原著論文
Original Paper

エージェントが動画を紹介する話のタイミングの違いによるユーザの動画閲覧の評価

User Evaluation of Video Viewing Based on Timing Differences in Agent Video Recommendations

本居 恒輝
Koki Motoi
慶應義塾大学理工学部
Faculty of Science and Technology, Keio University
motoi@ailab.ics.keio.ac.jp

高山 直之
Naoyuki Takayama
慶應義塾大学大学院理工学研究科
Faculty of Science and Technology, Keio University
takayama@ailab.ics.keio.ac.jp

三宅 芹奈
Serina Miyake
（同　　上）
miyake@ailab.ics.keio.ac.jp

今井 倫太
Michita Imai
慶應義塾大学理工学部
Faculty of Science and Technology, Keio University
michita@ailab.ics.keio.ac.jp, https://www.ailab.ics.keio.ac.jp/webpage_personal/michita/

keywords: timing, agent, evaluation

Summary

This paper investigates how the timing of an agent's speech when introducing events in a video affects user engagement and comprehension, aiming to optimize video viewing experiences. With the rapid expansion of online video usage in education, entertainment, and information dissemination, efficiently extracting relevant information while maintaining concentration remains a challenge.

Previous research has explored automatic generation of contextual explanations by integrating visual and audio information, as well as interactive agents that supplement video content through conversation. However, there has been limited evaluation of how the timing of an agent's explanations influences user understanding and interest.

To address this gap, this study proposes a system where an agent introduces video events at three different timing conditions: (1) delivering explanations continuously before the video starts, (2) providing explanations while users are watching the relevant video segment, and (3) speaking after a certain period of inactivity during video viewing. The study evaluates how each timing affects users ' comprehension and interest.

An evaluation experiment was conducted to examine the impact of these timing variations. Participants were instructed to find specific sections in the video corresponding to the provided explanations. The results indicated that the most effective timing was when the agent spoke during user inactivity, leading to increased comprehension and engagement. This suggests that strategic timing of verbal explanations can enhance the effectiveness of video-based information delivery.

1. は　じ　め　に

動画コンテンツの普及に伴い，閲覧者が効率的に映像の内容を理解し，適切な情報を得るための支援技術が求められている．特に，教育，情報提供，エンターテインメントの分野では，閲覧体験を向上させるための補助的な仕組みが必要とされている．人工知能を活用した対話型エージェントが注目され，ユーザに適切なタイミングで情報を提供することで，理解の促進や関心の維持を支援する試みが行われている．しかし，エージェントが動画の説明を行う際の発話のタイミングが閲覧行動に与える影響については明らかになっていない．

従来の研究では，動画説明生成技術やエージェントによる対話システムの開発が進められ，閲覧者の理解を補助するための情報提示手法が検討されてきた．例えばVideoBERT[Chen Sun et al. 19] や Dense Video Captioning[Luowei Zhou et al. 18] などの技術により，映像の内容を自動で解析し，適切な説明文を生成することが可能になった．また，対話型エージェントを活用した研究 [Kuang et al. 24] では，閲覧者の行動や感情に応じて情報を提供するプロアクティブ AI が開発され，ユーザーエクスペリエンスの向上が図られている．しかし，動画説明生成技術とエージェントによる対話システムの技術が統合されたときに，どのようなタイミングでエージェントが介入するのが最も効果的であるのかについての研究はやられていない．

本論文の優れているところは，エージェントの動作タイミングが閲覧体験に与える影響を定量的に評価できる点である．具体的には，エージェントによる説明を動画再生前にまとめて提供する方式，閲覧中に特定の場面に応じて発話する方式，閲覧者が一定時間操作を行わないタイミングで発話する方式の三つの条件について比較することができる．評価実験を通じて，それぞれの条件が閲覧者の理解度や興味に与える影響を測定し，より効果的な動画閲覧支援の方法を探る．

加えて，本論文は，エージェントの存在が閲覧体験にどのように影響を与えるかを探ることで，今後の研究における設計指針を示すことが期待される．閲覧体験を向上するためには，エージェントの動作タイミングや動画説明生成などの技術の一つだけが機能するだけでは不十分であり，適切な動作タイミングと適切な説明生成が必要である．本論文を通じて，閲覧体験の向上とエージェントの役割の重要性が明らかになることを目指している．

2. 従来の手法の問題点

2·1 動画説明生成

動画の内容を自然言語で説明する技術は，動画を閲覧する人が内容を理解するための重要な手段として注目されている．この分野では，マルチモーダルデータを活用したモデルが多く提案されており，映像や音声を統合的に解析し，より自然な説明文を生成する手法が発展している．

Sun らは，映像と音声を統合して学習するモデル「VideoBERT」を提案した.「VideoBERT」は，自然言語処理の進歩を活用し，動画フレームや音声から文脈に基づいた説明文を生成することに成功しており，動画と言語の統合的理解を目指す技術の基盤を築いた [Chen Sun 19]. 一方で，Zhou らは，動画内の個別イベントを検出し，それぞれに詳細な説明を付与する「Dense Video Captioning」の技術を発展させた．これにより，動画全体ではなく特定のシーンに焦点を当てた説明生成が可能となり，閲覧者の注意を引きやすくなることが示された [Luowei Zhou 18].

これらの技術は，単に動画全体を要約するだけでなく，特定のシーンやイベントに応じた詳細なキャプション生成を可能にし，動画理解の精度向上に寄与している．

2·2 エージェントが与える影響

エージェントの動作タイミングや感情表現，共同視聴，さらにはプロアクティブ AI の活用が，ユーザ体験や対話の質にどのような影響を与えるかについて，多くの研究が行われている．

まず，エージェントの動作タイミングに関する研究では，小林らが，エージェント集団のわずかな動作タイミングの違いがユーザに与える話しやすさの印象や雰囲気にどのような影響を及ぼすかを明らかにした [小林 14]. 具体的には，音声対話において，エージェントがユーザの発言に対して相槌を打つ，または頷くタイミングを制御することで，エージェント集団の雰囲気の良し悪しが変わることを示した．また，貴志らは，対話エージェントの頷きのタイミングがユーザの発話の長さに影響を与えることを分析し，適切なタイミングで頷くことでユーザの発話が長くなり，対話の質が向上することを示している [貴志 11].

エージェントの感情表現に関しては，西野らが，歌唱音楽の心象風景映像における前景エージェントが視聴体験に与える影響を調査した．研究の結果，エージェントの感情表現が視聴者の没入感や感情移入を高めることが明らかになった．これは，エージェントの存在が視聴者の感情的な関与を促進し，より深い体験を生み出す可能性を示している [西野 24].

さらに，エージェントの共同視聴による影響については，斉藤らが，エージェントとの共同視聴がユーザの孤独感の軽減に寄与する可能性を示した [斉藤 22]．エージェントが共に視聴することで，視聴体験を他者と共有している感覚を生み出し，孤独感を和らげる効果があることが報告されている．

また，プロアクティブ AI によるユーザビリティ評価支援については，Kuang らが，UX 評価プロセスにおいて AI アシスタントの役割を調査した [Kuang 24]．彼らの研究では，AI による自動的な提案のタイミングが評価者の分析パフォーマンスや主観的な効率性に与える影響を評価した結果，AI による「事後提案」が最も高い信頼性と効率性を示すことが明らかになった．このことから，適切なタイミングでの AI の介入が，評価作業の質を向上させる可能性が示唆されている．

これらの研究を総合すると，エージェントの動作タイミングや感情表現が対話の質や視聴体験に影響を与え，共同視聴によって孤独感の軽減が期待できること，さらにはプロアクティブ AI の適切な介入が UX 評価の信頼性と効率性を高めることが明らかになっている．

2·3 既存研究の問題点

既存研究では，動画説明生成，エージェントのタイミング，感情表現といった個別の要素が視聴体験や視聴行動に与える影響について，それぞれ重要な知見が得られている．しかし，動画説明生成，エージェントのタイミング，感情表現要素などを統合的に評価するアプローチが存在していない．視聴行動を促進するためには，動画説明生成，タイミング，感情表現といった要素が個別に有意な影響を与えるだけでは不十分であり，複数の要素を統合した設計指針を明らかにすることが，今後の研究における重要な課題である．

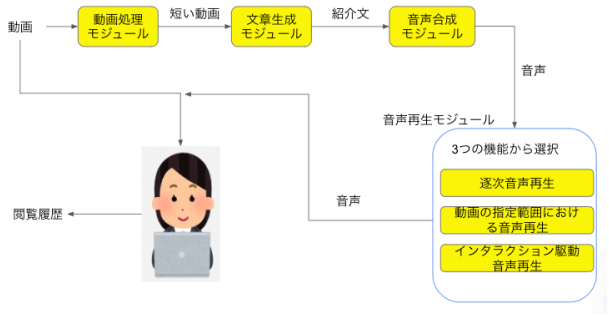

図 1 システム構成図

3. 今回の改良点

本論文では，エージェントが動画の説明を行う際の発話のタイミングが閲覧行動に与える影響について明らかにするために，エージェントが動画の出来事を紹介する文を 3 種類のタイミングで切り替えて発話するシステムを提案した．提案手法により動画説明生成とタイミングを統合したシステムが閲覧行動に与える影響を評価することができる．発話のタイミングは，開始時連続発話，閲覧中発話，非操作時発話の 3 種類である．図 1 にシステム構成図を示す．動画内の一部分を抽出し自然言語モデルを用いて紹介文を生成し音声合成し音声再生モジュールの中から一つを選択して音声を再生する．動画閲覧システムでは閲覧履歴を保存できるようになっており，ユーザは発話内容に該当する部分を動画から探す．

3·1 出来事紹介文生成

システムの前処理として 1 時間ほどの動画から 30 秒ほどの動画を 3 つ抽出しそれぞれの紹介文を生成する．

§ 1 動画処理モジュール

動画を GPT で処理できるようにするため OpenCV を用いて，動画のフレームを FPS(Frames Per Second) に基づき，2 秒に 1 フレームの間隔でフレームをスキップしながら，画像データを取得する．取得した画像は JPEG 形式でエンコードされ，Base64 形式に変換することで，外部 API とのデータ交換が容易になる．

§ 2 文章生成モジュール

OpenAI の GPT-4 [OpenAI 23]API を使用して，以下のような設定とプロンプトを GPT の入力とすることで，説明文を出力した．マルチモーダルの処理に高い性能を発揮するため GPT-4o を使用した．実際に使用した GPT のプロンプトを図 2 に示す．

- フレーム画像の Base64 データを入力として活用
- 生成される文章の内容とトーンをシステムプロンプトで制御．
- メッセージ長を 150 字以内に制限．

```
response = client.chat.completions.create(
    model=MODEL,
    messages=[
    {
        "role": "system",
        "content": (
            "文章は「この動画で起きていたことを説明するね」から始めて下さい"
            "あなたは、動画の内容を説明する興味を引く話を作る役割を担っています。"
            "動画内に登場するものを上げながら話を作成して下さい"
            "文章は日本語で、軽快かつ親しみやすいもので友達に話しかけるような文章にしてください。"
            "生成する文章は日本語で150字以内にしてください。"
        )

    },
    {
    "role": "user",
    "content": [
        "この動画の内容を理解して興味を引く体験談を書いてください。",
        *map(lambda x: {"type": "image_url",
                        "image_url": {"url": f'{x}', "detail":
"low"}}, base64Frames)
        ],
    }

    ],
    temperature=0,
)
```

図 2 GPT のプロンプト

図 3 動画閲覧システム

§ 3 音声合成モジュール

VOICEVOX を用いて生成された文章を合成音声に変換する．音声合成は以下の手順で実行される．

- 音声合成クエリを生成.
- 指定された話者 ID で音声データを合成.
- Unity で使用するために，WAV 形式で音声ファイルを保存.

3·2 動画閲覧システム

Unity を用いて動画再生をインタラクティブに制御し，ユーザの操作履歴を記録・保存するシステムを開発した．図 1 に動画閲覧システムを示す．本システムは，再生，停止，一時停止，シーク操作などの基本的な動画制御機能に加え，ユーザーが行った操作の履歴を時系列データとして保存する．これにより本論文の目的であるエージェントによる動画内で起きた出来事を紹介する話を話すタイミングの違いによってユーザの動画の閲覧の仕方にどのような影響を与えるかを分析することを可能にする．

§ 1 動画制御モジュール

動画の再生，停止，一時停止は Unity の VideoPlayer コンポーネントを利用して実現した．動画 URL はイン

スペクターで設定可能であり，ストリーミング動画にも対応している．以下のようなイベントが設定されている．

- Play ボタン：動画再生を開始
- Pause ボタン：再生中の動画を一時停止
- Stop ボタン：再生を停止し，履歴データを保存
- スライダー：ユーザがスライダーを移動させた際に動画の再生位置を変更
- 再生時間表示：再生時間を「分: 秒」のフォーマットで表示

§ 2 操作履歴記録モジュール

動画の閲覧履歴を収集するためにスライダーの値とシステムの開始時刻からの経過時間を記録する．履歴の記録は以下の 2 つのタイミングで行われる．

- スライダーが操作されたとき．
- 再生中に一秒間隔ごと．

記録された履歴は Stop ボタンが押された際に CSV 形式のテキストファイルとして保存される．

§ 3 動画紹介音声再生モジュール

本研究では，3 つの動画紹介音声の再生されるタイミングの違いがユーザーの動画閲覧にもたらす影響を調べるため 3 つの音声再生モジュールを作成した．

i. 逐次音声再生

このモジュールは実験条件 1 で使用する．音声リストに含まれている音声を順番に再生し，全ての音声が終了するまで自動的に次の音声を再生する機能を提供する．

ii. 動画の指定範囲における音声再生

このモジュールは実験条件 2 で使用する．指定された再生範囲に到達した際に，対応する音声クリップを再生する．再生後には，音声が重複して再生されないよう制御する仕組みを備えており，複数の音声範囲を効率的に管理する設計となっている．

iii. インタラクション駆動型音声再生

このモジュールは実験条件 3 で使用する．動画再生中にユーザーインタラクションの状況を監視し，一定時間操作がない場合に音声を再生する．

4. 実　　験

4·1 実 験 概 要

エージェントが動画の一部分を音声で紹介することで，ユーザーの動画閲覧に対する影響を評価する．異なるタイミングでエージェントに 3 つの紹介文を発話させて，ユーザの動画閲覧を比較した．

§ 1 実験準備

夜の渋谷を散策している youtube 動画 [YouTube 23] から 1 時間の動画を 3 つ切り取った．各 1 時間の動画から 30 秒の動画を 3 つ抽出して GPT でそれぞれの動画について紹介文を生成して音声合成を行い Unity 上に保存した．

§ 2 実験条件

3 つの条件で実験を行なった．

- ユーザが動画閲覧を始める前に 3 つの紹介文を連続して話す．
- ユーザが動画閲覧を始める前に 1 つの紹介文を流し，その後ユーザが紹介文に該当する動画の箇所を閲覧しているときに別の紹介文を話す．
- ユーザが動画閲覧を始める前に 1 つの紹介文を流し，その後ユーザが一定時間画面操作がない場合別の紹介文を話す．

§ 3 実験参加者

実験には，慶應義塾大学理工学部の 3，4 年生 10 人が参加した．実験参加者は男性が 8 人，女性が 2 人，平均年齢は 21.5 歳だった．

§ 4 実験手順

参加者には，エージェントによる動画の一部分の出来事について紹介する話を聞き，該当する箇所を 1 時間ほどの動画の中からスクロールバーを動かして探し出し，該当すると思われる箇所を閲覧するように指示した．参加者は動画とエージェントが紹介文を話すタイミングを変えた 3 つの条件で実験に参加し，実験の時間は最大 8 分間とした．エージェントによる発話は各条件で 3 回行われる．3 つの条件は全ての参加者で同じ順番で行い，動画と条件のペアは全ての被験者で同じ状態にした．8 分間経過する前に参加者が動画の閲覧をやめてしまった場合にはその時点で実験終了とした．一つの実験が終了するごとに実験に関するアンケートを行なった，3 つの条件は全ての参加者で同じ順番で行い，動画と条件のペアは全ての被験者で同じ状態で行った．3 つの条件の実験が全て終了した後，最終アンケートと動画アンケートを実施した．実験中のユーザの動画閲覧はスクロールバーの値で記録した．

§ 5 評価方法

参加者はそれぞれの条件の実験後に行なったアンケート内でエージェントが話した内容を覚えている範囲で記述してもらい，どのくらい興味深かったかを 5 段階でのリッカート尺度で取った．またシステムの使いやすさを評価するためにシステムユーザビリティスケール（SUS)[Brooke 96] を行なった．SUS は 5 段階でのリッカート尺度を使用して回答する 10 個の質問で構成されている．実験終了後には最終アンケートと動画アンケートを行った．最終アンケートでは，どの実験条件が一番エージェントの音声に該当する部分を探しやすかったかとその理由を聞き，動画アンケートでは実験に使用した音声の文章と元となった動画がどの程度一致しているか 10 段階でのリッカート尺度で取った．

4·2 実 験 結 果

§ 1 各実験後のアンケート

エージェントによる動画の紹介文の音声がどれだけ興味深かったかについてアンケートをとった結果の平均を以下の図 4 に示す．各音声における興味深さのアンケー

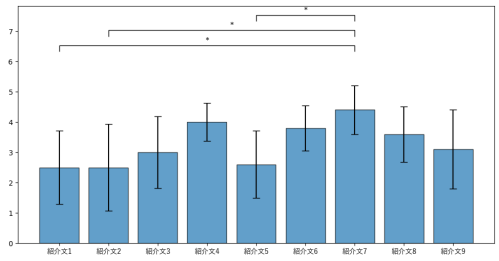

図 4 興味深さに関する結果

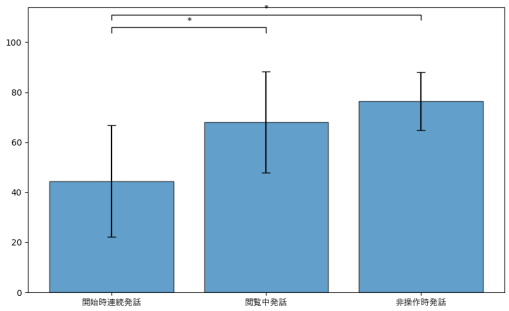

図 5 SUS の結果

ト結果について，$\alpha = 0.05$ とした ANOVA 検定を行なった．その結果 $p = 0.00065 \leq 0.05$ となり少なくとも 1 つのグループ間で有意差があると判断された．どのグループ間に優位差があるのか特定するために Tukey の HSD 検定を行なった．その結果，紹介文 7 と紹介文 1,2,5 の間に有意差があると判断された．

§ 2 SUS

SUS[Brooke 96] スコアと呼ばれる定量的な尺度を生成した．SUS の調査項目 10 項目の設問は，計算のために奇数設問と偶数設問に分かれている．奇数番号の設問の全ポイントの合計から 5 ポイント引いたものを x，25 から偶数番号の設問の全ポイント合計を引いたものを y として SUS スコアは $(x + y) \cdot 2.5$ で求められる．各実験条件で計算し，平均を取った結果を図 5 に示す．

また，各実験条件における SUS スコアに対し，$\alpha = 0.05$ とした ANOVA 検定を行なった．その結果，$p = 0.0034 \leq 0.05$ となり少なくとも 1 つのグループ間で有意差があると判断された．どのグループ間に優位差があるのか特定するために Tukey の HSD 検定を行なった．その結果，実験条件 1 と実験条件 2 間，実験条件 1 と実験条件 3 間には有意差があり，実験条件 2 と実験条件 3 間には有意差がないと判断された．

§ 3 動画の閲覧データ

記録したスクロールバーの値から各実験条件でのそれぞれの被験者の閲覧の様子をグラフにした．赤い帯状の部分は各紹介文の元となった部分を表している．以下の

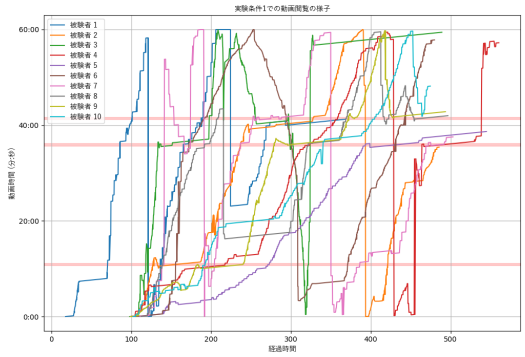

図 6 実験条件 1 の結果

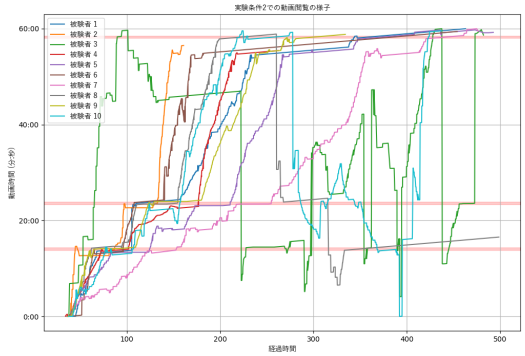

図 7 実験条件 2 の結果

図 6，図 7，図 8 に示す．すべての紹介文の該当する箇所を閲覧できた被験者は実験条件 1 では 0 人，実験条件 2 では 7 人，実験条件 3 では 9 人という結果となった．紹介文の該当する箇所を閲覧できた数の平均を図 9 に示す．紹介文の該当する箇所を閲覧できた数の平均は実験条件 1 では 1.6 個，実験条件 2 では 2.6 個，実験条件 3 では 2.9 個であった．すべての紹介文の該当する箇所を閲覧するのにかかった時間は，見つけられなかった場合を最大時間である 480 秒として平均値を出すと実験条件 2 は 357.5 秒，実験条件 3 は 318.5 秒であった．

また，実験条件 2 と実験条件 3 の全ての動画を閲覧するのにかかった時間について $\alpha = 0.05$ とした t 検定を行なった．$p = 0.172 \geq 0.05$ となり実験条件 2 と 3 で有意差はないと判断された．さらに，各実験場件における閲覧できた箇所の数について，$\alpha = 0.05$ とした ANOVA 検定を行なった．その結果 $p = 0.00038 \leq 0.05$ となり少なくとも 1 つのグループ間で有意差があると判断された．どのグループ間に優位差があるのか特定するために Tukey の HSD 検定を行なった．その結果，実験条件 1 と実験条件 2 間，実験条件 1 と実験条件 3 間には有意差があり，実験条件 2 と実験条件 3 間には有意差がないと判断された．

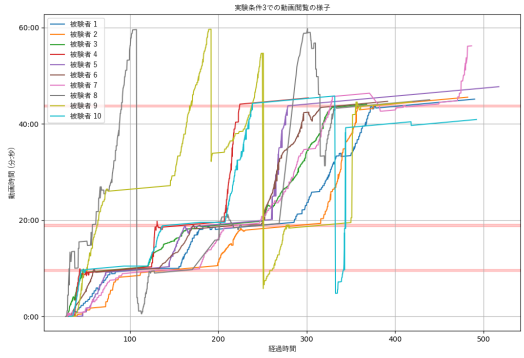

図 8　実験条件 3 の結果

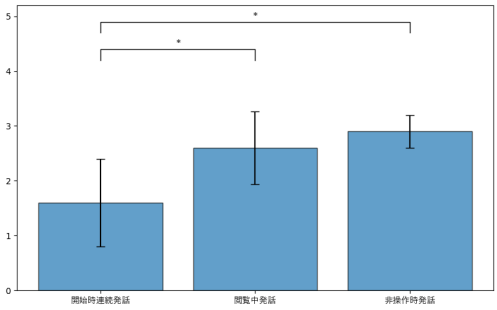

図 9　閲覧できた該当箇所の数

	条件 1	条件 2	条件 3
得票数	0	3	7

表 1　条件ごとの得票数

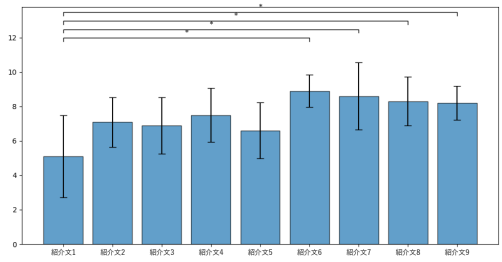

図 10　紹介文アンケートの結果

§4　最終アンケートの結果

どの条件が一番該当する部分を探しやすかったかのアンケートをとった結果を以下の表 1 に示す.

§5　紹介文アンケート

紹介文と実際の動画がどれほど一致しているかを評価し，平均を取った結果を以下の図 10 に示す．各紹介文と動画の一致アンケートの結果について，$\alpha = 0.05$ とした ANOVA 検定を行なった．その結果 $p = 0.00005 \leq 0.05$ となり少なくとも 1 つのグループ間で有意差があると判断された．どのグループ間に優位差があるのか特定するために Tukey の HSD 検定を行なった．その結果，紹介文 1 と紹介文 6,7,8,9 の間に有意差があると判断された.

4·3　考　　察

§1　システムについて

「どの条件が一番探しやすかったか」というアンケートの結果について考察を行う．まず，実験条件 1 の得票数が 0 だった理由として，最初に三つの紹介文をエージェントが話すことにより，何を言っていたかを覚えることができないということや，3 つの話していた内容が混ざってしまい内容を理解できていないまま動画を閲覧していることが原因として挙げられる．実際のアンケート結果でもエージェントが話した三つの話についてそれぞれ覚えている範囲で記述するように指示したが一つ目の話と二つ目の話が混ざってしまっている人や全く覚えていない人が数名見られた．また SUS スコア 0-100 のうち実験条件 1 の平均が 44.5 と非常に低い値であることからも実験条件 1 のシステムはユーザに対して非常に使用しづらいシステムであることがわかる.

次に実験条件 2 と実験条件 3 についてである．両条件ともエージェントによる紹介文が 1 つずつ紹介されるシステムである．SUS スコアでは両実験間に有意差はないと判断された．しかしながら最終アンケート結果では実験条件 3 の方が支持されており，ユーザにとって使用しや

すいシステムであったことがわかる．理由として実験条件 2 のシステムでは，該当する箇所を閲覧しているときに別の部分についての紹介文をエージェントが話し出してしまうため，動画の音声とエージェントの紹介音声が被ってしまい動画の閲覧を邪魔していることが挙げられる．一方で実験条件 3 のシステムでは一定時間スクロールバーを操作していない時にエージェントが別の紹介文について話し出すので，ユーザが動画閲覧に集中していない状況のため邪魔されたと感じず，エージェントの音声を聞き取りやすいと考えられる．

§2 紹介文について

まず，紹介文の音声がどれくらい興味深かったについてアンケートをとった結果，実験条件 1 で使用した話の平均が 2.7，実験条件 2 で使用した話の平均が 3.8，実験条件 3 で使用した話の平均が 3.7 であった．このように実験条件 1 の結果が他の実験条件に比べて低くなっているのは実験条件 1 の場合ではそれぞれの紹介文の内容について覚えていない場合があるため被験者が点数を低くつけたことが原因だと考えられる．またその他の理由としてスコアが低かった話 1,2,5 とその他の話を比べてみるとスコアが低い話では抽象的な話が多く具体的にどのようなシーンか想像できないことが原因であると考えられる．

次に，元となる動画と生成した紹介文がどれぐらい一致したかを評価したアンケートについて考察する．実験条件 1 では 3 つの話の内一致度の高い話ほど該当している箇所を閲覧している被験者の数が多かった．これは一致度が高い話ほどより具体的な情報を話しており，被験者の記憶に残りやすく該当する動画を探すのを容易にするからであると考えられる．

5. ま と め

本論文ではエージェントによる動画内で起きた出来事を紹介する 3 つの話を話すタイミングの違いによってユーザの動画の閲覧の仕方にどのような影響を与えるかを調査した．エージェントが動画の出来事を紹介する文を動画閲覧を始める前に連続して発話，紹介文に該当する動画の箇所を閲覧しているときに紹介文を発話，一定時間画面操作がないときに紹介文を発話の 3 種類のタイミングで切り替えて発話できるシステムを提案した．エージェントの話すタイミングの違いがユーザに与える影響を検証するため評価実験を行った．結果としてエージェントの発話タイミングが視聴体験に重要な影響を与えることが明らかになった．特に，ユーザが操作していない時に発話する条件が最も効果的であり，視聴者の理解度や興味を高めることが確認された．

謝 辞

本研究は，JST ,CREST,JPMJCR19A1 の支援を受けたものである．本研究を進めるにあたり，研究の機会及び貴重なご意見を頂きました，慶應義塾大学理工学部 今井 倫太 教授に深く感謝致します．

論文の査読をして頂き，細部にわたって御意見を頂きました理工学部研究科修士課程 1 年 高山 直之 氏 理工学部研究科修士課程 1 年 三宅 芹奈 氏に厚く御礼申し上げます．

実験に御協力頂いた被験者の方々に心より御礼申し上げます．

最後に日頃から御指導，御協力下さいました今井研究室の皆様に心より感謝いたします．

◇ 参 考 文 献 ◇

[Brooke 96] Brooke, J., et al.: SUS-A quick and dirty usability scale, *Usability evaluation in industry*, Vol. 189, No. 194, pp. 4–7 (1996)

[Chen Sun 19] Chen Sun, C. V. K. M., Austin Myers and Schmid, C.: VideoBERT: A Joint Model for Video and Language Representation Learning, *Computer Vision and Pattern Recognition* (2019)

[Kuang 24] Kuang, E., Li, M., Fan, M., and Shinohara, K.: Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing, in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16 (2024)

[Luowei Zhou 18] Luowei Zhou, J. J. C. R. S., Yingbo Zhou and Xiong, C.: End-to-End Dense Video Captioning with Masked Transformer, *Computer Vision and Pattern Recognition* (2018)

[OpenAI 23] OpenAI, : GPT-4 Technical Report, *arXiv preprint*, Vol. arXiv:2303.08774, (2023)

[YouTube 23] YouTube, : 【LIVE】”激しい喧嘩” リアルブレイキングダウン日韓戦 渋谷スクランブル交差点ライブカメラ クリスマス / クラブ街の報道・ニュース Shibuya Live Camera 12/23【LIVE】”激しい喧嘩” リアルブレイキングダウン日韓戦 渋谷スクランブル交差点ライブカメラ クリスマス / クラブ街の報道・ニュース Shibuya Live Camera 12/23 (2023)

[貴志 11] 貴志悠, 神田智子：対話エージェントのうなずきタイミングが発話長に及ぼす影響分析, HAI シンポジウム (2011)

[小林 14] 小林一樹, 湯浅将英, 片上大輔, 田中貴紘：わずかな動作タイミングの違いがつくるエージェントの雰囲気, 人工知能学会全国大会論文集 (2014)

[西野 24] 西野歩真, 石田真子, 米澤智子：歌唱音楽の心象風景映像への前景エージェントが音楽映像視聴体験に与える影響, HAI シンポジウム (2024)

[斉藤 22] 斉藤彰吾, 佐野睦夫, 大井翔：孤独感を軽減させる共同視聴エージェントの研究, 情報処理学会研究報告, Vol. HCI-197, No. 19, pp. 1–6 (2022)

〔担当委員：××○○〕

19YY 年 MM 月 DD 日 受理