力学モデル駆動による音声対話エージェントの動作生成

Behavior Generation for Spoken Dialogue Agent by Dynamical Model

中沢 正幸 [1*]　西本 卓也 [1]　嵯峨山 茂樹 [1]
Masayuki Nakazawa[1]　Takuya Nishimoto [1]　Shigeki Sagayama [1]

[1] 東京大学大学院 情報理工学系研究科
[1] Graduate School of Information Science and Technology,
The University of Tokyo

Abstract: For the spoken dialog systems with the anthropomorphic agents, it is important to give the natural impressions and the real presence to human. For this purpose, the head and gaze controls of the agent which are consistent with the spoken dialogs are expected to be effective. Our approach is based on the following hypotheses: 1) An agent performs the dialog concurrently with the intentional controls of the head and gaze to retrieve the information and to give signals. 2) The movement of the head and eyeballs is based on mathematical models. To achieve these purpose, we have adopt the mathematical model for movements of the agent. It are several merits to formulate by the mathematical model, a) the parameters can reflect the subjectivity which can generate various movement from this model, b) the movements of the agents can reflect the personality, c) the continuous movements of the agent can be controlled by the mathematics. In this paper, we propose a mathematics model by the second order system and perform comparison with the linear model and show the superiority.

1　はじめに

擬人化音声対話エージェントにおいて，対話の流れに応じて適切な情報の授受の制御を行うことは，対話相手である人間に自然な印象を与え，擬人化音声対話エージェントの実在感を高めるうえで重要である．我々は，音声合成の分野において声帯振動機構に基づいたモデルが合理的な方法で多様な基本周波数 (F_0) パターンを説明できることに着目し，様々な対話現象を力学現象に等価変換することで擬人化音声対話エージェントの心的状態と挙動を結びつけるモデルの構築を目指している．具体的には，次のような仮説に基づいて，擬人化音声対話エージェントの頭部・視線の動きを制御する手法について検討している．

1. 擬人化音声対話エージェントは相手に関する情報を得たり相手に合図を送ったりするために能動的に頭部・視線を動かしながら対話を行う
2. 擬人化音声対話エージェントの頭部・視線の動きは数理的な視線制御モデルに従う

我々は，擬人化音声対話エージェントの実在感を検証するプラットフォームとして，「擬人化音声対話エージェント基本ソフトウェア」[8] の基本通信プロトコルに適合し，腕・手の動作，表情，唇の形状，視線・頭部の動作をそれぞれ独立に制御することができる 3D CG で構築された擬人化音声対話エージェントを開発した．

この擬人化音声対話エージェントに音声認識・画像認識等の言語によるコミュニケーション能力を加えることで，視線や表情，身振り・手振り，感情などの非言語情報を表現でき，より実在感のある人間型の音声対話環境を構築することを目指している．

この目的のために，擬人化音声対話エージェントが実現すべきことは，人間と同じような動作を行い，自然な動きを実現することである．特に視線は，広い範囲で捉えると「表情」の一部と言われるように，自然な動きの実現と密接な関係がある [1, 2, 3, 4, 5]．

擬人化エージェントの振る舞いに個性（個人差）をパラメータとして含めることによって，擬人化音声対話エージェントにあたかも現実の人間のように振る舞わせることが本研究の目標である．

本報告では，対話エージェントの動作を制御する数理モデルの提案を行うとともに，2 次遅れ系標準形を用いて，擬人化音声対話エージェントの頭部・視線運動を制御する手法について述べる．さらに，数理モデルによって生成された動作の有効性の検証に関する予備実験について述べる．

*連絡先:東京大学大学院 情報理工学系研究科 システム情報学専攻
〒 113-8656 東京都文京区本郷 7–3–1 工学部 6 号館 140
E-mail: nakazawa@hil.t.u-tokyo.ac.jp

2 音声対話エージェント動作の数理モデル化

2.1 非言語情報の連続的な動作の生成過程のモデル化

人が向き合い対話を行う時，そこには種々の情報が表現され，相互にやりとりが行われる．その情報は藤崎 [6] によれば，1) 言語情報，2) パラ言語情報，3) 非言語情報に大別される．

ここでの言語情報とは，言語により規定され，主として文字による表記が可能な音声である．パラ言語情報とは，断定・疑問・勧誘・反論など，様々な意図が込められた音声である．また，丁寧・ぞんざい，改まった・くだけたなどの話者の態度の表現や，さらには，ゆっくり・早口，大声・小声などの話し方のスタイルも同様にパラ言語情報として音声に含まれる．すわなち，音声により表現される種々の情報は，伝えたい意味内容である言語情報に加え，話者の個人的な特徴や，年齢・性別・健康などの身体的な状態に関するもの，あるいは気質・感情などの心理的な状態に関するパラ言語情報で表現される．

音声の音響的特徴と，それらが担う各種の情報との関係を定量的に表現することは，音声言語の情報処理にきわめて重要であり，音声対話処理の技術的進歩に大きく貢献する．しかしながら，現実に観察されるのは，音声や動作といった最終出力であり，そこから溯って種々の情報を推定することは，複雑で困難な逆問題となる．

本研究では，言語情報及びパラ言語情報が表現する意味や意図に同期した非言語情報の生成において，頭部と視線の動作は，注視方向とその変動に分類可能であるという仮定をおき，対話エージェントの頭部・視線動作の連続的な生成過程を数学的にモデル化し，このモデルが対話時に現れる種々の動作に適用しうる一般性を持つことを実験的に確かめる．

2.2 2 次遅れ系標準形による連続的な動作の生成過程

我々は，人間型の擬人化音声対話エージェントの頭部及び視線の運動自由度を考慮した，数理モデルによる動作の自動生成を目的としており，これまで，視線運動モデルの予備的な検討を行なってきた [7].

本研究では，対話エージェントの動作生成に関して，従来のモーションキャプチャのようなデータ駆動でなく，一般性かつ，できるだけシンプルなモデルを想定し，入力としての言語・パラ言語情報からエージェントの動作出力につなげるための力学的なモデルを検討する．

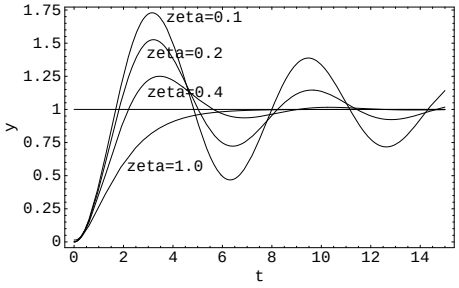

図 1: ステップ入力に対する時間応答 ($\zeta = 0.1, 0.2, 0.4, 1.0$)

対話エージェントの動作生成を 1 つの統合的なモデルと考えた場合，人間 (エージェント) の動作は，刺激 (言語入力) に対する応答として捉えることができる．その時の刺激は，ステップ入力が簡単で望ましく，エージェント動作の頭部や視線は質量と粘性を持つ系と捉えることができる．この時，人間は頭部や視線の動作を最適に制御していると想定され，例えば，PID 制御では，ダンパー係数が 0.8 前後の 2 次遅れ系標準形として再現でき，さらには臨界制動系で近似することも可能と考えられる．

人間の相槌などの振動的な動作は，早い相槌とゆっくりの相槌があるが，上下に周期的に力を加えて無理やり振動動作を行っているとすると，エネルギーのロスが大きく，このような動作をしているとは考えにくい．エネルギーのロスを抑えるには，筋肉の緊張を変える，すなわち，動作モデルのパラメータを変え，2 次遅れ系標準形の自由振動の動作を行っているという解釈も可能である．このような解釈ならば，対話エージェントの動作制御に都合がよいと言える．

本報告では，2 次遅れ系標準形を対話エージェントの動作モデルと仮定し，対話エージェントの頭部動作（3 自由度;Yaw, Pitch, Roll）及び視線動作 (2 自由度;Yaw, Pitch）を極座標系で定義し，対話における話者の状態や意図などを力学現象に置き換えることを目指す．

$$G(s) = \frac{\omega_n}{s^2 + 2\zeta\omega_n s + \omega_n^2} \tag{1}$$

2 次遅れ系標準形の伝達関数 $G(s)$ は，固有振動数 ω_n，減衰係数 ζ とし，式 1 で表される．この 2 次遅れ系標準形は，ステップ入力の応答に際して，$\zeta = 1.0$（臨界制動）を境界として振動現象の有無が変わる性質 (図 1) を持ち，質量，粘性摩擦，バネを持つ機械系の基本的な制御モデルとされる．この制御モデルを頭部・視線の運動に対応させるにあたっては，感情の変化，相槌，発話などが特定のインパルス入力やステップ入力に対応したり，モデルの質量やバネ係数を変化させるといった対応付けを行うことが必要となる．

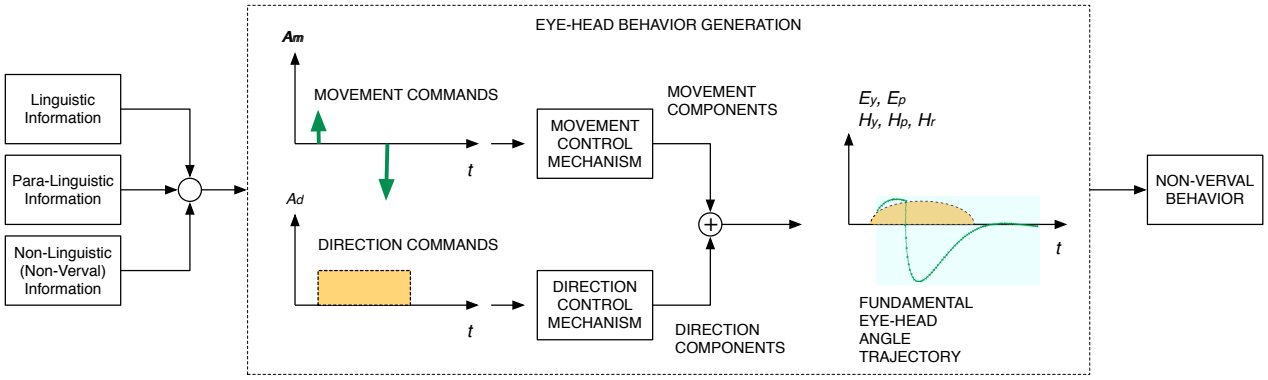

図 2: 音声対話エージェントの動作制御の流れ

$$g(t) = \frac{\omega_n}{\sqrt{1-\zeta^2}} e^{-\zeta\omega_n t} \sin\left(\omega_n\sqrt{1-\zeta^2}t\right) \quad (2)$$

$$y(t) = 1 - \frac{1}{\sqrt{1-\zeta^2}} e^{-\zeta\omega_n t} \sin\left(\omega_n\sqrt{1-\zeta^2}t + \phi\right)$$
$$\phi = \tan^{-1}\left(\frac{\sqrt{1-\zeta^2}}{\zeta}\right) \quad (3)$$

式 1 が $0 < s \leq 1$ の場合，2 次遅れ系標準形のインパルス応答 $g(t)$ は式 2 で表され，ステップ応答 $y(t)$ は式 3 で表される．言い換えれば，固有振動数 ω_n は運動の機敏さを表し，減衰係数 ζ は運動の仕方 (図 1) を表す．

自然な頭部及び視線の運動を単純な制御モデルで実現することが本研究の目的であり，得られる制御モデルは，ある場合は物理的に人間の身体運動に対応し，ある場合は心理現象を力学系に等価変換するためのモデルとして有用であると期待される．

2.3 頭部・視線運動モデルの定式化

対話エージェントの頭部及び視線動作の生成過程をモデル化しものを図 2 に示す．ここでは言語情報 (Linguistic Information) 及び，パラ言語情報 (Para-Linguistic Information) から，対話エージェントの頭部及び視線の動作を 2 次遅れ系標準形のインパルス応答及びステップ応答を用いて生成する流れを示している．

2 次遅れ系標準形のインパルス応答で表現した動作指令と動作制御機構及び，その出力を動作成分と定め，2 次遅れ系標準形のステップ応答で表現した方向指令と方向制御機構及び，その出力を方向成分と定めた．頭部及び視線を含む対話エージェントの最終的な動作出力 $F(t)$ (式 4) は，対話エージェントの視線に関する動作 (極座標系 2 自由度:yaw,pitch) と頭部に関する動作 (極座標系 3 自由度:yaw,pitch,roll) のそれぞれの角度成分の和として出力される．

$$F(t) = \sum_{i\in\{y,p\}} \Big\{\alpha E_i(t) + (1-\alpha)H_i(t)\Big\} + H_r(t) \quad (4)$$

$$\underset{m\in\{y,p\}}{E_m(t)} = \sum_{i=1}^{I} K_i^m g(t - T_{i_0}^m) + \sum_{j=1}^{J} K_j^m \Big\{y(t - T_{j_1}^m) - y(t - T_{j_2}^m)\Big\} \quad (5)$$

$$\underset{n\in\{y,p,r\}}{H_n(t)} = \sum_{i=1}^{I} K_i^n g(t - T_{i_0}^n) + \sum_{j=1}^{J} K_j^n \Big\{y(t - T_{j_1}^n) - y(t - T_{j_2}^n)\Big\} \quad (6)$$

式 5 は視線の動作出力を示し，視線動作指令の成分 $m \in \{y, p\}$ (それぞれの視線動作の 2 自由度;Yaw, Pitch 成分) を表す．同様に，式 6 は頭部の方向出力を示し，頭部方向指令の成分 $n \in \{y, p, r\}$ (それぞれの頭部動作の 3 自由度;Yaw, Pitch, Roll 成分) を表す．ここで，

$g(t)$: 動作制御機構のインパルス応答,
$y(t)$: 方向制御機構のステップ応答,
I: 動作指令の数,
J: 方向指令の数,
K_i^m: 視線動作成分 m の第 i 番目の動作指令の大きさ,
K_j^n: 頭部動作成分 n の第 j 番目の方向指令の大きさ,
$T_{i_0}^m$: 視線動作成分 m の第 i 番目の動作指令の生起時刻,
$T_{j_1}^n$: 頭部動作成分 n の第 j 番目の方向指令の開始時刻,
$T_{j_2}^n$: 頭部動作成分 n の第 j 番目の方向指令の終了時刻を示す．

なお，動作指令の大きさ K_i^m 及び K_j^n は，$-1 \leq K \leq 1$ の値を持ち，$E_m(t)_{m\in\{y,p\}}$ 及び，$H_n(t)_{n\in\{y,p,r\}}$ は，それぞれ極座標の角度 θ ($-\pi \leq \theta \leq \pi$) へと変換される．ただし，$K = 0$ 及び，$\theta = 0$ は正面とする．

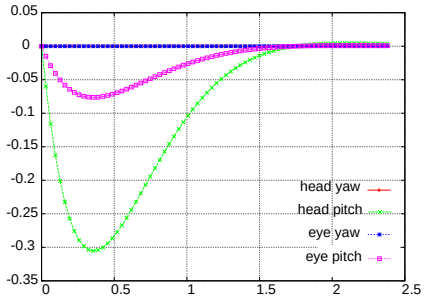

図 3: 受諾動作 (2 次遅れ系標準形モデル)

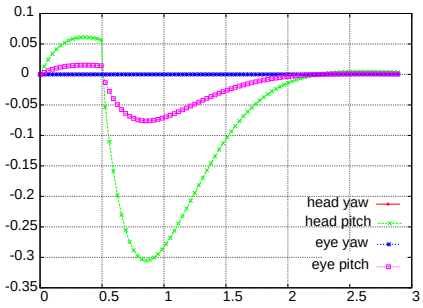

図 4: 相槌動作 (2 次遅れ系標準形モデル)

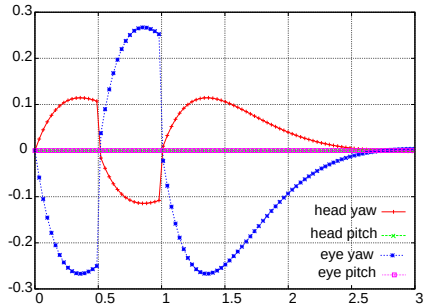

図 5: 否定動作 (2 次遅れ系標準形モデル)

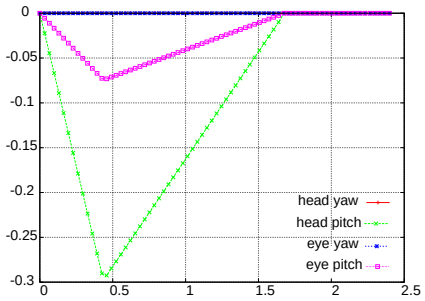

図 6: 受諾動作 (線型モデル)

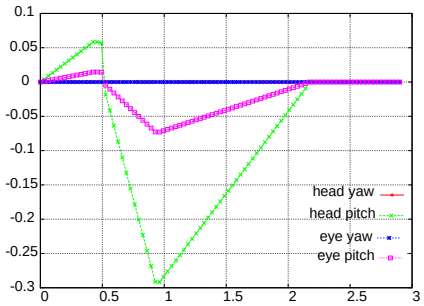

図 7: 相槌動作 (線型モデル)

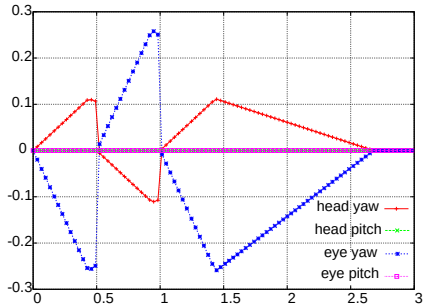

図 8: 否定動作 (線型モデル)

式 4 における α $(0 \leq \alpha \leq 1)$ は，頭部動作と視線動作の動作比率を表す．すなわち，$\alpha = 0$ の場合，頭部だけを動かすことを意味し，$\alpha = 1$ の場合，視線だけを動かすことを示す．$\alpha = 0.5$ の場合は，頭部及び視線がそれぞれ同じ割合で動作することを示す．この α は，値にかかわらず，対話エージェントの総合的な注視方向（頭部と視線を合わせた方向）が常に一定になるという特性を持ち，制御という面から非常に扱いやすい．

現段階では，あらかじめ発話テキストに言語情報とパラ言語情報が埋め込まれていることを想定している．例えば，ゆっくりと丁寧に「おはようございます」と挨拶を行う場合,「おはようございます＜意図:挨拶＞＜態度:丁寧＞＜スタイル:ゆっくり＞」のような形式でテキストに付加情報 (言語情報とパラ言語情報) として埋め込み，音声対話システムの頭部・視線動作生成モジュールに与えられることを想定している．

将来的には，言語情報からパラ言語情報を推定することが課題であり，藤崎モデルの F_0 パラメータ推定と同様に AbS 法を用い，対話の流れ，エージェントの個性・性格を考慮し，発話テキストのみからパラ言語情報を推定し，動作を自動的に出力する方法や，音声認識技術で用いられる HMM を利用し，言語情報 (発話テキスト) とパラ言語情報 (発話意図，態度，スタイル) を考慮した確率モデルから，言語情報が与えられた場合の事後確率最大となるパラ言語情報を推定するという方法も考えられる．

この場合，対話エージェントの個性・性格をどう定義し，推定したパラ言語情報との意味的な同期をどうとるかが課題となる．

3 2次遅れ系標準形と線形モデルによる視線運動・頭部運動の比較実験

3.1 実験の目的

本実験の目的は，対話エージェントの動作の生成方法 (線形，2 次遅れ系標準形) を比較し，2 次遅れ系標準形モデルによる生成が線形モデルによる生成よりも自然な (自発的，自律的) 動作を生成できる得ることを確認することである．

統制条件として，生成する対話エージェントの動作は，3 種類 (相槌，受諾，否定) の動作とし，あらかじめ動作指令 (インパルス指令)，方向指令 (ステップ指令) を与え生成した．被験者配置は被験者内配置とし，持越し効果を避けるため，カウンタバランス (ランダムな順番で提示) を行った．実験は，一対比較法 (線形モデルと 2 次遅れ系標準形モデルによる動作を続けて提示し比較) を用い，主観評価のアンケート方式で評価とした．このため，実験者効果はないと考える．

検討する各因子は，鈴木ら [9] がロボットのイメージを探るために考案した質問項目を参考に，7 種類（親しみやすい，感情をもつ様に感じる，冷たく感じる，単純な動きに感じる，生命がある様に感じる，心が通じない様に感じる，自分の意志で動作しているように感じる）とした．

集計方法は，間隔尺度 (そう思わない，あまりそう思わない，ややそう思わない，ややそう思う，かなりそう思う，そう思う) の累積頻度を集計し，各因子（親しみやすいなど）の累積値で評価を行った．

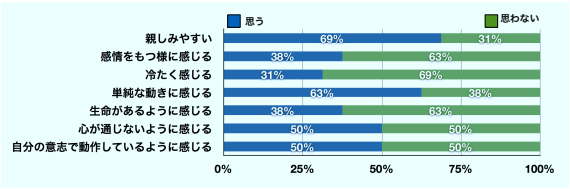

図 9: 受諾動作 (2 次遅れ系標準形モデル) の主観評価結果

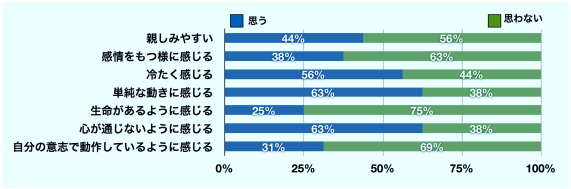

図 10: 受諾動作 (線形モデル) の主観評価結果

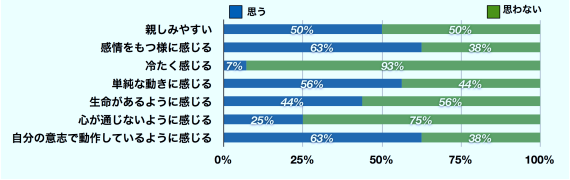

図 11: 相槌動作 (2 次遅れ系標準形モデル) の主観評価結果

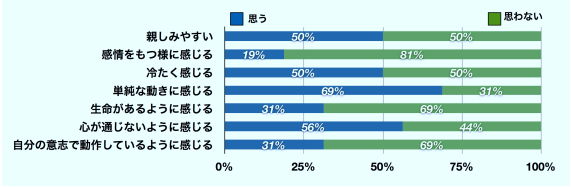

図 12: 相槌動作 (線形モデル) の主観評価結果

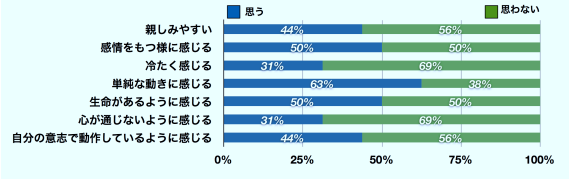

図 13: 否定動作 (2 次遅れ系標準形モデル) の主観評価結果

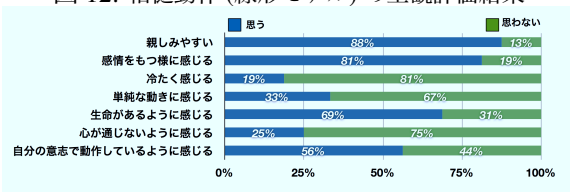

図 14: 否定動作 (線形モデル) の主観評価結果

3.2 実験方法

それぞれのグループにおいて，実験の目的，回答方法を伝え，実験対象として，受諾の動作 (図 3, 図 6)，相槌の動作 (図 4, 図 7)，否定の動作 (図 5, 図 8) について，線型モデル及び 2 次遅れ系標準形モデルによって生成した対話エージェントの動作ムービーを提示し，アンケートに答えてもらった．2 次遅れ系標準形モデルの動作指令 (インパルス指令) と方向指令 (ステップ指令) のパラメータは，$\zeta = 0.8, \omega_n = 3.0, K = 0.6, \alpha = 0.8$ (否定動作のみ $\alpha = 0.3$) とし，動作を生成し，複数のグループに分けられた被験者に対して，次のような手順で対話エージェントの動作ムービーを提示した．

- 線形モデルで生成された対話エージェントの動作ムービーを見せ，(動作繰り返しにより 30 秒)
- 2 次遅れ系標準形モデルで生成された対話エージェントの動作ムービーを見せ，(動作繰り返しにより 30 秒)
- 最後に，上記 2 つのムービーを左右に配置し同時に再生した．(動作繰り返しにより 30 秒)

被験者には，あらかじめ 3 種類の動作に対するアンケート用紙が渡されており，最後のムービーを見終わるまでアンケート記入を待つ必要はなく，自由に記入してもらった．

今回は，評価の差を明らかにするため，アンケート票の間隔尺度として「どちらでもない」は除外し，設問に対して「そう思う」もしくは,「そう思わない」のどちらかの尺度に応じた回答を記入してもらった．また，強い表現であればあるほど，被験者は選びにくいという知見に従い,「そう思う」,「そう思わない」は，強い表現（全く）はつけないままとした．一部被験者からは，強い表現をつけていない間隔尺度の順番がおかしいのではという意見が出たが，今回は,「全く」という語句を削除した理由を説明し，回答時には「全く」がつくものと等価と判断してほしいと伝えた．

なお，本実験で用いた線形モデルは，2 次遅れ系標準形モデルと比較するために，動作指令と方向指令を与える時刻は同じとした．これは，線形モデルにおいても，2 次遅れ系標準形と同じ時間タイミングで動作指令・方向指令を与えることを意味し，通常考えられる単純な線形モデル (動作の立上がりと立下がりが同じ時間タイミング）よりも，より精密なモデルを与え実験を行っている．

3.3 結果

対話エージェントの動作比較実験は，被験者は 16 名 (男性 14 名，女性 2 名) であった．

受諾動作 (図 9, 図 10) は 2 次遅れ系標準形モデルの方が，自発的・自律的動作であると比較的強く感じる傾向があったが，相槌動作 (図 11, 図 12) については，2 次遅れ系標準形モデルが自発的・自律的動作であると感じる傾向があきらかに強かった．ただし，否定動作 (図 13, 図 14) は，逆に，線形モデルの方が 2 次遅れ系標準形モデルよりも自発的・自律的動作であると感じる人が多いという結果になった．この理由としては，以下の考察が考えられる．

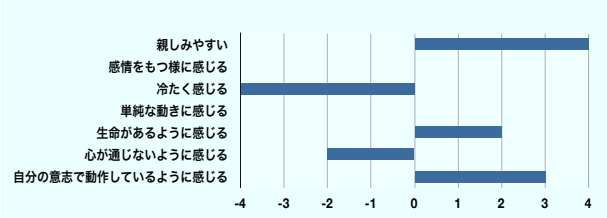

図 15: 2 次遅れ系標準形モデルによる受諾動作の相対評価 (線形モデルを基準)

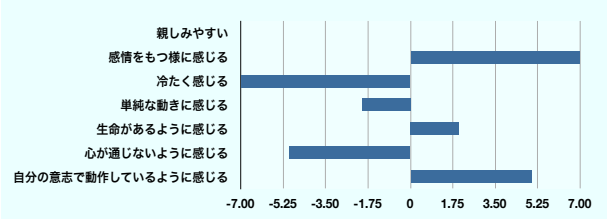

図 16: 2 次遅れ系標準形モデルによる相槌動作の相対評価 (線形モデルを基準)

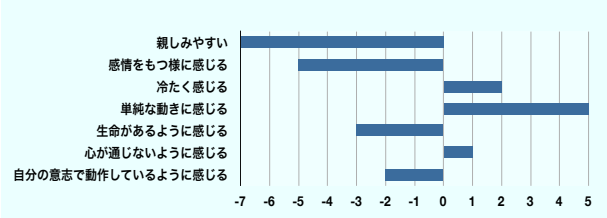

図 17: 2 次遅れ系標準形モデルによる否定動作の相対評価 (線形モデルを基準)

- 2 次遅れ系標準形は時間遅れの発生により初期動作が遅れている (線形モデルはすぐに動作を開始)
- 左右への連続的な動作を短時間で行うため，線形モデルの急激な動作切り替えが気にならない．逆に，レスポンスのよい動作に感じる．

図 15, 図 16, 図 17 は，それぞれ，2 次遅れ系標準形モデルによる各種動作 (受諾，相槌，否定) の評価の累積頻度から線形モデルの累積頻度を引いたものであり，線形モデルを基準とした 2 次遅れ系標準形モデルの各種動作の相対評価を表す．グラフが右側にあるほど，各因子（親しみやすい，感情をもつ様に感じる，冷たく感じる，単純な動きに感じる，生命がある様に感じる，心が通じない様に感じる，自分の意志で動作しているように感じる）に対する評価が高いことを示す．

これらの図によれば，受諾動作，相槌動作は，2 次遅れ系標準形モデルによる動作生成の評価が高いことが分かる．これに対し否定の動作は，2 次遅れ系標準形モデルによる評価は低く，逆に線形モデルによる動作生成の評価が高いことが分かる．これは，今回の実験で利用した対話エージェントの動作が動作生成のフレームレートの制約 (高フレームレートでの動作生成能力の低さ) により，速い動きを正確に再現できず，正しく評価できていないことが影響していると考えられる．

しかしながら，ゆっくりとした動作のように，その動作の様子がモデルにより異なる場合，2 次遅れ系標準形モデルは線形モデルより，その挙動に自然性（自発的，自律的）を与えることができるという点に加え，一般性かつ，シンプルなモデルから多様な動作を生成できるという点で優位であると言える．

4 まとめ

本報告では，頭部・視線動作の生成過程のモデル化の必要性を説き，言語情報とパラ言語情報と連動する非言語情報を再現するモデルを示し，頭部・視線制御モデルの基本的な数理モデルの定式化に向けて，2 次遅れ系標準形モデルと線形モデルの比較評価実験を行い，対話エージェントの動作生成に 2 次遅れ系標準形モデルが適用可能なことを示した．

今後は，発話テキストに内包される意図・態度・スタイルに関する情報からその言語情報に相応しいパラ言語情報の考察を通して，モデルの詳細化を行い，音声対話エージェントの頭部・視線動作の生成の適用範囲 (思考など知的活動によって誘起される頭部・眼球運動等) を広げていく予定である．

参考文献

[1] Argyle, M., and Dean, J.: Eye contact, distance and affiliation, Sociometry, 28, pp. 289–304, (1965).

[2] Argyle, M.: Bodily Communication, 2nd ed. Methuen Co., (1988).

[3] Duncan, S.D., Jr., and Fiske, D.W.: Face-to-face interaction: Research, Methods, and Theory, Hillsdale, N.J: Lawrence Erlbaum, (1977).

[4] Kendon, A.: Some Functions of Gaze-direction in Social Interaction, Aca Psychologica, Vol. 26, pp. 22–63, (1967).

[5] Lord, C., Haith, M.M: The perception of eye contact, *Perception & Psychophysics*, 16, 413–416, (1974).

[6] 藤崎博也: 韻律研究の諸側面とその課題, 日本音響学会 平成 6 年秋季研究発表会講演論文集, 1, 287–288 (1994).

[7] 中沢正幸，西本卓也，嵯峨山茂樹: 擬人化音声対話エージェントにおける視線制御モデルの提案, 人工知能学会 SIG-SLUD-A303，pp. 21–26, Mar, (2004).

[8] 嵯峨山茂樹, 他: 擬人化音声対話エージェントツールキット Galatea, 情処研報, IPSJ–SLP–45–10, pp. 57–64, (2002).

[9] 鈴木佳苗, 橿淵めぐみ, 坂元章, 長田純一，ロボットに対するイメージ尺度の作成とイメージ内容の検討 (1) – 三つ組み法によるロボットイメージ尺度の作成 –, 日本心理学会第 66 回大会発表論文集, 114, (2002).