Vol.2021-SLP-135 No.2
2021/2/18

ERATO 石黒共生 HRI から新学術領域研究「知能対話学」及びムーンショット「アバター共生社会」に展開する音声対話研究

河原達也 [1]

概要：今年度まで行われてきた ERATO 石黒共生ヒューマンロボットインタラクション(HRI)プロジェクトにおける音声対話研究の成果を報告するとともに、昨年度に開始された科研費新学術領域研究「人間機械共生社会を目指した対話知能システム学（知能対話学）」、及び今年度開始のムーンショット型研究開発事業に採択された「誰もが自在に活躍できるアバター共生社会の実現（アバター共生社会）」への展開について紹介する。

1. はじめに

音声対話システムは、スマートフォンアシスタントやスマートスピーカなどで身近なものとなった。これらで行われている対話は､情報検索や機器操作などのタスクであり、機械とのインタフェースという位置付けである。ユーザはシステムができるタスクを理解した上で、単純な文を明瞭に発声する必要がある。雑談の機能もあるが、基本的に一問一答形式で、データベースを検索して応答を生成するシステムが大半である。これは、人間どうしの対話とは大きな違いがある。人間どうしの対話では、１つのターンでたくさん話す一方で､聞き手は相槌をうつのが一般的である。

我々は､JST ERATO 石黒共生ヒューマンロボットインタラクション(HRI)プロジェクト（2014～2020 年度）において､アンドロイド ERICA を用いて、人間どうしが行うような長く深い音声対話に取り組んできた[1,2]。ここでは、音声対話を手段でなく、雑談でもなく、具体的な社会的状況を想定し、それ自体が明確な目的を有したものとして捉えている。状況・タスクは設定するが、ユーザは何を話してもよいオープンドメインの条件で、人間レベルのコミュニケーション能力の実現を目指したものである。

昨年度から、より多くの研究者と協力して、本研究を発展させるために、科研費新学術領域研究「人間機械共生社会を目指した対話知能システム学（知能対話学）」（2019～2023 年度）を開始した。本研究では、様々な観点から音声対話ロボットに関する研究に取り組んでいる。

さらに今年度、ムーンショット型研究開発事業の目標１において､「誰もが自在に活躍できるアバター共生社会の実現（アバター共生社会）」（2020～2025 年度予定）が採択され、参画することになった。本プロジェクトでは、システムだけで対話を行うのではなく、人間がアバターを用いて（同時に複数かつ遠隔で）行う対話をシステムが補助するという立場を想定している。全体としては、人間が行うインタラクションの実現を目指すが、システムはすべてを行う必要はないという点で、従来とは異なる枠組みである。

2. JST ERATO 石黒共生ヒューマンロボットインタラクション(HRI)プロジェクトの成果 a

2.1 タスク

2014～2020 年度において行った本プロジェクトにおいて我々は、アンドロイドに適した具体的な社会的状況を想定し、以下のタスクを設定した。このように状況・タスクは設定するが、ユーザは何を話してもよいオープンドメインの対話を目指した。これらの比較を表 1 に示す。

- 傾聴 [3,4]

高齢者の方に、印象に残った旅行や最近食べたものなどの話題について話してもらう。ERICA は、相槌や聞き返しを含めて、的確にフィードバックを行うことで、相手に円滑に話し続けてもらうことを目指している。

- 就職面接（練習）[5,6]

ERICA は面接官の役割をし、志望動機やスキルなどについて、相手の応答をふまえて掘下げ質問を生成して、情報を引き出すとともに、実際の面接のシミュレーションとなることを目指している。

- 研究紹介 [7]

研究室を訪問した様々な来訪者に対して研究紹介を行う。相手の理解度や興味に応じて、説明の内容や話し方を切り替えるようにできることを目指す。

- お見合い（練習）[8]

伝統的なお見合いでなく、婚活イベントなどでの対話を想定し、ERICA は一方（女性）の参加者の役割を担う。趣味や好きな食べ物などの話題について､ユーザに質問したり、ユーザの質問に答えるとともに、対話に応じたフィードバックを適宜行う。実際のお見合いのシミュレーションとなることを目指す。

表１　対話タスクの比較

	傾聴	就職面接	研究紹介	お見合い
システムの役割	聞く	尋ねる	話す	すべて
対話の主導権	ユーザ	システム	システム	混合
発話の大半	ユーザ	ユーザ	システム	両方
相槌の大半	システム	システム	ユーザ	両方
発話権交替	少ない	明確	明確	複雑

1 京都大学　情報学研究科
Kyoto University, School of Informatics

a https://www.jst.go.jp/erato/ishiguro/

ⓒ2021 Information Processing Society of Japan

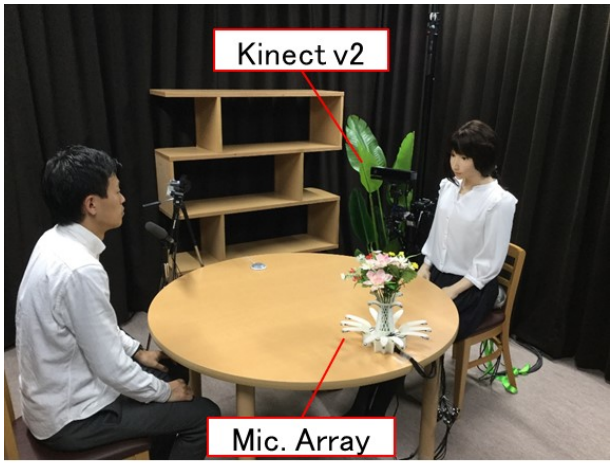


図 1　アンドロイド ERICA との対話

2.2 主要技術

これらのシステムを実現する上で、本研究で取り組んだ主な技術は以下である。

- 高速・高精度の話し言葉音声認識 [9]

クライアント PC で動作する単語単位の End-to-End 型のモデルにより、実時間の百倍近い速度の音声認識を実現した。これは 10 秒の音声を 100ms で処理できることに相当する。

- 相槌生成 [10,11]

「うん」「はい」や「ふん」「へー」などの相槌は、「聞いている」「理解している」といったフィードバックの機能・効果がある。したがって、長い発話が行われる対話では、聞き手（システム）は話し手（ユーザ）の発話中に適切な相槌をうつことが重要である。本研究では、多様で自然な相槌を的確なタイミングで生成するモデルを実現した。

- 円滑なターンテイキング [12,13]

一問一答型の対話システムと違って、長い発話が行われる対話ではターンの区切りは明確でないが、発話の終了をすばやく検出して、円滑にターンを切り替えることが重要である。本研究では、TRP（対話移行適格場）を定義した上で、再帰型ニューラルネットワークと確率的な状態遷移モデルを組み合わせて実現した。

- エンゲージメント検出 [14]

システムにより長い対話を実現する上で、ユーザがどの程度対話継続の意欲があるかを示すエンゲージメントを検出することが必要になる。相槌・笑いや視線などのマルチモーダルな振る舞いに基づいて検出するモデルを実現した。

2.3 傾聴対話システム [3,4]

傾聴対話システムは、聞き手として相槌などのフィードバックを的確に行うことで、「理解」「共感」を示しながら、話し手の語りを促進することを目指す。具体的な応用例として、高齢者や入院患者に対する傾聴が想定される。人間のような相槌を生成する上で ERICA のようなアンドロイドが自然である。

本システムでは、ユーザ発話中の 100 ミリ秒毎に、その時点から 500 ミリ秒以内に相槌をうつか否かを韻律情報を用いてロジスティック回帰モデルにより予測している。これに加えて、音声認識・言語理解を用いて、繰り返し、掘り下げ質問、評価応答、語彙的応答、語彙的応答（極性）、話題提供などの多様な聞き手応答を生成する。

応答生成のための言語理解として感情分析と焦点語の抽出を行う。感情分析は、発話中の単語について既存の極性辞書を用いてポジティブ／ネガティブを判定する。この結果に基づき評価応答や語彙的応答（極性）を生成する。評価応答は、ユーザ発話がポジティブの場合には、「いいですね」または「素敵ですね」、逆にネガティブの場合には、「大変ですね」または「残念でしたね」となる。語彙的応答（極性）は、「そうですか」などの定型表現の韻律パターンをポジティブ／ネガティブに応じて変化させる。焦点語は、文末により近い名詞または形容詞とし、これに基づいて繰り返しや掘り下げ質問を生成する。繰り返しは、オウム返しのように焦点語を繰り返すものである。掘り下げ質問は、焦点語に「どんな」、「どの」、「なんの」、「どこの」、「いつの」、「だれの」などの疑問詞と組合せ、N-gram 確率が閾値以上かつ最大のものを出力する。対話例を示す。

対話例（U: ユーザ、E: ERICA）
E1: これまでで一番印象に残っている旅行についてお話を聞かせてください。
U1: 去年の春に京都にある公園に行きました。
E2: どの公園ですか。（掘り下げ質問）
U2: 八坂にある円山公園です。
E3: 円山公園ですか。（繰り返し）
U3: 有名な枝垂れ桜の周りを散歩しました。
E4: 散歩ですか。（繰り返し）
U4: とても綺麗で楽しかったです。
E5: いいですね。（評価応答）

多数の高齢者による対話実験において、５分程度の対話を破綻なく実現している。また、主観評価においても、話しやすさや話を聞いていた印象は WOZ（オペレータによる対話）と同程度の結果となっている。

3. 新学術領域研究「知能対話学」[b]

科研費・新学術領域研究「知能対話学」は、上記の ERATO プロジェクトをさらに発展させ、人間と知能システムやロボットとの共生社会を実現するための新たな学術領域を創成するために提案された。2019 年度に採択され、５年間の予定で研究を行っている。

b https://www.commu-ai.org/

ⓒ2021 Information Processing Society of Japan

以下の４つの班から構成される。

A01 人間との対話継続及び関係構築のための対話知能システム

A02 モジュール連動に基づく対話システム基盤技術の構築

A03 人と社会的に共生する対話システムのための行動決定モデル基盤技術の確立

A04 対話知能システムの研究開発及び社会実装のための法社会規範の研究

A01 班では、対話を継続し、関係を構築するための方法論について研究を行っている。人間が日常的に対話する際には、何か目的があるよりも、対話そのものあるいはそれによるコミュニケーションや関係の維持が目的であることが多い。コミュニケーションを行うロボットにもこの実現が求められる。本研究では、特に高齢者とのコミュニケーションを想定して、対話を継続させるためのメカニズムの解明と実現を目指す。高齢者との対話においては、音声認識に高い精度は期待できず、また意図理解も困難な場合がある。それでも、相槌・頷き・表情などのモダリティを活用したり、複数の人間・ロボットで会話の場を形成することで、対話を維持することは可能である。本研究では、このような観点から以下の課題に取り組む。

A) 複数モダリティの応答と複数ロボットのターンテイキングによる対話継続機構

B) 複数人対話における対話継続による人間関係構築支援の実現

C) 対話継続・人間関係構築支援に基づく対話サービスの構築と実証実験

これにより、「どのような対話プロトコルによれば、対話内容が理解できなくても対話継続が可能か、あるいは、意図理解が不十分であっても対話継続を通じた関係構築が可能か」を明らかにし、「このような頑健な対話プロトコルによって、どのような対話サービスが実現できるか」を実証する。

ERATO 石黒共生 HRI プロジェクトの成果を発展・統合し、さらに複数人を相手とする会話に適用することで、より一般的な対話メカニズムの解明と実現を目指す。

4. ムーンショット「アバター共生社会」

ムーンショット型研究開発事業は 2020 年度に設定され、目標 1 では、「2050 年までに、人が身体、脳、空間、時間の制約から解放された社会の実現」が掲げられた。

これに対して、石黒教授の提案「誰もが自在に活躍できるアバター共生社会の実現」が採択され、2020 年 12 月から開始した。本プロジェクトにおいても音声対話研究が重要な位置を占めている。

コロナ禍で「遠隔〇〇」が普及したように、空間的な制約からの解放だけであれば、テレビ会議システムやアバターを使えばある程度可能である。しかし、時間的な制約を打破するには、同時に複数のタスク・対話を行う必要がある。そのためには一部（多く）を自動化する必要がある。これを、音声対話システムの研究から捉えると、すべてをシステムで処理・応答する必要はなく、人間でないと難しい部分は人間に頼ってもよい設計になる。すなわち、自律でできる部分（定型的な紹介や質問）は自律で行い、難しい対応や人間関係の構築は人間が遠隔で行う半自律システムの実現を目指す。例えば、単純計算でも、80%を自動化できれば､同時に５人と別々に対話を行える可能性がある。ここで重要なのは。すべて人間が操作しているのと同等の性能・満足感を実現することである。

グループ２では、このような「自在遠隔音声対話の研究開発」を行っている。本研究開発課題では、音声認識・対話技術を発展させて、ホスピタリティの感じられる自律的な音声対話システムを傾聴対話や面接・面談・接客等で実現するとともに､アバターが操作者の意図や状況に応じて、遠隔操作対話と自律対話を臨機応変に切り替えられる自在対話技術を開発する。

そのためにまず、店舗・病院・教室や展示場・ホールのように多数の人がいる環境（実環境）で、目的となる話者の音声を強調・検出し、自動認識する技術の研究に取り組む。また、目的話者が２名以上の場合も想定し、話者の分離・同定を含めた枠組みについても研究を行う。

次に、人間のようにホスピタリティの感じられる自律的な音声対話システムの研究に取り組む。アンドロイド ERICA を対象に実装してきた相槌生成に加えて、共感を示す反応や共有笑いの生成などについて研究を行う。対話相手が２名以上の複数話者会話も想定して、これらの相槌・反応の生成に加えて、ターンテイキングなどの対話制御についても研究を行う。

さらに、自律対話と遠隔操作対話を融合した自在対話の制御機構の研究に取り組む。定型的な紹介や受け答えは自律で行い、人間関係の構築や自律での応答が難しい部分は人間が遠隔で行うことで、１人の操作者が複数のアバターを用いた自在対話を実現する。そのために、音声理解・自律対話の破綻検出と対話継続の意欲（エンゲージメント）の認識に基づいた切替えとともに、自律で行っていた対話の要約提示を行う。

高齢者の傾聴対話、就職面接対話、相談・カウンセリング、接客・営業、案内・プレゼンテーション、教育などの応用場面を想定して、システムを開発する。被験者実験による評価を行うことで、すべて人間が操作しているのと同じ性能・満足感の実現を目指す。システムの構成を図２に示す。

ⓒ2021 Information Processing Society of Japan

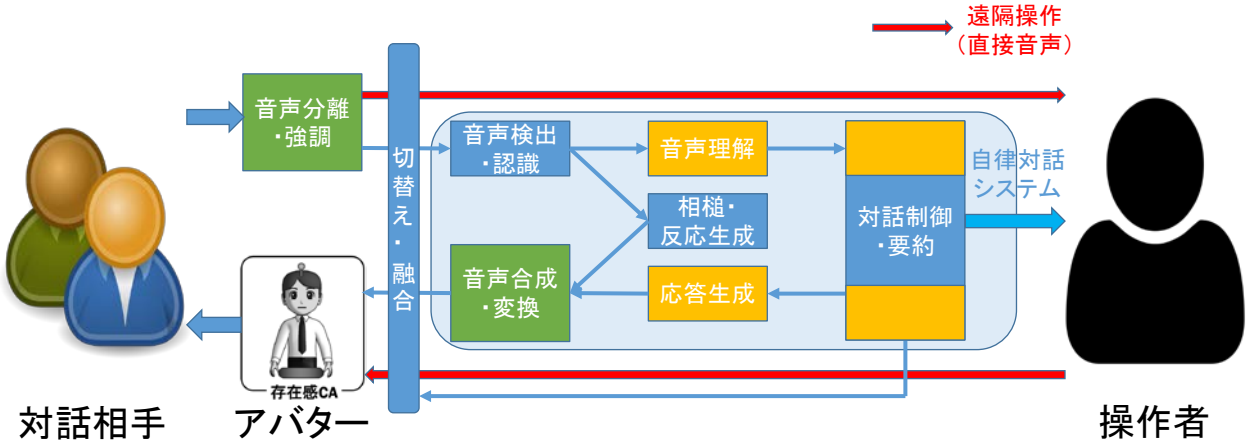


図２　自在遠隔音声対話システムの構成

5. おわりに

ロボットとの音声対話は、十年ほど前には、発話検出できない、音声認識できない、ターンテイキングできない（発話が衝突する）といった問題で、絶望的な様相であった。Push-to-Talk やマジックワードを用いることなく、自然に対話ができるようになったのは大きな進歩と考える。

本学のカウンセリングルームの杉原保史教授に傾聴システムを体験してもらったところ、2018 年は「子どもと話しているようだ」という評価であったが、今年度は「子ども感はなくなったが、少し変な人」ということで、人間レベルに近づいていると思われる。しかしながら、対話していて、とても話を理解しているようには思えず、「共感」も感じられない。高齢者の主観評価でも同様であった。

その反面、人間とは違って人工物であるロボットだから話しやすいという人が一定程度いるのも事実で、ロボットだからあえて空気を読む必要はないという意見もある。このように、ロボットによる対話研究は新たなステージに入っている。

謝辞: これらの研究に貢献して頂いた皆様に感謝します。

参考文献

[1] 井上昂治, 河原達也. アンドロイドを用いた音声対話研究. 日本音響学会誌, Vol.76, No.4, pp.236--243, 2020.
[2] T.Kawahara. Spoken dialogue system for a human-like conversational robot ERICA. In Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS), (keynote speech), 2018.
[3] 井上昂治, ララディベッシュ, 山本賢太, 中村静, 高梨克也, 河原達也. アンドロイドERICAの傾聴対話システムにおけるWOZとの比較評価. 人工知能学会研究会資料, SLUD-C002-21, 2020.
[4] K.Inoue, D.Lala, K.Yamamoto, S.Nakamura, K.Takanashi, and T.Kawahara. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In Proc. SIGdial Meeting Discourse & Dialogue, pp.118--127, 2020.
[5] 井上昂治, 原康平, ララディベッシュ, 山本賢太, 中村静, 高梨克也, 河原達也. 掘り下げ質問を行う就職面接対話システムの 自律型アンドロイドでの実装と評価. 人工知能学会論文誌, Vol.35, No.5, pp.D--K43_1--10, 2020.
[6] K.Inoue, K.Hara, D.Lala, K.Yamamoto, S.Nakamura, K.Takanashi, and T.Kawahara. Job interviewer android with elaborate follow-up question generation. In Proc. ICMI, pp.324--332, 2020.
[7] K.Inoue, D.Lala, K.Yamamoto, K.Takanashi, and T.Kawahara. Engagement-based adaptive behaviors for laboratory guide in human-robot dialogue. In Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS), 2019.
[8] 田中滉己, 井上昂治, 中村静, 高梨克也, 河原達也. 多様なふるまいからの好感判定に基づき発話構成要素を選択する お見合い対話システム. 人工知能学会研究会資料, SLUD-C002-28, 2020.
[9] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In Proc. IEEE-ICASSP, pp.5804--5808, 2018.
[10] 山口貴史, 井上昂治, 吉野幸一郎, 高梨克也, NigelG. Ward, 河原達也. 傾聴対話システムのための言語情報と韻律情報に基づく 多様な形態の相槌の生成. 人工知能学会論文誌, Vol.31, No.4, pp.C--G31_1--10, 2016.
[11] T.Kawahara, T.Yamaguchi, K.Inoue, K.Takanashi, and N.Ward. Prediction and generation of backchannel form for attentive listening systems. In Proc. INTERSPEECH, pp.2890--2894, 2016.
[12] D.Lala, K.Inoue, and T.Kawahara. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In Proc. ICMI, pp.226--234, 2019.
[13] K.Hara, K.Inoue, K.Takanashi, and T.Kawahara. Turn-taking prediction based on detection of transition relevance place. In Proc. INTERSPEECH, pp.4170--4174, 2019.
[14] K.Inoue, D.Lala, K.Takanashi, and T.Kawahara. Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. APSIPA Trans. Signal & Information Process., Vol.7, No.e9, pp.1--16, 2018.

ⓒ2021 Information Processing Society of Japan