3. エージェントによるマルチモーダルインタフェース

中津良平[†]

1. ま え が き

エージェントは、自律性を持ち人間の知的な活動を支援してくれるコンピュータプログラム、コンピュータシステムと考えることができるだろう。人間は種々の知的な活動を行っているが、その代表的なものは人間どうしのコミュニケーションであろう。通常は、人間のコミュニケーションに関わるエージェントがマルチモーダルなインタフェースを持つことは自明のこととして研究が進められることが多いようである。一方、マルチモーダルインタフェースの実現は困難な研究課題である。ここでは原点に帰って、人間のコミュニケーションに関わるエージェントの本質は何かを考察し、その結果としてマルチモーダルインタフェースの必要性を指摘する。さらに、マルチモーダルインタフェースの研究における重要な問題点を指摘し、今後の研究方向を明らかにする。また、マルチモーダルインタフェースを持つエージェントの研究例を、我々が進めている研究を中心として紹介する。

2. エージェントと擬人性

人間とコミュニケーションでき、また人間どうしのコミュニケーションを支援するコンピュータエージェントを考える場合、我々はどうしてもそこに人間らしさを備えた存在、すなわち実際はコンピュータソフトウェアであっても、それが人間的な姿・形をしており、人間のようにふるまう存在を期待

する。これはなぜだろうか。いろいろな理由が考えられるが、代表的なものとして以下の点を指摘することができる¹⁾。

(1) エージェントの行動の認知性
人間は我々が日常対応する機会の最も多い存在である。したがって、我々にとって人間の行動は予測・理解することが容易である。コンピュータの内部のデータをそのまま提示されるより、擬人化されたエージェントが解釈して説明してくれる方が我々にとって理解が容易である。

(2) 人間のふるまいの容易性
我々の意図をコンピュータに伝える場合も、人間を相手に自分の意図を伝えるのと同様の方法で行った方が楽である。すなわち、キーボードを操作したりマウスを操作するより、音声・表情・身振り・手振りをを用いた方が、自分の意図を容易に表現・伝達できる。

すなわち、コンピュータが相手の場合でも、我々が日常慣れ親しんでいるコミュニケーションの方法論を使うことが望ましいというのが、擬人化エージェントを用いる最大の理由であろう。いいかえると、「人間のような存在である方が我々が対応しやすいから」ということになる。もちろんこれは、人間を中心に考えた場合、正しい考え方である。しかしながら一方、以下に述べるように、慣れ親しんでおり使いやすいインタフェースを用いる場合、我々はそこにコンピュータにとって扱いにくい種々の無意識の情報を乗せてしまうという問題点を含んでいる。

3. エージェントとマルチモーダルインタフェース

擬人化されたエージェントと人間の

表1 コミュニケーションで用いられる情報

分類	種類	具体的内容、属性など
バーバル	音声	意味内容
ノンバーバル	身体動作	表情、身振り、手振り、視線など
	周辺言語その他	感情、声質など 身体的特徴、対人距離など

インタラクションを考えてみよう。人間どうしのコミュニケーションでは、音声の他に、顔の表情・身振り・手振りなどによってコミュニケーションを行っている。したがって、擬人化されたエージェントとのコミュニケーションを行う場合にも、人間どうしの場合と同じように、音声・表情・身振り・手振りをを用いたコミュニケーションを行なうことが望ましい。これがマルチモーダルインタフェースが必要とされる理由である。これまで主として研究が行われてきたのは、音声によるコミュニケーション(いわゆるバーバルコミュニケーション)をベースとして、顔の表情・身振り・手振りなどの、いわゆるノンバーバルな情報も同時に用いたインタフェースを実現することである。バーバルおよびノンバーバルな情報をを用いたインタフェースで用いられる各種の情報をまとめて表1に示す²⁾。

また、マルチモーダルインタフェースにはこれとは別の分類法がある。すなわち、人間どうしのコミュニケーションにおいてやり取りされる情報における論理面と非論理面に着目する考え方である。これまで音声を用いたコミュニケーションにおいて着目されてきたのは、音声に含まれる意味情報すなわち論理情報の取り扱いである。しかしながら、容易に理解されるように、

[†] 株式会社 ATR 知能映像通信研究所
“Multimodal Interface of Human-like Agents” by Ryohei Nakatsu (ATR Media Integration & Communications Research Laboratories, Kyoto)

音声に含まれる論理情報は音声の情報の一部にすぎない。そのことを説明するために人間のコミュニケーションのモデルを考えてみよう³⁾。

図1は人間のコミュニケーションのモデルを示したものである。このモデルが人間の脳の構造に似ていることに注意されたい。表層、すなわち脳における新皮質に相当する部分には、言語の使用に基づくコミュニケーションを司る層が存在する。AIの分野における研究者は、この層のメカニズムを研究してきたといえる。音声認識や文字認識がその代表例であるが、他にもプランニング、推論などAIの中心分野は、いずれもこの分野を扱ってきた。たとえば、音声認識の分野では、長年にわたり高い認識性能を得るアルゴリズムの研究が行われてきた。研究は着実に進歩しているものの、人間の音声認識能力に比較するとまだ十分な認識性能を得るには至っていない。従来、これは認識アルゴリズムが不十分であるからと考えられがちであった。しかしながら、これは本質的な問題点を見過ごした見方ではないだろうか。

音声認識の取り扱ってきた対象は、音声に含まれる論理情報のみであるといえる。しかしながら、論理情報は音声を構成している情報の一部に過ぎない。感情情報、感性情報などの他のもっと豊かな情報が音声には含まれている。このような情報は、より深いレベルにある層、すなわち図1におけるインタラクション層や、リアクション層によって生成されたものであると考え

ることができる。インタラクション層は、発話のリズムや発話の番の交代など、コミュニケーションチャンネルを維持するための行動を司る層である。この層は、人間のコミュニケーションをスムーズに行うために重要な役割をしている。インタラクション層の下にはリアクション層がある。リアクション層は、人間のより基本的な動作を司る層である。例えば、音の来た方向に顔を向けたり、急に光が来ると目を閉じるといった機能を持つ層である。このような機能は、人間が動物であった太古の時代に獲得されたものといえる。すなわち、脳の旧皮質のように、すべての動物に共通の機能がこのリアクション層でコントロールされているといえる。

このように、人間のコミュニケーションは、論理的な行動や情報の取り扱いの他に、より深い層の働きが重要な役割をしており、感情・感性などの非論理情報は、これらの層の働きにより生成・理解されるものと考えられる。したがって、論理情報以外の情報の送受を含めた総合的なコミュニケーション機能を実現するためには、これらのインタラクション層、リアクション層のメカニズムを研究し、それらの結果を最上部のコミュニケーション層の機能と統合することが必要である。これによって、真に人間的なふるまいのできるエージェントを実現することが可能になると考えられる。

以上述べたように、マルチモーダルインタフェースという場合、これまで

は単に音声・表情・身振り・手振りなどの複数のモダリティを用いることを意味していたが、むしろ今後はそれらのモダリティを用いてやり取りされる情報の内容に注目する必要があるといえる。またその際、人間のコミュニケーションに着目すると、感情に代表される感性情報の取り扱いが重要な課題となる。したがって、以下では、まず感情を取り扱う技術の例を述べる。また、具体的なマルチモーダルインタラクションの例をATRにおける研究を中心として述べる。

4. 感情認識技術の例

4.1 音声に含まれる感情の認識⁴⁾⁵⁾

音声に含まれる意味内容を認識する音声認識研究の歴史は長い。感情の認識に関してはこれまでほとんど研究がなされてこなかった。これは、音声認識自身が困難な研究課題であるため、感情の認識を行う余裕がなかったためというのが最大の理由と考えられる。一方で、感情の含まれた音声の性質を分析する研究はいくつか行われてきた⁶⁾⁷⁾。そこでは4〜8種の種々の感情を含んだ音声の性質が調べられている。

感情の認識を行う際、何種類のどのような感情を認識対象とするかは、重要な問題である。上記の研究内容を参考にすると共に、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、ここでは以下の8つの感情を扱うこととした。

「普通」、「喜び」、「驚き」、「悲しみ」、「怒り」、「恐れ」、「からかい」、「軽蔑」

(1) 感情認識処理

図2は処理の流れのブロック図である。全体の処理は2つの部分、すなわち、音声特徴抽出部および感情認識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

図1 コミュニケーションのモデル

図2 感情認識処理のブロック図

図4 表情認識処理のブロック図

図3 感情認識部の構成

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。感情認識部の構成を図3に示す。第1段階ではニューラルネットワークが用いられる。第1段階は、8つの感情のそれぞれを認識できるよう学習された8つのニューラルネットワークが並列に置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ワーク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ワークの出力を論理処理して、最終的な認識結果を得る。

(2) 学習と認識

感情認識を行うためには、上に述べたニューラルネットワークをあらかじめ学習させておく必要がある。我々は日常の会話において、話者が代っても、また

同じ単語や文章に異なった感情が乗っていても、それらを認識している。したがって、不特定話者、コンテキスト独立型の感情認識を実現することが望ましい。その実現のため、以下のような音声サンプルを学習データとして用意し、ニューラルネットワークの学習を行った。

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：上記の8種類の感情

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。

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うち、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。

4.2 表情認識⁸⁾

人間の顔の表情は、人間どうしのコミュニケーションに重要な役割をしている。したがって、表情の認識を行うことは、画像通信において送り手の表情を少ない情報量で送る場合や、人間とコンピュータ間の高度なインタフェースを実現するために重要である。表情認識は、統計的手法、ルールベースの手法などがこれまで各種試みられてきたが⁹⁾¹⁰⁾、最近HMMを用いる方法が提案され高い認識性能を得ている。HMMは本来、音声認識のために開発された手法であり、時間変化するパターン情報を統計的に処理するのに適した方法である。ここではHMMを用いた表情認識の手法を紹介する。

(1) 処理フロー

表情処理の全体の流れを図4に示す。処理は画像処理、認識処理から構成され、画像処理は、入力された動画画像から、表情の動きに応じた特徴ベクトル列を出力する。認識処理では、あらかじめ認識対象の表情カテゴリ毎にHMMを学習により作成しておく。

図5 移動ベクトルの例

図6 Left-to-right 型 HMM の構成

計測された特徴ベクトル列を生成する確率を各 HMM を用いて産出し、産出された確率が最大となる HMM に対応する表情カテゴリーを認識結果とする。また、表情カテゴリーとしては、「怒り」、「嫌悪」、「恐れ」、「悲しみ」、「幸福」、「驚き」の合計 6 種類を取り扱う。

(2) 特徴抽出

特徴抽出は、まず動画像の連続する 2 枚の画像から勾配法により求めたオプティカルフローを時間的に積算して移動ベクトルを得る。図 5 に移動ベクトルの例を示す。次に、右目および口をそれぞれ中心とする 2 つの領域に対して、移動ベクトルの各成分にフーリエ変換を施して得た特徴量から、右目を含む領域から 5 個、口を含む領域からは水平方向の動きを考慮して 6 個のフーリエ変換係数の低周波成分を選択

する。

(3) HMM による認識

認識処理では、HMM を用いて特徴ベクトルの生成確率を産出する。HMM の構成は図 6 に示したように left-to-right 型とする。HMM は以下で定義できる。

状態： $s = \{s_1, s_2, \dots, s_N\}$

出力シンボル： $O_1, O_2, \dots, O_M$

状態遷移確率： a_{ij} (状態 s_i から状態 s_j に遷移する確率)

出力確率： $b_j(O)$ 状態 s_j においてベクトル O を出力する確率。ここで、出力確率は、多次元空間上の連続分布とし、複数個の多次元正規分布の荷重平均、すなわち混合分布により近似する。複数の正規分布を用いることにより、同一表情における個人差、同一人物の表情表出の変動等に起因する異なったパターンを精度良く近似することができる。

次に、あらかじめ O の学習データが与えられたとき、最適な状態遷移確率、および出力確率を求めておく。認識時には、出力されたシンボルの系列が与えられたとき、これを出力する最適の状態遷移およびその場合の出力確率が得られる。したがって、カテゴリー毎に学習した HMM を用意しておくと、観測されたシンボル確率が与えられたとき、それを出力する最適の HMM が決定される。

(4) HMM の学習

HMM の各パラメータ a_{ji} と $b_j(O)$ の推定は Baum-Welch アルゴリズムを用いる。これは、出力列 $O = O_1, O_2, \dots, O_T$ に対する期待値が最大になる確率として求められる。

(5) 認識実験

4 人の被験者 (男性 3 人、女性 1 人) から基本 6 表情の動画像を 10 シーケンス収集し、その中の 3 シーケンスを HMM の学習用に、残りを認識評価用に用いた。同一人物で学習と認識を行った場合には、98% の認識率が得られた。さらに、不特定多数人物の表情認識への拡張として、3 人のデータで学習を行い、残りの 1 人のデータを認識する実験を行い、約 65% の認識性

能が得られた。

5. マルチモーダルインタフェース研究の例

5.1 これまでの研究例

マルチモーダルインタフェース研究の従来の例を概観しよう。

- (1) Human Reader¹¹⁾：末永らによって提案された初期のマルチモーダルインタフェースの例。単語音声認識、画像認識機能を持った擬人化エージェントである。
- (2) TOSBERG-II¹²⁾：竹林らによって研究されたマルチモーダルインタフェースを持ったエージェント。ファーストフード店における店員との対話をタスクとして、音声認識・合成機能を持ったアニメーション画像によるエージェントを構築した。
- (3) 社会的エージェント¹³⁾：長尾らによって研究された。複数の人間の会話を認識し、会話の流れを把握することによって会話の進行をサポートする機能を持つエージェントを構築した。

これらの研究はいずれも、主としてコミュニケーションにおける論理的な情報に着目しており、非論理的な情報の取り扱いは行っていない。

5.2 感性エージェント「MIC」⁴⁾⁵⁾

非論理的な情報の代表例である感情を用いたマルチモーダルインタフェースの研究を行った例として、ATR における研究例を述べる。

我々の声を使ったコミュニケーションでは、感情は非常に重要な役割を果たす。時には音声の中に含まれている論理情報より重要な役割を果たすこともある。このことは、赤ん坊が母親の声の中に含まれる意味情報を認識できるようになる前に、感情情報の認識ができるようになるという事実からも確認される。大人の場合でも、音声の中に含まれる意味情報と感情情報を認識し、それらを統合することによって相手が何を言おうとしているかを、深いレベルで認識できるからこそコミュニケーションが円滑に進むと考えられ

図7 感性エージェントの処理の流れ

図8 感性エージェントMICの反応例 (By Naoko Tosa)

る。したがって、本当に人間的な振る舞いをするエージェントを生成するためには、音声に含まれる感情を認識し、それに反応する機能を持たせる必要がある。

このような考え方に基づいて、アーティストと工学者の協力により、音声に含まれる感情を認識し、それに応じて顔の表情や動作で反応するコンピュータキャラクタ「MIC」を開発した。全体的な処理の流れを図7に示す。MICは以下の特徴を持っている。

(1) MICは人間の声に含まれる8種類の感情を認識し(感情認識の内容については4.1節参照)それに反応するキャラクタである。

(2) 全身像のCGを作り、単に顔の表情による反応だけではなく、全身でのアクションにより反応である感情表現をすることにした。また、背景にも感情に対応した映像を用いることによって、容易にキャラクタの感情を観客に理解させるようにした。これらの感情表現は、アーティストがデザイ

ン・作成することにより、人間との自然なインタラクションを実現することを狙った。

(3) MICは人間の声に含まれる感情に反応する他に、人間の動作に反応する。カメラから取り込まれた人間の全身像から、頭および両手の先の位置を認識し、人間の、しゃがむ、立つ、両手を上げる、左右に動くなどの動作に追従して動作するモードと、反対の動作をするモードを持っている。

代表的なMICの反応パターンを図8に示す。

6. むすび

人間の知的な活動、特にコミュニケーションを支援するエージェントにおけるマルチモーダルインタフェースについて論じた。従来は音声・表情・身振り・手ぶりなどの複数のモダリティを使うことのみに研究の重点がおかれていた。それに対し、人間どうしのコミュニケーションにおいて重要な役割を果たしている感情や感性の認識・表

現が、今後のマルチモーダルインタフェースの研究において重要な課題であることを示した。また、感情を取り扱ったマルチモーダルインタフェース研究の例を述べた。(1997年12月22日受付)

(文 献)

- 1) 間瀬健二: “擬人化エージェントの生成”, 信学会専門講習会「知的エージェント研究の最前線」, pp. 9-15 (1996)
- 2) 黒川隆夫: “ノンバーバルインタフェース”, オーム社 (1994)
- 3) 中津良平: “アーティスティックな要素を取り入れたバーチャルリアリティ技術—アートと工学の融合をめざして—”, 日本バーチャルリアリティ学会論文集, 1, 1, pp. 1-9 (1996)
- 4) N. Tosa and R. Nakatsu: “The Esthetics of Artificial Life-Human-like Communication Character ‘MIC’ & Feeling Improvisation Character ‘MUSE’”, Proceedings of A-Life V Workshop, pp. 122-129 (1996)
- 5) N. Tosa and R. Nakatsu: “Life-like Communication Agent-Emotion Sensing Character ‘MIC’ and Feeling Session Character ‘MUSE’”, Proceedings of ICMCS '96, pp. 12-19 (1996)
- 6) S. Mozziconacci: “Pitch Variations and Emotions in Speech”, ICPHS 95, 1, p. 178 (1995)
- 7) K. R. Scherer: “How Emotion is Expressed in Speech and Singing”, ICPHS 95, 3, p. 90 (1995)
- 8) 大塚尚宏ほか: “連続出力確率密度分布を用いたHMMによる動画像からの複数人物の表情認識”, 信学論, J80-D-II, 8, pp. 2129-2137 (1997)
- 9) I. A. Essa: “Analysis, Interpretation, and Synthesis of Facial Expressions”, MIT Media Laboratory, Perceptual Computing Group Report, No. 303 (1994)
- 10) Y. Yacoob and L. S. Davis: “Recognizing Human Facial Expressions from Long Image Sequences Using Optical Flow”, IEEE Trans. on PAMI, 18, 6, pp. 636-642 (1996)
- 11) 末永康仁, ほか: “Human Reader: 人物像と音声による知的インタフェース”, 信学論, J75-D-II, 2, pp. 190-202 (1992)
- 12) Y. Takebayashi, et al.: “Noisy Spontaneous Speech Understanding Using Noise Immunity Keyword Spotting with Adaptive Speech Response Cancellation”, Proc. IEEE pp. II115-II119 (1993)
- 13) K. Nagao and A. Takeuchi: “Social Interaction: Multimodal Conversation with Social Agents”, Proc. 12th AAAI (1992)

なかつ りょうへい
中津 良平

1969年、京都大学工学部電子工学科卒業、1971年、同大学院修士課程修了。同年、日本電信電話公社(現NTT)武蔵野電気通信研究所入所。1980年、横須賀電気通信研究所。主として音声認識の基礎研究、応用研究に従事。1990年、NTT基礎研究所研究企画部長、1991年、NTT基礎研究所情報科学研究部長。1994年よりATRに移り、現在、(株)ATR知能映像通信研究所代表取締役社長。マルチメディア要素技術の研究およびマルチメディア技術を応用した通信方式の研究などに従事。工学博士。