顔情報を活用した二段階推論に基づくリアルタイム対話システムの構築と評価

Development and Evaluation of a Real-time Dialogue System Based on Two-stage Inference with Facial Information

三輪 大翔 [1]　　亀谷 由隆 [1]

Yamato Miwa[1]　　Yoshitaka Kameya[1]

[1] 名城大学 情報工学部

[1]Faculty of Information Engineering, Meijo University

Abstract: 本研究では，マルチモーダル情報を活用し，よりユーザに寄り添う対話システムの実現を目的とする．具体的には，ユーザの発話内容と表情・視線情報を統合し，客観的視点から適切な対話方針を推定・出力した上で応答を行う二段階推論によるリアルタイム対話システムを構築した．評価実験では，対話方針の生成による効果を検証し，会話の自然さや友好性には課題が見られた一方，共感性において評価が高い傾向が確認された．

1. はじめに

近年，大規模言語モデル（LLM：Large Language Model）の発展により，対話システムの応答品質は飛躍的に向上した．それに伴い，マルチモーダル処理能力の拡張や音声によるリアルタイム対話システムの構築などが進められており，より人間らしく高度な対話の実現が期待されている．その中でも，マルチモーダル情報の活用は，高度なインタラクションを実現する上で重要な要素の１つである．人間同士のコミュニケーションにおいては，発話内容といった言語情報のみならず，声のトーンや表情，視線，仕草といった非言語情報を踏まえて対話相手の意図を統合的に理解し，適切な応答を行っている．しかし，LLM を利用した現状の対話システムの多くは，依然としてテキスト情報のやり取りが主流である．システムは，表情などのマルチモーダル情報を活用して相手の状態を汲み取り，対話を円滑に進める必要がある [1].

そこで本研究では，マルチモーダル情報を活用し，よりユーザに寄り添った対話を行うシステムの実現を目的とする．この目的を達成するために，本研究では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 Remdis [2] と顔特徴量分析ツール OpenFace 3.0 [3] を用いて，ユーザの表情・視線情報を活用した対話システムを構築し，その有効性を検証するための評価実験を行う．本システムは，ユーザの発話内容と表情・視線情報を統合し，客観的な視点から適切な対話方針を推定・出力した上で応答を行う二段階推論を行う．これにより，マルチモーダル情報を活用してユーザの感情や意図を深く理解し，よりユーザに寄り添った高度なインタラクションを行うシステムの実現を目指す．

2. 関連研究

システムのより高精度なユーザ状態推定を実現するため，話者が発するマルチモーダル情報を活用する研究はこれまで進められてきた．岡田ら [4] は，話者の発話内容に加え，表情・音声の音響的特徴を利用し，各情報源から推定される感情が食い違うような複雑な心理状態をリアルタイムで検出するシステムを構築した．また，髙鍋ら [5] はマルチモーダル感情推定の精度向上のために，オンラインカウンセリング中のクライアントの，感情分析のためのマルチモーダルデータセットを構築した．このように，言語情報と非言語情報の双方を活用し，より正確にユーザ状態を推定するための研究が進められている．

人とシステムのより良い対話の実現に向け，マルチモーダル情報の活用のみならず，LLM を活用してユーザの心理や潜在的な意図を汲み取ろうとする研究も進められている．飯田ら [6] は，対話型生成 AI と他者モデルの統合により相手の意図を読む AI の実現を試み，これらの研究アプローチの可能性を示唆した．長澤ら [7] は，認知的共感性と情動的共感性という人間の共感プロセスをモデル化し，共感性

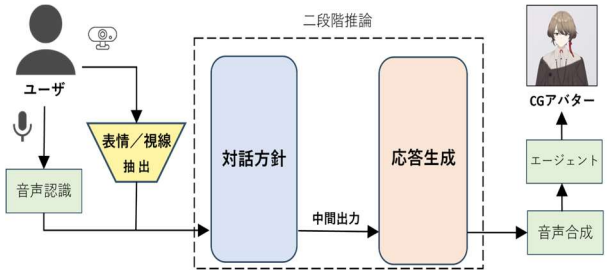

図 1　提案システム概要図

の高い応答生成対話システムを構築した．

以上の背景を踏まえ，本研究ではユーザが表出するマルチモーダル情報を活用し，相手の意図や気持ちを考慮したより良い対話を実現するリアルタイム対話システムについて検討する．

3. 提案システム

3.1 システム概要

本研究で提案するリアルタイム対話システムの特徴は，「表情・視線情報の利用」と，「対話方針・応答生成の二段階推論」にある．本システムの概要図を図 1 に示す．

システムは，Remdis を開発基盤とし，顔特徴量抽出モジュールとして OpenFace 3.0 を組み込むことで構築した．対話方針および応答の生成を行う二段階推論部には LLM を利用し，対話インタフェースには CG アバター「うか」[1] を使用した．また，使用する LLM については，東京科学大学の tokyotech-llm/Llama-3.1 Swallow-8B-Instruct-v0.5 を採用した．

本システムは，マイクから入力されたユーザ発話に対してシステムが音声応答を返す，ターン制の音声対話システムである．ユーザはマイクを通じて，画面上の CG アバターとの対話を行う．これに対して，CG アバターは音声応答に加え，表情やジェスチャを伴うマルチモーダルな応答を表出する．また，本システムはユーザの顔情報を利用して応答生成を行うため，対話中は Web カメラを用いてユーザの顔画像を逐次収集する．取得した顔画像は，OpenFace 3.0 によりリアルタイムで逐次解析され，表情・視線情報が抽出される．なお，これらの顔画像は解析終了後直ちに破棄するため，システム上に顔画像データを保存・蓄積しない．

[1] CG-CA Uka (c) 2023-2024 by Nagoya Institute of Technology, Moonshot R&D Goal 1 Avatar Symbiotic Society

3.2 顔情報の活用

本システムでは， OpenFace 3.0 を用いてユーザの顔画像から表情・視線情報を取得し，ユーザの発話内容とともに入力プロンプトへ組み込む．図 2 に入力プロンプトの例を示す．本研究においては，表情情報を「ユーザの表情」，視線情報を「ユーザ関心」としてプロンプトへ反映する．以下に，表情・視線情報の活用方法の詳細を述べる．

```
こんにちは
＝＝＝
ユーザの表情: Happy
ユーザ関心： 高
```

図 2　表情・視線情報を活用した入力プロンプト例

3.2.1 表情情報

表情情報は，ユーザの表情から推定される次の 8 つの感情カテゴリ [Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, Contempt] を活用し，これらの中から予測結果として出力された 1 つの感情カテゴリを「ユーザの表情： [選択された感情カテゴリ]」といった形式でプロンプトに記載する．また，表情の抽出タイミングについては，岡田ら [4] のシステムを参考に，ユーザの発話終了時点の表情を用いる．これらの表情情報プロンプト化の流れを図 3 に示す．

図 3　表情情報プロンプト化の流れ

3.2.2 視線情報

視線は，対話相手への関心を示す重要な指標となり得る．そこで本システムでは，直前の CG アバターの発話区間において，ユーザがアバターに視線を向けた割合を「注視率」として算出する．この割合を元に，ユーザの対話に対する関心度を「ユーザ関心」としてプロンプトへ反映する．

具体的な手順について述べる．OpenFace 3.0 の視線推定機能では，Web カメラに対する水平・垂直角度の視線ベクトル（例 : [0.0234, 0.0459]）が得られる．

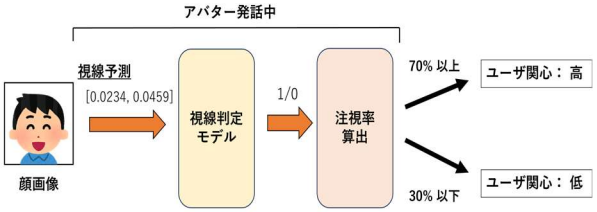

図 4　視線情報プロンプト化の流れ

本システムでは，この視線ベクトルを用いてアバターへの注視を判定する．はじめに，実験環境にて事前に収集したアバター注視時の視線データを用いて，入力された視線ベクトルに対してアバターを「注視している (1)」または「注視していない (0)」に分類する「視線判定モデル」を構築する．次に，アバターの発話中に取得された視線情報を逐次分類モデルへ入力し，取得された総フレーム数と，「注視している (1)」と判定されたフレーム数をカウントする．これらを基に，アバターの発話終了時，以下のように定義した注視率を算出する．

$$\text{注視率 (\%)} = \frac{[\text{アバター発話中の注視フレーム数}]}{[\text{アバター発話中の総フレーム数}]}$$

これらの結果，注視率が 70 %以上なら「ユーザ関心 高」，30 %以下なら「ユーザ関心 低」として入力プロンプトへ追記し，それ以外の場合はプロンプトへ反映しない．図 4 に，視線情報のプロンプト化の流れを示す．

3.3　二段階推論

本システムでは，「対話方針の決定」と「応答文の生成」からなる二段階推論を行う．本手法は，中間出力の生成により処理時間が約 1.5 秒増加する一方，ユーザやプロンプトに対するより深い推論を行う．まず，対話方針モジュール (LLM) は「対話方針アシスタント」として入力プロンプトに対して以下の手順で分析を行い，その結果を中間出力する．

手順 1： 文の感情

ユーザの発話内容を ＜ポジティブ，ややポジティブ，ニュートラル，ややネガティブ，ネガティブ＞ のいずれかに分類する．

手順 2： ユーザ心理

手順 1 の結果や入力プロンプトの情報を踏まえて，予測されるユーザの感情や意図などを 20 文字以内で言語化する．

手順 3： 対話方針

これまでの推論結果を踏まえて，次に CG アバターが取るべき対話方針を 20 文字以内で決定する．

文の感情：	ポジティブ
ユーザ心理：	挨拶をして、会話を始めたい
対話方針：	明るい口調で挨拶を返す

図 5　図 2 の入力プロンプトに対する分析結果

例として図 5 に，図 2 に示した入力プロンプトに対する分析結果を示す．

これらの分析結果を参考に，応答生成モジュール (LLM) が CG アバターの「うか」として，ユーザとより良い信頼関係を築けるような応答生成を行う．

3.4　対話例

本システムがどのような対話を行うかの具体例を示す．このとき，顔情報は入力するが二段階推論を行わず，応答を直接生成するシステムとの比較も行った．よって，提案手法を「二段階推論 あり」，比較手法を「二段階推論 なし」と表記する．

図 6 は，ユーザが笑顔でラーメンの話題を振った際の対話例である．このとき，ユーザは食べたラーメンが美味しかったことが推測される．出力結果を確認すると，比較手法ではユーザの嬉しそうな様子は捉えられているものの，ラーメンに対する感じ方までは推論できておらず味を尋ねる応答を生成した．一方，提案手法では，ユーザがラーメンについて話したいという期待を汲み取った上で話題を深掘りするなど，ラーメンを美味しいと感じたことを前提とした応答を生成した．この結果，ユーザの意図をより的確に反映した対話が確認された．

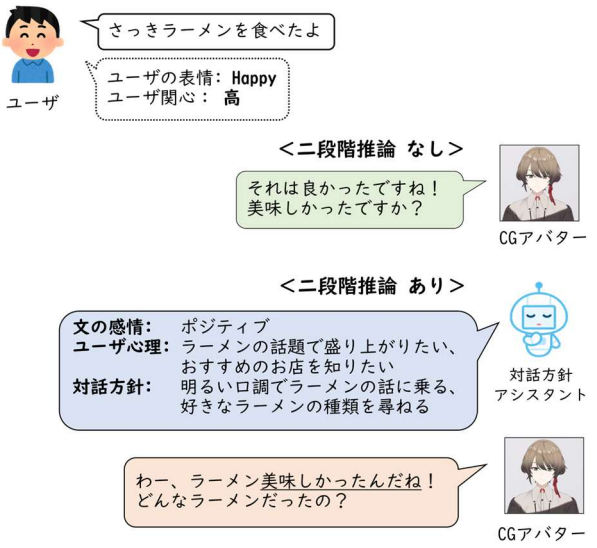

図 6　相手の意図を読んだ対話例

図 7 の対話例では，ユーザは忙しいことを報告する一方で，アバターへの視線が少なく，会話に集中できていない状況を想定した．このとき，比較手法は状況を直接尋ねるような応答を生成したが，提案手法は相手を気遣う対話方針に基づき，共感的な応答を生成した．この対話例では，応答の適切さはユーザの主観や状況に依存するものの，少なくとも相手の立場や感情を汲み取ろうとする「思いやり」のある応答が観測された．

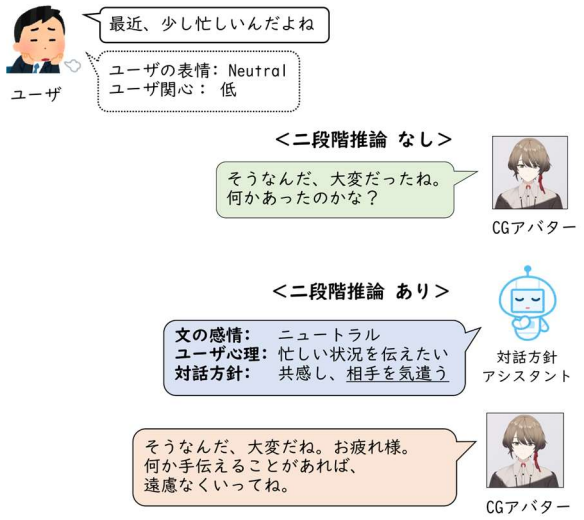

図 7　相手を気遣う対話例

4. 評価実験

4.1 実験設定

本実験では，提案システムの有効性を検証するために被験者 18 名を対象とし，3.4 節で述べた「二段階推論 なし」のシステムとの比較実験を行った．比較手法では，顔情報を入力として用いるが，二段階推論は行わず直接応答を生成する．

実験の流れについて述べる．本実験では，被験者は比較手法（二段階推論 なし）と提案手法（二段階推論 あり）の両システムと，話題を指定しない自由対話形式でそれぞれ 2 分間の対話を行った．このとき，実験結果への影響を考慮し，両システムの詳細は被験者には伝えず，内部的に異なる 2 種類の「システム A」「システム B」と伝えた．また，Web カメラを用いた顔画像の収集・応答生成のための一時的な利用については事前に同意を得た．本実験ではシステム A を比較手法，システム B を提案手法で固定し，被験者はシステム A から対話を始める．

これらの対話終了後，被験者は Google フォームを用いたアンケートに回答し，両システムの主観評価を行った．アンケートでは，システム A およびシステム B のそれぞれに対して，以下の 3 つの評価指標について 5 段階のリッカート尺度（1: 全くそう思わない ～ 5: 非常にそう思う）による評価を求めた．

（1）共感性
このシステムは，自分に対して真摯に向き合ってくれた（共感的だった）と思う

（2）自然さ
このシステムとの会話は，自然だったと感じた

（3）友好性
このシステムと，機会があればまた話したいと思う

また，これらの評価項目に加えて，各システムの印象について尋ねる自由記述式の設問も設けた．これにより，5 段階評価の数値データのみでは捉えきれない，ユーザが感じた応答の差異やシステム間の体験的な違いについても調査した．

4.2 実験結果

表 1　各評価項目の 5 段階評価の平均点

手法	共感性	自然さ	友好性
比較手法（二段階推論 なし）	4.278	**4.056**	**4.111**
提案手法（二段階推論 あり）	**4.389**	3.778	3.667

表 1 に本実験の結果を示す．表中の太字の数値は，各評価項目においてより高い評価を得た手法を示している．提案手法は，共感性の観点では比較手法を上回った一方，自然さ・友好性の観点では比較手法に劣る結果となった．

次に，得られた結果に対して差の有意性を検証するために，ウィルコクソンの符号付き順位検定を行った．まず，顔情報を活用した本システム（提案手法および比較手法）が，各評価項目において肯定的に評価されたかを検証した．有意水準 5%の片側検定により，5 段階評価の中間値（3）を基準とした 1 標本ウィルコクソンの符号付き順位検定を実施した結果，両手法のすべての項目で中央値 3 を有意に上回ることが確認された．これにより，本システムはいずれの項目においても，被験者から肯定的な評価が得られたといえる．

表 2　提案システムへの印象の自由記述で言及された提案手法におけるポジティブ／ネガティブ評価のコメントと件数

ポジティブ評価　［共感性］

カテゴリ	コメント（要約）	件数
共感的な会話内容	聞き返してくれて話しやすい，雑談を楽しむ印象	4
表情・ジェスチャ 表出	にこやかで好印象，動きがあって温かみがある	3

ネガティブ評価　［自然さ］

カテゴリ	コメント（要約）	件数
応答の不自然さ	機械感があった，論理的で堅苦しい印象，距離があった	4
応答遅延	答えるまでが長い，認識できてないのか不安に感じた	3
会話内容のズレ	質問を質問で返してきた，質問に答えてくれなかった	3
聞き間違い	少し聞き間違いがあった	1
その他	不自然だった （理由の記載なし）	1

続いて，両手法間の評価結果に統計的な差異があるかを検証するため，各評価項目について有意水準5%の片側検定にて2標本ウィルコクソンの符号付き順位検定を行った．この結果，「共感性・自然さ・友好性」のいずれの項目においても，統計的な有意差は認められなかった．

さらに，各システムに対する印象を問う自由記述式の設問について，提案システムに関する回答内容を分析した．具体的には，得られた記述内容を精査し，ポジティブ評価とネガティブ評価に分類した上で，カテゴリ別に件数を集計した．

これらの集計結果を表 2 に示す．提案手法に関する記述は 19 件確認され，そのうち 7 件がポジティブな評価，12 件がネガティブな評価であった．ポジティブな評価では共感性に関する肯定的意見が多く，「共感的な会話内容」や「表情・ジェスチャ 表出」が評価された．一方でネガティブな評価では自然さに関する課題が指摘され，「応答の不自然さ」や「応答遅延」が挙げられた．

4.3　考察

本実験では，「共感性・自然さ・友好性」の観点から提案システムの有効性を検証した．その結果，提案システムは共感性の観点では高い評価を得た一方で，自然さ･友好性には課題が残ることが分かった．この結果に基づき，本システムに対する考察を行う．

まず全体的な傾向として，顔情報を活用した両システムともに，それぞれの評価項目において，評価の中央値である 3 を有意に上回っていることが確認された．これは，本システムが表情・視線情報を適切に処理し，応答生成に活用できたことを示唆している．

次に，比較実験の結果に基づき，提案システムの評価について考察する．

まず，共感性の評価に着目すると，表 2 に示したような，「共感的な会話内容」や「表情・ジェスチャ表出」に関する肯定的な意見が見られた．これらは，話しやすさ・雑談を楽しむといった対話姿勢や，CG アバターのマルチモーダル表出の豊かさを評価したものだった．このことから，対話方針モジュールによる入力プロンプトの分析により，ユーザの気持ちや意図を汲み取った対話や，CG アバターの表情・ジェスチャといったマルチモーダル情報を考慮した応答生成ができたと考えられる．

一方で，表 2 で課題として挙げられた自然さの評価に着目すると，二段階推論による中間出力の課題であった「応答遅延」について，3 件の言及があった．比較手法と比べて応答までの時間が長く，これらの応答遅延が会話の自然さを下げる一要因になったとみられる．しかし，それ以上に多く挙げられた課題として「応答の不自然さ」があった．コメントでは，機械的な印象や，論理的で堅苦しい応答が指摘された．これらの原因としては，分析結果の付与によるプロンプト過多や，ユーザ心理の過度な推論が考えられる．これにより，応答が分析的・論理的になりすぎ，自然さが低下した可能性がある．

また，応答遅延と同頻度で報告された課題として「会話内容のズレ」があり，質問を質問で返してきた・質問に答えてくれなかったなどの質問無視が挙げられた．しかし，これらの課題は比較手法においても見られたことから，音声認識の段階における誤認識が原因であると推察する．これは，「聞き間違い」という課題についても同様の原因であると思われる．

以上の点から，本システムは「共感性」の観点では一定の効果が見られたが，応答内容の不自然さ・応答遅延などの「自然さ」が課題となり，結果として「友好性」が低下したと考えられる．

5. おわりに

本研究では，よりユーザに寄り添った対話を行うシステムの実現を目的に，顔情報を活用した二段階推論に基づく対話システムを構築した. 具体的には，ユーザの表情・視線情報を入力として活用し，対話方針の決定と応答生成という二段階の推論過程を導入した.

また，本システムの評価を行うため，被験者実験による比較実験を行った．結果としては，「共感性」の観点では高い評価を受けた一方,「自然さ･友好性」に課題が残る形となった．特に,「応答内容の不自然さ」と「応答遅延」については，今後改善が必要な課題となった.

今後の展望としては，ユーザの声のトーンや姿勢など，システムが取得できるユーザの非言語情報の拡張が考えられる．また，異なる LLM を使用した場合の比較実験の実施も今後の課題として挙げられる．本研究の結果を踏まえて，共感性と自然さの両立した，より高度なインタラクションを行う対話システムの実現に取り組みたい.

謝辞

本研究では，ムーンショット型研究開発「アバター共生社会」プロジェクトにおいて設計・開発された CG アバター「うか」を使用しました．また，本システムの実装においては，多くのオープンソースソフトウェアやライブラリの恩恵を受けました．これらの開発・公開に尽力されたすべての方々に深く感謝いたします.

参考文献

[1] 東中竜一郎：AI の雑談力，KADOKAWA，2021.

[2] 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，人工知能学会 言語・音声理解と対話処理研究会 (SIG-SLUD)，2023.

[3] J. Hu, L. Mathur, P. P. Liang, L. P. Morency: OpenFace 3.0: A Lightweight Multitask System for Comprehensive Facial Behavior Analysis, arXiv preprint, 2025.

[4] 岡田敦志，上村譲史，目良和也，黒澤義明，竹澤寿幸 : 表情・音響情報・テキスト情報からのリアルタイム感情推定システム, 第 31 回人工知能学会全国大会，1B1-OS-25a-4in1，2017.

[5] 髙鍋俊樹，松本和幸，木内敬太，康シン，西村良太，篠山学：感情分析のためのカウンセリングマルチモーダルデータセットの構築および評価，情報処理学会 第 86 回全国大会，5ZE-02，2024.

[6] 飯田愛結, 阿部将樹, 奥岡耕平, 福田聡子, 大森隆司, 中島亮一, 大澤正彦 : 意図を読む AI の実現に向けて: 対話型生成 AI と他者モデルの統合を例に, HAI シンポジウム 2024，G-28，2024.

[7] 長澤尚武, 萩原将文 : 日本語大規模言語モデルにおける発話意図の習得と共感対話システムの構築，日本感性工学会論文誌，Vol. 24，No. 4，pp. 345-353，2025.