

# ユーザの身体動作から興味を推定する マルチモーダル対話システム

小関健<sup>†1</sup> 小坂哲夫<sup>†1</sup>

**概要：**近年、音声認識・音声合成の技術向上に伴い、ヒトとマシンが音声によってインタラクションを行う対話システムが急速に普及してきた。しかし、利便性の低さや魅力の無さを理由として、これらの対話システムの使用を避けるユーザが存在している。本研究ではこのような使用欲求低下の問題を解決することをねらい、ユーザの音声のみでなく、身体動作も考慮する対話システムを開発してきた。本稿では、ユーザの顔面の向きから興味を推定することで対話エージェントが適切に話題選択を行うシステムを構築し、対話実験を通じてその機能の効用について検証する。実験より、エージェントがユーザの興味を推定して話しかけを行うことは対話をより行き詰まりにくくする効果があることが示唆された。

## 1. 概要

近年は音声認識技術の向上に伴い、Siri や Google Home のように音声対話をインタフェースとしたシステムが普及してきている。しかし、システムの音声認識の精度の低さ、応答の機械っぽさ、魅力の無さなどを理由として、ユーザがシステムの使用を避けることがある。本研究ではこれをユーザが音声対話システムに対して感じている心理的距離と名付け、それを解消することを目標としている。

心理的距離の改善のためにはシステムをより人間らしいものにすることが有効であるという知見[1][2]から、本研究でも擬人化エージェントの使用、および身体動作による内部状態表現を検討してきた。しかし、対話エージェントがユーザに身体動作を示すシステムが検討される一方、ユーザがシステムに対して示す身体動作を取り扱う仕組みの有用性については、まだ十分に検討されていない。現在はCeVIO[3]が開発する対話型デジタルサイネージ「CeVIO Vision」や、Live2D 社[4]が提供するタッチパネルアプリなどにおいても Kinect センサーによる簡易的なジェスチャ認識機能が実装されており、音声とジェスチャの両方を利用したヒトーマシン間インタラクションへの需要・関心は高まってきているといえる。またユーザのノンバーバル行動を巧みに利用した対話を実現することは、システムとの会話の自然性の向上や、情報・意思伝達の円滑化などによって、ユーザのシステムに対する心理的距離を軽減することが期待できる。

前研究では、ヒトとコンピュータのより自然な（ヒト同士のコミュニケーションに近い）音声対話を実現することをねらい、対話中のユーザの身体の動きをヒトーマシン間インタラクションに導入するためのシステムを構築した[5]。本稿ではこのシステムを利用することで、音声インタラクションに身体動作を導入することの効用を、実用的な

局面で検証する。

## 2. 提案システムの構築

本研究が提案する、ジェスチャと音声によるマルチモーダル対話システムの概要を図 1 に示す。提案システムは MMDAgent によって構築される音声対話アプリケーションと、Kinect によるジェスチャ認識アプリケーションによって構成される。これら 2 つのアプリケーションは並列的に動作しつつ、プロセス間通信によってジェスチャ認識の結果を文字列として送受する。

### 2.1 音声対話アプリケーション

音声対話アプリケーションは提案システムの根幹となる部分であり、MMDAgent[6]を用いて構築される。MMDAgent は名古屋工業大学国際音声研究所によって開発された、音声インタラクションシステム構築のためのオープンソースツールキットである。MMDAgent によって構築されるシステムでは画面上に表示される 3D キャラクターをエージェントとして音声インタラクションを行うことが可能で、音声入力のための大語彙連続音声認識エンジン Julius、合成音声出力のための生成エンジン OpenJTalk が搭載されている。ユーザの音声を認識したり、各種の入力に応じてシステムの応答を制御することが音声対話アプリケーションの役割となる。

### 2.2 ジェスチャ認識アプリケーション

ジェスチャ認識アプリケーションでは、Xbox One Kinect センサー(以下単に Kinect センサーと呼ぶ)を用いてユーザの身体情報を取得する。Kinect ではセンサーに映った人体から骨格情報などを取得し、その姿勢や連続的な動きをジェスチャという形で認識することができる。またユーザの顔面を認識し、口角の位置から笑顔であるかどうかを判別したり、センサーに対する Yaw, Pitch, Roll 角を取得することができる。

<sup>†1</sup> 山形大学  
Yamagata University

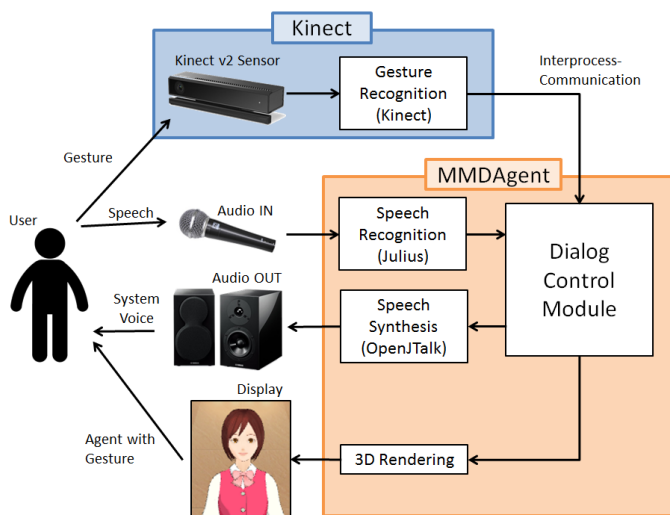


図 1 提案システムの概要図

センサーから得られた骨格・姿勢などの体幹に関わる情報は、予め作成されたジェスチャ認識器を通してジェスチャが行われたかどうかを判定する。また、ユーザの顔面の情報は数値化された状態で出力され、これらの数値を基にユーザがどの方向を向いているかなどを判定する。

ジェスチャ認識アプリケーションによる判定の結果は音声対話アプリケーション内の対話制御モジュールに文字列として送信される。音声対話アプリケーションが文字列の受信をトリガーとして対話を制御することによって、ユーザは音声とジェスチャの両方を通じて提案システムとインタラクションを行うことができる。

#### ユーザの身体動作からの興味推定

本研究では、対話システムを使用中のユーザが示す身体動作を「ジェスチャ」と呼ぶことにする。これにはユーザが情報提示やコマンド入力のために意識的に行うものを含んでいることはもちろん、うなずきや腕組み、視線移動のための頭部動作などの無意識的に実行されるものも含む。本研究では特に無意識に行われるジェスチャに注目する。

対話中の人間のジェスチャからは、多くの情報が読み取れると期待している。例えばシステムの前で腕組みをして首をかしげるユーザがいた場合、システムに対してどう話しかければいいのか迷っているのかもしれないと推測できる。また、システムがユーザに対して何かを説明しているとき、ユーザが何度もうなずくようであれば、ユーザはその話を理解できていると考えられる。

ユーザの身体動作から関心を推定する手法については、山倉の研究[7]でその知見が述べられている。山倉によれば、人対人の対話においては聞き手の顔の向き・うなずきの有無・上半身の動きを捉えることで、話題への関心の有無は70.5%の精度で推定が可能である。うなずきが多く検出されたり、顔が話し手や話題に関連した物体の方向を向いている場合に、聞き手は話題に対して関心があると述べられ

ている。

本研究では提案システムの機能のうち顔認識の機能を活用し、山倉が述べたような知見をヒト対システムのインタラクションにおいて検証する。山倉や千葉[8]など人間同士の対話を分析することによってヒト対システムのインタラクションについての展望が得られたという報告はあるが、人対システムにおいてそれらの知見が有用であるかを検証していきたい。

ユーザの視線を利用した興味推定の概要は図2に示す。システムは対話中のユーザの顔情報を取得し、センサーに対するユーザの顔面の角度を測定する。センサーと物体の位置関係を予め設定しておくことで、ユーザがどの物体に視線を向けているかを算出し、ユーザがどの話題に関心を示しているかを推定するのである。エージェントが発話する間、ユーザが話題に関連する物体かエージェントを見ているようであれば、ユーザは現在の話題に興味があるか、エージェントと対話する意思を持っているとみなすことができる。逆にエージェントが話す内容と関係のない物体を見ているようであれば、ユーザは現在の話題に関心がないといえる。提案システムによる視線検出は、ユーザがセンサーの右上・左上・右下・左下のうちどの方向を向いているかを80%超の精度で認識可能である[5]。これらを踏まえ、後述するような実験を設計・実施した。

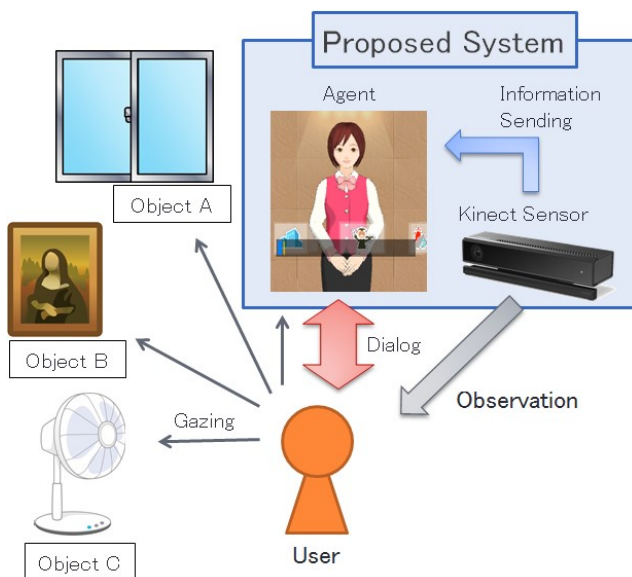


図 2 提案システムによるユーザの興味推定の概要

### 3. 対話実験

提案システムの評価、またヒト対システムの対話におけるユーザの身体動作を考察するために、提案システムを用いた対話実験を実施した。実験に使用されたシステムおよび実験室の構成を図3、図4に示す。

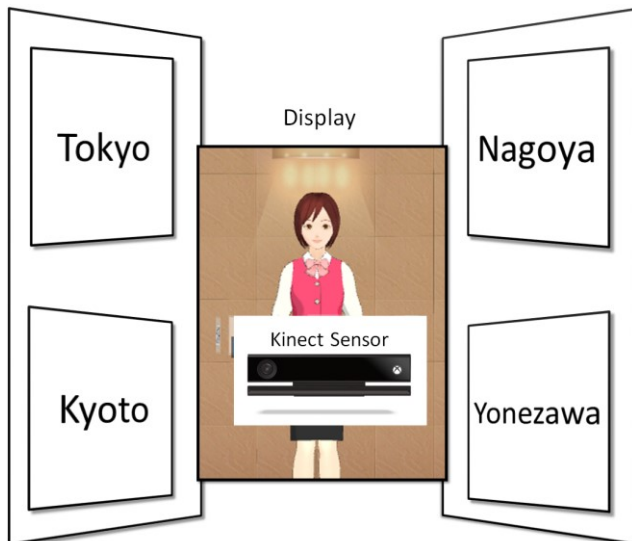


図 3 実験で使用する対話システム

### 3.1 実験の概要

対話実験では、被験者とシステムは「旅行へ行くとすればどこか？」について議論する。エージェントは架空の旅行代理店の店員を演じ、被験者に対して様々な旅行先の情報提示や推薦を行う。システムが話題として扱うことのできる旅行先の情報は、実験室に配置される4枚のポスターにそれぞれ記載されている。被験者はポスターを眺めたり、システムとの対話を通じて、自身が最も興味のある土地一つを旅行先として決定する。

### 3.2 実験室の構成

実験室は対話エージェントを表示するためのディスプレイと、被験者の音声・身体情報を入力するためのマイク・Kinect センサー、さらにそれぞれ2枚のポスターが掲示された左右2つのパネルによって構成される。センサーは被験者の胸ほどの高さに設置し、ポスターはセンサーと同じ高さに2枚、それよりも高い位置に2枚が左右に分けて配置される。また、実験の様子を収録するために後方にカメラを設置している。

### 3.3 対話システムの機能

実験で用いる対話システムでは、被験者の顔面がセンサーに対してどの角度を向いているかを Kinect センサーで計測することにより、被験者がどのポスターを見ているかを推定することができる。また、エージェントを眺めている状態（センサーに対して正面を向いている状態）も取り扱う。

被験者が一定時間同じポスターを眺め続けているような場合にはその旅行先に興味があるものとみなす。もし被験者が音声入力をせず、いずれかのポスターを眺め続けているのならば、エージェントが被験者に対して「〇〇に興味があるのですか？」のように話しかけ、「〇〇は△△な場所です。いかがですか」と旅行先の推薦を行う。エージェントによる推薦の途中で被験者が他のポスターを眺めている

ような場合には、エージェントは「それとも××のほうがお好きですか？」のように話題を切り替えることができる。

また、被験者がエージェントを眺めているような場合には、エージェントは「気になる旅行先の名前を教えてください」のように被験者に話しかける。

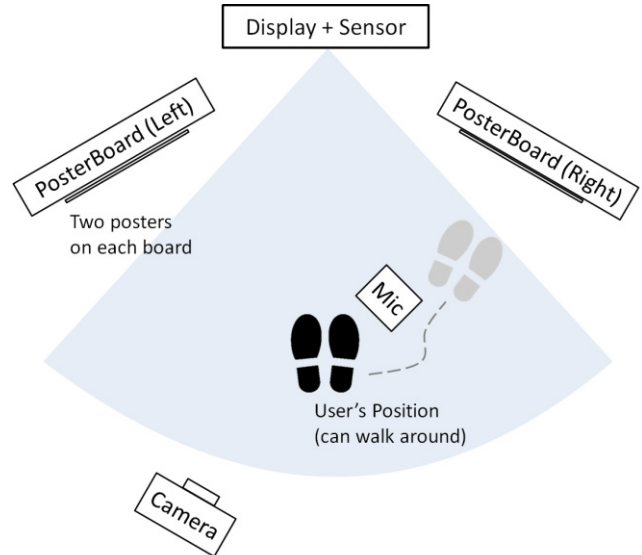


図 4 実験室の俯瞰図

### 3.4 評価方法とその尺度

提案システムを評価するために、Kinect センサーを用いた顔面角度測定を行うもの（視線推定あり）と行わないもの（視線推定なし）の2条件を比較する。「視線推定なし」における対話の分岐は、「視線推定あり」において被験者がエージェントを眺めている状態と同一のものが使用される。したがってエージェントによる旅行先の推薦は行われず、その代わり一定時間ごとに「旅行先を教えてください」と被験者に話しかける。

システムを評価するために設定した尺度を表1に示す。「話しやすさ」、「対話の充足度」、「継続使用欲求」、「情報提供の貢献度」は被験者による主観評価項目であり、エージェントによる適切な話題の切り替え、興味にあった旅行先の推薦などによって、システムの使用に対する印象を問う項目である。2条件のシステムのうち1つと対話を終えるたびにアンケートによって聴取し、1~10の10段階で評価する。「対話時間」は実験開始から旅行先の決定までに要した時間であり、タスク解決のスムーズさを問う客観評価として用意された項目である。被験者には「ポスターを眺めたりエージェントとの対話を通して旅行先を1つだけ選び、そこを旅行先として決定する旨をエージェントに伝える」ことを課題として対話に臨んでもらい、『旅行先を決定しました』とエージェントが発話すること」を対話の終了条件とした。また、被験者には「急いで課題を解決しようとせず、自身の満足のいくように旅行先を決定してもら

いたい」と伝えている。

2条件のシステムは4種×2セットのポスターを被験者ごとに使用順とポスターの配置を変更しながら用いた。また、被験者には2条件のシステムについて、ポスターが変更されたことのみを伝えている。

表 1 実験で用いる評価尺度

| 評価項目     | 内容                               |
|----------|----------------------------------|
| 話しやすさ    | この対話システムは話しやすさを感じさせるものであったか？     |
| 対話の充足度   | 興味のある旅行先について十分に対話することができたか？      |
| 継続使用欲求   | 今後同一のタスクを行うとしてこのシステムを引き続き使用したいか？ |
| 情報提供の貢献度 | エージェントによる情報提供は旅行先の決定に貢献したか？      |
| 対話時間     | 対話開始から、旅行先決定までに要した時間             |

表 2 実験から得られた各評価項目のスコア

|       |      | 話しやすさ | 対話の充足度 | 継続使用欲求 | 情報提供の貢献度 | 対話時間 |
|-------|------|-------|--------|--------|----------|------|
| 被験者 A | 推定あり | 5     | 4      | 4      | 6        | 3:16 |
|       | 推定なし | 6     | 7      | 4      | 7        | 4:22 |
| 被験者 B | 推定あり | 4     | 4      | 7      | 7        | 4:25 |
|       | 推定なし | 3     | 4      | 5      | 6        | 2:58 |
| 被験者 C | 推定あり | 5     | 5      | 5      | 4        | 3:37 |
|       | 推定なし | 5     | 3      | 5      | 5        | 4:31 |

## 4. 実験結果と考察

20代の男性3名を対象として実験を実施した。結果を表2に示す。またアンケートへの回答によれば被験者は全員、両条件において「自身の意思通りに旅行先を決定することができた」としている。

まず実験全体を通して、被験者はほとんどポスターを眺めたり読みふけるようなことがなく、地名を確認する程度に視線を向けたのち、すぐに対話システムに音声入力をした。実験のために用意したシステムでは3秒程度の持続的な視線を検出することを想定していたが、それほど長時間の視線をポスターに向ける被験者がおらず、今回の実験においてはエージェントの推薦発話はほとんど実行されなかった。視線検出機能の有無によってシステムを十分に比較するためには、より短い時間の視線を検出するように修正するか、長時間の視線を誘発するようなポスターを設置する必要がある。

主観評価については、どちらのシステムに対して高い評価をするのかは被験者ごとに異なり、明確な傾向は観察で

きなかった。被験者の自由記述から得た意見としては、「その土地に関するより詳細な情報（『グルメ』など）を訊ねようとしたが、システムはそれを教えてくれなかった」というものがあつた。実際、すべての被験者は「どのくらい遠いの？」など、エージェントが返答することのできない入力を行っていた。話しやすさ・対話の充足度の評価はこれによって低くなっているものと考えられる。

対話継続欲求は1名の被験者のみが評価に差をつけた。その被験者はどのように入力すべきかためらい、エージェントを眺めていた際に「気になる地名を教えてください」とエージェントから話しかけられたことが好印象であると回答した。また、ポスターを眺めて地名の読みがわからずに困っていた際にエージェントから「〇〇に興味がありますか？」と話しかけられ、対話が滞りにくい印象があつたと回答している。

情報提供の貢献度についてはすべての被験者が2条件間で異なる評価をした。エージェントによる旅行先についての説明の内容は2条件間で変わらないが、対話時間が長い場合に情報提供が役に立ったと評価される傾向があつた。被験者が旅行先の決定により長い時間をかけるようなとき、既に説明を受けた旅行先について、再度エージェントに説明を求めている。旅行先をよく吟味しようとする時にエージェントの説明が役立ったと感じられたのだと考察する。

対話時間は2条件の間で1分以上の差が生じていた。長くなるのか、短くなるのかは使用順とは関係性がないように思えるが、ポスターの内容が変更されることで被験者が地名の読み方に戸惑ったり、意思決定に要する時間が変動したと考えている。

総評として、実験中では被験者によるポスターへの視線はほとんど出現しなかったものの、被験者の視線をきっかけとした話しかけは被験者のシステムに対する印象に影響を与えたと考える。用意した評価尺度では有意な影響や明確な傾向は観察できなかったものの、被験者からの聴取によって条件間で印象に差があることが示されたため、評価尺度として適切なものを再検討していきたい。

## 5. まとめ

本稿では、ユーザの音声のみでなく身体動作をも入力として取り扱うことができるようなマルチモーダル対話システムを提案し、その構築と評価実験について報告した。提案システムは音声対話システム構築ツールキット MMDAgent とジェスチャ認識デバイス Kinect センサーを用いることで構築され、Kinect によって取得されたユーザの身体動作を対話制御モジュールに伝送することによって実現した。

ユーザの視線から話題に対する興味を推定し、システムが応答をリアルタイムに制御するようなシステムの印象評価実験を実施したが、被験者は主にシステムの方を眺め続

け、話題に関連する物体に向けられる視線が発生しにくかった。今後、視線が重要となる対話についての構想を深めていく必要がある。被験者の視線に基づくエージェントの話しかけ機能の有無は主観評価に有意な差を生じさせるものではなかったが、ユーザがポスターに書かれた漢字の読みがわからずにいたり、システムに対してどのように話しかければよいか悩んでいるような状況では対話を滞らせない効果があることが示唆された。

今後はユーザが持続的な視線を生じさせやすい環境を構築し、さらにより掘り下げた話題に対応できるような対話を設計した上で再度同様の実験を実施したい。また、今回は視線のみを用いたが、ユーザのうなずきや上半身の動作を認識することによる興味推定機能の開発を検討していきたい。

## 参考文献

- [1] 深山篤ほか. 視線分析に基づく擬人化エージェントのユーザビリティ評価の検討. 電子情報通信学会技術研究報告.HIP, ヒューマン情報処理, 2004, vol. 103, no. 743, p. 53-58.
- [2] 大澤博隆. 人工知能はどのように擬人化されるべきなのか?. 人口知能, 2014, vol. 29, no. 2, p. 182-189.
- [3] “CeVIO Official Web”. <http://cevio.jp/>.
- [4] “Live2D”. <http://www.live2d.com/ja/>.
- [5] Takeru, K. and Tetsuo, K.. Multimodal Spoken Dialog System Using State Estimation by Body Motion. IEEE 6th Global Conference on Consumer Electronics. 2017, p. 348-351.
- [6] “MMDAgent - Toolkit for building voice interaction systems -”. <http://www.mmdagent.jp/>.
- [7] 山倉和樹. 身体動作に基づいた対話時の話題に対する聞き手の関心の推定. 千葉工業大学情報科学部卒業論文. 2016. <http://www.imai.cs.it-chiba.ac.jp/files/2015/yamakura.pdf>.
- [8] 千葉祐弥, 伊藤彰則. ユーザの対話意欲推定のための人対人対話データの分析と WOZ システムの検討. 電子情報通信学会技術研究報告:信学技報, 2015, vol. 115, no. 346, p. 117-122.