公開

研 究 終 了 報 告 書

「マルチモーダルなふるまいに基づく音声対話の人間目標型評価」

研究期間: 2021 年 10 月～2024 年 3 月
研 究 者: 井上 昂治

1. 研究のねらい

大規模言語モデルが登場し、高度な対話応答の生成が一般の開発者にも容易に利用できるようになった。これにより、従来の音声対話システムの研究開発において対象とされてきた対話タスク・ドメインにおける対話の精度は飛躍的に向上し、会話ロボットや CG エージェントなどへの応用が加速している。このような背景をふまえ、今後の音声対話システムの研究開発においては社会的な対話の実現を見据えるべきである。社会的な対話とは、従来のシステムで扱われてきた対話タスク・ドメインとは異なり、対話の目的は共有されているが、そのゴール（達成）が明確には定義しづらいものである。例えば、傾聴やインタビューがこれに相当する。これらの対話は、現在の社会では、人間がその役割を担っているが、人手不足が今後深刻化していくなかで、音声対話システムで代替することが現実的な解として徐々に認知されていくと思われる。

社会的な対話を音声対話システムに実現させるためには、その研究開発において使用することができる適切な評価指標を確立することが重要である。しかし前述のとおり、社会的な対話はそのゴールが明確ではないため、客観的な評価指標を設定することは困難である。そのため、っこれまでの研究開発では、主観評価を多用せざるを得なかった。しかし、主観評価のみでは、研究の再現性や信頼性を担保するのは難しく、結果として研究成果の広がりが限定的になるという問題を分野全体において抱えていた。

そこで本研究では、音声対話システムを客観的に評価するための指標を確立することを目指した。具体的には、対話中に観測可能なユーザのふるまいに着目し、これに基づいて、システムを間接的に評価する。例えば、ユーザの発話量（どの程度話したか）は客観的に計測することが可能であり、この多寡はその音声対話の質を反映しているともいえる。ただし、どのようなふるまいがどのような場合に評価の手がかりになるかは自明ではない。また、評価したい項目については、従来の主観評価だけでなく、システムの「人間らしさ」も検討し、人間らしい自然な対話を最終的な目標とする評価指標を「人間目標型評価」と呼ぶことにする。以上を踏まえ、音声対話システムの人間らしさや主観評価と関係するユーザのふるまいを明らかにし、音声対話システムを客観的に評価するための新たな枠組みを提唱することを本研究のねらいとした。

2. 研究成果

（1）概要

本研究の最初の成果は、音声対話システムの「人間らしさ」と関係するユーザのふるまいを明らかにしたことである。まず、収録した傾聴対話データに対して、システムの「人間らしさ」のラベルを作成した。その後、この「人間らしさ」のラベルを目的変数、ユーザのマルチモーダルなふるまいを説明変数として、相関分析を実施したところ、発話量、単語の種類数、内容後の種類数、視線配布回数、平均交替潜時（システムが話し終えてからユーザが話し始めるまでの時間間隔）が、システムの人間らしさの程度と関係することがわかった。

公開

本研究の第二の成果は、上記の枠組みを複数の対話タスクへ拡張したことである。傾聴に加えて、就職面接、初対面会話も対象とし、対話タスクの特性に応じて、主観評価と関係するユーザのふるまいがどのように変化するかを明らかにした。SHAP（SHapley Additive exPlanations）を用いて分析したところ、傾聴と就職面接では、発話数、単語数、単語の種類数が主観評価と関係することがわかった。また、フィラー（「えっと」や「まー」などの場つなぎ的表現）や言い淀み（「ぐ、具体的には」などの言い直し表現）の数も関係することがわかり、よりフォーマルな場面である就職面接ではこれらが特に重要となることがわかった。一方で、初対面会話では、平均交替潜時のような相互作用的な要素が重要になることが明らかになった。

本研究の副次的な成果は、提案する評価指標の評価対象となる音声対話システムの要素技術についてである。まず、同調笑いの生成に取り組んだ。同調笑いとは、ユーザが笑ったときに、それに対してシステムも一緒に笑うことを指す。初対面会話のデータをもとに、同調笑いを生成するモデルを学習し、人間らしさ・自然さ・共感・理解といった評価項目が向上することを確認した。次に、ターテイキングシステムについても取り組んだ。ターンテイキングシステムとは、ユーザの発話権終了を的確に予測し、システムの発話を円滑に開始をするための技術である。本研究では、Transformer に基づく最新のモデルを拡張し、日本語・英語・中国語のマルチリンガルに対応するモデルを実現した。さらに、このモデルが CPU のみでリアルタイムで動作することも検証した。

上記の成果は、ACT-X における領域会議またサイトビジットでのコメント・議論を積極的に取り入れた結果得られたものである。特に、提案する評価の枠組みの位置づけや音声対話システムにとどまらない過去の研究との関係性を整理する際に大いに参考となった。加えて、これらの成果は、海外の研究者（主にスウェーデン王立工科大学）とも連携しながら得たものでもある。

（2）詳細

研究テーマ A 「音声対話システムの人間らしさを評価するためのユーザのマルチモーダルなふるまいの分析」

本研究の最初の成果は、音声対話システムの「人間らしさ」と関係するユーザのふるまいを明らかにしたことである。そして、本研究テーマを通じて、図1に示すユーザのマルチモーダルなふるまいに基づき音声対話システムを客観的かつ間接的に評価する枠組みを提唱した。

はじめに、本研究において使用する対話データの収録システムを構築した。このシステムでは複数のマイク・カメラを同期させながら録音・録画することができる。これにより、マルチモーダル対話データを容易に収録することが可能になった。そして、傾聴対話データ（70 対話）を収録した。

次にこの対話データに対して、システム側の映像・音声を隠蔽し、ユーザ側のふるまい（反応）からシステムが「人間」か「システム」のいずれであるかのラベルを複数の第三者に付与してもらった。そして、「人間」と判断された比率をそのシステムの「人間らしさ」の数値とした。なお、上記の対話データを1分程度の短い区間に切り出し、各区間に対して「人間らしさ」の数値を算出した。その結果、合計で 900 サンプル以上から成るデータセットを構築した。

その後、各サンプルにおいて、「人間らしさ」の数値とユーザのふるまいとの関係を分析した。その結果、発話・言語・視線のふるまい特徴とやや相関することがわかった。発話特徴は合計発話時間や平均ターン長、言語特徴は単語数やその種類数、視線特徴は視線配布回数などである。また、上記のふるまいから「人間らしさ」の数値を推定するモデルを構築・検証したところ、平均絶対値誤差は 0.146 となり、数値の刻み幅（0.200）よりも小さい誤差で推定できることがわかった。さ

公開

らに、上記の「人間らしさ」の数値とユーザによる主観評価の結果との関係性を調べたところ、「人間らしさ」の数値は、「システムはユーザの話を理解していた」や「ロボットは裏で人間が操作していたと思う」といった主観項目とやや相関することがわかった。

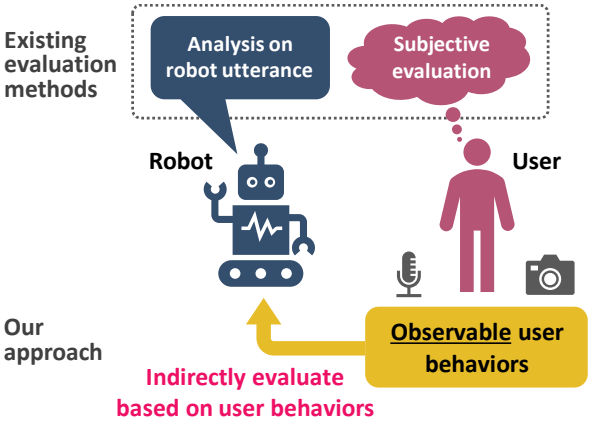

図1　本研究で提案した評価指標の枠組み

研究テーマ B 「音声対話システムの主観評価と関係するユーザのふるまいの分析」

本研究の第二の成果は、研究テーマ A の枠組みを複数の対話タスクへ拡張したことである。傾聴、就職面接、初対面会話といった性質が異なる対話タスクのそれぞれにおいて、主観評価と関係するユーザのふるまいを明らかにした。これにより目的の対話タスクの性質に応じて、比較すべきユーザのふるまいを効率的に選択することが可能になる。

まず、傾聴、就職面接、初対面会話のそれぞれにおける主観評価の平均値を正解ラベル、対話中のユーザのふるまいを特徴量として、XGBoost モデルを学習した。この学習したモデルに対して SHAP（SHapley Additive exPlanations）を適用した。この分析では、特徴量ごとに SHAP 値と呼ばれる指標を計算する。これは、ある特徴量を用いる場合の出力と用いない場合のモデルの出力の差分を、その特徴量を除くすべての特徴量の部分集合について平均したものであり、SHAP 値が大きいほどその特徴量がモデルに対して大きな影響を与えると解釈することができる。

表1　各対話タスクにおけるふるまいごとの SHAP 値

ふるまい	傾聴	就職面接	初対面会話
1. 発話時間/分	0.098	**0.212**	0.043
2. 発話数（IPU数）/分	**0.199**	**0.124**	**0.126**
3. 発話単語数/分	**0.162**	**0.145**	0.070
4. 発話単語の種類数/分	**0.242**	**0.183**	0.068
5. 発話内容語数/分	0.048	0.087	0.016
6. 発話内容語の種類数/分	0.066	0.042	0.024
7. 相槌数/分	0.050	0.067	0.062
8. フィラー数/分	**0.111**	**0.187**	0.047
9. 笑い数/分	0.093	0.042	0.016
10. 言い淀み数/分	**0.129**	**0.224**	**0.192**
11. 平均交替潜時	0.078	0.097	**0.126**

SHAP により表1の結果を得た。傾聴と就職面接では似たような傾向を得た。例えば、発話数、単語数、単語の種類数の SHAP 値が大きかった。ユーザの発話がメインとなる傾聴と就職面接においては、このような発話の量やバリエーションに関するふるまいが重要になることは妥当であるといえる。フィラー(「えっと」や「まー」などの場つなぎ的表現)や言い淀み(「ぐ、具体的には」などの言い直し表現)の SHAP 値も高いが、就職面接のほうがより顕著である。これは就職面接のフォーマルさに起因していると推測される。一方で、初対面会話では異なる傾向がみられ、特に平均交替潜時(システムが話し終えてからユーザが話し始めるまでの時間間隔)の SHAP 値が高くなった。初対面会話は頻繁に発話権や対話の主導権が入れ替わるため、交替潜時のような相互作用的なふるまいが重要になると考えられる。以上より、対話タスクの性質に応じて、異なるユーザのふるまいがその主観評価と関係することがわかった。

研究テーマ C 「音声対話システムによる共感表出のための同調笑い生成」

本研究の副次的な成果として、音声対話システムによる同調笑いの生成を実現した。これは音声対話システムの要素技術と位置付けられ、本研究において提案する評価指標の評価対象となることを想定したものである。

以前の研究において、傾聴対話システムの研究開発を実施してきたが、そこで明らかになったのが「共感性の不足」である。つまり、相手(ユーザ)の話を表面的には聞いているが、共感のような深いレベルで聞くことはできていなかった。そこで、共感を表出するふるまいとして「笑い」に着目した。ただし、すべての笑いを対象とすると難易度が高くなるため、本研究ではユーザが笑ったときにシステムも一緒に笑う「同調笑い」に焦点を定めた。

同調笑いの生成を実現するために、図2に示す3つのモジュールで構成されるシステムを提案した。これらは、ユーザの笑いの検出、同調笑いの有無の予測、システムの笑いの種類の選択である。この各モジュールを、初対面会話 82 対話分のデータから学習し、生成された同調笑いについて第三者による評価を実施したところ、同調笑いを使用することで、人間らしさ・自然さ・共感・理解といった項目について、システムに対する評価が向上することがわかった。また、このシステムはリアルタイムで動作することができ、対話デモ動画が公開されている。

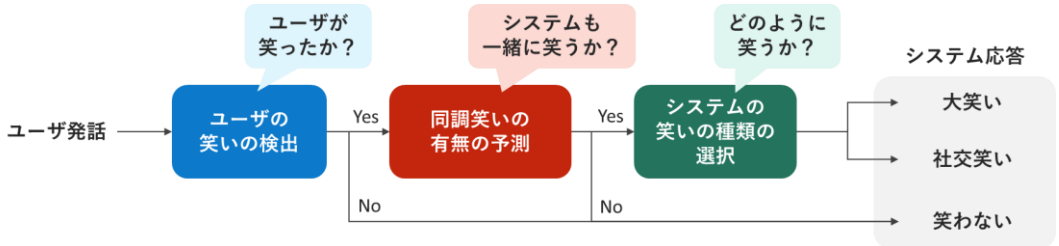

図2　同調笑いを生成するシステムの構成

研究テーマ D 「Voice Activity Projection を用いたマルチリンガルターンテイキングシステム」

さらなる副次的な成果として、ターンテイキングシステムにも取り組んだ。ターンテイキングシステムとは、ユーザの発話権終了を的確に予測し、システムが円滑に発話を開始するための技術である。本研究では、Transformer に基づく Voice Activity Projection(VAP)モデルをベースとした。VAP は図3に示すように、2つの Transformer が Attention を交差させるように構成されており、こ

れを Cross-attention Transformer と呼ぶ。このような構成にすることで、音声対話に参与している2名の交互作用がモデル化されることが期待される。また、VAP は連続的に次の音声アクティビティを予測するモデルであるため、従来の発話単位（ユーザの発話が終了してから、発話全体に対してターンテイキングの予測を行う）に比べて、より高速にかつ柔軟にターンテイキングの予測を行うことができる。

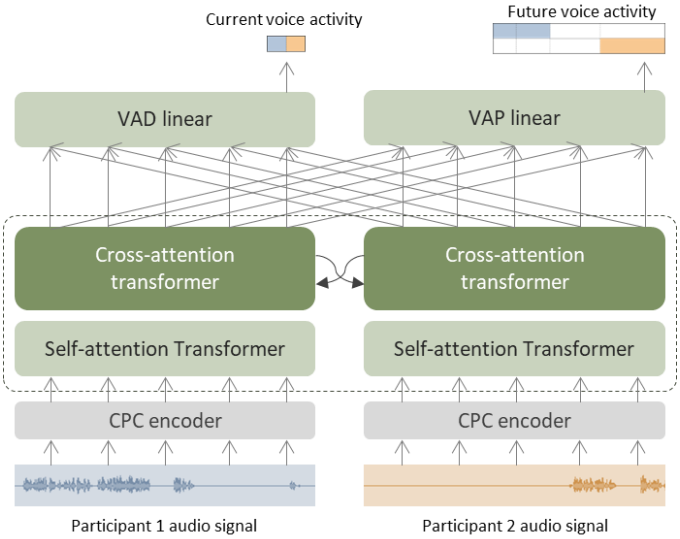

図3 Voice Activity Projection（VAP）モデルの構成

この VAP モデルは、従来は英語のみの音声対話で動作していたが、これを日本語でも動作するように拡張した。具体的には、単一の VAP モデルに対して、英語・日本語・中国語の混合データを用いて学習することで、すべての言語において頑健に動作することを確認した。図4に VAP の日本語での動作例を示す。この例では「青」から「橙」へ交互に発話権が移行しており、下部には各時間における約 600 ミリ秒後の話者についての VAP モデルによる予測を表している。この例から、発話権の移行が生じる少し前にそれらを適切に予測できていることがわかる。

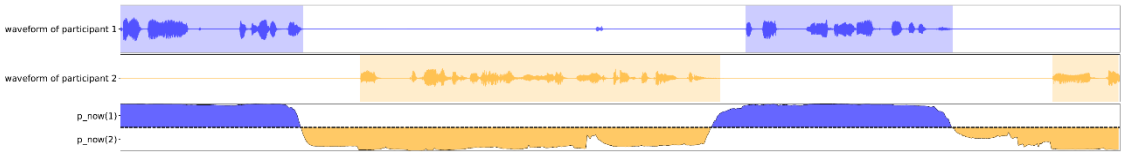

図4 VAP モデルの日本語音声対話での動作例

さらに、この VAP モデルが CPU のみでリアルタイムで動作することを検証した。具体的には VAP モデルの Transformer に入力する系列長とターンテイキングの予測精度の関係性を調査した。その結果、入力長を 1 秒程度（VAP の標準は 20 秒）にしても、予測精度が低下することなく、CPU 環境でもリアルタイムで動作することを発見した。

3. 今後の展開

まず今後1～2年において、本研究で提案した音声対話システムの客観的な評価の枠組みを、他の研究開発においても適用し、その有効性を検証していく。その際に、対象とする対話タスク、

公開

主観評価項目、ユーザのふるまいを多様化させ、汎用性を高めていく必要がある。その後、3～5年のスパンにおいて、対話タスク・主観評価項目の組合せを網羅する形で、客観評価に有効なユーザのふるまいを選定するためのガイドラインを策定し、提案する評価の枠組みを研究分野全体に浸透させていく。

上記と並行して、音声対話システムの要素技術として、Voice Activity Projection（VAP）モデルの発展にも取り組む。このモデルを基盤として、音声対話における様々なシステムのふるまいを生成することが期待される。そのためにも大規模な音声対話学習データの構築ならびにアノテーションが必要であり、これに今後1～2年で取り組む予定である。その後、様々なふるまいの生成を検証し、5年後には実環境において頑健に動作する基盤技術を整備することを目指す。

4. 自己評価

音声対話システムの人間らしさを客観的に評価するという本研究の目的は達成することができた。また、評価指標を汎用的なものにするために、対話タスクの性質に応じた評価の手がかりとなるユーザのふるまいも明らかにすることができた。さらに、当初は想定していなかったが、音声対話システムの要素技術についても成果を得ることができた。

本研究での取り組みは音声対話システムの評価指標についての新たな枠組みを研究分野の内外に提唱するものである。国内外の学会発表ならびに招待講演を通じて、多くの研究者から同意ならびに発展的なコメントを得ることができた。また、分野外の人々からも主にACT-Xでの領域会議ならびにサイトビジットにおいて多角的なフィードバックを得ることができた。さらに、国際的な賞の受賞や報道発表を実施するなど研究内容について社会に広くアピールすることができた。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：7件

1. Koji Inoue, Divesh Lala, Keiko Ochi, Tatsuya Kawahara, Gabriel Skantze, “Towards Objective Evaluation of Socially-Situated Conversational Robots: Assessing Human-Likeness through Multimodal User Behaviors,” International Conference on Multimodal Interaction (ICMI) Companion Proceedings, 2023, pp. 86-90.
音声対話システムを客観的に評価するための枠組みとして、ユーザのふるまいに着目する手法を提案した。本論文では、傾聴対話データを用いた会話ロボットの「人間らしさ」のアノテーション、およびふるまいとの関係性についての分析結果を報告した。
2. Koji Inoue, Divesh Lala, Keiko Ochi, Tatsuya Kawahara, Gabriel Skantze, “An Analysis of User Behaviors for Objectively Evaluating Spoken Dialogue Systems,” International Workshop on Spoken Dialogue Systems Technology (IWSDS), 2024.
論文 1 の発展として、複数の対話タスク（傾聴、就職面接、初対面会話）において、ユーザの主観評価と関連するふるまいについての分析結果を報告した。
3. Koji Inoue, Divesh Lala, Tatsuya Kawahara, “Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue,” Frontiers in Robotics and AI, 2022.
ユーザが笑った時にシステムも一緒に笑う「同調笑い」を生成するシステムについて提案し

公開

た。提案システムを傾聴対話システムに統合し、システムの人間らしさや共感性といった評価項目が向上することを確認した。

（2）特許出願

研究期間全出願件数：0 件

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

- 招待講演
  - 国際会議　1件 （IEEE RO-MAN 2023 Workshop）
  - 国内研究会　1件 （日本音響学会 音声研究会 兼 電子情報通信学会 ヴァーバル・ノンヴァーバル・コミュニケーション研究会）
- 受賞
  - NETEXPLO Innovation 2022 Award Winner
    - NETXEPLO は世界中のデジタル技術に関する独立調査組織であり、UNESCO のパートナーでもある。本賞は、2022 年に世界で発明されたもののうち、未来を変える可能性がある最も刺激的な 10 件を選定するものである。本研究における成果のうち、同調笑いの生成（研究テーマ C）が受賞の対象となった。
  - Frontiers in Robotics and AI, 2022 Outstanding Article
    - 本賞は、国際論文誌 Frontiers in Robotics and AI において 2022 年に投稿された論文のうち顕著な成果をあげたものを表彰するものである。本研究における成果のうち、同調笑いの生成（研究テーマ C）に関する論文が受賞の対象となった。
- プレスリリース
  - 京都大学「人と一緒に笑う会話ロボットを開発―人に共感し、人と共生する会話 AI の実現に向けて―」
    - 本研究における成果のうち、同調笑いの生成（研究テーマ C）に関する内容であり、その後、国内外のメディアにおいて多数報道された。