# 屋外利用に向けたAR対話エージェントの歩行機能実装

# Implementation of Walking Functionality for Outdoor AR Conversational Agents

前土佐勇仁[1*]　南泰浩[1]
Yuto Maetosa[1]　Yasuhiro Minami[1]

[1] 電気通信大学大学院
[1] The University of Electro-Communications

**Abstract:** 本研究では，拡張現実（AR）を用いて，対話型仮想エージェントが屋外でユーザと一緒に歩行する機能を提案する．提案システムは，タブレット型などの端末に実装する．これにより，ユーザが仮想エージェントと移動しながら対話できる事を目指す．また，本システムでは，端末の角度情報に依存せず，ユーザの移動方向を推定する歩行手法を実装した．実験では，屋外の定められたルートを歩行しながら仮想エージェントと対話を行い，歩行軌跡への影響を検証する．

## 1　はじめに

音声対話システムの発展に伴い，仮想エージェントを活用した音声対話システムが実用化され，施設案内などのコンシェルジュとして運用されている．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 [2] や仮想現実（VR），拡張現実（AR）などがある．例えば，Jens らが開発した AR 用の音声対話システム [3] や，Zhu らが開発した VR，AR 用の音声対話システム [4] がある．特に後者は大規模言語処理モデル（LLM）を採用しており，知能や対話性能が高水準である．しかし，これらシステムは室内での利用が想定されており，またユーザは仮想エージェントの前に着席あるいは立ち止まって対話する．

屋外での利用を想定する場合，仮想エージェントがユーザに追従しながら対話する事が求められる．例えば，吉田らの仮想エージェントを用いた広告システム [5] では，幅 6[m]、高さ 3[m] の大型ディスプレイに仮想エージェントを投影する．仮想エージェントはユーザの歩行を追従し，ユーザの歩行速度に応じて仮想エージェントの歩行速度が変化する．人間型と矢印型の仮想エージェントを用いた実験では，人間型の方が広告の注目を集めた他，仮想エージェントが広告の前で停止や減速をすると，歩行者も歩行速度を低下させ広告に注目した事が示唆された．この事から，屋外で人間型の仮想エージェントが違和感なく追従するためには，ユーザの歩行速度に合わせた歩行速度で移動動作を実現する事が重要である．

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．ユーザの直進やカーブといった複雑な移動やユーザが仮想エージェントに振り向く動作などに対応するため，端末の角度情報に依存しない歩行機能の実装を行う．加えて，エージェントがユーザの真横あるいは前方を目標に仮想エージェントを配置し，実際に相手と並んで対話している状況を再現する．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，エージェントは屋外での活動が可能になる．また，ユーザの移動に仮想エージェントが追従する事で，親しい距離感での対話が実現される．

## 2　提案システム

提案する音声対話システムの構成を図 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，ユーザの発話内容を取得する他，仮想エージェントとユーザとの距離が離れていた場合に，仮想エージェントをユーザの周囲に向かって歩行させる．サーバは，クライアントで取得したデータからシステムが話す文章を生成し，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．ユーザが「バイバイ」，「さよなら」,「さようなら」のいずれかを発話する事で，システムとの対話を終了する．

*連絡先：電気通信大学大学院情報理工学研究科
〒 182-8585 東京都調布市調布ケ丘 1 丁目 5-1
E-mail: yuto.maetosa@uec.ac.jp

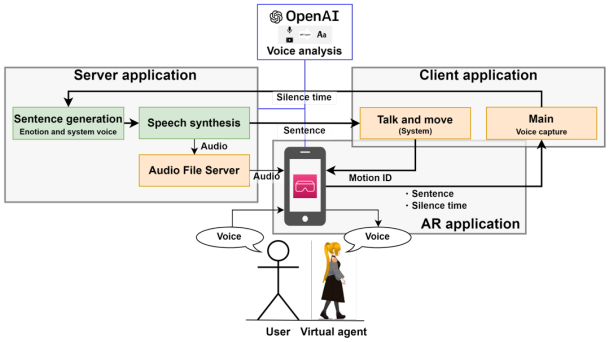


図 1: 音声対話システム.

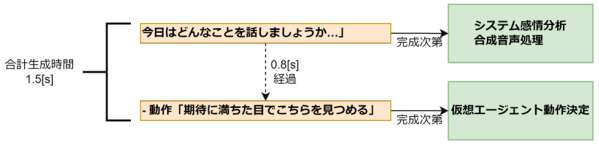


図 2: 生成例と対応する関連処理.

端末では入力音声から音量値を取得し，音量値が閾値を超えた場合に録音を開始する．録音中に音量値が閾値より下回ると，1 秒追加で録音を行う．この間に再度閾値を超えた場合，別音声として録音を再開する．閾値を下回って 0.5 秒後，提案システムは録音結果をサーバに送信し，音声認識モデル「kotoba-tech/kotoba-whisper-v2.0-faster[6]」を使用してユーザの発話内容を書き起こす．長時間ユーザが発話しなかった場合，一時的に音量計測を停止して仮想エージェントが話し返す [2].

対話文の生成は，LLM の「Aratako/calm3-22b-RP-v2[7]」を 5bit 量子化して実装した．対話文の例を図 2 に示す．ストリーミング処理を行い，仮想エージェントのセリフ，動作内容を順次生成する．動作内容の文章生成中には，サーバ側で仮想エージェントの感情分析と音声合成を並列で処理する．なお，感情分析には日本語の感情分析モデル「koshin2001/Japanese-to-Emotions[8]」を使用した．また，必要に応じてユーザの発話内容から検索エンジン用のクエリを生成し，検索結果を参照して対話文を生成する．

## 2.1 AR アプリケーション

AR アプリは，スマートフォン端末への実装を目的としたアプリケーションである．ユーザの音声を iOS や Android 端末で読み取り，クライアントに送信する他，クライアントやサーバで生成した動作内容を出力する役割を持つ．AR 環境の構築は Unity[9] で行った．

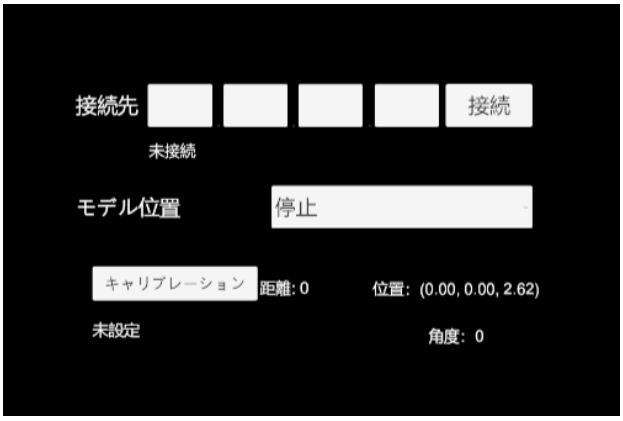


図 3: AR アプリの設定画面.

Vroid Studio[10] で作成したモデルを Unity ワールド内に表示し，仮想エージェントの動作や発話を処理する．

AR アプリ内の設定画面を図 3 に示す．Unity のワールド座標におけるカメラ位置は,「位置」で確認できる．図 3 の場合，カメラはアプリを起動した地点から $z$ 方向に 2.62[m] 移動している．ただし，実際はカメラの性能に左右されるため，現実，Unity のそれぞれの座標間には誤差が生じる．

## 2.2 移動アルゴリズム

移動アルゴリズムは，Unity ワールド座標の $xz$ 平面上でユーザと仮想エージェントの距離を計算し，仮想エージェントがユーザに追従するように設計した．ユーザが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に並んで歩行する．なお，AR アプリの設定で仮想エージェントがユーザの真横や前方に配置するかを選択する．

起動時の仮想エージェントは，Unity ワールド座標内のユーザの右方向 0.8[m] の場所にユーザと同じ方向を向いて停止している．図 5 の停止時のように，AR アプリを処理している端末カメラの画角内に仮想エージェントが収まると，仮想エージェントの角度 $\theta_A$ を以下の通り導出し，仮想エージェントが振り向く．

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

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．ただし，仮想エージェントを $A$，ユーザを $U$，端末を $M$ とする．仮想エージェントの移動先は，ユーザの左右いずれか最短距離の 1 つである．$t$ フレーム時，端末の Unity ワールド座標値 $M_t = (M_{t,(x)}, M_{t,(z)})$ を取得する．端末とユーザの距離を表すパラメータ $r$ の

初期値は，0.8[m] である．$r$ は，対話開始前に図 3 内の「キャリブレーション」ボタンを押し，ユーザがその場で回転する事で再設定する．仮想エージェントは，$r$ の範囲外に端末が移動するまで歩行を待機する．そのため，ユーザが仮想エージェントの方向に振り向いても，仮想エージェントは移動する事なく留まり続ける．

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 $\theta$ を以下の通りに導出する．

$$\theta = \arctan\frac{z}{x} = \arctan\frac{U_{t,(x)} - U_{(t-1),(x)}}{U_{t,(z)} - U_{(t-1),(z)}} \quad (2)$$

2 点 $U_tU_{t+1}$ の距離を $d_u$ とする．$d_u$ の決定は，使用している端末や仮想エージェントの配置設定（前方，真横）を踏まえて設定する．位置設定を「前方」と指定した場合，$d_u = r + 0.3$ で，「真横」の場合は $d_u = r$ である．以上より，ユーザの移動先 $U_{t+1}$ を以下の通りに予測する．

$$U_{t+1} = (U_t^x + d_u \cdot \sin\theta, U_t^z + d_u \cdot \cos\theta) \quad (3)$$

$A_{t+1}$，$A'_{t+1}$ は，$U_{t+1}$ を起点とする極座標 $Z(x,z)$ 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．なお，$Z(0,1) = 90°$ である．

$$\begin{aligned} Z^{U_{Temp}}(x,z) &= Z(r\cdot\cos(\theta+\frac{\pi}{2}), r\cdot\sin(\theta+\frac{\pi}{2})) \\ Z^{A_{t+1}}(x,z) &= Z^{U_{Temp}}(x,z)\cdot Z(0,1) \\ A_{t+1} &= U_{t+1} + Z^{A_{t+1}} \end{aligned} \quad (4)$$

$A'_{t+1}$ は，式 4 内の $Z(0,1)$ を $Z(0,-1) = -90°$ に変換して導出する．以上より，$U_{t+1}$ と $A_{t+1}$ との距離 $d_u$ は以下の通りに導出する．

$$\begin{aligned} T &= A_t - U_{t+1} \\ d_u &= \sqrt{(T_x - Z_{A_{t+1,(x)}})^2 + (T_z - Z_{A_{t+1,(z)}})^2} \end{aligned} \quad (5)$$

同様に，$A'_t$，$A'_{t+1}$ を用いて $(A'_{t+1}, d')$ との距離 $d'_u$ を導出する．

提案システムは，導出した $(A_{t+1}, d_u)$，$(A'_{t+1}, d'_u)$ のうち，$d_u$，$d'_u$ が小さい方を選択し，その座標を次フレームの移動目標点とする．これにより，仮想エージェントがユーザにぶつかる事なく，ユーザの移動に追従する事が可能となる．

## 2.3 仮想エージェントの動作表出

仮想エージェントが利用する動作数は全部で 114 種類あり，それぞれに判別に必要な文章（以下，判別文）と

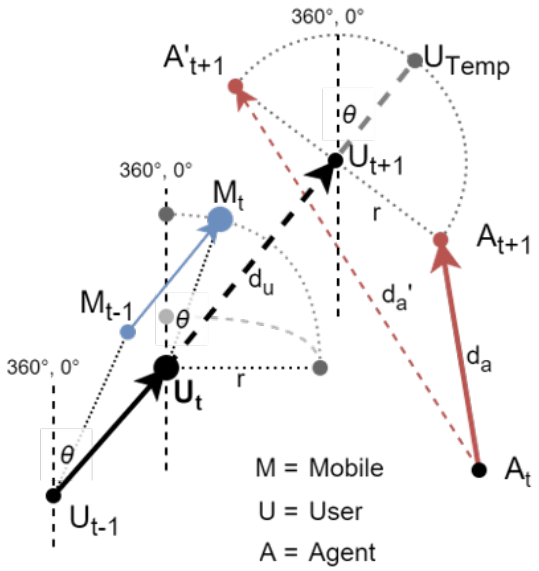


図 4: 移動アルゴリズム．

番号が割り振られている．対話開始時，仮想エージェントは 4 種類の初期待機動作からランダムで 1 つ決定し，準備が完了するまで待機動作を継続する．また，ユーザの音声入力処理終了後には 2 種類の思考動作の内，ランダムに交互に選択し，対話文生成内容の応答まで繰り返す．

仮想エージェントが音声応答する場合，提案システムは事前に設定した「キーワード動作」,「ランダム動作」,「感情別動作」で構成されている優先順位を元に動作を決定する．ただし，仮想エージェントを動作しない選択を取る場合もある．キーワード動作では，LLM で生成した動作内容と参照文となる判別文を，分散表現モデルである「pkshatech/GLuCoSE-base-ja-v2[11]」によってベクトル空間に埋め込む．そして，動作内容と判別文のベクトル間のコサイン類似度を計算し，この値を用いてキーワード動作の判定を行う．ランダム動作は 13 種類あり，一様分布の確率乱数で決定する．感情別動作は 11 種類で，仮想エージェントの感情情報から決定する．

### 2.3.1 移動時の動作

移動時の歩行動作は，移動アルゴリズム内の状態と図 4 内におけるユーザ $U_t$ と仮想エージェント $A_t$ との距離 $D$ を考慮して処理する．$D$ が大きくなる事で指定した範囲 $r$ を超過すると，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．$D > 1.65$ の場合，仮想エージェントは走行動作となり，

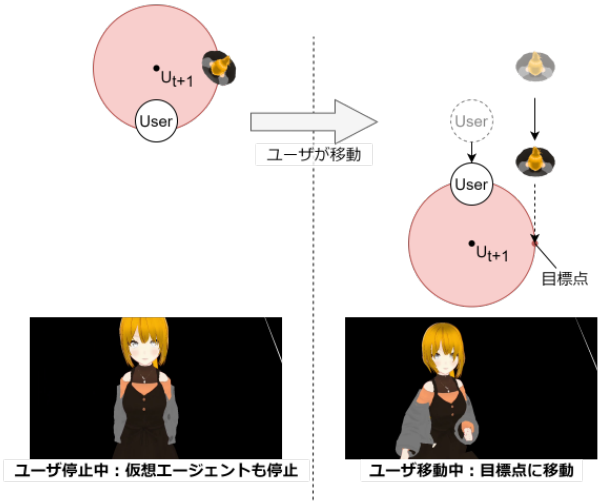


図 5: 仮想エージェントの移動.

図 6: 移動動作，発話，発話時動作の並行処理例.

$D < 1.5$ になるまで継続する．仮想エージェントの移動処理が終了すると，歩行動作も終了し待機動作に戻る．

応答文の発話と移動動作が重なった際の仮想エージェントの挙動を図 6 に示す．この場合，発話と同時に行う動作は上半身のみ，下半身は歩行動作を継続する．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．

### 2.3.2 疑似接触を伴う動作

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 [12]．本研究では，ハグや撫でるなどの仮想エージェントが行う動作を疑似接触を伴う動作として組み込んだ．領域内にいるユーザに対して，仮想エージェントが疑似接触を行う事で，ユーザとの距離感を縮める事ができる．疑似接触を行う範囲は，パーソナルスペース [13] を参考に設定した．俯瞰図を図 7 に示す．一般的に，ユーザと

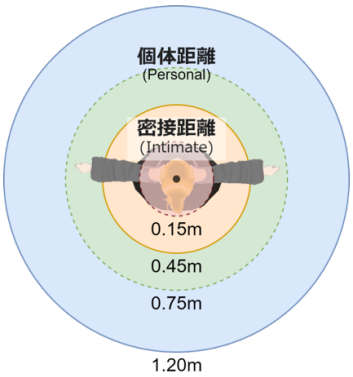


図 7: パーソナルスペースの俯瞰図 (抜粋).

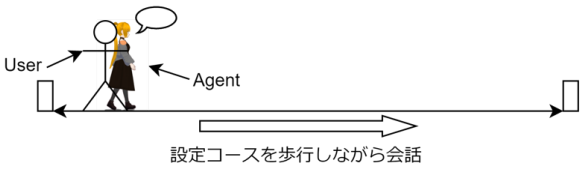


図 8: 実験イメージ.

仮想エージェントとの距離が 1.2[m] の範囲を下回ると，親しみのある距離感として認識される．

仮想エージェントを中心とした半径 0.5[m] の円の範囲内にユーザがいる場合，仮想エージェントは疑似接触を行う．ただし，範囲外にいる場合は手招き動作してユーザに接近を促す．また，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 実験

提案システムの応答速度や仮想エージェントの歩行挙動を検証した．まず，クライアント，サーバの両アプリの準備完了後に AR アプリを起動する．次に，図 3 内のキャリブレーションボタンをタッチし，音量閾値とデバイスとユーザの位置を把握する．キャリブレーション後に接続ボタンを押して実験を開始する．対話中には被験者，仮想エージェント両方の現在位置 $(x, z)$，発話内容，発話時間を記録する．これらの情報を元に，提案システムの歩行性能を評価する．

実験は，全て屋内で歩行した場合と屋外で歩行した場合の 2 パターンで行った．使用機器は iPad Mini（第 5 世代）で，携帯回線を用いてクライアント，サーバと通信しながら仮想エージェントと対話する．

仮想エージェントとユーザとの距離感を評価するため，$(x, z)$ 座標からユークリッド距離，DTW（動的時間

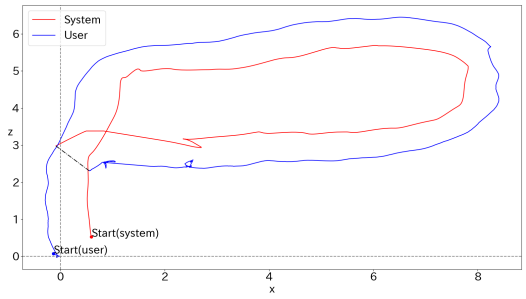


図 9: 屋内の移動軌跡 (点線は末端番号のペア).

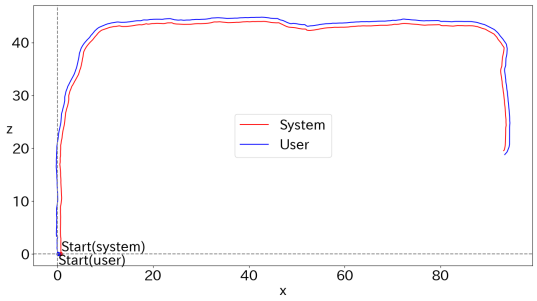


図 10: 屋外の移動軌跡.

表 1: 各種距離計算の結果.

| 距離計算手法 | 屋内 | 屋外 | 屋外 (再計算後) |
|---|---|---|---|
| ユークリッド距離 (線形) | 44.912 | 62.622 | 62.622 |
| ユークリッド距離 (各ペア平均) | 0.743 | 0.833 | 0.833 |
| DTW | 2570.894 | 4525.138 | 4525.138 |
| Derivative-DTW | 13.172 | 38.883 | 38.883 |
| フレシェ距離 | 0.968 | 1.685 | 0.897 |

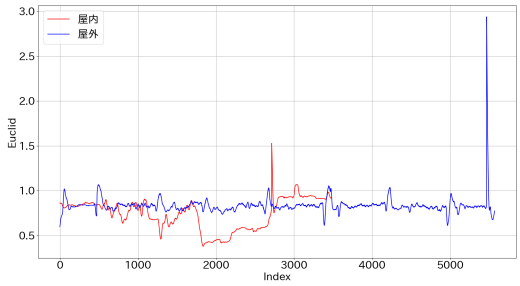


図 11: 要素番号ごとのユークリッド距離の推移.

圧縮法）距離，フレシェ距離を導出したユークリッド距離は線形全体の評価と各座標ペアの平均値，DTW 距離はDTW のみの手法と事前に微分する手法（Derivative-DTW[14]）に分けて評価した．

## 3.1 結果

対話中のユーザ，仮想エージェントの移動軌跡を図 9，図 10 に示す．Unity ワールド座標におけるユーザの歩行範囲は，屋内で 6.47[m] × 8.70[m]，屋外で 44.81[m] × 94.72[m]（いずれも縦×横）となった．仮想エージェントの移動軌跡は，ユーザの移動に沿って滑らかに推移しており，かつユーザとの距離感が保たれている．

各距離計算の結果を表 1 に，要素番号ごとのユークリッド距離の推移を図 11 に示す．Derivative-DTW 距離を用いた事で，屋内で 13.172，屋外で 38.883 となり，DTW 距離と比較して大きな差が生じた．各ペアの平均値を用いたユークリッド距離では，屋内で 0.743[m]，屋外で 0.833[m] となった一方で，屋外におけるフレシェ距離は 1.685[m] となり，屋内と比較して大きな差が生じた．図 11 から，屋外コースの最後の区間においてユークリッド距離が急激に増加したため，要素番号の範囲を 0 から 5000 に限定して再計算した結果，フレシェ距離は 0.897[m] となり，屋内との差が縮まった．

## 3.2 考察

本実験では，屋内，屋外のどちらでもユーザと仮想エージェントの距離感を保ちながら歩行する事が可能である事を示した．特に，各ペアごとのユークリッド距離平均やフレシェ距離では，1.2[m] を下回っており，仮想エージェントは多くの区間でユーザと親しい距離感を保って歩行できていた．また，仮想エージェントの歩行の安定性は，屋内より屋外の方が高いという結果が歩行軌跡やユークリッド距離の推移から読み取れる．これは，AR における平面認識の性能に左右されるためで，屋内と比較して障害物が少ない屋外の方が仮想エージェントとの歩行に適していると考えられる．

DTW 距離と Derivative-DTW 距離の比較から，歩行軌跡には加速，減速の要素が含まれている事が示唆された．図 11 より，ユークリッド距離が急激に増加した箇所が見られたため，仮想エージェントが停止してしまい，ユーザとの距離感が離れてしまった可能性が

ある．加速は，この状態から速やかに解決するために生じていると考えられる．

## 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．端末の角度情報に依存しない歩行機能を実装し，仮想エージェントはユーザが違和感を感じる事なく歩行する事が可能である事を示した．また，屋外で活動した場合でも，ユーザと親しい距離感を保ちながら対話できた事が実験にて示された．

歩行処理では，端末の角度情報を用いずに仮想エージェントの移動先を決定した．しかし，調節処理では正確に AR 端末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 端末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．また，屋外における仮想エージェントとの対話中にユーザが仮想エージェントにどの程度注目しているかを調査する事で，対話の質を向上させる事が期待される．

## 参考文献


[1] Annalena Aicher, Klaus Weber, Elisabeth Andr?, Wolfgang Minker, Stefan Ultes: The Influence of Avatar Interfaces on Argumentative Dialogues, *The 23rd ACM International Conference on Intelligent Virtual Agents, No.24, pp.1-8 (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*, 2022-AAC-20, No.4, pp.1-7 (2022)*

*[3] Jens Reinhardt, Luca Hillen, Katrin Wolf: Embedding Conversational Agents into AR: Invisible or with a Realistic Human Body?, The 14th International Conference on Tangible, Embedded, and Embodied Interaction, pp.299-310 (2020)*

*[4] Jiarui Zhu, Radha Kumaran, Chengyuan Xu, Tobias H?llerer: Free-form Conversation with Human and Symbolic Avatars in Mixed Reality, 2023 IEEE International Symposium on Mixed and Augmented Reality, pp.751-760 (2023)*

*[5] Naoto Yoshida, Sho Hanasaki, Tomoko Yonezawa: Attracting Attention and Changing Behavior toward Wall Advertisements with a Walking Virtual Agent, The 6th International Conference on Human-Agent Interaction, pp.61-66 (2018)*

*[6] Hugging Face: kotoba-tech/kotoba-whisper-v2.0, https://huggingface.co/kotoba-tech/kotoba-whisper-v2.0-faster (*参照*: 2025* 年 *1* 月 *17* 日*)*

*[7] Hugging Face: Aratako/calm3-22b-RP-v2, https://huggingface.co/Aratako/calm3-22b-RP-v2 (*参照*: 2025* 年 *1* 月 *9* 日*)*

*[8] Hugging Face: koshin2001/Japanese-to-Emotions, https://huggingface.co/koshin2001/Japanese-to-emotions. (*参照*: 2025* 年 *1* 月 *9* 日*)*

*[9] Unity Technologies: Unity, https://unity.com/ja (*参照*: 2025* 年 *1* 月 *16* 日*)*

*[10] Pixiv: Vroid Studio, https://vroid.com/studio (*参照*: 2025* 年 *1* 月 *22* 日*)*

*[11] Hugging Face: pkshatech/GLuCoSE-base-ja-v2, https://huggingface.co/pkshatech/GLuCoSE-base-ja-v2 (*参照*: 2025* 年 *2* 月 *13* 日*)*

*[12] Keishi Tainaka, Tetsuya Kodama, Isidro Mendoza Butaslac, Hiroya Kawase, Taishi Sawabe, Masayuki Kanbara: TSUNDERE Interaction: Behavior Modification by the Integrated Interaction of Cold and Kind Actions, The 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp.153-156 (2021)*

*[13] Edward T. Hall: The Hidden Dimension, Knopf Doubleday Publishing Group (1990)*

*[14] Eamonn J. Keogh, Michael J. Pazzani: Derivative Dynamic Time Warping, The 2001 SIAM International Conference on Data Mining, pp.1-11 (2001)*