Vol.2017-EC-43 No.10
2017/3/10

多人数会話での自然な視線・動作を実会話事例の学習により実現するエージェント

葛島健人[†1]　三武裕玄[†1]　長谷川晶一[†1]

概要：ロボット技術や VR 技術の発展によって会話エージェントが，日常生活に入り込み，人々の交流を豊かにすることが期待されている．会話エージェントが人同士の会話に自然に入り込むためには，人同様に視線やジェスチャー等の非言語的なふるまいを適切に行わなければならない．本研究では，多人数会話の実会話事例から機械学習を用いて行動決定モデルを獲得することで，自然な視線・動作を行う会話エージェントを実現する． 具体的には，まず実会話での話者と聴者の視線・動作・発話の時系列を記録し HMM を学習する．次に，得られた HMM に話者の行動を入力し聴者の行動を推定することでエージェントの視線・動作を生成する．

Conversational Agent Learning Natural Gaze and Motion of Multi-Party Conversation from Example

KENTO KUZUSHIMA[†1]　HIRONORI MITAKE[†1]
SHOICHI HASEGAWA[†1]

1. はじめに

近年，キャラクタ性を持ったロボットの登場や，没入 VR の発達によって人とキャラクタを同じ世界に立たせることが可能になってきた．人はこれまで，ゲーム等の仮想世界においてキャラクタに様々な役割を与え，コントローラを通してインタラクションを行ってきたが，キャラクタと同じ世界に立つことが可能になった今，キャラクタに会話等の直接的な社会的コミュニケーションを含むインタラクションを求めている．中でも会話は，社会的コミュニケーションの重要な要素であり，人はキャラクタと自然な会話を行えないと相手に社会性や知性を感じられずコミュニケーションへの期待を削がれてしまう．そのため，現在，自然な会話を実現するための会話エージェントの研究が盛んに行われている．

会話エージェントとの自然な会話を実現するためには，話し言葉や文字といった言語インタラクションだけでなく視線やジェスチャーといった非言語インタラクションを実現することが必要である．

また，自然な会話の実現に必要なこととして会話のインタラクティブ性の再現が挙げられる．会話エージェントは話者の発話・視線・動作といった会話行動に対してリアルタイムに反応動作を決定，行動しなければならない．従来の会話エージェントは人手で作成した行動決定モデルに従って動作するものが多いが，人手で作成するためには実会話を注意深く観察し，法則性を発見，アルゴリズム化する必要があり，容易な作業ではない．

そこで本研究では，会話エージェントが話者の会話行動に応じてリアルタイムに反応動作を決定するための行動決定モデルを，実会話事例から HMM を用いた機械学習によって獲得することを目的とする．実会話における会話行動を記録し，それを再現するような行動決定モデルを機械学習により獲得することで，従来の人手で行動決定モデルを作成する方法よりも自然な会話を実現することが期待できる．

会話行動の中でも視線は，話題の対象への注意を共有するために使用されたり，受話者によって「次話者になるための発話準備中であること」を伝えるために使用されたり，発話者によって「発話権を譲ろうとしていること」を伝えるために使用されたりと、多くの場面で相手に情報を伝える手段として用いられており，自然な会話の実現において重要なものである[1]．そこで本研究では，会話時の視線の動かし方に着目して学習や再現を行う．

2. 関連研究

非言語情報も用いてコミュニケーションを行う会話エージェントの研究は数多くなされている．渡辺らの研究[2]では話者の発話の切れ目を検知するモデルを作成し，それを用いることでエージェントが話者に頷きを返すタイミングを適切に取得できることを示した．ただし，このエージェントは発話にのみ反応し，話者の視線や動作には反応しない．

DeVault らの研究[3]では話者の発話や視線，表情や動作といった多くの会話行動に応じて適切な動作やタイミングを返すエージェントを作成した．この研究では，行動データや行動決定モデルを人手で精巧に作成することでジェス

†1 東京工業大学
Tokyo Institute of Technology

Vol.2017-EC-43 No.10
2017/3/10

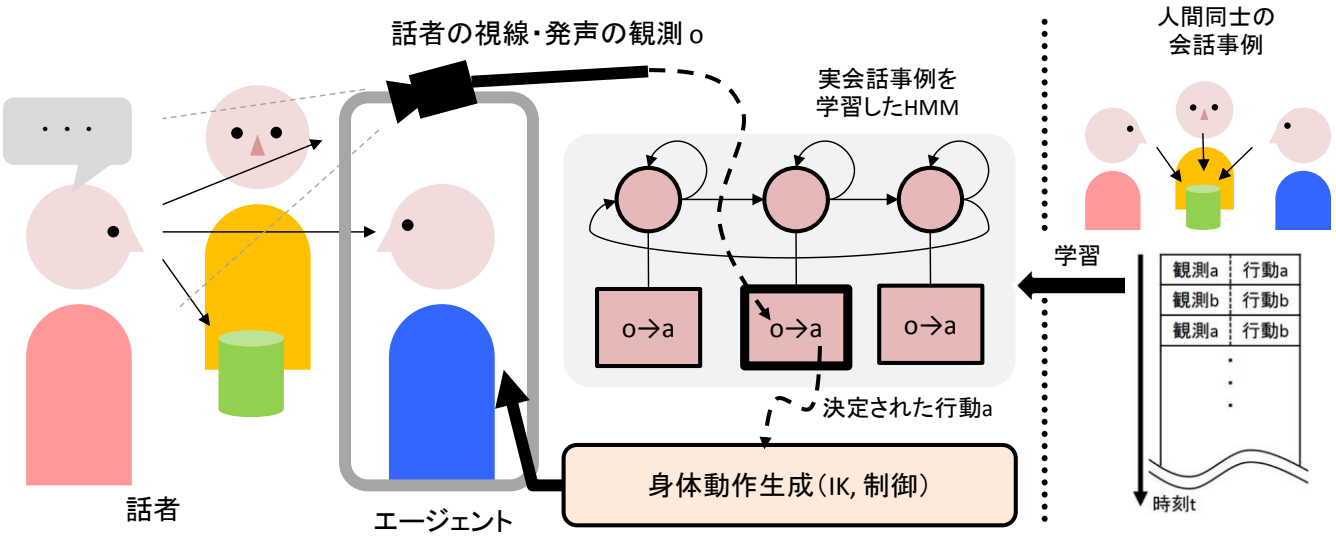

図 1 システム全体像

チャーの選択やそれを話者に返すタイミングについては適切である一方，視線についてはモデルの作成が十分ではなく，エージェントが話者から視線を動かさない，といった不自然さを感じることがある．

会話時の視線の重要性に着目し，それを再現する研究として益子らの研究[4]が挙げられる．この研究では，話者の発話，視線，動作と，エージェントの発話情報から，エージェントの会話状態を同定し，その状態に適切な視線と頭部の動作を決定，再現した．ただし，この研究ではエージェントの行動決定モデルは人手で用意したものであるため，自然な会話を十分に再現できない可能性がある．

3. 提案手法

人間同士が会話で行う自然な視線移動の振る舞いを再現するため，会話エージェントの行動決定モデルを実会話事例から学習することで得る．

今回の研究では 3 人会話で，話題の対象となる物体が存在し，エージェントが聞き役となるような会話を対象とした．まず，実会話事例における視線移動や発話等の時系列データを記録し，次にそれを，学習に利用できるように「何を見ていたか」「喋っていたか」等の離散的なデータに変換する．その様にして得られた学習データから隠れマルコフモデル（HMM）を学習し，獲得した HMM を行動決定モデルとして用いることでインタラクティブに会話行動を決定する．決定に基づきキャラクタ身体物理モデルを駆動することで会話動作を実現する．全体像を図 1 に示す．

3.1 実会話事例の収集

会話エージェントの行動決定の規範となる，人間同士の会話の様子を記録する実験を行った．2 名の話者とエージェント役として 1 名の聴者に話題の対象となる物体を挟んで対面会話を行ってもらい，その様子を記録した．今回の実験では，参加者の頭部の位置・向きや物体の動きを計測するためにモーションキャプチャーを，頭部を基準とした視線を計測するために装着型のアイトラッカーを，発話を計測するためにマイクを使用した．

3.1.1 実験環境

会話中に関係のない物体への注意を極力減らすために物が少ない部屋を用意した．部屋の中には机が 2 つ置いてあり，そのうち 1 つを囲むように参加者に座ってもらった．参加者間の視線移動だけでなく，物への視線が生じるように机の上に話題の対象とする物体を 2 つ置いた．参加者にはモーションキャプチャーのマーカーとアイトラッカー，マイクを身に着けてもらい，物体にもマーカーを着けた．（図 2）

モーションキャプチャーは「OptiTrack」[5]，アイトラッカーは「Pupil」[6]，マイクは「HYP-190H」[7]を使用した．

図 2 参加者に装着してもらう装備

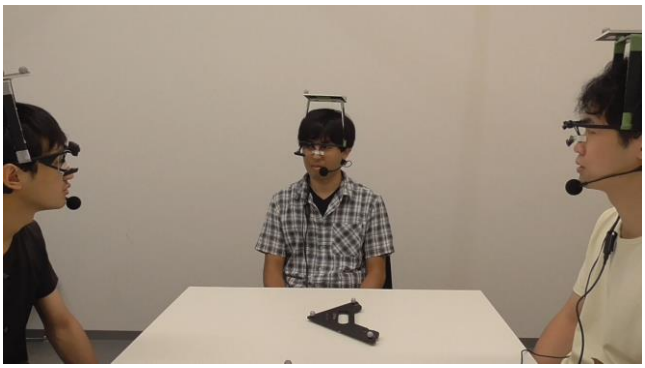

図 3 会話中の様子

3.1.2 会話

会話中の様子を図 3 に示す．図の左右にいる参加者が話者で中央奥にいるのがエージェント役である．エージェント役は発話せずに話者の会話を聞き，自然に視線移動や頷きといった会話動作を行ってもらった．話者 2 人はエージェント役に話しかけ，また，話者同士で会話を行った．会話内容にはタスクを設定せずに自由に会話を行ってもらったが，その際に，机の上に置いてある物を何かに見立てながら会話をしてもらうことでエージェント役の視線移動を引き出した．

3.1.3 収集データ

541.6 秒の会話を約 0.02 秒毎に記録し，47 次元のベクトル 27080 個の時系列データを得た．会話参加者に関しては頭の位置と傾き，視線の方向と発声の音量を記録した．机の上に置いてある物体は位置と傾きのみを記録した．それぞれの対象の記録内容を表 1 に示した．位置は三次元座標，傾きはクォータニオン，視線は方向ベクトルをそれぞれ記録した．参加者 1 人あたり 12 次元，物体 1 つあたり 7 次元記録した．参加者 3 人と物体 2 つで合計 47 次元のデータとなった．

表 1 記録内容

記録対象	記録内容
位置	(x, y, z)
傾き	(w, x, y, z)
視線	(x, y, z)
音量	volume

3.2 HMM の学習

記録したデータを基に、会話での視線のやりとりを認識・予測するモデルを学習することで、会話エージェントの行動決定手法を構成することを考える。会話での人の行動は，状態依存性とある程度の確率性があると考えられる．そこで，会話インタラクションの時系列データによって HMM を学習し，得られた HMM に，ユーザの会話行動を入力し対応する聞き手の行動を推測することで会話エージェントの行動を決定する．

3.3 HMM によるインタラクティブ行動決定

前節で獲得した HMM の出力記号は注視状態の組（o, a）であるが，行動決定手法として用いる際は o を対話相手の注視状態（=観測），a を対応するエージェントの注視状態（=行動）と見なす．すなわち，本手法によって観測 o を入力として行動 a を決定する行動決定アルゴリズムを作る．

HMM の遷移行列を T，状態集合を S，エージェントが取りうる行動の集合を A とし，時刻 t に状態 s にいる確率を p_s（t），状態 s の出力記号が（o, a）となる確率を $p_{(o,a)}$（s）とする．

まず, 遷移確率に基づき時刻 t+1 での状態確率分布 P(t+1) を得る．

$$P(t+1) = T^t P(t) \quad (1)$$

次に任意の状態 s で観測 o と行動 a の組が出力される確率を足し合わせ，時刻 t+1 において行動 a が選ばれる確率 p_{t+1}（a）を計算する．

$$p_{t+1}(a) = \frac{\sum_{s \in S} p_s(t+1)\, p_{(o,a)}(s)}{\sum_{a \in A}\sum_{s \in S} p_s(t+1)\, p_{(o,a)}(s)} \quad (2)$$

最後に，確率 p_{t+1}（a）に基づき行動 a を選択し，観測 o と合わせて（o, a）を受理したとみなして状態確率分布を Forward アルゴリズムを用いて更新する．

以上の処理を学習に用いた時系列データと同じ時間刻み（今回は 0.1 秒）ごとに実行することで，インタラクティブに行動を決定する．本手法の概要を図 4 に示す．

3.4 注視対象の選択

前節で決定した行動に応じて，エージェントの注視対象を選択する．HMM の出力する行動は注視対象の指示であり，1:発話者，2:発話者の注視物体，3:非発話者，4:非発話者の注視物体，5:誰も見ていない物体，6:それ以外，の 6 種類が存在する．それぞれの指示に対し，選択するアルゴリズムを表 2 に示す．ただし該当する対象が存在しなかった場合は直前の注視物体を選択する．

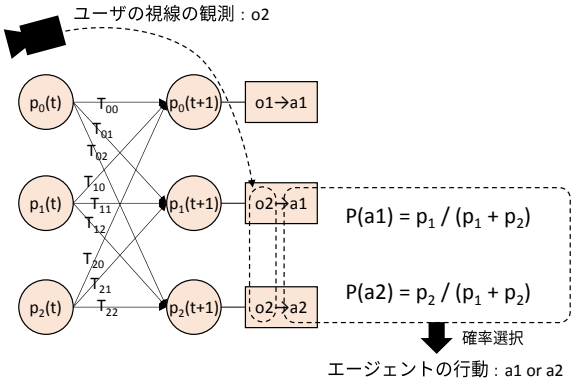

図 4 提案手法概要

表 2 注視対象選択方法

指示	選択方法
1: 発話者	発話者の中に，直前の注視対象がいる場合はその者を，そうでない場合は発話者の中からランダムで選ぶ.
2: 発話者の注視物体	発話者の注視物体の中に，直前の注視対象がいる場合はその物体を，そうでない場合は発話者の注視物体の中からランダムで選ぶ.
3: 非発話者	非発話者の中に，直前の注視対象がいる場合はその者を，そうでない場合は非発話者の中からランダムで選ぶ.
4: 非発話者の注視物体	発話者の注視物体の中に，直前の注視対象がいる場合はその物体を，そうでない場合は発話者の注視物体の中からランダムで選ぶ.
5: 誰も見ていない物体	発話者も非発話者も注視していない物体の中に以前見ていた物体がある場合はその物体を，そうでない場合は発話者も非発話者も注視していない物体の中からランダムで選ぶ.
6: それ以外	壁を注視対象とする

3.5 身体動作生成

注視対象選択結果に基づき，キャラクタ身体物理モデルの眼球・頭部・胴体を動作させ注視動作を実現する．眼球・頭部が注視対象の方を向くような姿勢を目標姿勢とし，各関節角度を逆運動学により決定する．目標関節角度を実現するよう関節トルクを PD 制御により与える．
注視対象が変化した際は目標姿勢間を滑らかな加減速で補間する．逆運動学における関節の重み付け，PD 制御の係数を調整し自然な身体動作を得る．

4. 実装

Unity[9]を使用し，提案手法による聞き手エージェントを交えて VR 空間内で会話を行うデモを作成した．会話データ収集時と同様に机と 3 体のキャラクタモデルを配置し，そのうち 1 体に前章で作成した HMM を使ったインタラクションシステムを組み込んだ．残りの 2 体はアバターとし，ユーザが入り込んで会話を行う．ユーザへの映像の提示には Oculus Rift DK2[10]を使用した．ユーザの視線，発声，頭の位置と傾き，手の動きを計測し，エージェント，アバターの動作の生成に利用した．計測では視線は HMD に内蔵したアイトラッカー，発声はマイク，頭の位置と傾きは Oculus Rift 内部のセンサー，手の動きは Kinect センサー[11]を使用した．デモの様子を図 5，6 に示す．


図 5 体験の様子


図 6 デモの様子

2 名のユーザ同士が視線やジェスチャーにより自然に会話できるようにするため，ユーザの視線，発声，頭の位置と傾き，手の動きをアバターで再現した． また，提案手法の他にも会話時の動作を生成する有用なモデルとしてボトムアップ注意による動作とうなずき動作を実装した．
なお，デモにおいてキャラクタモデルとしてプロ生ちゃんを，ご利用ガイドライン[12]に従って使用した．

4.1 ボトムアップ注意による注視

人間は注視対象を，意思に基づくトップダウン注意だけでなくボトムアップ注意によって選択することがある．会話においては,「話者を見る」「話者が見ているものを見る」といった行動はトップダウン注意によって得られるものと言える．一方，ボトムアップ注意とは感覚刺激の著しい変化によって払われる注意のことで「急に動いたものを見る」「大きな音がした方向を見る」といった行動はボトムアップ注意によって得られるものである．本研究では，会話状況を観測し，それに応じた注視対象の選択を HMM を用いて行っているが、これはトップダウン注意による注視を生成していると考えられる．一方，会話では相手のジェスチャー等の動作に反応して視線が動くなどボトムアップ注意

に基づくふるまいも起きる．益子ら[4]もこのことに注目して会話におけるバーチャルヒューマンの動作に取り入れている．しかし，今回の HMM ではこのような動作は生成されない．そのため，エージェントにボトムアップ注意についての実装をルールベースで行った．ユーザの頭，胴体，手の動きに対して動きの速さに応じた注意量を割り振り，決めておいた閾値を超えたものを注意対象とする．実装に関しては三武らの研究[13]を参考に行った．

4.2 うなずき動作

本研究で作成したエージェントは発話を行わない聞き手エージェントである．そのため，聞き手エージェントとして振る舞いをより自然にするためにうなずき動作の実装をルールベースで行った．アルゴリズムは渡辺らが行った研究[2]を参考にし，独自に簡易にしたものを実装した．

5. 評価

5.1 実現した振る舞い

作成したエージェントが実会話から学習できた会話動作を確認するために話者 2 名とエージェントとで会話を行った．会話中に見られたエージェントの振る舞いとその時の話者 2 名とエージェントの行動をそれぞれ表に起こしたものを図 7~9 に示す．表の横軸は会話が始まってからの時刻を表し，縦軸は視線の対象を表す．表中の塗りつぶされた点は発話していたことを表す．会話を行うと図 7 の様に「基本的には発話者を見ながらも，受話者も時折ちらちらと見る」，図 8 の様に「発話者がエージェントを注視しながら喋ると，エージェントも発話者のみを注視する傾向が強まり，話者がエージェントを注視するのをやめるとエージェントも話者以外を見るようになった」，図 9 の様に「発話者が何らかの対象を見ながら発話すると，発話者だけでなく，発話者が注視している対象を同時に注視する共同注視が行われる」といった振る舞いを見ることができた．

5.2 アンケートによる印象の評価実験

提案手法によって作成した注視対象決定アルゴリズムを評価するために，アンケートによる印象調査を行った．

5.2.1 評価方法

まず，本研究で作成したエージェントの他に「常に発話者を注視対象に選ぶ」という単純な聞き手エージェントを用意し，また，人間の行動を観察して，「発話者を見る」「発話者が見ている相手を見る」「発話者が見ている物体を見る」といった，いくつかの視線に関する振る舞いをルールベースで再現する聞き手エージェントを約 3 週間かけて用意した．加えて，実際に人間が操作して動作する WOZ 法で動くエージェントも用意した．人間はエージェントに注視対象のみを指示し，他のエージェントと視線以外の部分で挙動が異なることのないように視線・動作の生成を行う．前章で作成したデモを用いて，それぞれのエージェントと話者 2 名とで机上に置いてあるものを説明する

図 7 発話者/受話者を注視

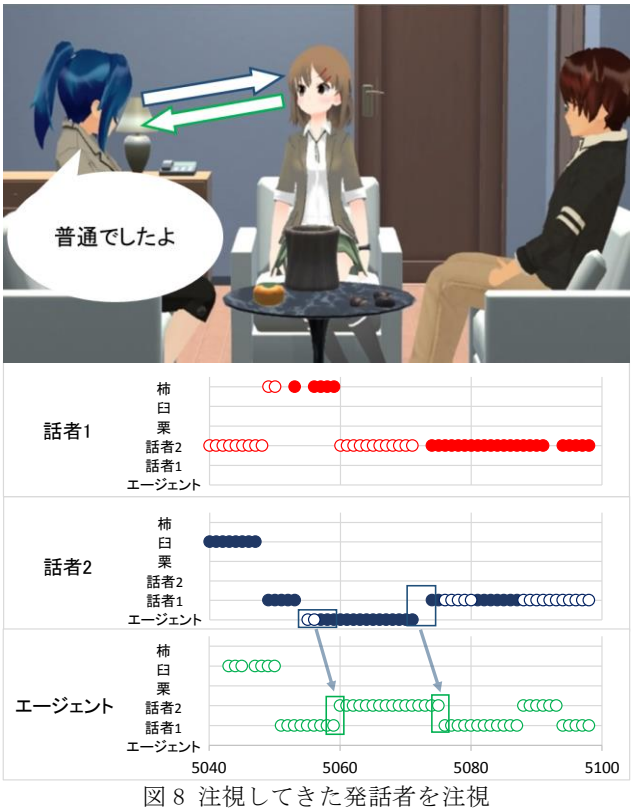

図 8 注視してきた発話者を注視

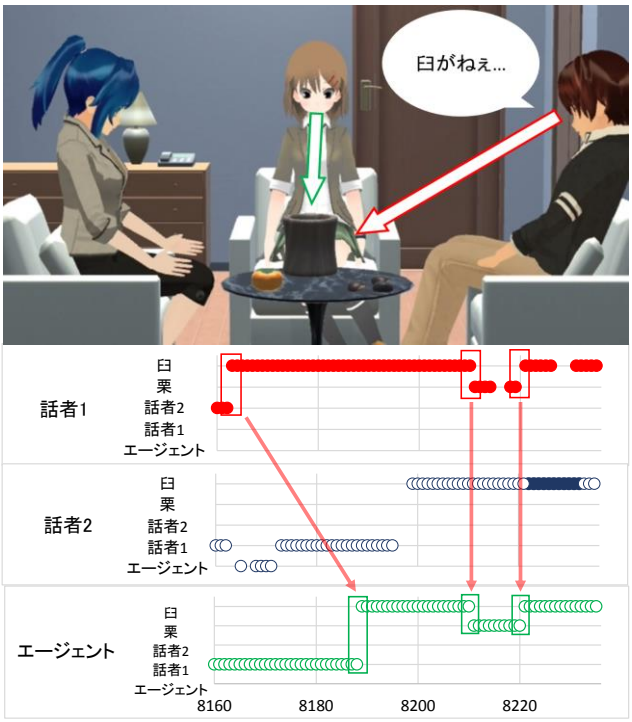

図 9 共同注視

タスクの会話を行い，会話の様子をビデオに記録した．この際，提案手法を正しく評価するために，ボトムアップ注意による注視対象決定機能及び，うなずき動作機能は使用しなかった．

次に，作成した 4 つのビデオを被験者に見せ，それぞれのビデオの聞き手エージェントの印象を以下の 8 つの評価軸について 10 段階で絶対評価をしてもらった．

- 不自然
- いきいきしている
- 大げさ
- 人間くさい
- 機械的
- 話を聞いている
- 興味を持っている
- 話についてきている

実験では，被験者にビデオのエージェントがどのような手法で動いているかは一切伝えず，また見せるビデオの順番はランダムに入れ替えた．

5.2.2 結果

インターネット上で実験参加者を募り，23 名の被験者から回答を得られた．エージェント毎に各評価軸での評価値の平均を図 11 に示す．なお，「常に発話者を注視対象に選ぶエージェント」を WAT，「人間の行動を観察してルールベースで作成したエージェント」を ATT，「提案手法で作成したエージェント」を HMM，「人間が WOZ 法で操作したエージェント」を WOZ と呼ぶ．また，表の項目の表記は省略しているが前節で述べた通りである．

図中に「提案手法（HMM）とそれ以外のエージェント」

図 10 アンケート画面

と「人間が動かしているエージェント（HUM）とそれ以外のエージェント」で、関連二組の差の t 検定を行い，$p<0.05$ で有意に差が認められた箇所と $0.05<p<0.1$ で有意な傾向が見られた場所を示した．

5.3 考察

5.1.節より提案手法を用いることで人同士の多人数会話から，いくつかの視線についての振る舞いを学習し，再現できたことが分かる．

5.2.節の図 12 を見ると，常に発話者を注視するエージェント（WAT）は人間（WOZ）よりも「不自然」という印象が強いことが分かり，ルールベースのエージェント（ATT）は人間（WOZ）よりも「不自然」という印象が強い可能性がある．一方，提案手法（HMM）のエージェントは人間（WOZ）よりも「不自然」な印象が強いとは言えないことが分かった．また，全てのエージェントは人間（WOZ）より「話を聞いている」「話しについてきている」という印象において有意差がないことが分かる．これは，被験者が「話を聞いている」「話についてきている」という印象を視線以外の情報が無い場合は相手の「発話者を見る」という行為から認識しているからだと考えられる．また，アンケート実験前は「話についてきている」印象は「発話者を見る」という行為だけでなく「話題の対象を見る」行為の有無や，その自然さによって大きく変化するのではないか，という考えがあったが，常に発話者を見るエージェント（WAT）が人間（WOZ）と有意に差がないことから必ずしもそうでないことがわかった．また，提案手法（HMM）は人間（WOZ）と比べ，「大げさ」という印象が強く，「いきいきしている」「興味を持っている」という印象が強い傾向があることが分かる．これは，提案手法で作成したエージェント（HMM）の学習がうまくい

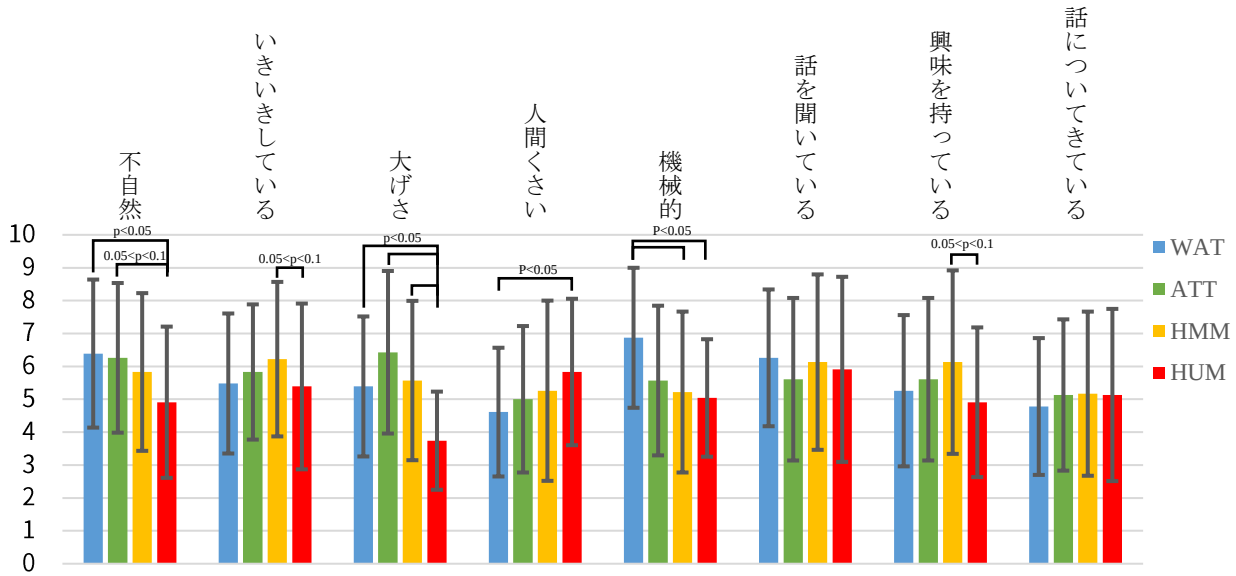

図 11　スコア平均値

かず，視線対象の切り替えを人間（WOZ）より頻繁に行ってしまった，もしくは，切り替えのタイミングが不自然に早く行われたために，被験者が「大げさ」と捉えたと考えられる．同様に，それをポジティブに捉えて「いきいきしている」「興味を持っている」という印象を持ったと考えられる．また，提案手法のエージェント（HMM）及び人間が動かしたエージェント（WOZ）の印象はエージェント役の性格やその時の傾聴態度，調子に影響される．そのため，実会話データ収集時とビデオ撮影時のエージェント役の印象の差が反映されたとも考えられる．また，ルールベースで作成したエージェント（ATT）と比較すると，全ての項目で有意な差が無いことが分かる．これは期待した結果とは異なったが，提案手法を利用することで，ルールベースで作成したエージェントと同等のエージェントを作ることができたとも考えられる．さらに発話者を見るだけのエージェント（WAT）と比較すると，「機械的」だという印象が弱いことが分かる．これは，両エージェントができる振る舞いの差を考えると，被験者が「受話者を見る行為」や「発話者の自分への視線に反応して見つめ返す行為」「話題に対象物を発話者と一緒に見るといった行為（共同注視）」を行うことができないエージェントは機械的であると認識したと考えられる．

6.　まとめと今後の展望

評価を行った結果，提案手法を用いることで人間同士の会話で発生する視線を使った振る舞いの一部を学習し，再現することができることが分かった．また，時間をかけルールベースで作成したものと比べても印象評価のスコアには有意差が生じなかったことから，提案手法を用いることで，約 3 週間かけてルールベースで作成したものと同等のエージェントを 1～2 時間の計測と 3～6 時間程の学習から作ることができた可能性を示唆した．

また，本研究の提案手法により作成したエージェントは人間が中に入ったキャラクタより「いきいきしている」，「興味を持っている」印象が強かったことを受けて，エージェント役を変えたり，演技したりして学習データを作成することで，印象の違うエージェントを簡易に作成できることが期待できる．本研究の手法を応用することで，ルールベースで印象の違うエージェントを作る際と比較して作成コストの削減や，作成したエージェントの印象の改善について検討を行っていきたい．

本研究ではエージェントの注視対象を決定するために，実会話データを離散的な学習データに変換を行ったが，連続的なデータを学習することができれば,「首をあまり動かさずに視線のみを注視対象へ向ける」といった振る舞いも再現することが期待できる．また，本研究では会話時の視線に着目したが，今回の手法は頷き等，視線以外の非言語的な振る舞いの再現にも利用できることが期待できる．そこで今後は，連続値を使った学習方法の検討や，本手法の視線以外の会話行動の生成への応用を行っていきたい．

7.　謝辞

本研究（の一部）は国立研究開発法人科学技術振興機構（JST）の研究成果展開事業「センター・オブ・イノベーション（COI）プログラム」の支援によって行われました．

参考文献

[1] Kendon, Adam. "Some functions of gaze-direction in social interaction." *Acta psychologica* 26 （1967）: 22-63.

[2] 渡辺富夫. "身体的コミュニケーションにおける引き込みと身体性-心が通う身体的 コミュニケーションシステム E-COSMIC の開発を通して." ベビーサイエ

ンス 2 （2003）: 4-12.

[3] DeVault, David, et al. "SimSensei Kiosk: A virtual WOZan interviewer for healthcare decision support." *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[4] 益子宗：利用者注意に基づく能動的 CG, 筑波大学修士論文 2004

[5] SPICE.「OptiTrack」, [http://www.mocap.jp/optitrack/]

[6] Pupil Labs.「Pupil」, [https://pupil-labs.com/pupil/]

[7] オーディオテクニカ,「HYP-190H」, [https://www.audio-technica.co.jp/mi/show_model.php?modelId=2531]

[8] http://accord-framework.net/

[9] Unity Technologies.「Unity」. [http://japan.unity3d.com/]

[10] Oculus.「Oculus Rift DK2」. https://www3.oculus.com/en-us/dk2/

[11] Microsoft.「Kinect Sensor v2」, [https://www.microsoft.com/en-us/]

[12] プログラミング生放送.「ご利用ガイドライン」. [http://pronama.azurewebsites.net/pronama/guideline/]

[13] 三武 裕玄, 青木 孝文, 長谷川 晶一, 佐藤 誠：精緻なフィジカルインタラクションにおいて生物らしさを実現するバーチャルクリーチャの構成法, 日本バーチャルリアリティ学会論文誌, Vol.15, No.3, pp.449-458, 2010.