拡張現実を用いた歩行型音声対話エージェント

前土佐勇仁 [1]　南泰浩 [1]
[1] 電気通信大学大学院　情報理工学研究科
yuto.maetosa@uec.ac.jp　minami.yasuhiro@is.uec.ac.jp

概要

本研究は，仮想エージェントと歩行しながら対話する音声対話システムを提案する．AR グラスなどの持ち運びのしやすい端末に実装し，仮想エージェントと自由に移動しながらの対話を目指す．対話文生成では仮想エージェントの発話内容，動作内容の順で生成する．対話中にユーザと仮想エージェントの歩行軌道を記録した結果，仮想エージェントがユーザと横並びで対話できる事が確認された．

1　はじめに

音声対話システムの発展に伴い，施設案内などのコンシェルジュを用途とした製品が実用化され，運用されている．音声対話システムで音声以外で内容を表現する場合，テキストのみで表現する手法やキャラクタ型の仮想エージェントで表現する手法がある．一般的に，仮想エージェントによる表現はテキストのみと比較して，ユーザの対話のしやすさや没入感が向上する効果がある [1]．

仮想エージェントによる表現手法には，ディスプレイ投影や仮想現実（VR），拡張現実（AR）などの手法がある．ディスプレイ投影式では，プロジェクタやテレビなどの映像機器にキャラクタと背景を出力し，ユーザと向かい合う形で対話する．また，VR では VRHMD をに映し出された仮想空間内で対話する一方で，AR では，現実世界に仮想エージェントを投影してユーザと対話する．ディスプレイには，まるで仮想エージェントが現実にいるような世界が投影される．従来より，これら 3 種の表現手法の音声対話システム [2][3][4] や表現手法 [5] が提案されてきたが，これらシステムは室内での利用が想定されており，また着席あるいは立ち止まって対話する．

そこで本研究では，拡張現実を用いて仮想エージェントと歩行しながら対話する音声対話システムを提案する．実際に人間同士が並んで対話する状況を再現するため，エージェントはユーザの真横あるいは前方で対話する．また，ユーザが移動した際のエージェントの歩行では，直進やカーブといったユーザの複雑な動きにも対応できる処理を構築する．

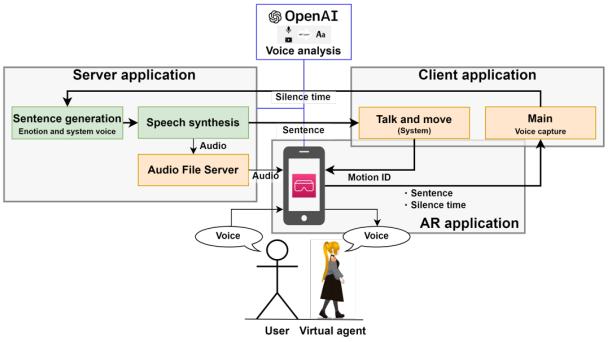

図 1　音声対話システム．

本研究により，仮想エージェントの行動範囲は疑似的に拡張され，エージェントは建物外での活動が期待される．また，歩行処理を加える事でユーザはエージェントが移動しても違和感ない対話が実現される．

2　提案システム

システム概要図を図 1 に示す．提案システムは，主にデスクトップパソコンで運用する Server application（以下，サーバ）と Client application（以下，クライアント），端末内で起動する AR application（以下，AR アプリ）で構成されている．AR アプリは，ユーザの発話内容の取得と，仮想エージェントの動作表出や発話処理を行う．クライアントでは，サーバで使用するユーザ・仮想エージェントの人物情報をユーザ側で設定する処理が組まれている．サーバでは，クライアントで取得したデータを基にシステムが話す文章（システム応答文）を生成し，合成音声ソフトで作成した音声データと共に AR アプリに送信する．

提案システムは一定の沈黙が発生した場合に仮想

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

エージェントが何らかの応答を行うため，ユーザが無言であっても仮想エージェントが半永久的に待機・停止しない．そのため，ユーザが「バイバイ・さよなら・さようなら」のいずれかの文章を発話する事で対話終了する．

2.1 対話文生成

対話文生成は，シナリオベース，LLM による通常の生成処理と拡張検索生成（RAG）で構成されている．

シナリオベースは，ユーザが無言だった場合や音声の認識不良（noise）の場合に適用される．この処理は，大規模言語処理モデルの対話文生成に先行して行われ，1 回目のみ行う．無言・認識不良時に出力されるユーザ文章は，シナリオファイルの中からランダムで 1 つ選択した文章をシステムの対話文とする．無言の場合は 30 通り，認識不良の場合は 10 通りから選択する．沈黙や認識不良が 2 ターン連続で発生すると大規模言語処理モデルの対話文生成で処理する．この時，無言時のユーザの発話内容は「……」として処理される．

使用した LLM は，5bit 量子化済みの「Aratako/calm3-22b-RP-v2[6]」である．クライアント側で設定した人物情報を基に，プロンプト文を構成する．また，ユーザと仮想エージェントの過去 3 ターンの対話内容を配置する．

あらすじ文は，直近 10 ターンに行われたユーザと仮想エージェントとの対話記録を基に，文の末尾に「- 序盤: 」を追加した．これにより，出力形式が「序盤・中盤・終盤」，あるいはそれに近い形式で生成される．

出力文の例を図 2 に示す．応答形式はセリフ，仮想エージェントの動作内容の順番で文章生成する．ストリーミング出力形式を採用し，動作内容の文章生成中はサーバ側で感情分析，合成音声処理を並列で実施する．

RAG は，LlamaIndex[7] と Gemini Flash 1.5 API[8] で構成されている．文章生成処理の開始と同時に，Gemini API による入力文の検索キーワード抽出の処理を実施する．図 3 のプロンプト文で検索クエリを出力し，クエリがない場合は「なし」と回答して通常の生成処理を続行する．検索クエリが見つかると，まず検索エンジン「Google」でクエリを検索し，Google サービス関連を除いた Web サイトを 3 個取得する．次に，これら Web サイトの文章をスクレイピングし，ベクトル化処理を実施する．同様に，ユーザ発話結果もベクトル化を行い，次に Llama Index によるベクトル検索を行い，ユーザ発話内容である質問内容に近い内容の Web サイト文章を 3 文抽出する．これら文章をプロンプトのメモリにコンテキスト文として追加し，対話文生成する．

2.2 仮想エージェントの歩行

仮想エージェントの移動アルゴリズムの概要を図 4 に示す．起動後，起動地点を原点とした 3 次元の拡張空間が生成される．端末を仮想エージェントのいる位置に向けると，画面内に仮想エージェントが表示される．ユーザが歩行すると，仮想エージェントはユーザの真横あるいは前方並ぶように歩行する．仮想エージェントの目標位置は，AR アプリの設定内容によって変化する．

t フレーム時，端末の AR アプリケーション上の座標値 $M_t = (M_{t,(x)}, M_{t,(z)})$ を取得する．なお，端末の移動と M_t の変化は連動する．U_t と M_t との距離 r を常に計測し，設定した閾値を超えると U_t の移動を開始する．r の閾値は 0.8[m] としたが，対話開始前にユーザがその場で回転する事で閾値を調整できる．

端末の移動方向 $\overrightarrow{M_{t-1}M_t}$ をユーザの移動方向 $\overrightarrow{U_{t-1}U_t}$ と推定し，ユーザの $(t+1)$ フレーム後の移動方向 θ を以下の通りに導出する．

$$\theta = \arctan\frac{z}{x} = \arctan\frac{U_{t,(x)} - U_{(t-1),(x)}}{U_{t,(z)} - U_{(t-1),(z)}} \tag{1}$$

ここから，ユーザの移動先 U_{t+1} を以下の通りに予測する．

$$U_{t+1} = (U_t^x + 0.4 \cdot \sin\theta, U_t^z + 0.4 \cdot \cos\theta) \tag{2}$$

A_{t+1}，A'_{t+1} は，U_{t+1} を極座標 $Z(x,y)$ に変換したのち，進行方向を起点に左右 90 度の位置に配置される．図 4 の通り，A_{t+1}，A'_{t+1} は，U_{t+1} の半径 r の円周上にある．以下は，A_{t+1} の導出方法である．

$$\begin{aligned} Z^{U_{Temp}}(x,y) &= Z(r \cdot \cos(\theta + \frac{\pi}{2}), r \cdot \sin(\theta + \frac{\pi}{2})) \\ Z^{M_{t+1}}(x,y) &= Z^{U_{Temp}}(x,y) \cdot Z(0,1) \\ A_{t+1} &= U_{t+1} + Z_{A_{t+1}} \end{aligned} \tag{3}$$

A'_{t+1} は，式 3 内の $Z(0,1)$ を $Z(0,-1)$ に変換して導出する．以上より，U_{t+1} と A_{t+1}，A'_{t+1} との距離 d，d' は以下の通りに導出する．

$$\begin{aligned} T &= A_t - U_{t+1} \\ d &= \sqrt{(T_x - Z_{A_{t+1,(x)}})^2 + (T_z - Z_{A_{t+1,(z)}})^2} \end{aligned} \tag{4}$$

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

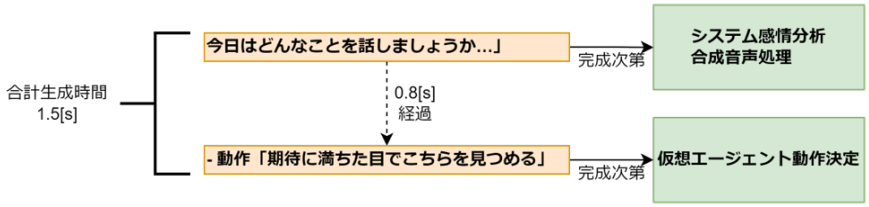

図 2　LLM による文章生成結果の例.

入力文から、ユーザーが何かを検索したいと考えているかどうかを判断し、以下の json 形式で検索クエリを作成せよ。
{"検索文": 作成した検索クエリ}

このタスクは検索エンジンの利用頻度を減らし、より効率的に情報にアクセスするために使用している

補足:
* 入力文の主要なキーワードが含まれている事
* 検索結果から、入力文に対する答えが直接的に得られる可能性が高い事
* 上記の条件を満たすクエリが作成できない場合、またはあなたの知識で入力文の内容に直接回答できる場合は、検索クエリを「なし」と回答

図 3　プロンプトのメモリ例（検索クエリ探索）.

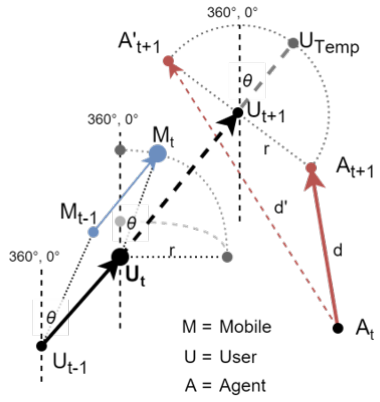

図 4　仮想エージェントの移動アルゴリズム.

図 5　疑似接触を伴う動作の例.

提案システムは，導出した (A_{t+1}, d)，(A'_{t+1}, d') を比較し，距離が近い点を移動先として判断し，仮想エージェントが移動し始める．

移動時の歩行動作は，移動アルゴリズム内の状態と図 4 内における U_t と A_t との距離 D_t を考慮して処理する．D_t が設定した r の範囲から離れると，仮想エージェントは歩行状態となり，通常の待機動作から歩行動作に移行する．$D_t > 1.65$ の場合，仮想エージェントは走行動作となり，$D_t < 1.5$ になるまでその動作を継続する．仮想エージェントの移動処理が終了すると，歩行動作も終了する．

歩行動作中にユーザが仮想エージェントの方に振り向き，端末のカメラ画角に収まった際は，仮想エージェントはユーザに顔をやや振り向き，目線を合わせる．また，応答文の発話と同時に動作をする際は，上半身のみ動作し，下半身は歩行動作を継続する．

2.3　動作表出

動作表出は，対話文生成時に生成される文章を基に決定する．まず，動作内容文をベクトル化し，表 1 のような表形式ファイル内の「MotionName」の全要素のベクトル値とのコサイン類似度を判定する．MotionName で判定する動作数は 71 種類である．コサイン類似度で最高スコアだった MotionName が 0.6 以上だった時，その MotionName と紐づけられている ARid が AR アプリの動作内容となる．動作の最高スコアが 0.6 未満の場合，40[%] の確率で 13 種類のランダム動作を指定する．それ以外の場合は，仮想エージェントの感情に関連する 11 種類の動作のいずれか 1 つ実行する．

没入感を高める工夫として，従来より仮想エージェントとの被接触を再現した報告がある [9]．本研究では，抱きしめるや撫でるなどのユーザに直接接触する動作は，図 5 のような疑似接触を伴う動作

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　動作内容と AR 用動作 ID 紐づけ表の例.

MotionName	TargetWords	ARid	TargetWords
大きく手を振る	手を.*振	54	o-i.vmd
落ち着かせる	大丈夫(、｜。)|落ち着いて	55	nadameru1.vmd
喧嘩や対立を仲裁する。待ったをかける	待った待った｜待って待って	56	nadameru2.vmd
ユーザーの成功を願う。ユーザーを祈る	お願い｜祈(る｜り｜った)|心配	57	inori.vmd
照れ笑いする	え(っ)?へへ｜エ(ッ)?へへ	58	ehehe.vmd

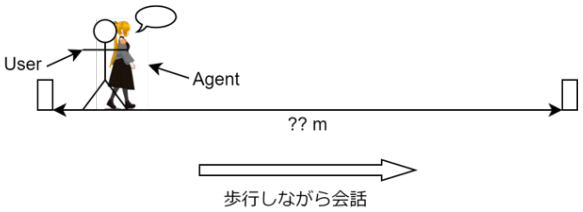

図 6　実験イメージ.

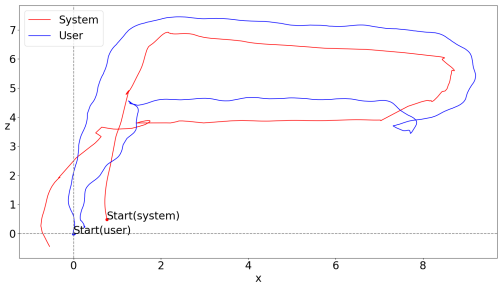

図 7　A コースにおけるユーザ（青）と仮想エージェント（赤）の歩行軌道．xz 面上で歩行した．「Start」の点は，ユーザ及び仮想エージェントの対話開始時点での座標点である．

として別途組み込んだ．パーソナルスペース [10]（個体距離）の概念を基に，提案システムでは疑似接触の判断基準を仮想エージェントを中心としたの半径 0.5[m] の円領域内と定めた．範囲外にいる場合は，手招き動作で誘導し，範囲内になると疑似接触を開始する．疑似接触は，ユーザが接近しない限り一時待機する．ただし，ユーザが接近せずに歩行を始めた場合や次の本音声応答が開始した場合は，疑似接触を中止する．

3　評価実験

提案システムを用いて，歩行しながら対話できるか検証実験を実施した．実験イメージを図 6 に示す．被験者は，仮想エージェントと対話しながら，あらかじめ設定した A と B の 2 種類のコースを歩行する．これを対話終了するまで行う．対話中には，対話時間，被験者のカメラ位置や仮想エージェントの位置を取得し続ける．これらデータを分析して評価した．

表 2　各手法による軌道一致率 [%].

近似手法	コース A	コース B
ユークリッド距離	0.8459	0.8740
動的時間伸縮法	0.5615	0.5376
フレシェ距離	0.9170	0.9020

評価の結果，平均応答時間はフィラー音声応答で 1.62[s]，本音声応答 3.21[s] となった．ユーザと仮想エージェントの歩行軌道では，軌道座標を 2 次元プロットし，観察評価を実施した．A コースの歩行軌道を図 7 に示す．その結果，大半の区間でユーザと仮想エージェントがなだらかな軌道を描きながら等間隔に歩行して対話していた事が示唆された．また，軌道一致率の結果を表 2 に示す．いずれも 0〜1 の範囲で正規化している．フレシェ距離やユークリッド距離で一致率が高い評価となっている事から，仮想エージェントはユーザが曲がった際もユーザが辿った軌道通りの歩行を行っている事を示唆しており，歩行しながら対話できた事を裏付けている．

4　おわりに

本研究では，拡張現実を用いて仮想エージェントと歩行しながら対話する音声対話システムを提案した．音声対話システム本体では，対話文応答の判断に使う過去会話内容の削減や RAG の仕様検討などの際に LLM を活用した．加えて，仮想エージェント自体がユーザに並んで歩行する事で，仮想エージェントは屋内外での活動範囲が拡大した．特に，歩行処理はユーザの複雑な動きに対応している事が実験にて示された．

仮想エージェントの移動アルゴリズムでは，図 4 内の r を調整して仮想エージェントの歩行先を決定していた．しかし，調節処理では正確に AR デバイスとユーザの距離を捉えられず，仮想エージェントが目的の歩行ができない場面もあった．今後の課題として，AR デバイスとユーザとの距離を正確に捉え，ユーザの邪魔にならない歩行処理の構築が挙げられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Annalena Aicher, Klaus Weber, Elisabeth André, Wolfgang Minker, and Stefan Ultes. The influence of avatar interfaces on argumentative dialogues. In **The 23rd ACM International Conference on Intelligent Virtual Agents**, 2023.

[2] 前土佐勇仁, 三枝亮. Cg キャラクターの行動表出によるユーザ無言時の話者交替の明確化. Technical report, 情報処理学会 研究報告アクセシビリティ, 2022.

[3] Reinhardt Jens, Hillen Luca, and Wolf Katrin. Embedding conversational agents into ar: Invisible or with a realistic human body? In **The Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction**, 2020.

[4] Jiarui Zhu, Radha Kumaran, Chengyuan Xu, and Tobias Höllerer. Free-form conversation with human and symbolic avatars in mixed reality. In **2023 IEEE International Symposium on Mixed and Augmented Reality**, 2023.

[5] Naoto Yoshida, Sho Hanasaki, and Tomoko Yonezawa. Attracting attention and changing behavior toward wall advertisements with a walking virtual agent. In **The 6th International Conference on Human-Agent Interaction**, 2018.

[6] Aratako. Aratako/calm3-22b-RP-v2, 8 2022.

[7] Jerry Liu. LlamaIndex, 11 2022.

[8] Google AI for Developers. Gemini モデル, 2024. https://ai.google.dev/gemini-api/docs/models/gemini?hl=ja.

[9] Keishi Tainaka, Tetsuya Kodama, Isidro Mendoza Butaslac, Hiroya Kawase, Taishi Sawabe, and Masayuki Kanbara. Tsundere interaction: Behavior modification by the integrated interaction of cold and kind actions. In **The 2021 ACM/IEEE International Conference on Human-Robot Interaction**, 2021.

[10] Edward T. Hall. **The Hidden Dimension**. Knopf Doubleday Publishing Group, 1990.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　付録

A.1　応答文生成のプロンプト文

<|im_start|>system
{仮想エージェント名}の設定
{仮想エージェント名}のロールプレイ
{仮想エージェント名}の設定
あだ名
ユーザとの関係性
一人称
現在の感情
言葉遣い
仮想エージェントの口調
ユーザの設定
名前，性別，あだ名
現在の感情
ロールプレイの補足
敬語禁止: 常にタメ口で，ユーザーを馴れ馴れしく呼ぶ
感情表現: {仮想エージェント名}の感情を激しく揺さぶり，その変化をリアルに表現
現在時間
現在時刻，日付
あらすじ
ユーザとの累計対話回数，最終対話時間
あらすじ本文
応答形式
以下の形式で順番に出力
{仮想エージェント名}のセリフ「1 文で描写」
{仮想エージェント名}の動作「結論のみ」
<|im_end|>
<|im_start|>user
ユーザのセリフ「{ユーザの発話内容}」<|im_end|>
<|im_start|>assistant
{仮想エージェント名}のセリフ「

図 8　プロンプト例（通常の応答文生成）.

<|im_start|>system
{仮想エージェント名}の設定
あだ名
ユーザとの関係性
一人称
ユーザの設定
名前
性別
あだ名 <|im_end|>
<|im_start|>user
小説家として、以下の文章を 3 点でまとめる
{ユーザとの過去 10 ターンの対話内容}<|im_end|>
<|im_start|>assistant
- 序盤:

図 9　プロンプト例（あらすじ文生成）.

A.2　コース B の歩行軌道

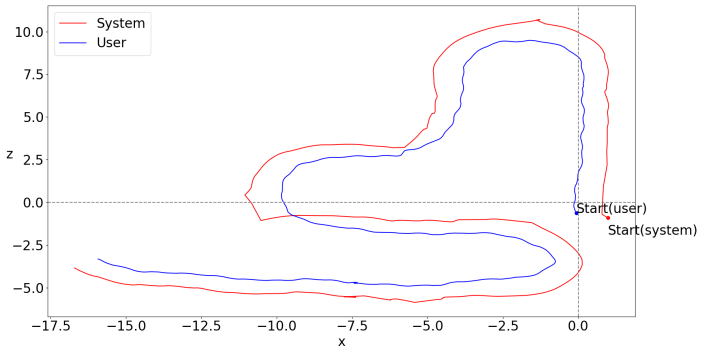

図 10　B コースにおけるユーザ（青）と仮想エージェント（赤）の歩行軌道.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).