第 26 回日本バーチャルリアリティ学会大会論文集（2021 年 9 月）

会話エージェントにおける視線とフィラーを組み合わせた非言語的な質問認識開始手法の研究

郭正雄 [1)]，葛岡英明 [1)]，吉田成朗 [2,3)]，川口一画 [4)]，鳴海拓志 [1)]，雨宮智浩 [1)]
Masao KAKU, Hideaki KUZUOKA, Shigeo YOSHIDA, Ikkaku KAWAGUCHI, Takuji NARUMI and Tomohiro AMEMIYA

1) 東京大学 情報理工学系研究科（〒 113-0033 東京都文京区本郷 7-3-1, {kaku, kuzuoka, narumi, amemiya}@cyber.t.u-tokyo.ac.jp）
2) 東京大学 先端科学技術研究センター（〒 153-8904 東京都目黒区駒場 4-6-1, shigeo@star.rcast.u-tokyo.ac.jp）
3) 国立研究開発法人科学技術振興機構 さきがけ (同上)
4) 筑波大学 システム情報系（〒 305-8573 茨城県つくば市天王台 1 丁目 1-1, kawaguchi@cs.tsukuba.ac.jp）

概要: スマートスピーカのような音声操作型の会話エージェントは, 特定の単語（ウェイクワード）を最初に言うことで質問の認識を開始する. 一方で, この方法では対話の構造が不自然になることや, エージェントの数だけウェイクワードが必要になるといった問題が考えられる. 本研究では, 視線や会話の間に挟む音声表現であるフィラーといった非言語的手がかりを組み合わせた自然なウェイク手法を提案し, その有効性を検証した.

キーワード： 会話エージェント, スマートスピーカ, 視線, フィラー

1. はじめに

スマートスピーカのような音声操作型のインタフェースでは, ウェイクワードと呼ばれる特定の発話をトリガーとして質問を認識し, ハンズフリーで操作ができることが特徴である. 実際にこうしたデバイスは日常生活の中で家電の操作や情報検索など様々な用途で利用されており, 広く普及が進んでいる [1].

その一方で, ウェイクワードをトリガーとして質問と応答を行う対話構造は自然ではないといった指摘がされている [2]. また, ウェイクワードを何度も繰り返す必要があること, プラットフォームの数だけウェイクワードを記憶する必要があることがユーザの負担を大きくするという問題も考えられる. そのため, より自然に操作できて, ユーザにとって負担が少ないトリガー手法を検討することが会話型インターフェースの普及にとって重要である.

筆者らはそうした自然なインタラクションの手段として, 非言語的な表現が有効ではないかと考えている. 我々は特に視線と, 「えーと」や「あのー」といった音声表現であるフィラーに注目した. そこで, 本研究では非言語的な表現を利用することでより自然で負担の少ない質問のトリガー手法を開発することを目的とし, 視線とフィラーを組み合わせた手法を提案した上で, 評価のために予備実験を行った.

2. 関連研究

2.1 スマートスピーカに関する研究

Sciuto et al. は Amazon 社の Alexa を利用している家庭の履歴ログの分析やインタビューを通して, 日常生活でスマートスピーカどのように利用されているかを調査した. その結果, 天気などの情報検索, タイマーなどのツール機能, 家電やメディアの操作といった様々な用途でスマートスピーカを利用していることがわかった [1]. Alexa などの市販されているスマートスピーカのほとんどは, 質問の始めに「アレクサ」などのようなウェイクワードをつけることで質問の聞き取りを始める手法をとっている.

一方で, Vtyurina et al は, 会話エージェントが指示しながら人間が作業を行い, その中の会話について調査した研究の中で, (トリガーワード, 質問, 返答) という枠組みが制約となり, 自然なインタラクションの障害になっていると指摘している [2]. また, ウェイクワードを採用したシステムはテレビやスマートフォンなど多くのデバイスに搭載されており, 様々なプラットフォームが乱立することによって複数のウェイクワードを記憶する必要が生じ, ユーザの負担が大きくなることも問題として挙げられる. そのため, より自然な操作性で負担の小さいトリガー手法を開発することが会話型インターフェースの普及にとって重要な課題であると考えられる.

2.2 非言語的表現の活用

より自然なインタラクションを検討するにあたり, 筆者らは人間がコミュニケーションの中で言語表現以外に様々な

非言語的表現を利用していることに注目した. そして, そうした非言語的表現を会話型インターフェースに取り入れることを検討した.

非言語的表現として，川口らは人間が会話の開始時に相互注視を行うことに注目し，会話エージェントを注視しお互いに目を合わせることで質問を開始する手法を開発した. これにより，システムの操作性や対話感が向上することを示した [3]. しかし, 多人数会話で評価した研究では, ユーザが質問中に注視をし続けず失敗する場合があることや, 誤認識を防ぐための視線が逸れた場合のタイムアウト処理をどれくらい時間で行うか調整が難しいといった指摘がされている [6].

これに関して, Goodwin は話し手は話している最中に聞き手を見続ける必要はなく, 視線を向けることはあくまで応対可能性の確認が目的であると述べており [7], 質問のトリガーとして注視されているかどうかという情報だけでは不十分であるということが示唆されている.

また, Pomykalsk et al は質問をトリガーする手法として指を鳴らすなどの 5 種類のジェスチャを比較し, その結果, 指を鳴らす場合と手を振る場合の評価が高かったとしている. この中で, ジェスチャは場合によっては別の意味に受け取られてしまう可能性があるため, 社会的な意味の薄いものが好まれたり, 音声と使い分けられるような複数のモダリティを備えた手法が望ましいといったユーザの意見も紹介されている [4].

さらに, 非言語的な表現は質問のトリガー以外の目的でも活用されている. Yan et al はスマートスピーカが誤った反応をしてしまった場合に人間の表情が変化することに着目し, 顔をしかめることによってスマートスピーカの誤った応答を中断する手法を開発し評価を行った. その結果, ボタンを押したり特定のワードを言う場合に比べて, より素早くかつ直感的に中断ができたと述べている [5].

ここまで, 非言語表現を取り入れた会話型インターフェースの研究について述べたが, いずれも単一のモダリティを利用したものであり, 複数のモダリティを組み合わせた例は少ない. そして, 前で述べたように, 単一のモダリティでは意図が曖昧であったり多義的な解釈が可能になってしまうことから, 複数のモダリティを組み合わせることの有効性が示唆されている.

そこで本研究は, 実際にユーザビリティを改善することが示された視線を用いる手法に対して, さらに他の非言語的表現と組み合わせることを検討し, フィラーに注目した. フィラーとは会話の始まりや途中で見られる「えーと」や「あのー」といった言い淀み表現であり, 会話の開始や会話の順番の維持を示す手がかりとして用いられることもある [8].

3. 提案手法

本研究では前章で述べたように, 複数のモダリティを備えた手法が有効であると考え, 視線を応対可能性の確認, フィラーを質問の意志の確認という 2 段階で質問のトリガーを行う以下のような手法を考案した.

1. アイドル状態時にユーザが視線を向けると, 応対可能状態となる.
2. 応対可能状態時にユーザが「え〜」と発話すると, 聞き取り状態になる.
3. 聞き取り状態で質問を聞き取り, 返答を行う.
4. 返答が終了後, 再びアイドル状態となる.

これにより, フィラーを検知後はユーザは視線を常に向け続ける必要がなくなる, 意図しない視線が検知された場合でもフィラーを言わない限り誤って聞き取ることを抑制できるといった利点を自然な操作感を損なわずに得ることができると考えられた.

4. 評価実験

4.1 装置

実験装置は, Raspberry Pi 4 を中心に, 音声の入出力のためのマイクとスピーカを備えたスピーカフォンと視線認識用のカメラ, 状態を示すための LED で構成し, 筐体は 3D プリンタで作成した. また, フィラー認識はオープンソースのウェイクワード認識ライブラリ[1]を使用し, 質問の聞き取りと応答は Google Assistant Service[2]を利用した.

4.2 タスク

会話型インターフェイスを備えたシステムが普及していく中で, 複数のシステムが存在する状況を想定した評価が必要であると考え, 3 台のスマートスピーカに連続して質問を行うタスクを設定した. この時, 質問の種類として「廊下の明かりをつけて」といったスマート家電の操作を意図したもの, 「今日の午後の気温は」といった情報検索を意図したもの, 「明日の 9 時にアラームを設定して」といったツール機能を想定したものの 3 種類を用意し, 3 台のスマートスピーカにもそれぞれ対応する役割を与えた.

そして, 参加者には 30 個の質問を 10 分以内にそれぞれの内容に対応する役割のスマートスピーカにできるだけ多く質問するよう伝えた. この時, 各スマートスピーカの役割及び質問の種類をわかりやすくするために, 役割に対応したシンボルを設定し, それぞれのスマートスピーカと参加者に渡す質問リストの質問に表示した.

3 台のスマートスピーカは参加者の頭部の位置を中心とした半径 40cm, 中心角 120 度の扇型状に左右対称になるように 60 度ごとに配置した. ただし, 実際の実験では 3 台のうち参加者の右側に配置した 1 台に問題が発生したため, 質問の役割の対応を修正した上で配置は変えず左と中央の 2 台のみで実験を行った.

実験は被験者内計画で実施され, 大学院生の 20 代男性 6 名が参加した. 参加者にスマートスピーカの使用経験を確認したところ, 6 人中 5 人が使用した経験があると答えた.

[1]https://github.com/MycroftAI/mycroft-precise
[2]https://developers.google.com/assistant/sdk

図 1: 実験中の様子（システムの前方に役割を示すシンボルが置かれている）

4.3 条件

実験条件として以下の 3 条件を設定した．なお, 視線の検出は顔の向きをカメラで自動検出することにより行った．

- ウェイクワード条件：決められたフレーズを最初に発話することでトリガーを行う条件．ウェイクワードのトリガーは実験者が手動で行った．実験ではウェイクワードは「ヘイ, サトウ」,「お〜い, スズキ」,「タカハシさん」の 3 種類を使用し, これらのウェイクワードは質問リストに記載した．
- 視線条件：視線によってトリガーする条件．視線を向けたら聞き取りを始め, 質問の発話終了後のポーズが検出されれば返答を行い, その前に視線が 1 秒以上逸れた場合は返答の処理をキャンセルするよう設定にした．
- マルチモーダル条件：視線とフィラーを組み合わせた前述の提案手法に対応する条件．フィラーの検出は前述のライブラリを利用した．応答可能状態において視線が 1 秒以上逸れた場合はアイドル状態に戻るように設定した．

なお, 参加者にスマートスピーカの状態をフィードバックできるように, 聞き取り状態の時は青, 返答中は緑, マルチモーダル条件の応答可能状態の場合は黄色に LED を光らせた．

4.4 評価方法

各条件について, システムの操作感を System Usability Scale[9], ユーザの作業負荷を NASA-TLX[10] によって評価を行った．

4.5 結果

はじめに, System Usability Scale の結果について述べる．各項目のアンケート結果に基づいて評価点を算出した結果, 図 2 のような結果となった．この結果について, 有意水準 $\alpha = 0.05$ で一要因分散分析を行ったところ有意傾向 ($p < 0.1$, $F = 3.79$) が見られた．下位検定として Bonferroni 法による多重比較を行った結果, ウェイクワード条件とマルチモーダル条件の間で有意差 ($p < 0.05$) が見られた．

次に, NASA-TLX についても同様に評価を行った．その結果, 評価点は図 3 のようになり, 分散分析の結果, 有意傾向

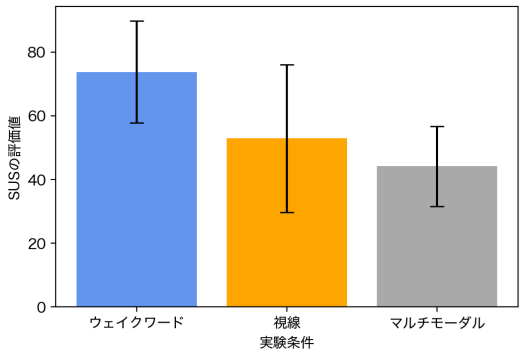

図 2: System Usability Scale によるユーザビリティの評価点（エラーバーは 95%信頼区間を示す．以後も同様.）

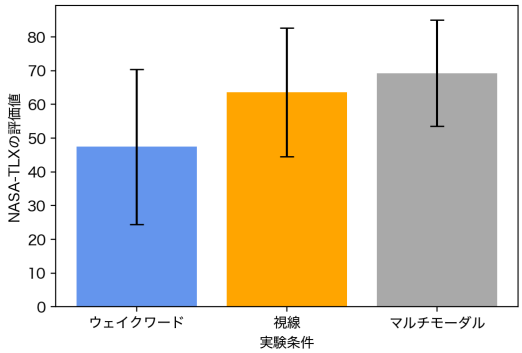

図 3: NASA-TLX による作業負荷の評価点

($p < 0.1$, $F = 3.93$) が見られた．また, 多重比較ではウェイクワード条件とマルチモーダル条件の間で有意差 ($p < 0.05$) が見られた．

5. 考察

5.1 ウェイクワード条件が高評価となった要因

本研究の実験の結果, SUS によるユーザビリティの評価と NASA-TLX による作業負荷の双方でウェイクワードが最も高く評価される傾向が見られ, 川口らの研究に反する結果となった．その原因として幾つかの要因が考えられる．

まずはじめに, 今回の実験ではウェイクワードは手動による操作であったため, 自動検出で精度が劣る視線検出とフィラー検出を利用した条件よりも高く評価されたことが考えられる．また, 今回の実験の設定では各スマートスピーカ間の距離が狭く, 複数台が同時に参加者の顔を認識してしまうことが頻繁にあったため, 確実に単独でトリガーできるウェイクワードが評価され, 視線を用いる手法の評価が下がったことも一因として考えられる．そのため, 視線やフィラーの自動検出の精度を向上させ誤作動を減らす必要があると言える．

また, 機材の不調により操作する対象のスマートスピーカが 3 台から 2 台になったことや, 実際の使用場面とは異なり質問用紙にウェイクワードが記載されていたために, ウェイクワードを使用することの負荷があまり大きくなかったことも考えられる．これについて, 実験中に間違ったエージェ

ントに呼びかけた失敗と 「ヘイ, …」や「お〜い, …」といったウェイクワードの呼びかけ部分を混同してしまう失敗が 1 度ずつ見られたことから, 負荷が大きくなった場合にウェイクワードの評価が下がる可能性が示唆されている.

こうした実験上の制約による影響が考えられる一方で, ウェイクワードがユーザにとって十分に使いやすい手法であるということも実験後のアンケートの自由記述コメントから伺える. 実際に, 6 人の参加者のうち 3 名がウェイクワードは身体的な動きを伴ったり複数の作業を続けて行わなければならない他の条件に比べ, 手間が少なく最も使いやすかったという趣旨のコメントをしていた. しかし, 川口らの研究とは異なり今回の実験ではほとんどの参加者がスマートスピーカを使用した経験があったことから, 参加者がある程度ウェイクワードを使う手法に対して慣れていた影響の可能性も考慮する必要がある.

5.2 視線条件とマルチモーダル条件の比較

次に, 視線条件とマルチモーダル条件を比較した場合の考察を述べる. 両者間の比較はウェイクワードの場合とは異なり精度の面での差は小さく, 実際の操作感の影響が強いと考えられる.

この 2 条件間では大きな差は見られないが, SUS と NASA-TLX の双方でやや視線条件の方が評価が高い傾向が見られ, SUS の評価点を見ると 6 人中 5 人とほとんどがマルチモーダル条件よりも視線条件の方が評価が高かった. 自由記述コメントで 2 名が視線を向けた上でさらにフィラーを言うことが必要以上の負担に感じられたと言うコメントをしており, 手順の多さの影響が現れた可能性が考えられる.

その一方で, SUS は 6 人中 3 人と半数が視線条件よりもマルチモーダル条件の方が NASA-TLX の評価点が小さい, つまり作業負荷が小さいと言う結果になっていた. これに関しては, 視線条件において視線を向け続けるのを忘れて質問を聞き取ってもらえない場面が数回見られたり, マルチモーダル条件は顔を向け続ける必要がないのが良かったという自由記述コメントがあったことを踏まえると, 当初の想定通りマルチモーダル条件は視線を向け続けなくても良いことが作業負荷の軽減につながった可能性があると考えることができる. しかしながら, 実際の使用場面では質問用紙を見ながら質問する場面は少ないと考えられ, 実生活に組み込まれた際に人間がスマートスピーカに対してどのような視線の振る舞いをするのかということも慎重に検討する必要があると考えられる.

6. 結論

本研究では, スマートスピーカのような会話型インターフェースにおいて, 視線やフィラーといった非言語的表現を活用することでより自然で負担の小さいトリガー手法を開発することを目指した. そして, 視線を向けた後に応対可能状態となり, その上でフィラーを発話して聞き取りを始めるマルチモーダルな手法を考案し評価を行なった. その結果, ウェイクワードを用いた手法がユーザビリティと作業負荷の双方で高く評価される傾向が見られ, マルチモーダルな手法の評価は低かった. この結果は, 実験上の制約や設定によってウェイクワードが有利となったことが影響したと考えられるが, 視線のみを用いる条件と比べた場合にマルチモーダルな手法は視線を向け続けなくても良いと言う特徴が作業負荷の面では評価されたことも示唆された.

今後は, 実験条件の見直しや提案手法の改良を行った上で, 本実験を行うことを予定している. また将来的な展望として, ウェイクワードや視線, フィラーといった表現を段階的に扱うのではなく総合的に判断してトリガーする手法や, フィラーや視線以外の非言語的表現の活用した手法の開発を検討している.

参考文献

[1] Sciuto, Alex, et al. ""Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage" Proceedings of the 2018 Designing Interactive Systems Conference. 2018.

[2] Vtyurina, Alexandra, and Adam Fourney. "Exploring the role of conversational cues in guided task support with virtual assistants." Proceedings of the 2018 CHI conference on human factors in computing systems. 2018.

[3] 川口一画, and 葛岡英明. "スマートスピーカーにおける注視の入出力を用いたインタラクションの効果." ヒューマンインタフェース学会論文誌 21.3 (2019): 269-278.

[4] Pomykalski, Patryk, et al. "Considering Wake Gestures for Smart Assistant Use." Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.

[5] Yan, Yukang, et al. "FrownOnError: Interrupting Responses from Smart Speakers by Facial Expressions." Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.

[6] McMillan, Donald, et al. "Designing with Gaze: Tama–a Gaze-Aware Smart Speaker Platform." Proceedings of the ACM on Human-Computer Interaction 3 (2019).

[7] Goodwin, Charles. "Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning." Sociological inquiry 50.3-4 (1980): 272-302.

[8] Skantze, Gabriel. "Turn-taking in conversational systems and human-robot interaction: a review." Computer Speech & Language (2020): 101178.

[9] Brooke, John. "SUS-A quick and dirty usability scale." Usability evaluation in industry 189.194 (1996): 4-7.

[10] 芳賀繁, and 水上直樹. "日本語版 NASA‐TLX によるメンタルワークロード測定 各種室内実験課題の困難度に対するワークロード得点の感度." 人間工学 32.2 (1996): 71-79.