Vol.21 No.3, 2019

原著論文

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

川口 一画 *1　葛岡 英明 *2　ドナルド マクミラン *3

The Effect of Interaction Using Gaze Input / Output on Smart Speaker

Ikkaku Kawaguchi*1, Hideaki Kuzuoka*2, and Donald McMillan*3

Abstract – **In this study, we focused on the problem of existing smart speakers that natural conversation is not accomplished due to restrictions of the dialogue structure. In order to solve the problem, we proposed the method using gaze instead of wake-word based on sociological knowledge about "Initiation talk". We developed the smart speaker "Tama", that have the cameras to detect the user's gaze and the robotic head to show the gaze for users. To evaluate the effectiveness of the proposed system, we conducted an experiment using the "Tama". For the evaluation, we compared three conditions: wake word condition, gaze detection condition , and mutual gaze condition. Experimental results showed that the usability of the system and the sense of conversation are improved when using gaze input / output.**

Keywords : **Smart Speaker, Gaze Interaction, Gaze-Detection, Mutual Gaze**

1. はじめに

Amazon 社の Amazon Echo [1] や Google 社の Google Home [2] に代表されるスマートスピーカーは、ウェイクワードと呼ばれる特定の発話（「Alexa」,「OK Google」等）をきっかけに利用者の音声を処理し、認識されたコマンドを実行する。そして、その結果を音声で提示する。このような音声の入出力を用いる対話型インタフェースは、ハンズフリーで情報検索や予定確認等の様々なタスクを実行出来ることから日常生活との親和性が高く、近年急速に普及が進んでいる [3]。

HCI 分野では、スマートスピーカーの利用実態に関する調査が行われている [4], [5]。これらの研究では、家庭に設置されたスマートスピーカーの利用ログの収集・分析や、スマートスピーカー利用場面の会話分析を行った。その結果として、スマートスピーカーは日常生活のルーチンの中で活用されているものの利用用途が限られていること、そして対話構造の制約から自然な対話が実現されていないことを指摘した。

これに対して Vtyuria らは、Wizard of Oz(WOz) 法によって柔軟な応答を行うスマートスピーカーシステムを想定した実験をおこない、ウェイクワードを用いない自然な対話ができる場合に、ユーザの満足度が高くなることを示した [6]。そこで本研究では、ウェイクワードによらないスマートスピーカーの対話開始手法の提案を目的とする。

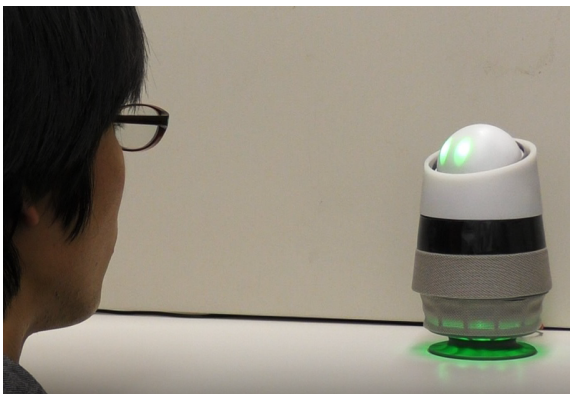

図 1　“Tama” とのインタラクションの様子
Fig. 1　Interaction with “Tama”

このような目的に対して、本研究では人間同士のインタラクションにおける会話の開始（Initiation talk）に関する社会学的知見に着目した。Heath [7] によれば、会話の開始時点では発話だけでなく非言語的情報も重要な役割を果たす場合がある。その中でも「注視」は相手の活動を喚起・促進し、インタラクションの開始および進行に関わる。そこで本研究では、利用者とのインタラクションに注視の入出力を用いるスマートスピーカー “Tama” を提案する（図 1）。

本論文では、関連研究に基づく Tama の設計指針・システム構成について述べた後、検証のため実施した評価実験について説明を行う。その後、提案手法の効果についての考察および今後の展望について述べる。

*1: 筑波大学 システム情報系
*2: 東京大学大学院 情報理工学系研究科
*3: ストックホルム大学
*1: Faculty of Engineering, Information, and Systems, University of Tsukuba
*2: Graduate School of Information Science and Technology, The University of Tokyo
*3: Stockholm University, Sweden

2. 関連研究

本章では、初めに既存のスマートスピーカーに関する関連研究について説明したのち、人間同士のインタラクションにおける「会話の開始」に関する社会学的知見について説明する。その後、本研究で着目する「注視」を用いたシステムに関する関連研究について述べ、最後に本研究での設計指針について述べる。

2.1 スマートスピーカーに関する研究

Sciuto らは[4]、Amazon Echo の利用者を対象とした使用ログ収集と、利用中の家族に対するインタビュー調査により、システムが日常生活にどのように取り入れられているかの実態を示した（用途、利用場所等）。特に、利用者は天気の確認やスマート家電の操作、音楽再生等のコマンドを日常生活のルーチンに組み込み利用する一方で、それ以外のコマンドを新規に利用することは少ない。

Porcheron らは[5]、Amazon Echo を利用している 5 組の家族を対象に、1 か月間利用中の音声の録音を行い、会話分析により利用実態の調査を行った。収集されたデータより、スマートスピーカーが複数人の会話の中でどのように利用されているかを示した。また、現在のスマートスピーカーはあらかじめ決められた対話の構造に沿った限定的な応対しか出来ないことから、「対話型インタフェース」という表現は誤った呼称であることを主張した。

Vtyuria[6] らは、既存の音声アシスタントにおいては一般的に「ウェイクワード-質問-回答」という枠組みの制約があることを指摘した。そして、WOz 法を用いて実験者がシステムの対話内容を制御し、より自然な対話を実現した場合に、利用者がどのようなインタラクションを行うかの調査を行った。調査の結果、利用者は WOz 法によって柔軟な応答を行うシステムに対して、より自然な方法で話しかけ、満足度も高くなることが示された。そして、WOz 法により会話開始時のウェイクワードが不要となった点がこのような結果が得られた要因の一つであると主張した。

Sciuto らが示したように、既存のスマートスピーカーは日常生活の中で活用されているものの、その利用用途は限定的である。これは、Pocheron らおよび Vtyurina らが指摘したように、対話構造の制約により自然な対話が実現されていないことから、単純な音声入力インタフェースとしてしか認識されていないためであると考えられる。このような課題を解決するためのアプローチとして、本研究では Vtyuria らの知見をもとに、会話開始時のウェイクワードを省略する方策の検討を行うこととする。

なお、ウェイクワードの省略については音声アシスタントサービスの提供元においても検討が行われており、一度ウェイクワードによりコマンドを実行した後、数秒間認識を継続する機能が実装されている[8], [9]。ただし、このような簡易的手法では、利用者の意図せぬタイミングで認識が継続されるという弊害が発生する。そのため本研究では、利用者が対話を開始しようとする意図を適切に検出し、自動的にコマンド認識を開始する方策の実現を目指す。

2.2 人間の会話に関する社会学的知見

対話開始時のウェイクワードを省略するためには、何らかの方法で利用者が対話を開始しようとする意図を検出する必要がある。そのため、本研究では人間同士のインタラクションにおける会話の開始（Initiation talk）に関する社会学的知見に着目した。

Schegloff によれば、人間が会話を開始する場合、初めに相手が対応可能であるか（availability）を確認する必要がある[10]。例えば、問いかけに対して回答があった場合、それは対応可能であることを強く示す証拠となる。また、対話を継続する上では、参加者が相互に対応可能であることを提示し続けることが必要であるとした。

また、Heath は診察場面における医師—患者間のインタラクションの開始について分析を行い、医師の定型的な質問が会話を進める上で果たす役割について考察した[7]。また、会話の開始時に非言語的表現が果たす役割にも着目し、注視が相手の行動を誘起･促進し、インタラクションを開始･進行させる役割を持つことを示した。同様な注視の役割については Kendon、Nielsen も説明しており[11], [12]、相互注視 (mutual gaze) が会話開始のきっかけとなることが示されている。

さらに Heath らは、会話の開始にあたって非言語的表現が果たす役割についてより詳細な分析を行うため、オフィスでのインタラクションの様子の観察を行った[13]。そして、人は会話を開始しようとする場合、他者が対応可能であるかを観察 (monitor) し、他者が何らかのタスクを終了したタイミングで発話を行うことを示した。さらに、Salvadori によれば、人は自身が対応可能であることを示すため、タスクの区切りとなるタイミングを何らかの行為（タイピングを止める等）によって他者に伝達する[14]。

これらの社会学的知見をまとめると、会話を開始する上ではまず相手が対応可能であるかを確認する必要があり、そのために観察が行われる。この際、話し手側からの注視は状況を観察し会話を開始しようとする意図を示す。ここで、話し手の注視に対して受け手が注視を返した場合、それは聞き手が対応可能であることを示す。そのため、相互注視が成立した場合、会話が開始される。

本研究ではこれらの知見に基づき、スマートスピーカーとのインタラクションを開始しようとする意図を

検出するために利用者の注視を用いることとする。

2.3 注視をインタラクションに利用するシステムに関する研究

HRI 分野では注視をヒューマノイドロボットとのインタラクションに用いる方策についての研究がなされている。例えば、Bohus らは利用者の顔方向や身体配置等に基づいて参与の意図が検出された場合に案内を開始するガイドロボットの開発を行った [15]。また Mutlu らは、ロボットを含んだ複数人会話において、ロボットの注視が人間の注視と同様に他者の参与構造に影響を及ぼすことを示した [16]。ただし、このようなヒューマノイドロボットを用いた研究では、観光地紹介のようにシチュエーションを限定することでシナリオに沿った高度な対話が実現されている。このような高度な対話機能と、高い擬人性を有するヒューマノイドロボットの身体が組み合わさることで、ロボットは人間に近い存在として捉えられ、注視を用いたインタラクションが成立した可能性がある。そのため、対話機能・身体性ともに限定的なスマートスピーカーにおいても注視が有効であるかについては別途検証が必要である。

一方で、Anas らが提案した CoffeePet はコーヒーメーカーに注視の入出力機能を付与したエージェント型システムであり、利用者との相互注視が成立した場合にコーヒーを淹れる機能を持つ [17]。ただし、CoffeePet はコーヒーを淹れる以外の機能を持たず、対話型のインタラクションは行われない。そのため、スマートスピーカーにおいて注視を用いた場合の効果は明らかになっていない。

2.4 本研究の設計指針

本研究では、既存のスマートスピーカーにおける課題として、対話構造の制約により自然な対話が実現されていない点に着目した。その解決のため、Vtyuria らの知見をもとにウェイクワードを用いないスマートスピーカーの対話開始手法の提案を目指す。そのための方策として、会話の開始に関する社会学的知見に基づき、注視を用いるシステムの提案を行う。

なお、人間同士のインタラクションにおいて、注視は一方的に提示されるのではなく、参加者が相互行為として互いに提示しあうことによってその役割を果たす [11], [12]。そのため、利用者の注視をシステムデザインに組み込もうとする場合、システムに利用者の注視を検出する機能を実装するだけでは不十分であり、システム側からも注視を提示する必要があると考えた。そこで、以下の 2 点の設計指針を設定した。

I 利用者が対話を開始しようとする意図を注視により判別し、コマンド認識を開始する

II 利用者の注視を検出した場合、対応可能であることを示すためシステム側からも注視を提示する

(1) (2)

図 2 システム外観

Fig. 2 Appearance of the system

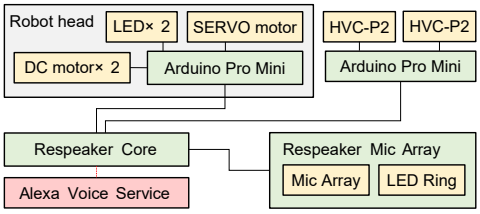

図 3 システム構成

Fig. 3 System configuration

本研究では、上記の設計指針に基づくシステム開発を行い、その効果の検証を行う。

3. システム設計

本章では、2.4 節で示した設計指針に基づき実装した提案システム “Tama” についての説明を行う。Tama は利用者の注視を検出するためのカメラと、システム側からの注視を提示する簡易的なロボット頭部を搭載し、注視の入出力を利用可能なスマートスピーカーである。システムの外観を図 2–(1) に示す。

3.1 ハードウェア構成

Tama のシステム構成を図 3 に示す。システム全体の制御基板として、Seeed 社の ReSpeaker Core [18] を用いた。ReSpeaker Core は CPU 内蔵型の Wi-Fi モジュールや音声 CODEC IC 等を有する多機能型モジュールであり、Linux ベースの OS である OpenWrt で制御される。音声認識用のマイクには、同じく Seeed 社の Respeakaer Mic Array を用いた [19]。Mic Array 上にはフルカラー LED リングが実装されており、視覚的なフィードバックを提示することが可能である。音声出力用のスピーカーは USB バスパワー型のスピーカーを用いた。

利用者の顔検出には、OMRON 社の HVC-P2(長距離検出タイプ) を用いた [20]。HVC-P2 は基板に接続されたカメラを用いて、顔・視線の検出および角度推定等の様々な機能をデバイス単体で実行することが出来る。HVC-P2 の制御には ARDUINO Pro Mini を用い、検出された顔の位置情報のみを数値情報として ReSpeaker に送信する構成とした。これにより、撮影された顔画像の処理は HVC-P2 内で完結し、ネットワーク上に顔画像が流出するリスクが防止される。な

おシステムでは、検出範囲拡大のため HVC-P2 を水平方向に 2 台並べて用いており、検出画角は水平約 100 度、垂直約 40 度であった。また、HVC-P2 の仕様により顔・視線検出可能距離は 3m であった。

注視の提示を行うロボット頭部は、サーボモータ・DC モータを用いて pan-tilt 各 1 自由度の動作が可能である。また、実験において視線を提示しない条件を設定するため（詳細は 4.1 節参照）、頭部モジュール全体を DC モーターにより回転させ筐体内に収納することが可能である（図 2–(2)）。また、目の部分にはフルカラー LED を搭載し、注視を提示することが出来る。モーター及び LED の制御には ARDUINO Pro Mini を用い、ReSpeaker Core からのシリアル通信により制御が行われる。

本体のデザインは、既存のスマートスピーカーを参考に円筒形のシンプルな形状とした。全高は Google home [2] とほぼ同等の約 15cm とした。頭部は直径 60mm の半球とし、透明素材の表面を白色に塗装して用いた。これにより、内部の LED の発光が透過して表面にロボットの目が提示される。なお、システム名 “Tama” は球状の頭部形状に由来する。本体の下部には ReSpeakaer Mic Array を配置し、開口部より周囲の音声を取得するとともに LED リングの発光を外部に提示する。頭部・胴体等の筐体は、Form 社の SLA 型 3D プリンタ Form2 で出力し、表面処理・塗装を行った上で使用した。

3.2　ソフトウェア構成

スマートスピーカーの核となる音声アシスタントには Amazon 社の Alexa Voice Service(AVS) [21] を用い、ReSpeaker Core 上で API を利用する制御プログラムを Python で実装した。制御プログラムは、2 つのシリアルポートを介して HVC-P2 およびロボット頭部を制御する Arduino Pro Mini とシリアル通信を行い、検出された利用者の頭部方向の取得とロボット頭部方向の制御を行う。また、SPI 通信により Mic Array 上の LED リングの制御を行う。

なお、実験では注視の検出を行わずウェイクワードを用いる条件も設定した（詳細は 4.1 節参照）。この条件においては、制御プログラムにおいて音声認識ライブラリ PocketSphinx [22] を利用し、オンボードでのウェイクワード検出を行った。

3.3　インタラクションデザイン

利用者の注視が検出された場合のシステムの挙動について説明する。なお今回実装したシステムでは、利用者の注視検出に HVC-P2 の視線角度認識機能ではなく、顔角度認識機能を用いている。これは、眼鏡着用者に対する検出角度の信頼性が不十分であったためである。利用者の注視が検出された場合、HVC-P2 より制御プログラムに対して検出された利用者の顔位置

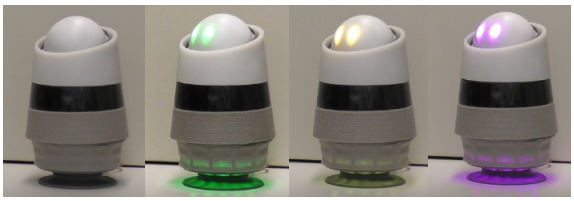

図 4　インタラクションのプロセス
Fig. 4　Process of interaction

（角度）が送信される。制御プログラムは、顔方向を受信した時点で API を介して Alexa のコマンド認識を開始させる。これは、通常のスマートスピーカーにおけるウェイクワード認識を注視の検出に代替したものである。コマンド認識の開始と同時に、ロボットの目は検出された顔の方向に回転する。これにより、システムが対応可能であることが利用者に示され、利用者との相互注視が達成される。その後利用者はコマンドの発話を開始する。制御プログラムは、利用者の発話の終端を検知した時点で取得した音声をクラウド上で処理し、認識したコマンドの内容に沿って合成音声での応答を返す。なお API の仕様上、システムが発話している際中の利用者の割り込みは許可されない。一連のインタラクションの中で、ロボットの目および本体下部の LED 発光色はシステムの状態をフィードバックする役割を持つ。図 4 に示す通り、注視が検出されコマンドを認識している最中は緑、コマンドをクラウド上で処理している最中は黄、システムが発話している際にはピンクに変化する。発話終了後、次に注視が検出されるまでは消灯状態となる。

4.　実験

本研究では、利用者とのインタラクションにおいて注視の入出力を用いる提案手法の効果を明らかにするため、実装した “Tama” を用いた実験室実験を行った。評価にあたっては、提案手法の効果として以下の 2 点の仮説を設定した。

H1.　ウェイクワードを省略し注視により認識を開始する場合、システムの操作性が向上する

H2.　システム側からも視線を提示することでインタラクションが人間との対話に近づき、印象が向上する

なお、H1. において着目する操作性とはシステムのユーザビリティのことを指し、評価に当たってはユーザビリティの定量的評価に広く用いられている評価指標 System Usability Scale(SUS) を用いることとする [23]。本章では、上記 2 点の仮説を検証するために実施した実験についての説明を行う。

4.1 実験条件

設定した仮説について検証を行うため、実験条件は以下の 3 条件とした。各条件の外観は図 2–(1),(2) に示された通りである。

C1. ウェイクワード条件 ”Alexa”という発話によりコマンド認識を開始する。ロボット頭部を用いた注視の提示は行わず、システムの状態遷移は本体下部 LED の発光色により提示される。(図 2–(2))

C2. 注視検出条件 利用者の注視を検出した場合コマンド認識を開始する。ロボット頭部を用いた注視の提示は行わず、システムの状態遷移は本体下部 LED の発光色により提示される。(図 2–(2))

C3. 相互注視条件 利用者の注視を検出した場合コマンド認識を開始する。ロボット頭部を用いて利用者の位置を追従する注視を提示し、システムの状態遷移はロボットの目の発光色および本体下部 LED の発光色により提示される。(図 2–(1))

実験は参加者内配置の実験デザインとし、実験参加者は各条件でタスクを実施した。なお、タスク開始前に各条件のシステムの挙動について説明を行うとともに、実際にコマンドを入力する練習を行わせた。順序効果を考慮し、参加者ごとの条件順序はカウンターバランスをとった。

実験参加者は大学関係者（大学生・大学院生・教員）より募集し、参加者数は 12 名とした（平均年齢 25.2 歳、女性 2 名/男性 10 名)。なお、事前にスマートスピーカー使用経験を確認したところ、スマートスピーカーを自身で所持・利用している参加者はいなかった。電通デジタル社が 2018 年に実施した調査によれば日本国内におけるスマートスピーカーの所持率は約 6 %であり [24]、本研究の実験参加者におけるスマートスピーカーの所持率は一般的な日本人を母集団とした場合に想定される範囲内であると考えられる。

4.2 実験タスク

本実験では、実験タスクとして「旅行計画を立てる場面を想定し、何らかの資料を見ながら断続的に複数のコマンドを実行する」というタスクを設定した。これは、ハンズフリーでコマンド実行が可能なスマートスピーカーの特性を活かし、ガイドブック等を見ながら情報を検索するというような並行作業を行う場面を想定したタスクである。なお本実験では、アンケートによる印象評価を実施することから、試行毎のインタラクション内容（質問回数・内容の複雑さ・エラー発生率等）を極力統一する必要があると考えた。そのため各試行ごとにタスクリストを配布し、実験参加者はそれを確認しながらスマートスピーカーを用いてコマンドを実行するという並行作業を実施した。タスクリストの内容は、旅行計画というコンテキストに沿った内容としつつ（目的地に関する情報や乗換経路の検索等）、Sciuto らの調査 [4] で示されたスマートスピーカーでよく使われる機能（天気予報・買い物リストの確認、家電操作等）を極力取り入れたものとした。1 試行ごとのコマンド総数は 14 コマンドで、うち 2 件は複数ターン対話が継続する形式のコマンドであった (件名・日時を複数ターンで登録するカレンダー機能等)。また、一部のタスクではシステムの回答をタスクリスト上にメモするよう指示した。用いたタスクの例を表 1 に示す。なお、タスクリストに沿ったタスク進行、および音声を用いたコマンド実行に慣れるよう、実験本番に先立ち本番同等のタスクリストを用いた練習を行った。

表 1 タスクの例
Table 1 example of tasks

No.	コマンド内容
1	卓上のライトをつける
2	現在時刻を確認する（時刻をメモする）
3	愛知県の県庁所在地を確認する
4	愛知県の人口を確認する
5	愛知県の知事を確認する
6	愛知県の名産品を確認する
7	愛知県の観光地を確認する
8	名古屋城について検索する
9	東京から名古屋までの経路と所要時間・料金を調べる（複数ターン対話を継続、結果をメモする）
10	買い物リストの中身を確認する
11	買い物リストに項目を追加する
12	土曜日の名古屋の天気を確認する
13	カレンダーに名古屋旅行の予定を追加する（複数ターン対話を継続）
14	卓上のライトを消す

実験環境を図 5 に示す。このような着席してインタラクションを行う環境は、Sciuto らの調査 [4] で示されたスマートスピーカーの設置場所のうち、リビングルームやホームオフィスを想定したものである。実験参加者正面の机上には左側約 45 度方向に実験システム、正面にタスクリストを配置した。このような配置により、注視検出条件・相互注視条件では、コマンド認識を開始するためにタスクリストからシステムの方に注視を移動させる必要が生じるようにした。これは、タスクリストを確認している最中の誤検出を防止することを目的とした設定である。

4.3 評価項目

本実験では、H1・H2 について検証を行うための評価項目を設定した。まず H1 については、システムの操作性を評価するための評価指標である System Usability Scale(SUS) [23] を用いた。SUS は 10 項目からなるアンケートであり、7 段階のリッカートスケールによって回答させた。

H2 については、ビデオを用いてインタラクション中の実験参加者の注視方向を分析するとともに、シス

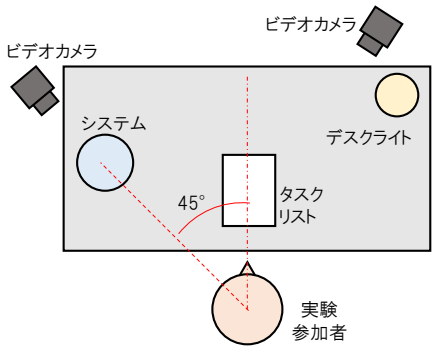

図 5　実験環境
Fig. 5　Experimental setup

テムの印象を評価するために ”Flow”（実行中の行為にどの程度興味・関心を持ち、精力的に集中しているかを示す心理学的概念）に関するアンケート [25] を実施した。注視方向の分析では、録画したビデオを用いて、タスク中の各時点で実験参加者がシステムの方向を注視していたかどうかをフレームごとに判別した。Flow に関するアンケートは、元の項目 [25] より本実験内容と無関係と判断した項目を除いた 8 項目からなり、SUS 同様 7 段階のリッカートスケールにより回答させた。なお除外したのは、元の研究 [25] における創造的な実験設定（ブロック型の提案システムを使って自由に何かを作る）に関係する「システムを使ってやりたいことが出来なかった」「システムは私の想像力を刺激した」という項目、および「次回もシステムを使った実験に参加したい」という項目である。

アンケートで用いた質問項目を表 2 に示す。質問項目 1〜10 は SUS に関する質問、質問項目 11〜18 は Flow に関する質問である。実験では各条件のタスク終了後にアンケートを実施した。また、3 条件終了後には各条件に対する意見や条件間の差分についてインタビューを行った。

表 2　アンケート項目
Table 2　list of questions

	質問項目
1	このシステムを頻繁に使用したいと思う
2	システムは必要以上に複雑であると感じた
3	システムは容易に使用することが出来た
4	システムを使えるようになるためには、技術者の支援が必要だと感じた
5	システムがもつ様々な機能はうまく統合されていると感じた
6	システムにはつじつまの合わない点がいくつもあるように感じた
7	ほとんどの人はシステムの使い方をすぐに理解すると考えられる
8	システムを使用するのはとても面倒であると感じた
9	自信を持ってシステムを使うことが出来た
10	システムをうまく使えるようになるまでに多くのことを学ぶ必要があった
11	システムを使っている時、うまくコントロール出来ていると感じた
12	システムを使っている時、他のことについて考えることがあった
13	システムを使っている時、他のことで気が散ることがあった
14	システムを使っている時、その作業に完全に没頭していた
15	システムを使うことは本質的に興味深いと感じた
16	システムとの対話は私の好奇心を刺激した
17	時間を忘れるくらいシステムに熱中した
18	システムを使うことについて肯定的な意見を持った

5.　実験結果と考察

本章では、実験の各評価項目についての結果を示し、考察を行う。

5.1　認識エラー率の確認

各評価項目についての分析を行う前に、各条件における認識エラー率について確認を行った。ここで扱う認識エラーは、コマンド開始エラーおよび Alexa の応答エラーの 2 種類である。

まずコマンド開始エラーについて、ウェイクワード条件では、ウェイクワードを発話したにも関わらず認識されなかった場合をエラーとした。また、注視検出条件と相互注視条件では、実験参加者がシステムの方向を見た後、2 秒以上認識されなかった場合をエラーとみなした。これは、HVC-P2 による顔検出のサイクルが約 500ms 毎であり、正常に検出される場合は概ね 2 サイクル（約 1 秒）以内で顔が認識されていたことから、正常値の 2 倍の 2 秒以上の時間がかかった場合をエラーとみなした。各条件のエラー率は、ウェイクワード条件 17.64%(SD=13.94)、注視検出条件 20.10%(SD=25.20) 、相互注視条件 12.19%(SD=12.12) であった。ここで、条件を要因とする一要因分散分析を行ったが、条件の効果は有意ではなかった（$F(2,22)=0.572$, $p=.572$）。

次に、各条件における Alexa の応答エラー率について比較を行った。ここで、Alexa の応答エラーとは、コマンド認識が正常に開始されたにもかかわらず、Alexa 側の音声認識エラー等により正しくコマンドが実行されなかった場合のことを示す。各条件のエラー率は、ウェイクワード条件 9.5%(SD=10.23)、注視検出条件 9.8%(SD=8.20)、相互注視条件 5.5%(SD=8.89) であった。ここで、条件を要因とする一要因分散分析を行ったが、条件の効果は有意ではなかった（$F(2,22)=1.562$, $p=.232$）。

以上に示した通り、各条件間でそれぞれの認識エラー率に有意差はなかった。このことから、次節以降で扱う各評価項目の結果に対して、エラー率の差による影響は生じていないと考えられる。

5.2　アンケートに関する分析

実施した SUS と Flow のアンケート結果について、クロンバックの α 係数を算出したところ、SUS では $\alpha=0.91$、Flow では $\alpha=0.75$ であり、内的整合性が示された。そこで、SUS と Flow それぞれについて、全設問の回答の平均値を算出し、分析に用いることとした。両項目について各条件の平均値を示したグラフを図 6 に示す。

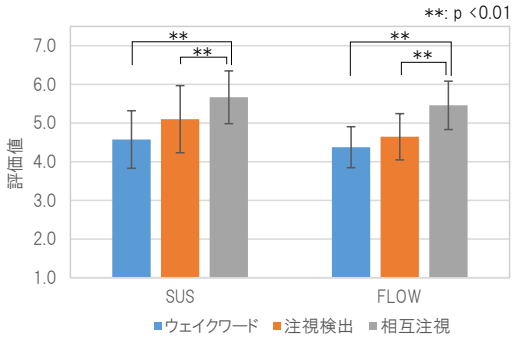

図 6　アンケート結果
Fig. 6　Result of the questionnaire

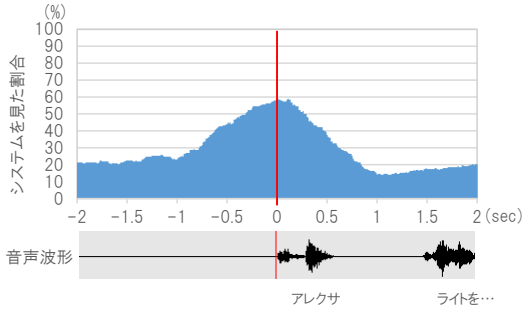

図 7　ウェイクワード前後でシステムを見た割合
Fig. 7　Ratio of participants who looking at the system around wake-word

5.2.1　SUS に関する結果と考察

SUS の結果について、各条件を要因とする一要因分散分析を実施した結果、条件の効果が有意であった（$F(2,22) = 7.675$, $p = .003$）。下位検定として Bonferroni 法による多重比較を行った結果、相互注視条件とウェイクワード条件・注視検出条件の間でそれぞれ有意差が見られた ($p = .005$, $p = .004$)。これらの結果より、SUS の評価値すなわちシステムの操作性は相互注視条件において他の 2 条件よりも有意に向上することが示された。一方でウェイクワード条件と注視検出条件の間には有意な差はなく、十分な操作性の向上が見られなかった。

このような結果と関連して、インタビューにおいて「何度もウェイクワードを言わなくてはならず煩わしく感じた」といった内容の意見が 12 名中 8 名から得られた。一方で、相互注視条件と注視検出条件の差分に言及して「相互注視条件ではロボットと対話している感覚で自然に視線を向けていたが、注視検出条件では何を見て話しているんだろうと感じた」という意見が得られた。これらのことから、ウェイクワードの発話は煩わしく感じられるものの、それを注視に置き換えるだけでは、注視を操作のための付加的行為として意識的に行う必要があり、注視検出条件における操作性の向上に繋がらなかった可能性がある。一方で相互注視条件では、注視を向けるという行為がインタラクションの中で自然に誘起されたため意識的に操作を行っている感覚が軽減され、操作性の向上に繋がった可能性がある。

5.2.2　Flow に関する結果と考察

次に Flow の結果について、各条件を要因とする一要因分散分析を実施した結果、条件の効果が有意であった（$F(2,22) = 25.202$, $p = .000$）。下位検定として Bonferroni 法による多重比較を行った結果、相互注視条件とウェイクワード条件・注視検出条件の間でそれぞれ有意差が見られた (いずれも $p = .000$)。これらの結果より、SUS と同様に Flow についても相互注視条件において他の 2 条件よりも有意に評価値が向上することが示された。

このような結果と関連して、インタビューにおいて「目で追ってくれるので、話を聞いてくれているのだと感じた」、「他の条件よりも会話をしている感覚が強かった」、「タスクに集中できた」、「可愛らしく楽しく感じた」等の意見が得られた。これらのことから、実験参加者は相互注視条件においてインタラクションがより対話的であると感じ、その結果として、タスクにより強く興味・関心を持ち集中できた可能性がある。

5.3　実験参加者の注視方向に関する分析

5.3.1　ウェイクワード発話時にシステムを見た割合に関する結果と考察

注視方向に関する分析項目として、まずウェイクワード条件において、ウェイクワード発話時に実験参加者がシステムの方を見たかどうかの判定を行った。これは、実験映像より、注視を用いないウェイクワード条件においても、実験参加者がシステムの方を見る様子が多く見られたためである。判定にあたっては、ウェイクワードを発話した時点を起点として次に発生するコマンドの発話時点までの区間を判定区間とし、その中で一度でもシステムの方向を見ていれば「見た」と判断した。なお、まれにウェイクワードの識別エラーが発生したが、その場合実験参加者は異常を察知してシステムの方向を見る傾向があったため、エラーが発生した試行は判定対象外とした。集計の結果、ウェイクワード発話時 71.3%の割合で実験参加者はシステムの方を見ていることが明らかになった。

このような傾向をより詳細に調べるため、ウェイクワードを発話する前後のどの時点で実験参加者がシステムの方を見ていたかの分析を行った。ここでは、ウェイクワードを発話した時点を起点（0 秒）として、前後 2 秒間の各時点で実験参加者がシステムの方を見た割合を算出した。結果を図 7 に示す。なお、グラフ下の波形は実験参加者の発話波形の例を参考に示したものである（厳密には、波形は発話毎に毎回少しずつ異なる）。図 7 に示された通り、実験参加者がシステムを見る割合は発話を開始した時点（0 秒）がピークとなり、6 割程度であることが示された。

上記の様に、多くの実験参加者がウェイクワード発

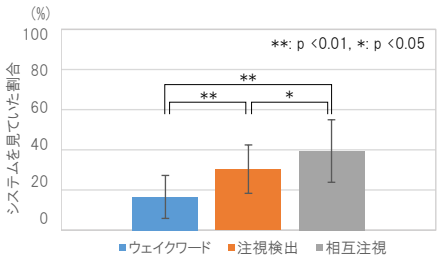

図 8　システムを見ていた時間の割合
Fig. 8　Proportion of time spent in looking at the system

話時にシステムの方を見た理由として、インタビューでは「システムが正しく検出しているか確認するためシステムを見た」という意見が得られた。一方で、図 7 に示された通り、システムの方を見た割合はウェイクワードの発話を開始した時点、すなわちシステムがフィードバックを返す前がピークとなっていることから、必ずしもフィードバックを確認するためにシステムを見ているわけではないと考えられる。このような結果に関連して、人間同士の会話においても、短い質問等の開始時点で相手の方を見る傾向があることが報告されていることから [11]、同様な行為がシステムとのインタラクションにおいても発生したと考えられる。

このように、注視を向けるという行為は通常のウェイクワード条件においても発生することが示された。そのため、注視をウェイクワードの代替手段として用いる本研究の提案手法は、人間の自然な行為を利用した妥当な設計指針であるように考えられる。一方で、SUS の結果より、注視検出条件では操作性の有意な向上は見られていないことから、例え人間の自然な行為を利用する場合でも、それが操作として意識されると心理的な負荷向上に繋がると考えられる。

5.3.2　インタラクション中にシステムを見ていた割合に関する結果と考察

次に、各条件においてインタラクション中にシステムを見ていた割合の比較を行った。分析にあたっては、実験参加者がコマンドを発話した時点からシステムからの音声応答が終了するまでの区間を評価区間として、各コマンドにおいて評価区間中にシステムを見ていた時間の割合を算出した。なお、ウェイクワード条件におけるコマンド発話前（ウェイクワード発話時）の注視については、ここでは評価対象外とした。また、Alexa の認識ミス等によりシステムが正常な応答を返さなかった場合、実験参加者は異常を察知してシステムの方向を見る傾向があったため、そのような場合は評価対象外とした。

全試行の平均割合を条件ごとにまとめた結果を図 8 に示す。結果について、各条件を要因とする一要因分散分析を実施した結果、条件の効果が有意であった（$F(2,22) = 23.633$, $p = .000$）。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注視検出条件・相互注視条件間、および注視検出条件と相互注視条件間でそれぞれ有意差が見られた（$p = .000$, $p = .000$, $p = .012$）。

上記結果より、インタラクション中にシステムを見ていた時間の割合について考察を行う。まずウェイクワード条件と注視検出条件・相互注視条件間の差分について、注視検出条件と相互注視条件ではコマンド認識を開始するために必然的にシステムの方を見ることから、その後のインタラクションにおいても惰性によりシステムを見る時間が長くなったと考えられる。次に注視検出条件と相互注視条件の差分については、5.2.2 節で示した Flow の結果と関連し、相互注視条件ではシステムとのインタラクションにより強く関心を持ったことから、システムを見る時間が長くなったと考えられる。このような結果に関連して、2 人の人間が会話を行う状況においては、相手を見る時間の割合が 28〜70%（平均 49%）であるという結果が報告されている [11]。このような知見と比較して考えると、相互注視条件では他の条件よりも人間同士の対話に近い注視行動が見られたといえる。

5.4　総合考察

5.4.1　仮説に対する検証と本研究の貢献

本研究では、以下の 2 点の仮説を設定し、検証を行った（再掲）。

H1.　ウェイクワードを省略し注視により認識を開始する場合、システムの操作性が向上する

H2.　システム側からも視線を提示することでインタラクションが人間との対話に近づき、印象が向上する

H1 については、SUS の結果より、ウェイクワード条件と比較して相互注視条件では有意に操作性が高くなることが示されたものの、注視検出条件では有意差は見られなかった。すなわち、注視を用いる手法は、注視の検出だけでなくシステム側からの注視の提示と組み合わせた場合にシステムの操作性向上に寄与することが示され、仮説は部分的に支持された。

H2 については、Flow の結果より、ウェイクワード条件・注視検出条件と比較して相互注視条件では有意に評価が高くなることが示された。また、インタラクション中に実験参加者がシステムの方を見た時間も他の条件と比較して有意に長く、より人間同士の対話に近い注視行動が見られた。これらの結果より、H2 は支持された。

上記の結果より、スマートスピーカーの操作性および対話感を改善するためには、注視の検出とシステム側からの注視を組み合わせる手法が有効であることが示された。

本研究により得られた知見は、スマートスピーカーを改善し、より対話的なインタラクションを実現するために活かすことが出来る。例えば、Sony Mobile Communications 社はロボット頭部を備えたスマートスピーカー Xperia Hallo [26] を販売しており、そのようなシステムのインタラクション設計に活かすことが可能であろう。

また、Amazon Echo Show [1] 等の様にディスプレイとカメラを備えたスマートスピーカーの普及が進んでおり、このようなデバイスにおいてもディスプレイ中でエージェントを提示する等の手段を用いることで、本研究で提案したような視線のインタラクションを疑似的に活用することが可能であろう。

5.4.2 本研究の制約と今後の課題

本研究の実験設定上の制約として、着席状態でシステムが目に見える位置に配置された場合のインタラクションのみを対象としている。一方でスマートスピーカーは家庭内の様々な場所で利用されることから、例えば環境内を歩きながらインタラクションを行う等、異なる状況下においても提案手法の効果を検証する必要がある。

次に実験参加者に関する制約として、本研究で募集した参加者 12 名はいずれもスマートスピーカーを自身で所持・利用していなかった。スマートスピーカーを自身で所持し日常的に利用している場合、提案システムに対する評価が変化する可能性がある。そのため、スマートスピーカーを日常的に利用しているユーザーを対象とした評価を行う必要がある。

また、インタビューにおいて「家で使う場合、システムが見えない位置から利用することも考えられるので、そのような場合はウェイクワード条件の方が良い」という主旨の意見が得られた。このような意見に代表されるように、注視をウェイクワードの代わりに用いる本研究の提案手法では、カメラの画角内に入らなければシステムを起動できないという制約が発生する。さらに、画角内であってもカメラに近すぎたり遠すぎたりする場合には検出精度が低下するという制約も発生する。これらの制約を考慮すると、実環境での使用を視野に入れた場合は注視検出とウェイクワードの両方を併用できる構成が望ましいと考えられる。そのため、ウェイクワードと注視を組み合わせる場合の適切なインタラクションデザインを検討する必要がある。

さらに、本研究では利用者が一名の場合のインタラクションに限定して評価を行っているが、実環境では複数名が同時にシステムとのインタラクションに参加する状況も発生する。これに対して、複数の参加者からの注視を検出した際にどのように対応するか等の動作指針を検討・実装する必要がある。

6. おわりに

本研究では、既存のスマートスピーカーにおける課題として、対話構造の制約により自然な対話が実現されていない点に着目した。その解決のため、「会話の開始」についての社会学的知見に基づき、利用者とのインタラクションに注視の入出力を用いるシステムの提案を行った。

提案手法の効果を評価するため、開発したスマートスピーカー “Tama” を用いた評価実験を実施した。実験では、着席状態でタスクリストに沿ってシステムとインタラクションを行う実験タスクを設定した。評価にあたっては、注視を用いないウェイクワード条件、注視の入力のみを用いる注視検出条件、および注視の入出力を用いる相互注視条件の比較を行った。実験結果より、注視の入出力を用いた場合に、システムの操作性および対話感が向上することが示された。

今後の方針として、本研究で評価を行った着席状態以外の状況（環境内を歩きながらインタラクションを行う等）においても提案手法の効果を検証する。また、ウェイクワードと注視の入出力の併用を検討するとともに、複数人の利用者を想定した動作指針の検討・実装を行う。

謝辞

本研究は JSPS 科研費 18H06473、および筑波大学研究基盤支援プログラム（タイプ A）の助成を受けた。

参考文献

[1] Amazon.com, Inc.: Echo & Alexa; `https://www.amazon.co.jp/b?ie=UTF8&node=5364343051`

[2] Google LLC.: Google Home; `https://store.google.com/jp/product/google_home`

[3] Koetsier, J.: Smart Speaker Users Growing 48% Annually, To Hit 90M In USA This Year (online article); `https://www.forbes.com`,Forbes Media LLC

[4] Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I.: ”Hey Alexa, What’s Up?” A Mixed-Methods Studies of In-Home Conversational Agent Usage; In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS ’18*, pp.857-868, ACM Press, (2018).

[5] Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S.: Voice Interfaces in Everyday Life; In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI ’18*, pp.1-12, ACM Press, (2018).

[6] Vtyurina, A., & Fourney, A.: Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants; In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Paper No. 208, ACM Press, (2018).

[7] Heath C.: Body Movement and Speech in Medical Interaction; *Cambridge University Press*, (1986).

[8] Amazon.com, Inc.: Follow-Up Mode; https://www.amazon.com/gp/help/customer/display.html?nodeId=202201630
[9] Google LLC.: Continued Conversation; https://support.google.com/googlehome/answer/7685981?hl=en
[10] Schegloff, E. A.: Sequencing in Conversational Openings; *American Anthropologist*, vol.70, No.6, pp.1075-1095, (1968).
[11] Kendon, A.: Some functions of gaze-direction in social interaction; *Acta Psychologica*, vol.26, pp.22-63, (1967).
[12] Nielsen, G.: Studies in self-confrontation; *Munksgaard*, (1964).
[13] Heath, C., Jirotka, M., Luff, P., & Hindmarsh, J.: Unpacking collaboration: the interactional organisation of trading in a city dealing room; Computer Supported Cooperative Work (CSCW), 3(2), pp.147165, (1994).
[14] Salvadori, F. A.: Open office interaction: Initiating talk at work (Doctoral dissertation); *King's College London*, (2016).
[15] Bohus, D., Chit W. Saw, & Eric Horvitz.: Directions robot: in-the-wild experiences and lessons learned; In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp.637-644, (2014).
[16] Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., & Hagita, N.: Footing in human-robot conversations: how robots might shape participant roles using gaze cues; In *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction*, pp.61-68, (2009).
[17] Anas, S.A. binti, Rauterberg, M., & Hu, J.: Designing Elements for a Gaze Sensitive Object: Meet the CoffeePet; In *Proceedings of the 5th International Conference on Human Agent Interaction – HAI ' 17*, pp.223231, ACM Press, (2017).
[18] Seeed Technology Co.Ltd.:Respeaker Core; https://www.seeedstudio.com/ReSpeaker-Core-Based-On-MT7688-and-OpenWRT-p-2716.html
[19] Seeed Technology Co.Ltd.:Respeaker Mic Array; https://www.seeedstudio.com/ReSpeaker-Mic-Array-Far-field-w-7-PDM-Microphone-p-2719.html
[20] OMRON Corporation: ヒューマンビジョンコンポ（HVC）シリーズ; https://plus-sensing.omron.co.jp/product/hvc-p2.html
[21] Amazon.com, Inc.: Alexa Voice Service; https://developer.amazon.com/alexa-voice-service
[22] Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., & Rudnicky, A. I.: Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices; In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1, pp.I–185–I–188, IEEE, (2006).
[23] Brooke, J.: SUS - A quick and dirty usability scale; Usability evaluation in industry, pp.189-194, (1996).
[24] 株式会社電通デジタル:スマートスピーカー利用実態調査調査レポート; https://pages.dentsudigital.co.jp/rs/859-LJN-010/images/電通デジタル_スマートスピーカー利用実態調査_20190218.pdf
[25] Zuckerman, O., & Gal-Oz, A.: To TUI or not to TUI: Evaluating performance and preference in tangible vs. graphical user interfaces;*International Journal of Human-Computer Studies*, 71(78), pp.803-820, (2013).
[26] Sony Mobile Communications inc.: Xperia Hallo G1209; https://www.sonymobile.co.jp/product/smartproducts/g1209/

（2019 年 2 月 14 日受付，5 月 31 日再受付）

著者紹介

川口　一画

1986 年生．2017 年筑波大学大学院システム情報工学研究科知能機能システム専攻博士後期課程修了．博士（工学），修士（感性科学）．現在，筑波大学システム情報系助教．平成 27 年度総務省異能 vation プログラム本採択．CSCW,HRI の研究に従事．情報処理学会，日本デザイン学会会員

葛岡　英明　（正会員）

1962 年生．1986 年東京大学工学部機械工学科卒．1992 年同大大学院情報工学専攻博士課程修了．博士（工学）．2006 年筑波大学大学院システム情報工学研究科教授．2019 年，東京大学大学院情報理工学系研究科教授，現在に至る．CSCW，HRI，バーチャルリアリティの研究に従事．ヒューマンインタフェース学会，日本バーチャルリアリティ学会，情報処理学会，ACM 各会員．

Donald　McMillan

He is an assistant professor at the Department of Computer and Systems Sciences, Stockholm University, Sweden. He received his PhD in Human Computer Interaction from the University of Glasgow, UK, in 2012. His research interests include combining conversational and speech based interaction with a long standing focus on how technology is used and interwoven into the ongoing collaborative, contextual, and complex lives of users.