音声言語情報処理 32−13
(2000. 7. 15)

音声対話システムにおける擬人化エージェントの挙動の数理的モデル

川本 真一　中井 満　下平 博　嵯峨山 茂樹

北陸先端科学技術大学院大学 情報科学研究科

〒923-1292 石川県能美郡辰口町旭台 1-1

URL http://www-ks.jaist.ac.jp/index-j.html

あらまし　人間らしい音声対話システムの実現を目指した擬人化エージェントの感情の動きなどの振舞の数理的モデルを提案する。提案するモデルは、人間の振舞を模擬することで、内面はわからないが個性を持っているように見えることを目指している。個性的なキャラクタに感じられる面白さを擬人化エージェントの個性を演出することで実現したい。そこで擬人化エージェントの振舞をHMMをベースとする数理的モデルで扱うことで、様々な個性の演出、個性的な振舞を効率的に扱う事が出来ると考えた。実際問題として、音声の韻律特徴によって振舞が変化するモデルを試作し、対話についての主観評価実験を行った結果、モデルが対話の面白さの演出能力を持つ可能性を示す結果が得られた。

キーワード　擬人化エージェント、音声対話システム、振舞のモデル化、隠れマルコフモデル

A Mathematical Model of Anthropomorphic Agent's Behavior in Spoken Dialog Systems

Shin-ichi Kawamoto　Mitsuru Nakai　Hiroshi Shimodaira　Shigeki Sagayama

School of Information Science

Japan Advanced Institute of Science and Technology

1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292

URL http://www-ks.jaist.ac.jp/

Abstract　A mathematical model of anthroponmoriphic agent's emotional behavior is proposed to realize a user friendly spoken dialog system. The proposed agent does not actually model a human mind but it just simulates the behavior of a human being so that the users feel as if it had own personality. Such agents with cyber-personalities would achieve more natural and friendly interface between human and computers. Hidden Markov Model is employed to realize such agents with various types of behavior. As the first step, an agent system whose behavior changes based on the prosody of the users speech is developed. Subjective evaluation experiments showed that the users feel more naturalness in the dialog of the proposed system than in that of the system that does not change its behavior.

Keyword　anthropomorphic agent, spoken dialog system, modeling of agent's behavior, hidden Markov model

1. はじめに

将来の人間と機械とのコミュニケーションにおいては、機械があたかも一個の人間のように話し、聞き、振舞うことの実現が必要であろう。例えば、聞き手が話し手の相談をじっくりと聞き、話し合う状況を想定する。人間同士の対話では、話し手は音声言語だけではなく、感情や声に込められた意図やジェスチャ等、人格や個性を反映したさまざまな情報が表出している。そして、聞き手も言語情報だけでなく、話し手の表出する情報を受け入れ、理解しようとする。しかし機械相手の対話では、人間は機械に合わせて、機械の受け入れてくれる情報だけを伝えようとする。これでは、人間が伝えたい情報を十分に表現できていない。逆に機械から人間へも同等の情報が伝えられることが望ましい。理想としては、機械との対話においても、人間が伝えたい情報を十分に引き出すためには、機械も人間のように存在感があり、人格を持ち、個性や感情を反映して振舞うような人間らしさが必要であると筆者らは考える。

このような機能の実現のためには、情報認識技術に関しても朗読音声の認識だけではなく、自然な対話音声の認識、発話意図や表情、ジェスチャ等の理解が必要であり、情報生成技術に関しても、感情や発話意図を反映した音声合成、自然な表情、ジェスチャの生成などが必要である。更に人間の表出する情報の理解・生成の基礎技術を活用・統合するためのメカニズムを始め、多くの課題がある。筆者らはその中の課題の1つとして、感情の動きのような人間の振舞のモデル化を通じて、個性をパラメトリックに扱うことについて検討する。

機械に個性を付与することは、難しい問題であり、対話をより人間らしいものにするために従来から指摘されている課題の1つである [1][2]。また筆者らは、人間の表出する振舞には個性が関与し、振舞から個性が見出されると考え、個性はコミュニケーションの制御要因の1つとして必要であると考えている。

近年、デジタルペットなど様々な個性豊かなキャラクタが登場し、話題となっている。デジタルペットが時に素直に、時に我儘に振舞うことで、これらのキャラクタに個性を感じ、キャラクタに愛着を持ってかわいがる。これは、愛情表現も含めたかわいがるという振舞をキャラクタが受け止めてくれていて、個性的な振舞いには、愛情が反映されていると錯覚することで、単に機械を扱うというのではなく、様々な情報を引き出すことが出来るのではないかと考えている。

現在、デジタルペットなど様々なキャラクタから感じられる個性はヒューリスティックに作られていたり、ルールとして定義されている事がほとんどである [3]。擬人化エージェントで様々な個性を演出しようと考える時、これらをヒューリスティクスやルールとして定義するのではなく、工学的、数理的なモデルとして扱う事が出来れば、様々な個性の演出、個性的な振舞を学習可能な問題として扱う事が出来ると考えられる。そこで筆者らは、擬人化エージェントの個性を演出するための個性的な振舞の数理的モデルを提案する。また、今回は限られた情報についての例題として、今回は音声の韻律特徴によって振舞が変化するモデルを試作し、提案したモデルに関して、限られた条件下で、その演出能力の可能性について検討する。

2. モデルによる個性の表現

本論文における個性とは、人間を比較する時に見られる性質の違いとして捉える。またこの個性は、人間から観測される挙動から見出されるものと捉える。本論文では、個性を見出すような人間や機械の挙動を個性的な振舞と呼ぶことにする。

個性を表現するためのアプローチとしては、人間の内面のメカニズムを解明し、それを忠実に機械化する事も考えられる。しかし筆者らは人間の外見、振舞を模擬することでも、内面はわからないが個性を持っているように見せることが出来ると考える。

モデルによって人間の振舞を模擬することを考える時、モデルは実際に人間から観測される振舞を学習できることが望ましい。また、人間の振舞が連続量の信号として表されるならば、この連続量の信号を扱えるモデルであることが望ましい。

そこで筆者らは、このモデル化の基礎として、連続量の非定常信号を扱うことのでき、効率的な認識・学習アルゴリズムが存在する隠れマルコフモデル (Hidden Markov Model; HMM) の利用を検討する。

直観的な例として、感情などによって異なる振舞が生成されることを想定し、HMM による個性の表現可能性について考えよう。例えば、各状態を怒っている状態や嬉しい状態などに割り当てるならば、モデルのそれぞれの状態で表出しやすい振舞があり、

各状態を表出する振舞に確率的な偏りを持った信号源と捉えることができる。また各状態の確率的な接続は、怒っている状態から嬉しい状態へ確率的かつ滑らかな変化を表現し、状態への遷移のしやすさによって怒りっぽいなどの性格付けを出来る可能性がある。つまり、この各状態の振舞生成の確率的な偏りと、各状態への遷移のしやすさにより、表出する振舞系列を変化させ、そのように捉えられるような様々な個性を表現できる可能性がある。

そこでHMMを適用することで、人間から観測される情報を模擬することで「個性」を持っているように見えるシステム、個性的な振舞を表現するシステムを実現する為の擬人化エージェントの挙動の数理的モデル化を試みる。

3. 対話における個性的な振舞のモデル化

3.1. 問題の定義

対話のモデル化にあたって、様々な要因を考える必要がある。例えば、対話人数や、アイコンタクトの有無、あいづちや会話への割り込み (参与) など、さまざまなイベントも要因となっている。

ここでは、最も基本的な対話の状況について考えるために、**1)** 2人の話者の対話とし、**2)** 一方が発話中に、他方の発話が割り込まず、**3)** 1つの発話に対して1種類の振舞しか現れないという制約条件を設ける。

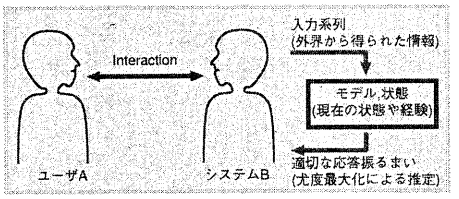

図 1: 対話時の振舞決定問題

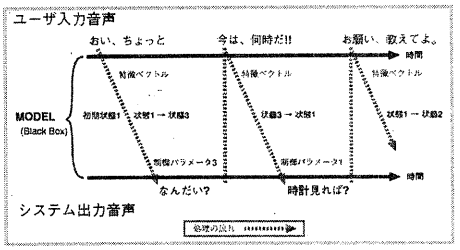

図 2: 想定する対話の流れ

2人の話者A(ユーザー、外部), B(システム) の対話において、Aの発話に対して、Bが応答・振舞を決めるとき、図1のように、

- 「**Bの現在の状態や経験**」を基に、
- 「**Aから得られた情報**」を利用して、
- 「**適切な応答・振舞を決定**」する

問題を考える。

これをHMMに適用する際は、

- 「**Bの状態・モデル**」を基に、
- 「**Aからの入力系列**」を利用して、
- 「**適切な応答・振舞を尤度最大の基準で推定**」する

問題ととらえる。

ここで想定する対話 (図2) では、システムが入力を受理し、それを反映して状態を変化させ、出力を与えると考え、HMMは確率的に入出力の変換関数を切替えるモデルととらえる。

HMMを対話のモデルとして適用することで、**1)** モデルの学習可能性や、**2)** 対話の履歴情報を反映した振舞の変化といったメリットを得られる可能性がある。

HMMは、大量のデータからの統計的な学習能力が高く評価されるモデルであり、効率的な認識アルゴリズムや学習アルゴリズムが存在することで知られる。このことは、大量の対話データによるモデルの学習可能性を示す事実であり、提案するモデルにとってのメリットととらえることができる。

また、HMMの尤度計算過程においては、直前までの尤度の蓄積を現在の尤度計算に反映させる。直前までの尤度は、対話履歴を示す情報と捉えることができ、直前の対話音声の判別だけでは表現できない振舞を、提案するモデルが表現できる可能性がある。例えば、以前に言ったことを根に持つといったことを表現するときは、履歴の情報は必要と考えられる。

3.2. 問題の定式化

長さTの入出力系列を

$$XY = (\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_T, \mathbf{y}_T) \quad (1)$$

とするとき、入力系列を受理し、その入力に対して適切な出力を生成するモデルを、入出力系列XYを生起するモデルMと捉える。但し、$\mathbf{x}_t \in \mathcal{R}^{D_x}$は入力特徴ベクトル、$\mathbf{y}_t \in \mathcal{R}^{D_y}$は出力特徴ベクトルと

する。また、入出力系列をモデル M が生起する確率を $P(XY|M)$ とし、初期状態の集合を S_I、最終状態の集合を S_F とする。

状態遷移系列を

$$Q = q_0, q_1, q_2, \ldots, q_T \tag{2}$$

とする。但し、時刻 t における状態 q_t、入出力 $(\mathbf{x}_t, \mathbf{y}_t)$、$q_0 \in S_I$、$q_I \in S_F$ とする。

初期状態 q_0 が i である確率 (初期確率) を π_i $(\sum_{i \in S_I} \pi_i = 1)$、状態 i から状態 j に遷移する確率 (遷移確率) を $a_{i,j}$ 、状態 j のときに入出力 $(\mathbf{x}, \mathbf{y})$ を生起する確率 (入出力同時確率) を $b_j(\mathbf{x}, \mathbf{y})$、状態 j のときに入力 $\mathbf{x}$ を生起する確率 (入力確率) を $b_j(\mathbf{x})$、状態 j で入力 $\mathbf{x}$ が与えられた時、出力 $\mathbf{y}$ を生起する確率 (条件つき入出力確率) を、$b_j(\mathbf{y}|\mathbf{x})$ とするとき、入出力系列 XY を生起する確率は式 (3) のように表現される。

$$\begin{aligned} & P(XY|M) \\ &= \sum_{allQ} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1},q_t} b_{q_t}(\mathbf{x}_t, \mathbf{y}_t) \\ &= \sum_{allQ} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1},q_t} b_{q_t}(\mathbf{x}_t) b_{q_t}(\mathbf{y}_t|\mathbf{x}_t) \end{aligned} \tag{3}$$

3.1 節で設定した問題において、与えられるものは入力系列 X

$$X = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T \tag{4}$$

のみで、出力系列を推定することになるので、$P(XY|M)$ を最大化する出力系列 Y

$$Y = \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T \tag{5}$$

を探索する問題となる。

この探索を簡単化するために、出力は現在の状態と入力によって一意に決定されると仮定する。つまり、式 (6) で示される変換関数を仮定する。

$$\mathbf{y}_t = f_{q_t}(\mathbf{x}_t) \tag{6}$$

この仮定により $b_{q_t}(\mathbf{y}_t|\mathbf{x}_t) = 1$ となり、このとき式 (3) は次式で与えられる。

$$P(XY|M) = \sum_{allQ} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1},q_t} b_{q_t}(\mathbf{x}_t) \tag{7}$$

したがって、Y に関する探索問題は、入力系列を生成 (受理) する尤度を最大化する問題に簡単化することができるため、HMM で通常用いられる学習・認識アルゴリズムが利用できる。

式 (6) の変換関数として、具体的に以下のような 2 通りの変換関数を考える。

$$f_k(\mathbf{x}) = \mathbf{b}_0^{(k)} \tag{8}$$

$$f_k(\mathbf{x}) = \mathbf{b}_0^{(k)} + B^{(k)}\mathbf{x} \tag{9}$$

式 (8) は、状態 k に割り当てられた代表的な出力ベクトル $\mathbf{b}_0^{(k)}$ への写像関数であり、入力には依存しない最も簡単な写像関数である。代表的な出力は、各状態における出力分布の平均ベクトルとする。

式 (9) は、状態 k に割り当てられた入力 $\mathbf{x}$ を説明変数として重回帰式により出力 $\mathbf{y}$ を線形近似した写像関数である。ここで、$B^{(k)}$ は重回帰係数行列である。

これにより、出力の尤度による探索を省き、変換関数による出力の推定によって出力系列の生成を実現する。

4. 振舞の数理的モデルの学習、および振舞生成のアルゴリズム

モデルを生成する手順の概略は、1) 平均ベクトル近似モデルの生成、2) 入力尤度最大化を基準としたモデルの学習、3) 入力と出力の変換関数の推定である。

ここで状態は急激に、頻繁に変化することはないと仮定し、発話のマクロ的な特徴 (例えば、基本周波数の平均、分散など) を考える。

4.1. 平均ベクトル近似モデルの生成

式 (8) によって出力を推定する最も簡易的なモデルを考える。ここではそのモデルを平均ベクトル近似モデルと呼ぶことにする。

平均ベクトル近似モデルの生成手順を次に示す。

1. 入力のクラスタリングにより状態数を決定
   (a) ベクトル量子化を用いてクラスタリング
   (b) クラスタ数を HMM の状態数とする
2. クラスタリングを基に入出力の確率分布を算出
   (a) 各クラスのデータを使って、出力確率分布を算出
   (b) 出力と直前の入力との関係より、入力確率分布を算出
3. 各状態に入出力変換関数を割り当てる
   (a) 各状態の出力確率分布の平均ベクトルを変換関数の出力とする

4. 直前の状態と現在の状態より遷移確率を算出

4.2. 入力尤度最大化を基準としたモデルの学習

入力系列 X をモデルが生成する尤度を最大化することを基準としたHMMの学習アルゴリズムを用いる。

4.3. 入力と出力の変換関数の推定

変換関数を決定するために、各状態に属する入出力データをViterbiアルゴリズムによるセグメンテーションによって近似的に割り当てる。

各状態に属する入出力データを用いて、式(9)の $B^{(k)}$ を重回帰法により推定する。

4.4. 出力系列の生成

与えられたモデルに対して、入力を受理することで、出力を生成する手順を考える。

1. 入力系列だけから最適状態遷移系列を生成
   - 時刻 t における入力 $\mathbf{x}_t$、状態 q_t、最適状態 q_t^*、全状態集合 Sのとき

$$q_t^* = \underset{s \in S}{\operatorname{argmax}} P\left(q_t = s | \mathbf{x}_t, q_{t-1}^*\right) \quad (10)$$

2. 生成した最適状態遷移系列を利用して、入力系列を出力系列に変換
   - 式(8)、もしくは式(9)を用いて変換

通常、HMMの状態遷移系列は、入力が全て決まってから決定される。しかし、このモデルを適用する場合には、ある時刻までの入力が与えられたとき、それに対する出力が逐次決定される必要がある。

実際の環境では、ある時点での尤度最大の状態を、その時点での最適な状態と近似することで、逐次的に状態系列を求めることを近似的に実現する。

5. 評価実験

モデルの生成する振舞によるユーザーへの影響について検討することを目的に、対話音声の聴取による主観評価実験を行った。実験条件を 表2 に示す。また、実験に用いた新聞購読勧誘の対話例を表1に示す。本実験の評価に際して、想定するタスクに適切かつ、学習や重回帰式の推定に十分なデータベースが整備されていないため、学習や重回帰式を用いた変換関数は使わず、簡易的に実現できる平均ベクトル近似モデルを実験対象のモデルとして選択した。

表 1: 対話例

User ごめんください。

System はい。

User フーリエ新聞のものですが、今、どちらか新聞をお読みになってますか。

System いいえ。でも、うちは新聞読まないんですよ。

User そこを何とか、3ヶ月で結構ですから。商品券をお付けしますよ。

System いままでも一度もとらなかったから、本当に結構です。

User 残念だなぁ。なかなかこんなサービスないんだけどなぁ。

System 本当に結構です。お帰りください！

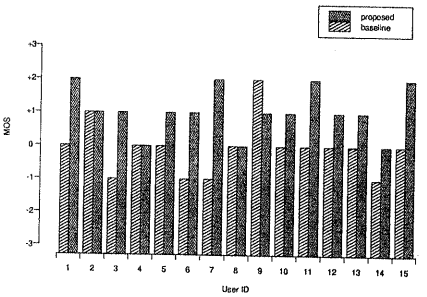

図 3: 評価項目「面白い」の主観評価結果(baseline:1種類の振舞(「普通」)に固定したモデルとの対話、proposed:提案モデルとの対話)

5.1. 振舞に変化を持たせることの効果についての検討

ユーザの発話音声について、1種類の振舞「普通」に固定したモデル(baseline)との対話と、提案モデル(proposed)との対話の2つのモデルとの対話を比較した。ここでのbaselineの応答は従来の音声対話を想定したものである。

幾つかの対話例を話者とは別の15名の男性被験者に聞いてもらうことで、評価実験を行った。評価項目は、自分がユーザとなってシステムと対話をすることを想定したときに、システムの応答が面白いかどうかについて、+3(面白い),+2,+1,0(どちらでもない),−1,−2,−3(面白くない)の7段階評価を行った。その主観評価結果の1つを図3に示す。

図3の結果から、モデルの生成する応答との対話の方が面白いと感じている人が多く、振舞に変化を

表 2: 実験条件

対話タスク	新聞購読勧誘の対話 (ユーザ:新聞販売員, システム:客)
対話内容	固定 (予め決まった対話音声を生成)
振舞	4 種類 (「普通に」「怒って」「機嫌良く」「面倒そうに」の気分を演技)
音声	16kHz, 16bit, linear サンプリング
モデルパラメータ	4 状態 (4 種類の振舞に対応)、経験的に設定
特徴量	ユーザの発話の韻律的特徴量、4 次元 (1 発話全体での F_0 と Power の平均・分散)

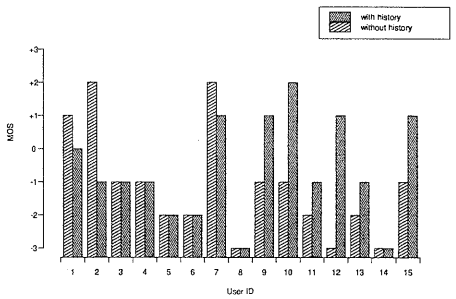

図 4: 評価項目「自然さ」の主観評価結果 (baseline:前の入力での尤度を初期化するモデルとの対話、 proposed:提案モデルとの対話)

持たせることで、システムとの対話の楽しさを演出できる場合が存在すると考えられる。

5.2. 履歴情報の効果についての検討

ユーザの発話音声について、前の入力での尤度を初期化するモデル (without history) との対話と、前の入力での尤度をそのまま利用するモデル (with history) との対話の 2 つのモデルとの対話を比較した。筆者らが提案するモデルは (with history) である。また、履歴情報を持たないモデル (without history) は、1 発話毎にモデルの各状態の尤度を一様な値に初期化することで実現した。

幾つかの対話例を話者とは別の 15 名の男性被験者に聞いてもらうことで、評価実験を行った。評価項目は、ユーザの話しかけに対して、システムが自然な応答をしているかどうかについて、+3(自然),+2,+1,0(どちらでもない),−1,−2,−3(不自然) の 7 段階評価を行った。その主観評価結果の 1 つを図 4に示す。

図 4の結果から、履歴情報を持ったモデルの応答の方が自然な対話に感じる人が若干多く、履歴情報を持つモデルの対話が自然に感じる場面が存在すると考えられる。

6. むすび

本論文では、擬人化エージェントの個性を演出するための個性的な振舞の数理的モデルを提案した。今後は、モデルの拡張、他の特徴量の導入、学習可能な音声対話データベースの調査、および実データにおけるモデルの学習とその評価を行っていきたい。

また現在、音声の韻律特徴によって振舞が変化するモデルを試作し、擬人化エージェントに組み込んだ擬人化音声対話システムを構築中である [4]。

謝辞

擬人化音声対話システムの試作において、音声認識に関して北陸先端科学技術大学院大学 情報科学研究科 松田茂樹 氏、加藤裕 氏、横山国宏 氏、対話音声の収録、および生成に関して同研究科 槻尾洋志 氏、 藤永勝久 氏、擬人化エージェントの顔画像生成に関して同研究科 井波暢人 氏、対話処理部に関して同研究科 山崎義人 氏に御協力頂いた。この場を借りて感謝する。

参考文献

[1] 八木 他, “音声対話におけるエージェントの態度と人間の発話の関連,” 情報処理学会 研究報告 95-SLP-7-14, 1995.

[2] 嵯峨山, “音声認識技術実用への課題,” 情報処理学会論文誌, Vol.36, No.11, pp.1047-1053, 1995.

[3] 牛田 他, “デジタルペット — 心を持った機械達 —,” 情報処理学会 会誌 Vol.41 No.2, pp.127-136, 2000.

[4] 川本 他, “確率的な振舞を伴う擬人化対話エージェント,” Interaction2000 論文集, pp. 61-62, 2000.