Vol.2025-1 No.2-3
2025/2/22

VR 空間内のオブジェクト認識を用いたマルチモーダル対話システム

関根 寿[†1] 細谷謙多[†1]
関戸陽士[†1] 小坂哲夫[†1]

人間とエージェントによる音声対話システムの開発が進んでいる．これに対し我々は音声以外のモダリティも併用するマルチモーダル対話システムを開発してきた．マルチモーダル対話では音声のみならず表情や身体動作など言語情報以外の方法を利用して対話を行なう．エージェントの表示としてはスクリーン上に表示する方法と，ヘッドマウントディスプレイを利用したＶＲ空間上での表現について検討を進めている．しかし, それらのシステムではユーザの身体情報を用いた対話の拡張が主であり, 物体を介した対話については実現できていなかった．本研究では,VR 空間上で複数の任意のオブジェクトを用いたテーマの伝達を行う対話システムを構築した．オブジェクトを利用することにより，より直観的で自然性の高い対話を実現した．

Multimodal Dialogue System Using Object Recognition in VR Space

Hisashi Sekine,[†1] Kenta Hosoya,[†1] Yoji Sekido[†1] and Tetsuo Kosaka[†1]

The development of human-agent spoken dialogue systems is in progress. In contrast, we have developed a multimodal dialogue system that uses modalities other than speech. In multimodal dialogue systems, the dialogue is conducted using not only speech but also facial expressions, body movements, and other non-verbal information. We have been studying the display of agents on a screen and in a VR space using a head-mounted display. However, these systems have mainly extended dialogue using the user's body information, and have not been able to realize dialogue through objects. In this study, we constructed a dialogue system that communicates themes using multiple arbitrary objects in a VR space. By using objects, we realized a more intuitive and natural dialogue.

1. はじめに

音声処理技術の発展により, 様々な音声対話システムが実用化されている. 特に, 大規模言語モデル (Large Language Model;LLM) の登場により, 多くの対話システムにおいて, 人間に近い自然な応答が可能な環境が実現した. 対話システムとは, 人間と機械が言語を用いてやり取りを行う仕組みであり, 音声対話システムはその中でも音声による入出力を主軸としたものを指す. しかし, 現在主流の音声対話システムの多くは音声や言語ベースの対話に限られており, 人間同士のコミュニケーションと比べて理解度や自然性に課題が残る.Birdwhistell によれば, 人対人の対話において言語情報が伝える情報量は全体の 35%程度にすぎず, 残りは非言語情報によって伝えられるとされている[1]. 以上を踏まえ, 我々は非言語情報を用いたマルチモーダル対話システムを開発してきた.

文献[2] では, ユーザの身体動作を用いた対話システムを構築した. この研究では, 視線を顔の向きで読み取り, その向きからエージェントの発話内容を変化させるシステムを実現し, 手を振るなどの身体動作を用いたコミュニケーションを可能とした. しかし, 人によっては顔の向きではなく視線で見てしまうことから正確な推定は行われていなかった. そこで, 文献[3] では画像認識システムから視線情報取得する対話システムを提案した. ユーザの視線情報を対話に取り入れることで、「話しやすさ」と「対話の円滑性」の向上が確認された.

また, 文献[4] ではユーザの表情を利用し, 表情からポジティブ/ネガティブの情報を取得することで, ユーザが笑顔になるとエージェントが笑顔で返すシステムを構築した. その後, 文献[5] において, トランスデューサ (FST) による手書きシナリオから生成 AI に変更することで, 非タスク型対話システムと呼ばれる雑談のような特定の目的を持たない対話を行うエージェントを構築した. これにより,「人間らしさ」や「継続的使用意欲」の向上が確認された.

以上[2]〜[5], ではスクリーン表示によるエージェント対話を行ってきた. しかし, スクリーン表示では, 環境要因の影響を受けやすく対話の自由度やエージェントの距離感において課題が残されていた. そこで, 文献[6] では VR 表示での対話システムを構築した. この文献は千葉ら[7] の研究による音声対話処理ツールキットと VR を統合したシステムであり,2D 表示のエージェントに比べ, パーソナルスペースの距離が近いことから, 親近感においてより人間との対話に近い感覚でインタラクションできることが示された. 一方で,VR 空間におけるマルチモーダル化はプラットフォームの提供に過ぎず, 自然性の向上において多くの課題

†1 山形大学
Yamagata University

が見られた.

音声/言語ベースの対話において, ゼロ照応と呼ばれる現象が存在する. これは, 発話者が特定の情報を省略して発話を行うことを指す. 例えば,「これは何ですか？」という内容は, 音声やテキストのみでは「これ」の特定が不可能である. この課題に対して, 石井ら[8] の研究では, 人対ロボットにおいて指差し、視線を用いた物体の特定を行うことで、曖昧な言語指示対象の同定及び対話を実現している. また, 稲邑ら[9] の研究では,SIGVerse[10] と呼ばれる VR プラットフォームを用いた VR 空間での指差しによる物体指定を行っている. 文献[8][9] はオブジェクトを用いた対話の有用性を示しているが, いずれも人とエージェントによるオブジェクトを用いた対話システムの構築や非タスク対話システムにおける複数オブジェクトの認識は出来ていない.VR 空間では任意のオブジェクトを容易に生成でき, またユーザが自由にインタラクトできる点から複数のオブジェクトを用いたテーマの伝達及び選択問題の実現が期待できる. 本報告では提案システムの構成や動作について述べる.

2. 関 連 研 究

リアルタイムマルチモーダル対話システムのプラットフォームとして,Remdis[11] がオープンソースとして公開されている.Remdis は VAP(Voice Activity Projection) と呼ばれる独自のターンテイキング技術により, ユーザとエージェント間の円滑な会話の切り替えを実現している. また応答生成には ChatGPT[12] を用い, マルチモーダル出力のために 2023 年 12 月より音声インタラクション構築ツールキットとして公開された MMDAgent-EX[13] と連携している. 文献[6] では,Remdis をベースとしたシステムを構築しており,MMD エージェントをゲームエンジン Unity[14] で開発したアプリケーションに置き換えることによって VR 空間上でのエージェントの表示を行っている. エージェントは MMD4Mecanim[15] というツールを用い,PMX ファイルで提供されている初音ミクの 3D モデル[16] を表示している. また,Remdis の音声合成部では日本語テキストの入力から音声を生成する HMM テキスト音声合成システムである OpenJTalk[17] を使用している. 本研究では, この Remdis とそれに基づいた VR 環境をベースにシステムを構築した.

3. システムの概要

3.1 システムの全体構成

マルチモーダル対話システムの全体構成を図 1 に示す. 矢印におけるオブジェクト認識及びオブジェクト情報送信が本システムの開発部分である. 本システムは, 文献[6] による VR

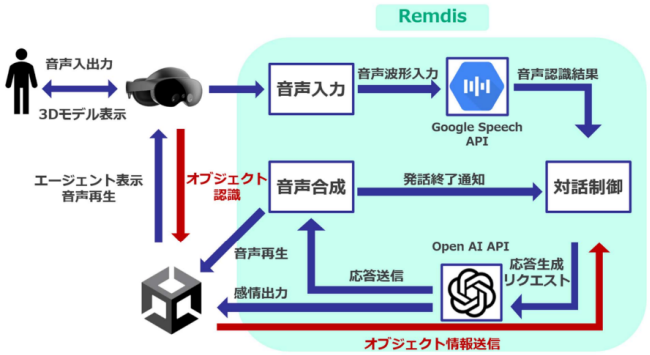

図 1 提案システムの全体図
Fig. 1 Overall view of the proposed system

対話システムをベースとして,Unity 内で任意のオブジェクトを複数配置し, ユーザのハンドアニメーションによるイベント生成, オブジェクト情報と Remdis の送受信を行うシステムを開発することで, 対話時にエージェントがオブジェクトの内容を認知するシステムを実装している.

3.2 本研究における Remdis の役割

本研究において,Remdis はユーザの音声入力, 音声認識,LLM による応答生成, 応答文の音声合成のために用いる. それぞれの機能がモジュール化して分けてあるため, 比較的コンパクトで拡張しやすいフォルダ構造となっている. また、各モジュール間の通信は RabbitMQ と呼ばれるメッセージングミドルウェアを採用しており,Exchange によってメッセージのルーティングを制御できる.Exchange を指定することで, 複数のモジュールと送受信が可能となっている. 本研究においてはオブジェクト送信用の Exchange を追加し, 音声認識結果がコミットされた際に送信することで, オブジェクト情報の通信を実現している.

Remdis では,Voice Activity Projection(VAP) を用いた適切なタイミングでの相槌生成やターンテイキングが行えるが, 本研究では開発環境のインターネット回線が遅く,ChatGPT の応答生成と VAP が判定したユーザの発話終了のタイミングがかみ合わないことにより, 高確率で対話が破綻してしまう原因となっていたため,VAP を使用せずに対話システムを構築した.

3.3 実 行 環 境

本研究では実行環境として,WindowsPC で VR アプリケーションを再生し,PC に接続し

た HMD(MetaQuestPro) で Unity で作成した対話システムの空間を表示させる. 同時に同じ PC で Remdis を実行し, マイクとスピーカーは HMD に標準搭載のものを用いる.

3.4 本研究における Unity の役割

Unity は Unity Technologies が開発したゲームエンジンであり, ゲーム開発だけでなく, 建築設計, 医療, 教育, シミュレーションなど, 産業におけるさまざまな分野で幅広く活用されているソフトウェアである. 本研究においては Remdis と HMD の連携および,VR 空間上での対話環境の構築に用いる.

VR アプリケーションの作成において,Unity の公式パッケージである XR Interaction Toolkit[18] を用いた.XR Interaction Toolkit とは,Unity で提供されている VR/AR 体験の作成を目的とするコンポーネントベースの高度なインタラクションシステムであり, ハンドトラッキングやテレポート, オブジェクト操作といったインタラクションを簡単に実装できる. このシステムを用いることで, クロスプラットフォームの VR アプリケーションを容易に作成できる. 本研究では XR Interaction Tool kit のサンプルに含まれているプレイヤーオブジェクトを 3D 空間上に配置することでエージェントを表示しており, ユーザのハンドトラッキングやアニメーション, エージェントに認識可能なオブジェクトの実装よって,VR での目的とする対話を実現している.

図 2 にコントローラの操作方法について示す.HMD コントローラはグリップボタン, トリガーボタン, スティックボタンを使用し, グリップボタンとトリガーボタンを同時に押すことにより掴むアニメーション, グリップボタンのみ押すことによりレイキャストを表示し, 指を指すアニメーションを行う. ボタンを押下中にユーザが発話をしたとき, オブジェクト情報がユーザ発話の末尾に送信される. また, スティックボタンによって VR 空間内の移動が可能である.

3.5 ChatGPT プロンプト

図 3 は本研究で用いた ChatGPT のプロンプトである. ベースは Remdis 及び文献[6] の研究によって使用されたプロンプトをそのまま採用し, オブジェクト情報の伝達において, 一文を追加している.

オブジェクトの伝達方法は左右, 両手の 3 つで条件分岐を行っている. 例えば, リンゴを右手で掴んだ場合, ユーザ発話の末尾に「/リンゴ,GrabRight」という情報がエージェントに送信される. 尚, 両手の伝達は掴む場合と指差す場合どちらにも対応しており, それぞれのモーションに依存はしていない.

図 2 コントローラの操作方法
Fig. 2 How to operate the controller

```
==
あなたはユーザと雑談するアシスタントです． 次のユーザ発話に対する気の効いたリアクションや一つの質問を作成し，句読点（、,。,！,？）で分割して出力してください．文章の区切りには空白を入れてください．ユーザー発話の末尾に"/"がある場合はユーザーが手に取ったオブジェクト情報と手の状態が記述されています.これらの情報は以下に基づいて解釈してください．
==
/オブジェクト情報,手の状態
GrabRight：右手で掴んでいる
GrabLeft：左手で掴んでいる
GrabBoth：両手で掴んでいる（先に書いてあるオブジェクトは左手で持っている）
==
最後にアシスタントの感情（0_平静,1_喜び,2_感動,3_納得,4_考え中,5_眠い,6_ジト目,7_同情,8_恥ずかしい,9_怒り）と動き（0_待機,1_ユーザの声に気づく,2_うなずく,3_首をかしげる,4_考え中,5_会釈,6_お辞儀,7_片手を振る,8_両手を振る,9_見渡す）を出力してください．出力は以下のフォーマットに従ってください
==
こんにちは。よろしくお願いします。/0_平静,2_うなずく
==
```

図 3 ChatGPT プロンプト
Fig. 3 ChatGPT Prompt

3.6 対 話 例

システムを動作させたときの対話例を図 4 に示す.

この発話と同時にユーザがりんごを右手でつかんだ場合，図のようにオブジェクト名と GrabRight が発話に付加されてエージェントに情報が伝達される.

4. 対話システムの有用性検証

本実験では VR 空間上で表示されるエージェントとオブジェクトを用い, 情報収集ミッショ

User: こんにちは.
Agent: こんにちは今日はいかがお過ごしですか?
User: これは何でしょうか?/りんご,GrabRight
Agent: それはおいしそうなりんごですね. 甘いですか?

図 4 システムの対話例
Fig. 4 Example of system interaction

ンをユーザに課した評価実験を行った.

4.1 実験方法

本実験では「この世に存在しない特定の役割」を持つオブジェクトを配置し, エージェントに質問を繰り返すことで、そのオブジェクトの役割と使い方を聞き出す対話システムを実装した. 従来システムとは異なり比較対象が存在しないため, 情報収集ミッションによる客観評価からエージェントに対する周囲の環境を使用した対話及び選択問題の有用性について検証する. また, 客観評価によるタスクの成功率からオブジェクトを用いた対話による今後の拡張性について検証する.

オブジェクトの内容は施験者があらかじめ考え, システム自体に一貫性のとれたものを採用する. 被験者が HMD を取り付け, 施験者が合図を出したタイミングでタイマーを開始し実験を開始する. 被験者は目の前にあるオブジェクトを手に取ったり, 指を指すことによってエージェントにそのオブジェクトの役割と使い方を問う. 被験者が理解できたと判断した時点でタイマーを止め, 最後にアンケートを回答してもらうことで実験を終了する. 尚, タイマーによる計測は各オブジェクトごとに行う. タスク達成率とタスク達成までの平均時間から有用性の検証を行う.

本実験では成人男性 11 名, 成人女性 1 名の計 12 名を対象に実験を行った. 順序効果を考慮してそれぞれの対話システムの順番はランダムで行った.

4.2 シナリオ内容

表 1 にシナリオの詳細について示す. 今回 4 つのオブジェクトを用い, 存在しないオブジェクトとして「光を放出あるいは吸収する機器」のシナリオを作成した. オブジェクト名は Unity 上の各オブジェクトに, オブジェクト内容は ChatGPT のプロンプトに記述する. 尚、エージェントは存在しないオブジェクトに対してすでに知っているものとして回答してもらうよう指示している.ChatGPT のプロンプトに付け足した情報を図 5 に示す.

表 1 シナリオ詳細
Table 1 Scenario Details

オブジェクト名	オブジェクト内容
青電池	情報送信装置に入れることで, 光を吸収する情報を送ることができる。情報送信装置に対する役割は青い電池と同様であるため, ユーザの質問内容として赤い電池との選択問題を想定している.
赤電池	情報送信装置に入れることで, 光を放出する情報を送ることができる。情報送信装置に対する役割は赤い電池と同様であるため, ユーザの質問内容として青い電池との選択問題を想定している.
送信機	電池を入れることで, 情報受信装置に情報を送ることができる.
受信機	情報送信装置から情報を受け取ることで, 光を放出あるいは吸収する. このオブジェクトのみ掴むことができず, 指差しによって回答することを想定している.

==
あなたはユーザと雑談するアシスタントです. 次のユーザ発話に対する気の効いたリアクションや一つの質問を作成し, 句読点（、 ,。 , ! , ? ）で分割して出力してください. 文章の区切りには空白を入れてください. ユーザー発話の末尾に"/"がある場合はユーザーが手に取ったオブジェクト情報と手の状態が記述されています.これらの情報は以下に基づいて解釈してください.
==
/オブジェクト情報,手の状態
GrabRight : 右手で掴んでいる
GrabLeft : 左手で掴んでいる
GrabBoth : 両手で掴んでいる（先に書いてあるオブジェクトは左手で持っている）
==
ユーザには,情報収集ミッションを行ってもらいます.ユーザにはS1,S2,M,Lの4種類のオブジェクトが提供されているため,質問されたらあなたはそのオブジェクトについて知っている前提で話をしてください.以下にオブジェクトのシナリオを示します.
S1 : Mに差し込むための電池.S1を取り付けたMをLに送信することで,Lは光を放出する.
S2 : Mに差し込むための電池.S1を取り付けたMをLに送信することで,Lは光を吸収する.
M : 情報送信装置.電池をMの中央に差し込むことで動作する.インタラクトすることで,LのON,OFFが可能となる.
L : 情報受信装置.S1またはS2を差し込んだMから情報を受け取ることで,光を放出または吸収する.
==
最後にアシスタントの感情（0_平静,1_喜び,2_感動,3_納得,4_考え中,5_眠い,6_ジト目,7_同情,8_恥ずかしい,9_怒り）と動き（0_待機,1_ユーザの声に気づく,2_うなずく,3_首をかしげる,4_考え中,5_会釈,6_お辞儀,7_片手を振る,8_両手を振る,9_見渡す）を出力してください. 出力は以下のフォーマットに従ってください
==
こんにちは。よろしくお願いします。/0_平静,2_うなずく
==

図 5 実験用 ChatGPT プロンプト
Fig. 5 Experimental ChatGPT prompt

4.3 実験環境

VR 空間での対話実験環境の俯瞰図を図 6 に示す.HMD に MetaQuestPro を用いて,Unity で作成した VR 空間内でエージェントを表示させる.VR 空間上での操作方法は図 2 と同様である. 表示するエージェントは初音ミクであり, オブジェクトの見た目は Unity に標準搭載の図形オブジェクトを用いた抽象的な表現を行っている.HMD で表示される環境を図 7

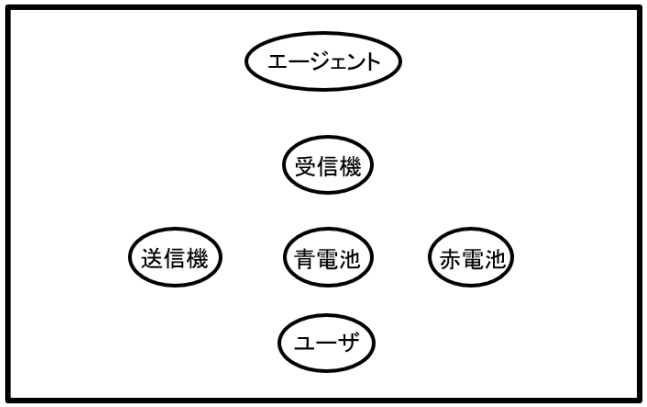

図 6 対話実験環境の俯瞰図
Fig. 6 Overhead view of interactive experimental environment

図 7 HMD で表示される環境
Fig. 7 Environment displayed by HMD

に示す.

5. 実験結果と考察

5.1 実 験 結 果

タスク成功率は 100%を確認した. また, 表 2 に本実験で計測したタスク達成平均時間を示す.

表 2 タスク達成時間
Table 2 Task accomplishment time

オブジェクト名	平均タスク達成時間 (MM：SS)
青電池	01:44
赤電池	01:53
送信機	01:08
受信機	00:52

5.2 考　　察

オブジェクト情報の理解を条件にタスク達成時間の計測を行ったため, 実験の結果としてすべての被験者がオブジェクト情報の理解を示したことを確認した. よって, 目的としていた複数のオブジェクト情報の伝達の達成を示せた. 一方で被験者の質問内容によっては返答に若干の変化が見られ, 目標とする応答の出力までに時間を要した面も見られた. 例えば,「このオブジェクトは何ですか？」という質問と「このオブジェクトの用途は何ですか？」という質問では前者はオブジェクトの名称からその役割までのすべてを説明するのに対し, 後者は対話制御部における推測により用途のみを説明する, つまりオブジェクト自身の名称を省略して説明を行う傾向が見られた. この現象に関しては対話制御部で完結する内容であるため,ChatGPT プロンプトの修正により改善できるものと考える. また, 被験者に共通する意見として「エージェントの話すスピードが速く, 内容を聞き取るのに時間を要した」という意見をいただいた. 話すテンポは経験則により,1 分間に 300 文字程度が適切であると言われており, 今回の実験で用いたエージェントはそれよりも速く且つエージェントの返答における会話の間が短い傾向が見られた. 本実験は雑談のような一時的な内容ではなく, 被験者の記憶力を問う内容が主体であったため, よりエージェントの話すスピードに注意が向いたと考える. 本実験では青電池から順に被験者への質問を行うことで, タスク達成時間の評価を行った. 計測結果から, 全体として 2 分弱でオブジェクト情報の理解ができている傾向が見られた. 送信機以降の計測時間が電池よりも短くなっている理由として, 全体のオブジェクトに一貫性を持たせたシナリオを提示していることが挙げられる. 赤/青電池を理解することで、他の二つのオブジェクトの役割がおおよそ推測できる場合もあるため, それによって回答時間の短縮が行われたものと推測する. 今回得られた情報は客観的なものであるが, 比較実験ではないため, この数値から推測する内容は全て主観的なものとなる. しかし, 全てのオブジェクトが 2 分以内に回答できていることからオブジェクトによるエージェントの解答方法

によって被験者が自然性に違和感を覚えることはあまり生じないという点において, 情報収集ミッションにおけるオブジェクト情報の伝達及び選択問題の実装が有効であると考える.

6. ま と め

6.1 結　　論

本研究では, エージェントとオブジェクトをヘッドマウントディスプレイ (HMD) を利用した VR 空間内に配置し,4 つのオブジェクトの用途を被験者が質問する対話実験を通じて, 周囲の環境下に存在するオブジェクトをテーマとしたマルチモーダル対話及び選択問題の実現を図ったシステムの構築, 評価を行った. 情報収集ミッションにおける客観評価の対話実験を行った結果, オブジェクトの情報伝達率 100%及びタスク達成時間 2 分弱の記録を得られた. これにより, 目的としていたユーザーが任意の複数のオブジェクトをインタラクトする行動を通じて, エージェントとの対話を促進するシステムが有用であることが示せた. 一方で, エージェントの会話スピードや質問の返答方法に対して改善が必要であることを確認した.

6.2 今後の展望

本研究では, 対話システムにおいて会話の自由度及びオブジェクトを用いたテーマの伝達, 選択問題の実現のために VR を用いた研究を行った. 実験結果から, オブジェクト情報の伝達はスムーズにできるものの, 返答方法や会話のテンポ感による理解度に差異が見られることを確認した. エージェントの応答に関しては対話制御部に関する内容であるため,ChatGPT のプロンプトを見直す必要があると考える. また, 今回 Remdis において VAP によるターンテイキングを利用せずに対話を行っている。今後の研究方針として,VAP を用いたターンテイキングの実現が挙げられる.

一方で, 今回のオブジェクトを用いたテーマの伝達はユーザがエージェントに向かって会話を行ったときのみを対象としている. しかし, 実際の人対人での対話では片方から話題を一方的に出すのは不自然であり, エージェント側からの返答も含めることが望ましいと考える. そのため, 視線制御や人間の行動原理に基づいて, オブジェクトを数秒間見つめた時などにエージェント側からユーザに向けて質問を行うシステムを構築することで, 対話における自然性が向上すると考える.

謝辞　本研究のベースとしてマルチモーダル対話システム開発プラットフォーム Remdis を使用した．開発メンバーに対し，ここに記して感謝する.

参 考 文 献

1) Ray L. Birdwhistell, ”Kinesics and context: Essays on body motion communication,” Univ. of Pennsylvania Press., 1970.
2) Takeru Koseki, Tetsuo Kosaka, ”Multimodal Spoken Dialog System Using State Estimation by Body Motion,” Proc. of IEEE GCCE2017, pp. 348-351, 2017.
3) 斎藤順平, 加藤正治, 小坂哲夫, ”ユーザの非言語情報を併用したマルチモーダル対話システムの開発とその評価”, 東北地区音響学研究会, 3-18, 2020.
4) 細田尚輝, 小坂 哲夫, ”ユーザの表情の感情情報を用いた音声対話システム,” 信学技報, HCS2023-52, 2023.
5) 関戸陽士, 小坂哲夫, ”生成 AI を用いたマルチモーダル対話システムの評価,” 情処研究報告, Vol.2024-SLP-152, No.48, 2024.
6) 細谷謙多, 関戸陽士, 小坂哲夫, ”生成 AI を用いた VR 空間内 3D エージェントとの音声対話システムの開発”, 信学技報, HCS2024-40, 2024.
7) 千葉祐弥, 光田航, 李晃伸, 東中竜一郎, ”Remdis: リアルタイムマルチモーダル対話システム構築ツールキット,” 言語・音声理解と対話処理研究会, pp.25-30, 2023.
8) 石井里奈, 宮森恒,, ”視覚情報と知識を利用したマルチモーダル対話における曖昧な言語指示対象の同定”, 第 16 回データ工学と情報マネジメントに関するフォーラム,2024.
9) 稲邑哲也, 進藤裕之, 松本裕治, ”対話型ロボットの学習効率化のためのクラウド型 VR プラットフォーム”, 人工知能, Vol.35, no.1, pp.72-78, 2020.
10) 稲邑哲也, ”社会的知能研究のためのシミュレーションプラットフォーム：SIGVerse”, 日本ロボット学会誌, Vol.31, No.3, pp.240-243, 2013
11) Remdis: Realtime Multimodal Dialogue System Toolkit, https://github.com/remdis/remdis
12) OpenAI,Chatgpt,https://openai.com/chatgpt
13) MMDAgent-EX, https://mmdagent-ex.dev/
14) Unity, https://unity.com/ja
15) MMD4Mecanim, https://stereoarts.jp/
16) 初音ミク@む〜ぶ Ver23, https://piapro.jp/t/R5Hj
17) OpenJTalk,https://open-jtalk.sp.nitech.ac.jp/
18) XRInteractionToolkit, https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@3.0/manual/index.html