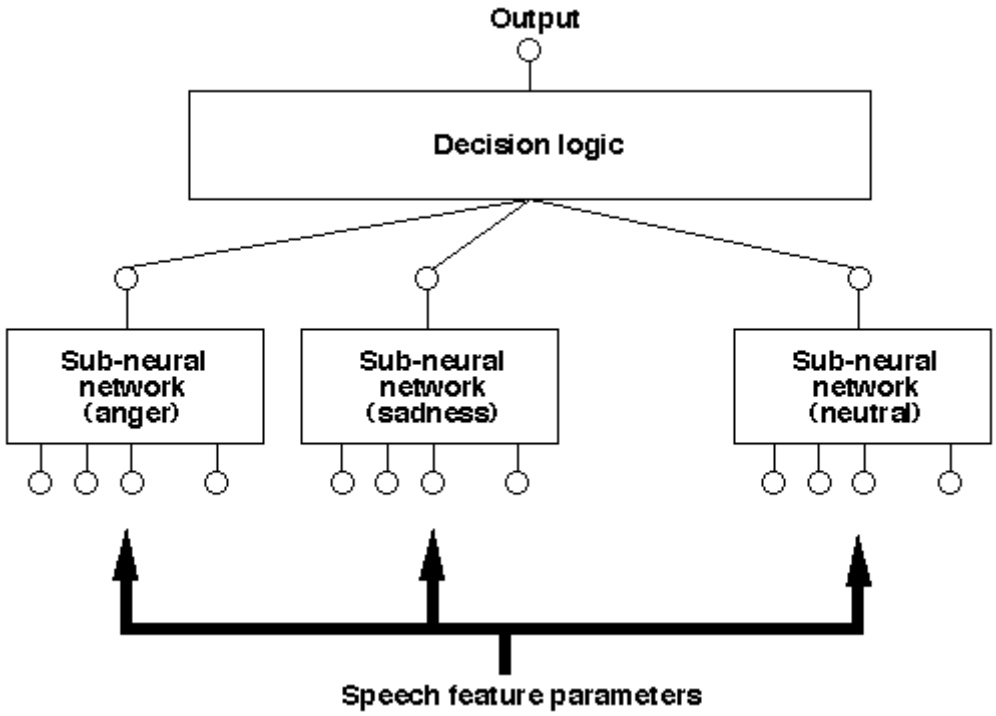

図３

各々のサブネットワークの構成を図４に示す。基本的には、これらのサブネットワークは同じ構造をしている。すなわち、３層構造であり、入力層は音声特徴量 の次元に対応した１５０個のノードよりなり、中間層は２０ないし３０のノードからなり、そして出力層は１個のノードから構成されている。感情認識の困難さは、認識すべき感情によって大きく異なっているため、１個のニューラルネットを用意するより、各々の感情に対応したニューラルネットを用意しておいて、これらをそれぞれの感情にチューンした方がいいと考えられるため、このような構造を採用した。このことは予備実験によって確かめられた。すなわち、怒り、悲しみのような負の感情が比較的認識しやすいのに対し、喜びのような正の感情の認識は困難であって、１つのニューラルネットを用いた場合は学習が収束しなかった。また、中間層のノードの数のような個々のサブネットワークの細部構造は、学習が収束しやすいように個々の感情毎に変えてある。

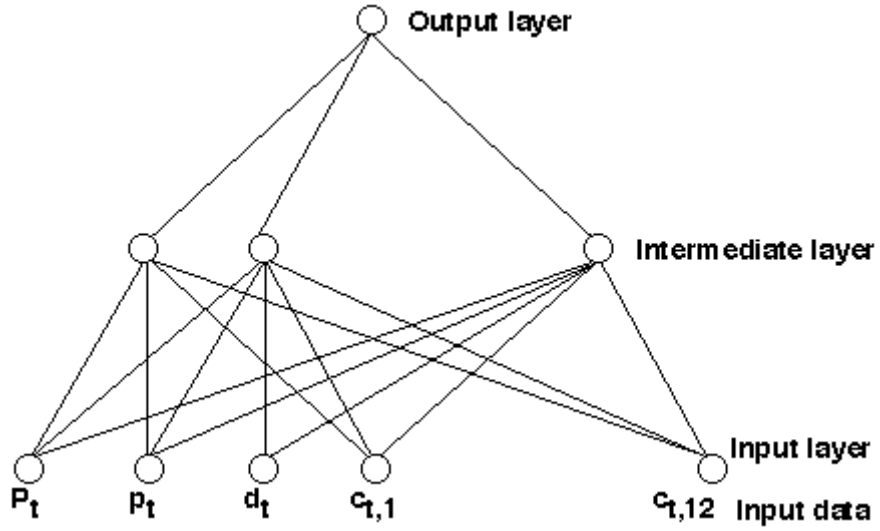

図４

８つのサブネットワークの出力を統合して最終的な感情認識結果を得るために、サブネットワークの後段に論理部をおいている。その内容については後述する。

(2) ニューラルネットの学習

感情認識を行うためには上に述べたニューラルネットをあらかじめ学習させておく必要がある。我々の目標は不特定話者、コンテキスト独立型の感情認識であるため、以下のような音声サンプルを学習データとして用意した。

単語: １００個の音韻バランスがとれた単語
話者: １０名（５名の男性と５名の女性）
感情: 普通、怒り、悲しみ、喜び、恐れ、驚き、愛想をつかす、からかい
音声サンプル: 各々の話者が８つの感情で１００個の単語を発声する。
したがって、全体で各話者毎に８００個の音声サンプルが得られる。

この学習データを用いて種々の予備実験を行った。その結果、男性、女性をまとめたニューラルネットを用意するより、男性、女性それぞれにチューンしたネットワークを用意する方が学習、認識共に有利であることがわかった。このことは男性と女性の感情表現が本質的に異なっており、まとめて扱うことが困難であることを示唆しているが、詳しい分析等は別の機会にゆずりたい。

(3) ニューラルネットによる感情認識

感情認識の際には、音声特徴抽出部で得られた音声特徴量が、上に述べた方法で学習が行われた８つのサブネットワークに入力されれる。その結果として、８つの実数値が得られる。これを、

V=(v1, v2, ..., v8)

と表現することとする。

感情認識部の性能を評価するための小規模の認識実験を行った。８つのサブネットワークの出力の内、最大の出力を与えるサブネットワークに対応した感情を認識結果とする簡単な論理を用いて認識実験を行ったところ、約６０%の感情認識結果が得られた。

(4) 感情平面への写像

上に述べたように、感情認識部の出力はベクトル、V=(v1, v2, ..., v8)で表現される。次にVから最終的な感情認識結果を得ることが必要である。我々のこれまでの研究では、感情認識結果は２次元の平面上の点として表現してきた(2)。これまでの研究における結果・経験に基づき、本研究では感情平面Eの上に８つの感情を図５、図６に示すように再配置することとした。

したがって、VからEへの写像を行うことが必要になる。この写像を行う際にもニューラルネットを用いるということが考えられるが、今回は時間の関係もあって、以下のような簡単な論理を用いて写像を行うこととした。　ｍ１およびｍ２を８つの実数値、v1, v2, ..., v8の最大値および第２位の値とする。また、(xm1, ym1), (xm2, ym2) をm1, m2に対応する感情位置であるとする。このとき、最終的な感情認識結果、 (x, y) を次式で求めることとする。

x = c*xm1 + (1-c)*xm2、y = c*ym1 + (1-c)*ym2　(c：定数)

４．１から４．３までの処理によってMICの感情認識が行われる。この感情認識法は主としてMICの行う音声中に含まれる感情の認識のために開発されたものであるが、今回はMUSEが行っている楽器から入力された音の認識の際にも同じ論理を用いている。

<-- Table of Contents

5　Future Work

音韻バランス単語を学習して、言葉を選ばず、声の抑揚を認知して感情表現を行う。

キャラクターが生存するサイバースペースを設計し、その中でキャラクター間でのコミュニケーションと、人間ともInteractionできる方法を研究する。

<-- Table of Contents

6　まとめ

本論では、新しいタイプの人工生命キャラクター"MIC" & "MUSE" を紹介した。基本的コンセプトと、このような生物の様なキャラクターの設計は、アートとテクノロジーの両方の視点から研究している。テクノロジーとアートがお互いに共通する問題を発見することは、技術的な実用性と人間的な意識が混合しながらシナジー効果を生み出す。便利さと娯楽的要素にあふれる先端技術は、ややもすれば人間の批判的探究性を奪い、生きる目的を物質的な豊かさにすり替えてしまう。しかし、アートは深遠で普遍的なものを目指す方向に破壊と創造を繰り返す。アートとテクノロジー、そして世界を見つめる畏怖の念、万物を包み込む宇宙といったものとのバランスを考える必要はある。それは、言葉を超えた別の感覚で認識しなければ実感できないことかもしれない。

<-- Table of Contents

参考文献

[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).

<-- Table of Contents

RETURN