

# 音声対話型検索におけるシステムの応答の長さが検索行動へ与える影響の分析

飛岡 憲<sup>†</sup> 山本 岳洋<sup>†</sup> 大島 裕明<sup>†</sup>

<sup>†</sup> 兵庫県立大学 大学院情報科学研究科 〒651-2197 兵庫県神戸市西区学園西町 8-2-1

E-mail: <sup>†</sup>af24h009@guh.u-hyogo.ac.jp, <sup>††</sup>t.yamamoto@sis.u-hyogo.ac.jp, <sup>†††</sup>ohshima@ai.u-hyogo.ac.jp

**あらまし** 本研究では、音声のみでシステムとユーザが検索を行う音声対話型検索における、システムの応答の長さとユーザの検索行動の関係について調査する。音声対話型検索では、長い応答はユーザの認知的負荷が高く、特に比較や意思決定を伴うような複雑な検索タスクにおいてはその影響は大きいと考えられる。そこで本研究では、システムが応答を端的に要約して応答する短応答型提示を提案する。実験のため、既存の API を利用し実世界における音声対話と近い形で音声対話型検索を可能とするシステムを構築した。実験では、通常の長い応答をユーザに返すシステムと、短い応答を返す提案システムの 2 つを用いて実験参加者に検索を行ってもらった。また、検索タスクの複雑性と応答の長さとの関係を明らかにするため、複雑な検索タスクと簡単な検索タスクで実験を行った。実験の結果、短い応答を返す提案システムの方が、検索タスクの複雑さによらずユーザの認知的負荷が減少する傾向にあり、その結果、タスクのサマリ結果にシステムから得た情報を多く含んでいることが明らかになった。一方で、ユーザのサマリ内容は、長い応答を返すシステムよりも考察や分析の質において劣る傾向にあった。ユーザのクエリを分析した結果、探索的検索を促すために、初期段階で広範囲に情報を提示するシステム応答戦略の必要性が示唆された。

**キーワード** 音声対話型検索, ユーザインターフェース

## 1 はじめに

ユーザ体験の向上と自然言語処理技術の進歩を成長要因として、人工知能を搭載した音声対話エージェントは、年平均成長率 28.5%、2033 年までに約 319 億米ドルに達すると予想されている<sup>1</sup>。今後、ChatGPT<sup>2</sup>、Gemini<sup>3</sup>、Copilot<sup>4</sup>のようなサービスが、音声対話エージェントのエンジンとして実装されると考えられる。現在、音声検索は、天気やニュースを調べるなど、答えが明確なタスクにのみ利用されている [1] [20]。しかし、技術の進歩によって、自然な対話が実現されるにしたがって、将来的には複雑な探索的検索にも音声対話型検索が利用されることが期待される [29] [16]。

この探索的検索は、学習や意思決定をするための検索である。つまり、単純に一問一答で答えがわかるようなタスクではなく、比較や分析が必要となるような複雑性が高いタスクである。ユーザが獲得した情報を比較する際に、システムの応答が長文だと、記憶することが難しく、有効な検索ができないことが想定される。また、複雑性が高ければ高いタスクであるほど、情報獲得が困難になると考えられる。

初期の対話システムは、主にルールベースと統計モデルに依存しており、特に、Eliza のようなシステムは、短く簡潔な応答を生成するように設計されていた。さらに、これらのシステ

ムはタスク指向で、単純な対話のために定義されたルールに基づいていた [28]。

現在の対話システムは、ChatGPT を例にとると、何も指示を与えない場合、ユーザのクエリに対して非常に長い応答を返す。例えば、ユーザが「アメリカの平均的なアルコールの消費量を教えてください。」と質問した場合、ChatGPT は図 1 に示すように非常に長い応答を返す。仮にこの文字を音声で読み上げるならば、ユーザにとって情報量が多すぎるため、記憶することが困難である。

このような背景から、応答提示方法を検討することは、音声対話型検索によるユーザの情報獲得において重要なテーマである。そこで、我々は問題を解決するために、ユーザの応答に対してシステムが短く的確な応答を返す「短応答型提示法」を提案する。この提案手法は、ユーザの認知的負荷を軽減できるため、満足度の高い情報検索の実現と得られる情報の質の向上に寄与することが期待される。

音声による情報検索結果の提示に関する先行研究では、音声対話型検索ではなく、音声ベースのウェブ検索インタフェースの設計に焦点を当てている [26] [27]。これらの研究では、テキストと音声という 2 つのメディアにおける検索結果の表現を比較し、ユーザの好みを分析している。音声対話型検索における応答提示方法を理解するためには、より現実的な環境下での実験が必要である。そこで、完全に音声のみでの検索が可能なデスクトップ Web アプリケーションを構築し、実験を行った。

また、音声対話型検索において、探索的検索へのアプローチを検証しようとする動きがあるものの、応答の長さが検索行動

1 : <https://scoop.market.us/ai-in-voice-assistants-market-news/>

2 : <https://chatgpt.com/>

3 : <https://gemini.google.com/app>

4 : <https://copilot.microsoft.com/>

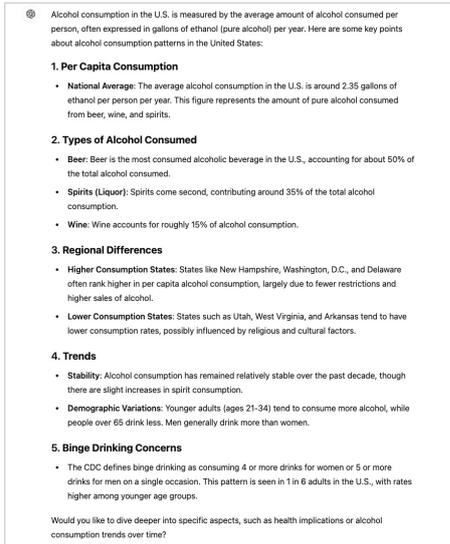


図 1: ChatGPT は箇条書きで長い応答を返答する (gpt-4o-2024-08-06).

や獲得した情報の質へ与える影響が明らかになっていない [12].

以上から、以下の研究課題を明確にする必要がある：

**RQ1:** 音声対話型検索において短応答型提示は有効か。

**RQ2:** タスクの複雑性と応答の提示方法の組み合わせは、ユーザの検索行動や獲得した情報にどのような違いを生み出すのか。

タスクの複雑性を鑑みて、人間同士の対話分析実験を行った Trippas [25] の研究を参考にトピックを選択した。また、OpenAI の ChatGPT を活用し、タスクの複雑性の高低を組み合わせ、2つのシステムを比較できるブラウザベースのウェブアプリケーションを構築した。ウェブアプリケーションを用いて、ユーザ調査 ( $N = 8$ ) を行った。ユーザ体験の満足度、認知負荷の観点から、短応答型提示が有効である傾向が見られた。

また、獲得した情報のナゲット再現率は、短応答型提示の方が高いが、ユーザのサマリについては、長応答型提示の方が、考察や分析が含まれた内容になっている傾向が見られた。

## 2 関連研究

### 2.1 音声対話型検索による探索的検索

人とシステム間のインタラクションのアプローチは、ユーザ主導かシステム主導かの大きく2種類に分けられる。現在の検索エンジンは、ユーザ側からクエリを発行し、検索エンジンがその結果を回答するというインタラクションであることからユーザ主導のシステムであるといえる。一方、対話型検索システムは、エージェントがユーザのニーズをくみ取って回答する中で、ユーザとエージェントが状況に応じて会話を主導していく混在型主導が求められるシステムであると定義されている [18].

Allen らが、この混在型主導の対話を研究し始め [3], その後の研究者たちによって、対話セッション開始時は、理想的にはシステムが会話をリードし、クエリの提案や、クエリの明確化を求めることで、ユーザにとって魅力的な対話型検索を提供で

きるとした [2] [19] [31].

音声対話型検索研究の分野では、Trippas らが、相互作用性、複雑性、仲介性を取り入れて対話型検索の定義を拡張した [23]. また、音声のみの対話型検索における個人間の相互作用 [24] [25] や、文脈を維持する対話型検索エージェントが人間の行動やパフォーマンスに与える影響 [7] が研究されている。また、Abhishek らは、音声対話エージェントの探索的検索への可能性を探求した [12]. 既存のディスプレイ画面を備えた Alexa が探索的検索タスクをサポートする能力を欠いており、ユーザの不満を引き起こしていることを示した。また、カスタムスキルを実装し、上位3件をディスプレイ上に表示したり、文章を要約したりすることによって、ユーザの理解度や満足度が向上することを示した。

しかし、音声対話型検索において、応答の長さが与える情報検索行動への影響は明らかになっていない。

### 2.2 システムの応答提示方法

応答の提示方法に関する研究は、主にデスクトップ上での視覚的表現に焦点を当てている。しかし、音声認識技術や自然言語処理技術の進歩に伴って、結果を音声のみで提示する方法を調査する研究がされ始めた。多くの研究者が、検索結果ページを単純に音声に変換する方法は理想的ではないと指摘した [14] [23] [27].

Vtyurina らは、テキスト結果をそのまま音声に変換することが、ユーザ体験を損ね、認知負荷を増加させる可能性があることを明らかにした [27]. Trippas らは、クラウドソーシングを用いて要約の長さに関するユーザの好みを調査した [26]. ユーザにクエリと検索結果を提示し、さまざまな長さの結果要約をテキストまたは音声で提示した。その後、ユーザに最も適切な結果を選んでもらう実験を行った。ユーザの選択は、テキスト形式では情報量の多い長めの要約が好まれる一方で、音声形式では短めの要約が好まれることが示唆された。また、クエリの種類 (単一側面か多側面か) によって結果が異なり、単一側面のクエリでは短い音声要約が好まれる傾向があった。さらに、音声応答が情報豊富であることが、聴取時間に関係なくユーザにとって有益である可能性が示唆された。

これらの研究は、音声の提示方法の好みを明らかにすることに焦点を当てている。また、音声対話型検索ではなく、ユーザがデスクトップ上で検索を行い、結果が音声とテキストで提示される形式であった。そのため、純粋な音声検索を対象とした実験環境ではない。

Gabriel らは、2つの短いキーワードベースの応答と1つの完全な文章 (ベースライン) で応答する音声アシスタントを開発した [9]. 短い応答はベースライン応答よりも効果的であると示唆された。さらに、特定のトピックについては、コマンド形式の応答が、理解しやすく、便利で好まれると明らかにした。そのため、カスタマイズ可能でユーザに適応した応答形式が望ましいと主張した。この研究は音声対話型検索であったが、探索的検索を目的としたものではなく、むしろ、音楽、天気、リマインダーなどの、数回のクエリを実行するための情報提示方

法に関するものであった。

### 3 実験

本章では、リサーチクエクションを明らかにするために実施した実験の詳細について述べる。本実験では OpenAI の API を用いて、デスクトップ対話システムを構築し、オンライン上で実験を行った。実験参加者は 2 種類の異なるシステムを用いて、与えられた 2 つのトピックに関する検索タスクをこなした。被験者内実験計画を用いて、実験参加者が全ての実験条件を経験できるようにした。2 つの検索タスクを行う順序については、実験条件およびタスクの両者を考慮した Graeco-Latin 方格 [13] に従い実験参加者を割り当て、順序効果を最小限に抑えた。

#### 3.1 トピックとシナリオ

本研究では、2 つのトピックとそれに対応するシナリオを作成した。参加者には、それぞれのトピックに関連する情報を多面的に検索し、理解したことを要約するように指示した。実験に使用したトピックとシナリオは以下の通りである：

**複雑性:** Remember (複雑性低)

**トピック 1:** シナモンの原産地

**シナリオ:** 先日、あなたはヨーロッパから輸入されたスパイスクッキーを食べていた。あなたはシナモンの原産地を調べようと思った。

**複雑性:** Analyze (複雑性高)

**トピック 2:** 一人当たりのアルコールの消費量の比較

**シナリオ:** あなたは最近、大きなパーティーに参加し、翌日二日酔いで目覚めた。そのため、アルコールの平均消費量についてもっと知りたいと思った。特に、一人当たりの消費量を報告する情報に興味があり、国、県、市町村レベルなど、グループ間で比較したいと考えた。

学習分類法 [4] における認知的複雑さのフレームワークに基づいて作成された Trippas [25] の研究のトピックとシナリオを参考にした。本研究においては、5 つの認知次元のうち、Remember (認識, 想起) と Analyze (区別, 整理, 帰属) の 2 つの次元を使用した。Remember は複雑性が低く、Analyze は複雑性が高いとされる。

#### 3.2 実験環境

リサーチクエクションを明らかにするため、2 つのシステムを準備した：

**長応答型提示 (ベースライン)** ChatGPT (gpt-4o-2024-08-06)。

ユーザのクエリに対して、長い文章で様々な観点から情報を提示したり、時には箇条書きで提示したりする方法。ユーザが知りたいと思うことに対して、さらに情報が追加される。

**短応答型提示 (提案手法)** ユーザのクエリに対して、ユーザが知りたい情報を 1 2 文で端的に提示したり、要約提示したりする方法。



図 2: 実験に用いたウェブアプリケーションのインターフェース

ChatGPT のモデルは gpt-4o-2024-08-06 を使用した。長応答型提示のプロンプトは「あなたは情報検索ができる優秀な日本人です。ユーザからの相談や質問に答えてください。」とした。一方、短応答型提示では、前述のプロンプトに、「必ず、とても短く、簡潔に、要約して伝えてください。」という一文を加えた。これらのプロンプトによって、表 1 に示したように、長応答型提示と短応答型提示を実現した。なお、共通のプロンプトは、ELYZA-japanese-Llama-2-7b のデフォルトプロンプトを参考としている<sup>5</sup>。

実験用のシステムのインターフェースは図 2 のとおりである。参加者は、このウェブアプリケーションを用いて、システムと対話を行う。システムとの対話方法は、以下の通りである：

1. ユーザは、マイクボタンを押して発言を開始する。
2. マイクボタンを再度押すと、音声は WhisperAPI でテキストに変換され、ChatGPT に入力される。
3. ChatGPT で応答が生成されると、テキストは text-to-speech API によって音声に変換され、ユーザの画面上で読み上げられる。なお、応答が読み上げられる間、ユーザは介入できない。
4. ユーザはマイクボタンを再度押すことで、必要に応じてシステムとの対話を続けることができる。

実験用のシステムは、フロントエンドは Next.js で実装し、バックエンドは Firebase で実装した。システムの応答には ChatGPT を使用した。ChatGPT を採用した理由は、応答生成の精度が高いことと、OpenAI が本システム構築に必要な speech-to-text と text-to-speech 機能を全て提供しているからである。これによって、必要な機能を実現しつつ、システムの複雑さを軽減することが可能となった。

また、Firebase を用いて、発話テキストデータ、ユーザとシステムのそれぞれがクエリ発行した時間、及びクエリ発行し終えた時間といった検索行動の分析に必要なログデータを取得した。

#### 3.3 実験の流れ

図 3 に実験の流れを示す。まず、実験参加者にシステムにログイン後に、実験の手順を説明した動画を視聴してもらった。

<sup>5</sup> : <https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b>

表 1: 長応答型提示（ベースライン）と短応答型提示（提案手法）の比較.

ユーザクエリ	長応答型提示（ベースライン）	短応答型提示（提案手法）
シナモンの原産地はどこですか？	シナモンの原産地は、スリランカやインド南部とされています。特に、スリランカで生産されるシナモンは「セイロンシナモン」として知られており、最も歴史が古く、品質が高いとされています。また、「カシア」と呼ばれる種類のシナモンは中国やベトナム、インドネシアなどでも広く生産されています。これらの地域はシナモンの生産と輸出で重要な役割を果たしています。	シナモンの原産地はスリランカやインド南部です。
日本人の平均アルコール消費量はどれくらいですか？	日本における一人当たりのアルコール消費量は、一般に年間で約 7~8 リットルの純アルコールに相当するとされています。これにはビール、ワイン、日本酒、焼酎、ウイスキーなどの異なる種類の酒類すべてが含まれています。これはあくまで全国平均であり、消費量は年齢層や地域、個々の生活習慣などによって異なることがあります。また、消費量は社会的なトレンドや政策の影響を受ける可能性もあるため、年によって若干の変動があります。具体的な数値や詳細なデータについては、厚生労働省や国税庁の最新の統計情報を確認するのが良いでしょう。彼らは定期的にアルコール消費に関する報告を提供しています。	日本人の年間アルコール消費量は一人あたり約 8 リットルです。



図 3: 実験の流れ

この動画<sup>6</sup>によって、研究目的、検索タスクの概要やシステムの使い方を理解してもらった。システムのトップページに、テキストデータや検索に関わるログをクラウド上に保存し、本研究の分析することについての説明が記載されており、実施内容について同意した実験参加者のみ、実際の実験に参加できることとした。

次に、実際の実験タスクに入る前に、使用する検索システムに慣れるために、訓練タスクを行ってもらった。訓練タスクでは、「ダークチョコレートの健康効果について」をトピックに設定し、自由に検索を行ってもらった。最大検索時間は 15 分とすること、情報が十分に集まったと感じたらいつでも検索を終了してもよいことが画面上に提示された。

訓練タスク終了後は、実験参加者に 2 つの検索タスクを実施してもらった。はじめに、3.1 節で示したトピックとシナリオのいずれかが、グループの割り付けに従って表示される。続けて、トピックとシナリオを読んでから、1 つ目の検索タスクを実行してもらった。実験参加者が 1 つ目の検索タスクを終了した後、システムに関するアンケートに回答してもらった。続いて、2 つ目の検索タスクも同様の手順を踏んで実施した。最後に、デモグラフィック属性に関するアンケートに回答してもらった。

### 3.4 アンケート

アンケートでは、検索で理解した内容を自由記述で回答する項目、認知的負荷を測定する NASA-TLX [10]、応答の長さが情報検索において有効であったかどうかを問う質問で構成した。

応答の長さに関わるアンケート項目は以下のとおりである：

- Q1 この音声対話型検索システムの応答の長さは適切だと感じた。
- Q2 この音声対話型検索システムの適切な応答の長さが、情報獲得するうえで役に立ったと感じた。
- Q3 この音声対話型検索システムは様々な情報を獲得するうえで役に立ったと感じた。
- Q4 意見をまとめるうえで必要な情報が十分に集められたと感じた。

5 段階のリッカート尺度 (1: まったくあてはまらない, 2: どちらかというあてはまらない, 3: どちらともいえない, 4: どちらかというあてはまる, 5: とてもよくあてはまる) を設定して回答してもらった。

### 3.5 実験参加者

著者が所属する大学のコミュニケーションツールを通して 8 名の実験参加者を募集した。5 名が男性、3 名が女性であった。参加者全員が情報学を専攻する学生であった。そのうち 2 名が学部生、残りの参加者は大学院生であった。参加者に音声対話型検索エージェントを使って情報を調べることがあるかどうかについてアンケートをとったところ、全員が「使ったことがある」と答えた。最も一般的な用途は天気予報の確認で、次いで音楽再生であった。また、音声対話型検索エージェントの使用頻度については、3 名が週に 1 回、1 名がほぼ毎日、残りの参加者はほとんど使用しないという結果であった。

実験の開始から終了までに最大 60 分の時間をとった。実験を最後まで完了した 8 名の実験参加者には約 1,500 円を支払った。どの実験参加者も問題なく実験をやり終えたため、すべての実験参加者の検索タスクに関するデータを分析に用いた。

## 4 結 果

本章では、1 章で述べた 2 つのリサーチクエスションに関する実験結果とアンケート結果を述べる。二元配置分散分析を用いて、システム間に統計的に有意差があるかを比較した。有意水準は 5% に設定した。

6 : <https://youtu.be/ETYonvNW-9Y>

## 4.1 RQ1: 音声対話型検索において短応答型提示は有効か

### 4.1.1 認知的負荷

音声対話型検索において、システムの長い応答はユーザへの認知的な負荷が高いため、適切ではないと考えられる。したがって、提案手法である短応答型提示によって、認知的な負荷の軽減が期待される。有効性を評価する指標として、認知的負荷を測定するために NASA-TLX を用いて評価を行った。

表 2 は、長応答型提示（ベースライン）と短応答型提示（提案手法）、そして、複雑性の高低を含めた NASA-TLX の平均値を示したものである。

Mental について、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.21, p = 0.66$ )。また、複雑性の主効果に有意差が見られなかった ( $F(1, 12) = 2.73, p = 0.12$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.01, p = 0.94$ )。この結果から、提示法の相違や複雑性が高いタスクは、必要な思考、判断、記憶、集中に関わる負荷に差がないことが示唆された。ただし、有意差はないものの、短応答型提示で複雑性が低い場合に、最も低いスコアとなっている。

Physical について、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.24, p = 0.63$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.02, p = 0.89$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.08, p = 0.78$ )。この結果から、提示法の相違や複雑性の高低において、身体的な負荷に対する度合いに差がないことが示唆された。有意差はないが、長応答型提示で複雑性が高い場合に、最も低いスコアとなっている。

Temporal について、二元配置分散分析の結果、提示法の主効果に有意差が見られた ( $F(1, 12) = 5.83, p = 0.03$ )。また、複雑性の主効果に有意差が見られた ( $F(1, 12) = 5.17, p = 0.04$ )。しかし、提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 1.37, p = 0.26$ )。この結果から、複雑性が高いタスクや長応答型提示の場合は、タスクを遂行するために感じる時間的なプレッシャーや急かされる感覚が大きいことが示唆された。

Performance について、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.65, p = 0.43$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.14, p = 0.71$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.75, p = 0.40$ )。この結果から、提示法の相違や複雑性の高低において、タスクの成功の達成度合いに差がないことが示唆された。有意差はないが、長応答型提示で複雑性が高い場合に、最も高いスコアとなっている。

Effort について、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.59, p = 0.46$ )。また、複雑性の主効果に有意差が見られなかった ( $F(1, 12) = 2.75, p = 0.12$ )。提示法と複雑性の交互作用効果に有意差は見られなかつ

た ( $F(1, 12) = 0.17, p = 0.68$ )。この結果から、提示法の相違や複雑性が高いタスクは、タスク遂行のために投入した精神的及び身体的努力の負荷に差がないことが示唆された。有意差はないが、短応答型提示で複雑性が低い場合に、最も低いスコアとなっている。

Frustration について、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 1.61, p = 0.23$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.01, p = 0.92$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.04, p = 0.84$ )。この結果から、提示法の相違や複雑性の高低において、タスク遂行中に感じたストレスや苛立ち、不安の度合いに差がないことが示唆された。有意差はないが、短応答型提示で複雑性が低い場合に、最も低いスコアとなっている。

Overall Score について、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 1.32, p = 0.27$ )。また、複雑性の主効果に有意差が見られなかった ( $F(1, 12) = 3.38, p = 0.09$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.11, p = 0.75$ )。この結果から、提示法の相違や複雑性の高低において、タスク遂行中に感じたストレスや苛立ち、不安の度合いに差がないことが示唆された。有意差はないが、短応答型提示で複雑性が低い場合に、最も低いスコアとなっている。

### 4.1.2 検索の満足度

アンケートは、情報獲得という点において、応答の長さをユーザがどのように主観的に評価しているかを明らかにすることである。表 3 に各システムのアンケート結果を示す。

Q1 では、ユーザが情報検索時にシステムの応答の長さがユーザの情報検索行動に与える影響を調査した。二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.71, p = 0.42$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.32, p = 0.58$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.32, p = 0.58$ )。有意差はないが、短応答型提示で複雑性が低い場合に、最も高いスコアとなっている。

Q2 では、情報検索時に、応答の長さによって、ユーザが様々な情報を収集できていると感じたかどうかを調査した。二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.16, p = 0.70$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.55, p = 0.47$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.26, p = 0.62$ )。有意差はないが、短応答型提示で複雑性が低い場合に、最も高いスコアとなっている。

Q3 では、ユーザが様々な情報を得るためにシステムを有用と感じるかどうかを調査した。二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.65, p = 0.44$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.65, p = 0.44$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.07, p = 0.79$ )。有意差はないが、長応答型提示で複雑性が高い場合に、最も高いス

表 2: 長応答型提示（ベースライン）と短応答型提示（提案手法）の NASA-TLX の平均値の比較。括弧内は標準偏差。スコアは加重スコアとなっている。低いほど優れていることを示す。ただし、パフォーマンスは高いほど優れている。最高値は太文字で示す。

		Mental	Physical	Temporal	Performance	Effort	Frustration	Overall score
長応答型提示	複雑性低	202.00(170.68)	4.00(8.94)	45.00(42.28)	127.00(105.10)	108.00(46.04)	131.00(181.60)	41.13(23.99)
	複雑性高	323.33(100.17)	<b>3.33(5.77)</b>	141.67(96.74)	96.67(40.41)	158.33(144.42)	110.00(186.21)	55.55(15.06)
短応答型提示	複雑性低	<b>175.00(99.87)</b>	5.00(8.66)	<b>10.00(10.00)</b>	123.33(153.32)	<b>60.00(36.06)</b>	<b>30.00(20.00)</b>	<b>27.00(3.46)</b>
	複雑性高	286.00(128.13)	7.00(10.95)	41.00(48.53)	<b>201.00(141.35)</b>	144.00(72.41)	37.00(48.43)	47.73(18.28)

表 3: タスク後のアンケート結果の平均。括弧内は標準偏差。最高値は太文字で示す。

質問事項	長応答型提示	短応答型提示	長応答型提示	短応答型提示
	複雑性低	複雑性低	複雑性高	複雑性高
Q1 この音声対話型検索システムの応答の長さは適切だと感じた。	3.00(1.58)	<b>4.00(1.00)</b>	3.00(1.73)	3.20(1.10)
Q2 この音声対話型検索システムの適切な応答の長さが、情報獲得するうえで役に立ったと感じた。	3.80(1.10)	<b>4.00(0.58)</b>	3.67(1.53)	3.60(1.14)
Q3 この音声対話型検索システムは様々な情報を獲得するうえで役に立ったと感じた。	4.00(1.22)	3.67(2.31)	<b>4.67(0.58)</b>	4.00(0.00)
Q4 意見をまとめるうえで必要な情報が十分に集められたと感じた。	<b>4.40(0.55)</b>	4.00(1.73)	3.67(1.53)	4.00(0.00)

表 4: クエリの平均発行回数。括弧内は標準偏差。

	複雑性低	複雑性高
長応答型提示	3.4(1.67)	7.0(3.00)
短応答型提示	6.3(5.86)	8.6(4.28)

表 5: クエリの発行に要した時間（秒）。括弧内は標準偏差。

	複雑性低	複雑性高
長応答型提示	19.22(16.99)	6.09(1.63)
短応答型提示	5.94(1.13)	13.63(6.85)

コアとなっている。

Q4 では、ユーザが自分の意見を要約するのに十分な情報を収集できたかどうかを調べた。二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.00, p = 0.95$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.51, p = 0.49$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.51, p = 0.49$ )。有意差はないが、長応答型提示で複雑性が低い場合に、最も高いスコアとなっている。

## 4.2 RQ2: タスクの複雑性と応答の提示方法の組み合わせは、ユーザの検索行動や獲得した情報にどのような違いを生み出すのか

### 4.2.1 ユーザの検索行動

**クエリ発行回数:** 短応答提示だと、十分な情報が一度のクエリで獲得できないことが想定されるため、クエリ発行回数が増加することが期待される。表 4 は、長応答型提示（ベースライン）と短応答型提示（提案手法）、そして、複雑性の高低を含めたクエリ発行回数の平均値を示したものである。

クエリの実行回数について、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 1.35, p = 0.27$ )。

また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 2.26, p = 0.16$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.12, p = 0.74$ )。

この結果から、提示法の相違や複雑性の高低において、クエリの実行回数に差がないことが示唆される。有意差はないものの、短応答型提示は、タスクの複雑性とは関係なく、クエリ発行回数が多い。また、複雑性の低いタスクはクエリ発行回数が多い傾向が見られる。

**クエリ発行に要した時間:** ユーザがクエリ発行に要する時間について、タスクの複雑性の高低に関わらず、短応答型提示の場合は、一度に得られる情報が少ないため、次のクエリの手がかりを得ることができずに、クエリ発行に要する時間が短くなることが想定される。表 5 は、長応答型提示（ベースライン）と短応答型提示（提案手法）、そして、複雑性の高低を含めたクエリ発行に要した時間（秒）の平均値を示したものである。

ユーザがクエリ発行に要した時間について、二元配置分散分析の結果、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.27, p = 0.61$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.44, p = 0.52$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 3.61, p = 0.08$ )。

この結果から、提示法の相違や複雑性の高低において、クエリの実行に要した時間に差がないことが示唆される。有意差はないものの、複雑性が高いタスクにおいて、長応答型提示の方が、短応答型提示よりも、クエリ発行に要した時間が短い傾向が見られる。

### 4.2.2 ユーザの獲得した情報

**ナゲット再現率:** 短応答型提示は、認知的負荷を軽減させることが期待できるため、長応答型提示と比較すると、検索後のユーザのサマリ内容には、システムから得た情報をより多く含んでいることが想定される。したがって、サマリ内の情報の違いを評価するために、ナゲット再現率を評価指標として用いて比較することにした。ナゲット再現率とは、システムの応答を

表 6: システム応答のナゲットの例

ID	ナゲット
S-P-CL-001	シナモンの原産地はスリランカやインド南部です.
S-P-CL-002	シナモンは日本では生産されておらず
S-P-CL-003	シナモンはヨーロッパでは一般的に生産されていません.

表 7: ナゲット再現率. 括弧内は標準偏差.

	複雑性低	複雑性高
長応答型提示	0.27(0.10)	0.28(0.11)
短応答型提示	0.54(0.36)	0.64(0.13)

分割不可能な単位にしたナゲットの総数をシステム応答のナゲット数とし、そのナゲットの総数のうち、実験参加者のサマリに含まれるナゲット数の割合を表したものである。ナゲット再現率は以下のように定義する：

$$\text{ナゲット再現率} = \frac{\text{ユーザのサマリ内のナゲット数}}{\text{システム応答のナゲット数}} \quad (1)$$

ナゲット再現率を評価するために、ナゲットに基づいてシステムの応答とユーザのサマリを著者が手動で抽出した。ナゲットは、情報検索・アクセス [5] [8] [21] や質問応答システムの評価 [11] [17] [6] だけでなく、対話システム [22] [32] の評価にも用いられている。ナゲットとは、評価者が、ある応答にそのナゲットが含まれているかどうかを二値的に判断できる事実と定義されている [15]。つまり、分割不可能な単位であることが求められる。表 6 は、著者が作成したシステム応答のナゲットの例である。システム応答のナゲット数は、長応答型提示かつ複雑性が低いタスクでは 18.6 件、長応答型提示かつ複雑性が高いタスクでは 27.3 件、短応答型提示かつ複雑性が低いタスクでは 7.7 件、短応答型提示かつ複雑性が高いタスクでは 13.0 件であった。

表 7 は、長応答型提示（ベースライン）と短応答型提示（提案手法）、そして、複雑性の高低を含めたナゲット再現率の平均値を示したものである。

ナゲット再現率について、二元配置分散分析の結果、提示法の主効果に有意差が見られた ( $F(1, 12) = 11.6, p = 0.01$ )。一方で、複雑性の主効果には有意差は見られなかった ( $F(1, 12) = 0.35, p = 0.56$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.15, p = 0.70$ )。

この結果から、短応答型提示方法は、ユーザの認知的負荷が下がり、システムからの応答を長応答型提示よりも多く獲得できたことが示唆された。

**サマリの比較：**探索的検索は、学習や意思決定をする検索であるため、ユーザが獲得した情報を、正確にかつ網羅的に収集できていること、情報を意味のある形で理解していること、情報を鵜呑みにせず、適切に批判、分析していることを評価する必要がある。したがって、最終的なサマリの質について、表 8

のように、Dqual（要約の質）、Dintrp（事実の関連性と構成）、Dcrit（批判的分析）の 3 つのパラメーターを用いて、著者が手動で評価した [30]。表 9 は、長応答型提示（ベースライン）と短応答型提示（提案手法）、そして、複雑性の高低を含めた各パラメーターの平均値を示したものである。

Dqual（要約の質）について、二元配置分散分析の結果、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.19, p = 0.69$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.19, p = 0.69$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.02, p = 0.89$ )。

Dintrp（事実の関連性と構成）について、二元配置分散分析の結果、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 0.03, p = 0.85$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.59, p = 0.47$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.00, p = 1.00$ )。

Dcrit（批判的分析）について、二元配置分散分析の結果、二元配置分散分析の結果、提示法の主効果に有意差は見られなかった ( $F(1, 12) = 2.76, p = 0.12$ )。また、複雑性の主効果にも有意差は見られなかった ( $F(1, 12) = 0.80, p = 0.639$ )。提示法と複雑性の交互作用効果に有意差は見られなかった ( $F(1, 12) = 0.41, p = 0.53$ )。

これらの結果から、提示法の相違や複雑性の高低において、ユーザのサマリ内容の質に差がないことが示唆される。有意差はないものの、Dcrit（批判的分析）については、長応答型提示の方が、短応答型提示よりも、パラメーターの値が高い傾向が見られる。

## 5 議 論

本章では 2 つのリサーチクエスションそれぞれについて議論する。

**RQ1:** 短応答型提示が有効であるかを検証するために、認知的な負荷の軽減が見られたかどうかを NASA-TLX を指標として用いた。提示法の相違や複雑性の高低において、Mental, Physical, Performance, Effort, Frustration, Overall Score について、有意差は見られなかった。一方で、提示法と複雑性の高低のそれぞれにおいて、Temporal の項目についてのみ、有意差が見られた。このことから、短応答提示は、時間的なプレッシャーからの解放につながることを示唆された (4.1 節)。システムが短い応答をするため、長応答型提示よりも、ユーザ

表 8: サマリ比較の評価のパラメーター

パラメーター	定義
Dqual (要約の質)	事実の質を 0~3 の尺度で評価。0 は要約内の事実が不適切、または関連性がない。3 は要約内に具体的に関連性の高い事実が十分に含まれる。
Dintrap (事実の関連性と構成)	事実の関連性を 0~2 の尺度で評価。0 は事実が互いに関連付けられていない。2 はすべての事実が互いに関連付けられ、流れや構成が整っている。
Dcrit (批判的分析)	批評の質を 0~1 の尺度で評価。0 は事実が列挙されている。1 は、その価値についての考えや分析がある場合や、事実の長所と短所の両方が述べられている場合を表す。

表 9: タスク後のサマリの評価の平均。括弧内は標準偏差。

パラメーター	長応答型提示	短応答型提示	長応答型提示	短応答型提示
	複雑性低	複雑性低	複雑性高	複雑性高
Dqual (要約の質)	2.40(1.34)	2.67(0.58)	2.67(0.58)	2.80(0.48)
Dintrap (事実の関連性と構成)	1.40(0.89)	1.33(0.58)	1.67(0.58)	1.60(0.58)
Dcrit (批判的分析)	0.60(0.55)	0.33(0.58)	1.00(0.00)	0.40(0.55)

表 10: システムが音声を読み始めるまでに要した平均時間 (秒)

	複雑性低	複雑性高
	長応答型提示	43.0
短応答型提示	10.0	18.6

側に主導権が早く移行し、聞きたいと思ったことを即座に聞けるからだと考えられる。また、複雑性が低いタスクは、複雑性が高いタスクと比較すると、比較や分析をするために網羅的に検索する必要性は低い。そのため、限られた時間でも十分な情報を獲得できたため、時間的プレッシャーを感じにくかったことが想定される。

有意性が見られなかったものの、提案手法が、複雑性の高低に関わらず、Physicalを除いた項目において、最高スコアを記録している(表 2)。特に、Performanceでは、複雑性が高いタスクにおいて、短応答型提示が最も良いスコアであり、検索タスクがうまくできたと感じていることが示唆される。しかし、短応答型提示によって得られた情報が有用であったかどうか不明瞭であるため、続けて、ユーザ応答の長さや獲得できた情報のアンケートの結果を考察する(表 3)。

有意差が見られなかったものの、Q1とQ2から短応答型提示は、特に、複雑性が低いタスクにおいて、応答の長さの適切性や情報獲得において有用であると評価されている傾向にある。

Q3から、長応答型提示は、タスクの複雑性に関わらず、様々な情報を網羅的に獲得していく上では、有用であると捉えられている。一方、Q4から、特にタスクの複雑性が高い場合、意見をまとめる上で十分に情報を集められたというスコアが低い傾向にあることから、システムの応答の中で、必要だと思ったが、取りこぼしてしまった情報があると感じている可能性がある。

これらの結果を踏まえると、提案手法である短応答型提示は、

認知的負荷を下げることは有用であるが、特に、複雑性の高いタスクにおいては、ユーザは長応答型提示の方が多様な情報を獲得ができたと感じていることが示唆された。ただし、獲得した情報をまとめていく上では、認知的負荷が高いため、重要な情報を聞き逃した可能性がある。したがって、タスクが複雑な場合は、基本的には、短応答型提示でユーザのクエリに対応しつつも、様々な情報を与える機会を設けて、負荷を下げつつ、多様な情報獲得ができるように、タスクに応じたシステムの応答戦略を設計する必要性が示唆される。

**RQ2:** まずは、検索行動に関して議論する。有意性は見られなかったものの、短応答型提示は、タスクの複雑性に関わらず、クエリの発行回数が増える傾向が見られた(表 4)。これは、長応答型提示よりも、十分に情報を獲得したと感ずるまでのシステムが提示する情報量が少ないために、クエリ発行回数が増えたと考えられる。また、クエリ発行に要した時間について、有意差はないものの、顕著な数値であったのは、長応答型提示で、複雑性が低い場合、クエリ発行に要した時間がかかっていることと、短応答型提示で、複雑性が高い場合にクエリ発行に要した時間がかかっていることである(表 5)。これは、長応答型提示において、複雑性が低い場合においては、得られた情報がすでに十分であるから、さらに何を聞けば良いのかと、クエリが思い浮かばず時間がかかってしまったのに対して、複雑性が高い場合は、得られた情報から、さらに別観点の情報を獲得できたため、クエリが思いつきやすくなった結果、クエリ発行に要する時間が減少した可能性がある。反対に、短応答型提示で、複雑性が高い場合は、手がかりになる情報が少ないため、何を聞けば良いか分からず、クエリ発行に時間を要したことが考えられる。

続いて、結果的に得られた情報について議論する。ナゲット再現率は、タスクの複雑性に関わらず、短応答型提示が有意に

高いことが明らかになった(表7)。認知的負荷が下がり、結果として、ユーザのサマりにシステムからの情報が多く記載されたと考えられる。正確に情報を得るという点において、短応答提示が有効であることが示唆される。

さらに、得られた情報の質について不明瞭であるため、サマリの質を評価した(表9)。有意差はなかったものの、長応答型提示かつ複雑性が高いタスクでは、考察や比較、分析がサマりに記載されている傾向にある。短応答型提示が、「日本の一人あたりのアルコール消費量は年間でおよそ8リットル前後です。」といった事実のみのサマリが目立つのに対して、長応答型提示では、「一番アルコール消費量が多い県は、秋田県だそうで、文化的な背景もありながら、気候的に他の件に比べて比較的気温が低いことも影響しているのではないかと考える。」というように、事実に加えて、考察、分析が見られた。短応答型提示で、複雑性が高いタスクにおいて、システムは「日本の一人あたりの年間アルコール消費量は、約8リットル程度(純アルコール換算)です。」と返すのに対して、長応答型提示で、複雑性が高い場合のシステムの応答では、「日本人の一人あたりのアルコール消費量は、年間おおよそ8リットル程度とされています。(省略)また、年齢や性別によっても消費量は異なることがあります。」と応答している。この場合、年齢や性別による比較対象を得たことによって、短応答型提示よりも、より深い考察に至りやすかったと考えられる。この比較対象の得やすさが、サマリの質に影響している可能性がある。

以上の議論から、短応答提示型が、情報を確実に得られるという点で有効である一方で、複雑性が高いタスクにおいては、ベースラインの長応答型提示の方が、早期に比較の軸を複数獲得できることが、考察や分析を深める結果に寄与した可能性がある。このことから、タスクに合わせたシステムの応答戦略を実装することが求められる。複雑性の高い検索の場合は、早期において、結果として長応答型となるが、複数情報提示を行い、それ以降は、短応答をベースにしつつも、適宜、情報を提示していくことが望ましいと考える。

**限界:** 最後に、本研究の限界点について述べる。

まず、短応答型提示の方がナゲット再現率が高いが、長応答型提示は、不要な情報が多かったため、ユーザがサマりに含めなかった可能性がある。これに対しては、アンケート結果から、長応答型提示において、意見をまとめる上で情報を十分に集められなかったと感じている傾向が見られることから、重要な情報を取りこぼしたと感じている可能性がある(表3)。この結果から、長応答型提示で、不要な情報がなかったとも取れる。しかし、この仮説を明らかにするために、「サマリをして、後から必要だと思った情報はるか。」といったアンケートを取る必要がある。

次に、本研究で用いたシステムでは、技術的な制約上、表10に示した通り、複雑性が高く、かつ長応答型提示だと、システムが音声を読み上げ始めるまでに時間がかかっていた。これが、NASA-TLXのTemporalに過度に影響を与えている可能性がある。したがって、テキストから音声に変更する時間を極

力抑えるために、リアルタイムAPI<sup>7</sup>を用いて、実装する必要がある。

最後に、本システムは、途中でシステムからの発話を止めることができない。仮にシステムがユーザの意図しない応答をした場合、ユーザが途中でシステムの応答を止められないことは、過度にストレスを与えることにつながる可能性がある。従って、システムの応答中にユーザが介入できるように実装をする必要がある。

## 6 ま と め

本研究では、音声対話型検索において、システムの応答の長さがユーザの検索行動へ与える影響を調査した。本研究における研究の主な発見は2つある。1つ目は、短応答型提示は、タスクの複雑性に関わらず、認知的負荷を軽減させるため、音声対話型情報検索には有効ということである。特に、複雑性が高いタスクにおいて、長応答型提示の方が、様々な情報を獲得できるものの、認知的負荷が高いため、重要な情報を忘れてしまったと感じることが示唆された。2つ目は、短応答型提示は、システムからの情報を正確に獲得できるものの、分析や考察を深める検索になりにくいことが示唆された。ユーザの検索後のサマリ内容では、長応答型の方が、考察や分析の入っている傾向に見受けられた。これは、長応答型提示によって、単一の情報でなく、複数の情報を提示していることが要因となり、ユーザの検索の幅が広がったからだと考えられる。これらの発見を踏まえると、音声対話型検索において、基本的には短応答型提示が認知的負荷を下げるため、特に複雑なタスクにおいては、早期に長応答型提示によって複数の点から情報を提示することが、情報獲得においては有用な戦略であることが示唆された。ただし、システムの応答生成に時間がかかってしまったため、システムがリアルタイムで応答できるようにすると同時に、得られた結果をもとにして、システムの応答戦略を設計し、大規模実験を進めていきたいと考えている。

**謝辞** 本研究はJSPS科学研究費助成事業JP24K03228, JP22H03905, JP21H03554, JP21H03775による助成を受けたものです。ここに記して謝意を表します。

## 文 献

- [1] Adobe. State of voice. *Adobe Digital Insights 2019*, p. 7, 2019.
- [2] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 475–484, 2019.
- [3] James F. Allen, Curry I. Guinn, and Eric Horvitz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, Vol. 14, No. 5, pp. 14–23, 1999.
- [4] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom. *A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, 2001.

<sup>7</sup> : <https://openai.com/index/introducing-the-realtime-api/>

- [5] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 659 – 666. Association for Computing Machinery, 2008.
- [6] Hoa Trang Dang and Jimmy Lin. Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 768–775. Association for Computational Linguistics, 2007.
- [7] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, and Sylvain Daronnat. Investigating how conversational search agents affect user's behaviour, performance and search experience. In *Proceedings of the 2nd International Workshop on Conversational Approaches to Information Retrieval*, 2018.
- [8] Matthew Ekstrand-Abueg, Virgil Pavlu, Makoto Kato, Tetsuya Sakai, Takehiro Yamamoto, and Mayu Iwata. Exploring semi-automatic nugget extraction for japanese one click access evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 749 – 752. Association for Computing Machinery, 2013.
- [9] Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. Keep it short: A comparison of voice assistants' response behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2022.
- [10] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Human Mental Workload*. 1988.
- [11] Makoto P. Kato, Virgil Pavlu, Matthew Ekstrand-Abueg, Tetsuya Sakai, Takehiro Yamamoto, and Mayu Iwata. Overview of the ntcir-10 1click-2 task. *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-10)*, 2013.
- [12] Abhishek Kaushik and Gareth Jones. Examining the potential for conversational exploratory search using a smart speaker digital assistant. In *VISIGRAPP*, pp. 305–317, 2023.
- [13] Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, Vol. 3, No. 1–2, pp. 1–224, 2009.
- [14] Jennifer Lai, Clare-Marie Karat, and Nicole Yankelovich. Conversational speech interfaces and technologies. In *Human-Computer Interaction*, pp. 71–82. CRC Press, 2009.
- [15] Jimmy Lin and Dina Demner-Fushman. Will pyramids built of nuggets topple over? In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 383–390. Association for Computational Linguistics, 2006.
- [16] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, Vol. 49, No. 4, pp. 41–46, 2006.
- [17] Teruko Mitamura, Hideki Shima, Tetsuya Sakai, N. Kando, Tatsunori Mori, Koichi Takeda, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, and Cheng-Wei Lee. Overview of the ntcir-8 aclia tasks: Advanced cross-lingual information access. In *NTCIR Conference on Evaluation of Information Access Technologies*, 2010.
- [18] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pp. 117–126, 2017.
- [19] Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2737–2746, 2018.
- [20] Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. Conversational query understanding using sequence to sequence modeling. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1715–1724, 2018.
- [21] Tetsuya Sakai, Makoto P. Kato, and Young-In Song. Click the search button and be happy: Evaluating direct and immediate information access. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, p. 621 – 630. Association for Computing Machinery, 2011.
- [22] Sakai Tetsuya, Zeng Zhaohao, and Luo Cheng. Evaluating helpdesk dialogues: Initial considerations from an information access perspective. *IPSJ Technical Report 2016-NL-228*, No. 1, 2016.
- [23] Johanne R. Trippas. Spoken conversational search: Audio-only interactive information retrieval. *ACM SIGIR Forum*, Vol. 53, No. 2, pp. 106–107, 2021.
- [24] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*, pp. 32–41, 2018.
- [25] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*, pp. 325–328, 2017.
- [26] Johanne R. Trippas, Damiano Spina, Mark Sanderson, and Lawrence Cavedon. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 991–994, 2015.
- [27] Alexandra Vtyurina, Charles L. A. Clarke, Edith Law, Johanne R. Trippas, and Horatiu Bota. A mixed-method analysis of text and audio search interfaces with varying task complexity. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pp. 61–68, 2020.
- [28] Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jing Zhou, Yufei Wang, and Kam-Fai Wong. A survey of the evolution of language model-based dialogue systems, 2023.
- [29] Ryen W. White and Resa A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan and Claypool Publishers, 2009.
- [30] Mathew J. Wilson and Max L. Wilson. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology*, Vol. 64, No. 2, p. 291 – 306, 2013.
- [31] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pp. 418–428, 2020.
- [32] Zhaohao Zeng, Cheng Luo, Lifeng Shang, Hang Li, and Tetsuya Sakai. Towards automatic evaluation of customer-helpdesk dialogues. *Journal of Information Processing*, Vol. 26, pp. 768–778, 2018.