2021年度 グループ２ 活動報告

研究開発項目２：自在音声対話の研究開発

研究開発課題１：自在遠隔音声対話の研究開発

（１）実環境下における音声検出・認識

CA の利用が想定される環境において、周囲の雑音や BGM などから、目的となる話者の音声を強調・検出した上で、自動音声認識を行う。本研究では、音声認識の頑健化に注力するとともに、音声強調・検出と認識を統合した End-to-End 処理系の深層学習による最適化を行った。話し言葉に BGM を-5dB から 5dB 重畳させた音声に対して、平均 84%の音声認識率を実現した。また、雑音を 0dB から 20dB 重畳させた音声に対して、平均 83%の音声認識率を実現した。ロボットとの傾聴対話で収集した高齢者の音声については、平均 66%の音声認識率を実現した。今後（コロナ禍終了後）、高齢者の対話データをさらに収集して改善に努める予定である。

（２）人間レベルの自律音声対話

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に向けた研究開発を行った。これまで、アンドロイド ERICA を対象に実装してきた自然で多様な相槌生成に加えて、相手への共感を示すための感情価（ポジティブ／ネガティブ）を含む反応（「へー」や「そうなんですね」など）や、場をなごませるための共有笑いの生成を実装した。傾聴対話における応答生成については、90%以上に問題がなく、80%以上でユーザから反応が示された。説明・プレゼンテーション対話の予備実験においても質問の 70%以上について対応可能と判断された。また、傾聴対話における主観評価においても、人間(WOZ)との対話に比べて概ね 90%以上の評価値となった。

（３）自律対話と遠隔操作対話の切替え・融合

自律対話と遠隔操作対話を融合した自在対話システムの制御機構の設計と実装を行った。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 CA を用いた自在対話を実現した。自律対話システムにおける音声理解・対話の破綻検出、エンゲージメントの認識に基づいた切替え、自律で行っていた対話の要約提示などについて検討を行った。本年度は、傾聴対話や説明・プレゼンテーション対話において、システムの実装及び動作検証を行った。研究開発課題２（猿渡）の音声変換、及び研究開発課題４（李）の CA 制御との統合を行った。１人の操作者で３名のユーザを同時に対話ができることを確認した。

課題推進者：河原達也（京都大学）。

研究開発課題２：音響情報処理・音声変換の研究開発

（１）多様な利用者に対応できる自律 CA 用音声合成

自律 CA が多様な利用者に対応するには、特定の CA 個性に限定されない多様かつ柔軟な音声合成技術を開発する必要がある。そのために、合成音声の自然性、再現可能な個性、再現に係る即時性のそれぞれの向上について研究開発した。本年度は、少数(100 人程度)の操作者の個人性を高精度に再現する音声合成技術を研究開発した。自然度・再現度に関する５段階主観評価の結果、4.0 前後のスコアを確認した。併せて、音声合成の高速化に向け、文全体の入力をまたずに合成可能な低遅延音声合成技術、合成誤りを訂正可能な End-to-End 音声合成技術、人間の音声のように、非流

暢な発声を可能にする音声合成技術を提案した。

（２）自律 CA 発話と遠隔操作発話を同化させる音声変換

CA 発話と遠隔発話をシームレスに融合するには、上記（１）で開発する自律 CA の合成音声と、遠隔操作者による遠隔音声を、違和感なく切り替える必要がある。この際、合成音声と遠隔音声の個人性の不整合により利用者に違和感を与えてしまうため、遠隔音声を合成音声に整合させるための音声変換技術を開発する必要がある。本年度は、当該年度のマイルストーンに優先して、音声変換ソフトウェア基盤の整備を進めた。通常の laptop PC で動作する基盤を構築し、それを MS プロジェクト全体に展開した。現在、複数の研究開発課題において利用されている。今後は、そこからのフィードバックを基に当該年度のマイルストーン達成に向かう。

（３）実環境下における音声分離・強調

雑音の多い実環境において、操作者は明瞭に CA を通して利用者の声を聞くと共に、操作者の声も明瞭に利用者に届けられなければならない。そのためには、まず利用者発話音声を高精度に分離・強調する信号処理を確立する必要がある。そこで自律 CA に取り付けられた複数のマイクロホン（これらは CA 各部に分散的に配置され位置未定かつ CA の動作に応じて時々刻々と位置が変動する）を想定し、それらで得られた多チャネル信号群に対してブラインド・セミブラインド音源分離を適用する。本年度は収録音声の強調・伝送処理に関して、申請者らが開発した独立低ランク行列分析等の数理アルゴリズムを基礎とした拡張理論を研究開発し、今後のベースラインとなる音声強調処理系を構築した。具体的には、話者方向から到来する拡散性雑音も抑圧できる、教師なし（ブラインド）・教師あり（深層学習ベース）音声抽出手法を開発した。教師なし・教師あり音声抽出手法両者に関して SDR 改善度は 8 dB 以上となっており、当該年度のマイルストーンを達成している。また、リアルタイム化のためのオンライン処理アルゴリズムの開発と、MS プロジェクト全体への展開に向けてソフトウェアを開発した。来年度中に展開予定である。

課題推進者：猿渡洋（東京大学）

研究開発課題３：対話知識処理の研究開発

（１）対話知識獲得

オープンソースソフトウェアをベースに、人間が状態遷移の記述を行い、話し相手、旅行の案内、および、プレゼンテーションを行うシステムを実装し、評価を行った。話し相手については、状態遷移とニューラルネットワークの対話モデルを組み合わせることで構築し、5 段階中 3.49 の満足度を得た。案内については、状態記述に従い、ユーザ情報を尋ね、観光データベースを参照して応答する観光案内システムを構築し、55%以上の話者に所定の観光地を推薦することができた。プレゼンテーションについては、プレゼンテーションコンテンツと状態遷移に基づき、プレゼンテーションと質疑応答を行うシステムを構築し、適切な回答が得られたかどうかのスコアとして 10 段階中 5.8 点を得た。いずれのシステムについても、達成目標を超えることができた。ここで、話し相手とプレゼンテーションについては、適切な応答を得られたという心象を重視し、主観評価値をタスク達成の尺度として用いた。システム構築に加え、今後のデータに基づくシステムの自動構築に向けて、人間同士の接客データ、および、プレゼンテーション映像データを収集した。

（２）対話状況理解および可視化

CA とオペレータの引継ぎに必要な情報を明らかにするため、昨年度に収集した、人間同士が引継ぎを行いながら対話を行ったデータを分析し、対話の引継ぎに有用な情報が、隣接ペアおよび属性値対であることを明らかにした。隣接ペアとは、質問・応答などの呼応関係にある発話のまとまりを指し、2 つ以上の発話からなる。扱ったすべての対話タスクにおいて、隣接ペアにおける発話同士は矢印でつながれ、話者同士がどのようなやり取りを行っていたのかを表す表現として頻出しており、対話の引継ぎに有効であることが示された。属性値対とは、テーブル構造のデータであり、特に接客や案内などのタスク指向型の対話タスクにおいて利用された。タスクが明確であるときは、そのタスクに関連する情報をテーブルの形で表現することが引継ぎに有効であることが分かった。

隣接ペアが有効という知見に基づき、対話データに対して、一問一答の隣接ペア形式（図１）で要約を行う要約器のプロトタイプを構築した。具体的には、まず、対話データに対して、一問一答の隣接ペア形式に要約したデータを学習用データとして作成した。そして、このデータから、対話を隣接ペア形式で要約するニューラルネットワークを用いた生成モデルを構築した。雑談（話し相手に対応）、案内、プレゼンテーションの対話データに対して、構築した要約モデルを適用し、モデルが出力した要約について「対話を理解できるか」という観点について評価したところ、いずれの対話タスクにおいても 10 段階で 6 点以上のスコアを得ることができた。このことから、達成目標を超えることができたと言える。作成された要約から対話状況が理解できたと感じられる心象の主観評価値は、対話状況理解の度合いと考えることができる。このことから、達成目標を超えることができたと言える。

加えて、雑談のデータを対象として、複数の対話要約のバリエーションについて、いずれの対話要約の形式が対話の引継ぎに有効であるかの分析を実施した。その結果、生成型要約と引継ぎ直前の 1 発話の組み合わせ、および、引継ぎ直前の 5 発話の抜粋型要約が、対話の引継ぎに有効であることが分かった。

（３）対話制御および制御インターフェース

状態遷移を記述することで構築した対話システムと人間のオペレータとの制御を切り替えできるインターフェース（図２）を作成し、CA および操作者の入れ替わりを実現できるようにした。このことから、達成目標を超えることができたと言える。加えて、CA とユーザの間で対話に問題が生じた際にタイムリーに介入できるようにするための対話破綻検出技術に着手し、最新のニューラルネットワークに基づくシステムを含む、複数の対話システムのデータについて対話破綻のアノテーションを行った。また、このデータを基に対話破綻検出器を構築した。

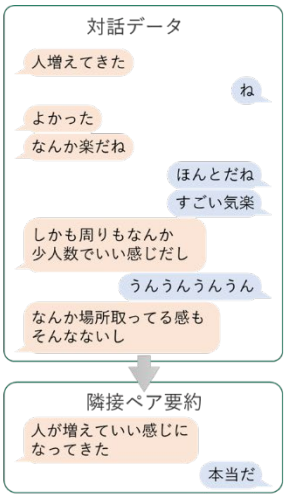

図１：隣接ペア要約の例

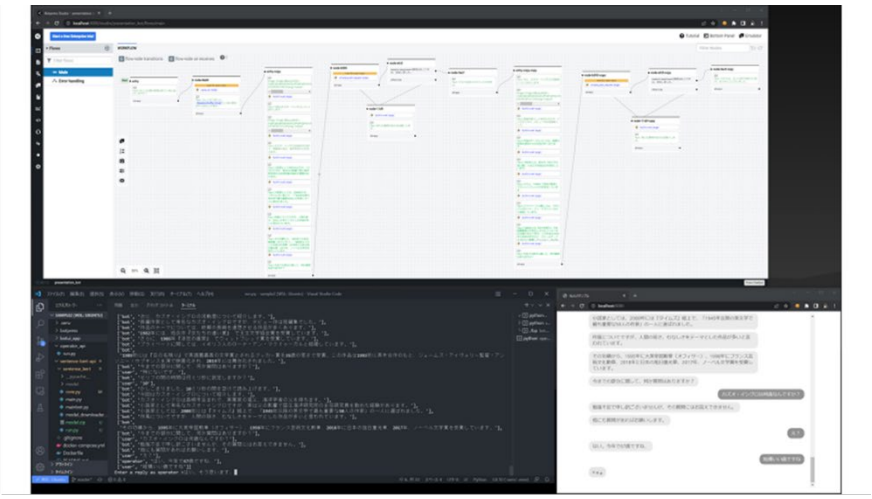

図２：構築した制御インターフェース

課題推進者：東中竜一郎（名古屋大学）

研究開発課題４：CG-CA 特有対話の研究開発

（１）CG エージェントの認知の研究の実施

（A）対話性認知の研究について、ユーザの音声と身体動作をリアルタイムに模倣する自己投影アバターを用いる手法を考案した。ユーザの動きに合わせてリアルタイムに動く分身アバターをエージェントと同じ CG 空間内に表示することで、ユーザの認知をシステム内へ引き込んで CG アバターとの会話のしきいを下げることを目標とする。20 代の大学生 22 名を対象として CG キャラクターを等身大で表示した据え置き端末を用いて複数の表示手法について実験した結果、自然な対話が行えるという認知は全員が獲得できたが、円滑さと持続性については分身アバターの有無の影響よりも自動応答生成システムと会話がうまく成立したかどうかが支配的な影響を与えていた。持続的な対話性認知は対話の内容と密接に関連付くことが分かったため、今後は研究開発項目２の他の課題推進者と協力して研究する必要があること、および、本課題としては会話中ではなく「会話を行う前の見た目から感じ取れる初回の話しかけやすさ」に的を絞った研究を行うべきであることが示唆された。

（B）存在感・生命感を伴う CG-CA の設計・デザインについて、複数の CG-CA モデルを設計し、その制作を進行した。対話 CG エージェントのデザインについては研究開発項目１研究開発課題１と連携しながら進めている。まず事例調査として、過去の対話システム研究の事例、テレノイドや ERICA といった対話アンドロイドやロボットの調査、それに加えて近年社会的に流行している VTuber における動向について調査した。これをもとにチーム内で議論を重ねた結果、特に CG キャラクターの個性と操演者の個性の関係性に着目して、３つの類型が定義可能であることを見出した。すなわち、（１）キャラクター自身の性格や個性の表現をできるだけ抑え、誰でも自分のアバターとして利用可能な「ジェネリック型」、（２）キャラクター自身が見てすぐ分かる強く尖った個性を持ち、アバターとして使う際にも操作者はそのキャラクターの個性に合致した操演が強く求められる「キャラクター型」、（３）弱いキャラクター性を持ち、操作者もそのキャラクター性を尊重した振る舞いが求められるが、比較的自由度が高く、操演の許容範囲が広い「着ぐるみ型（VTuber 型）」の３つである。

次にこの３種それぞれについてリファレンスとなる CG-CA の設計開発に着手した。このうち「ジェネリック型」については優先的に開発を進め、リアルな 3D-CG モデル「Rubina」（図 1）とアニメ調の 2D-CG モデル「ジェネ」（図 2）の２種類の CG-CA を年

度内に完成させた。それぞれ図１，図２のように、キャラクター性を控えめにしつつ性別を感じさせない中性的な見た目を重視したデザインとした。100 以上の表情パターンやフルボディの全身骨格を持ち、対面会話において高い生命感・存在感を表出するキャパシティを持つ。3D-CG モデルは Unreal Engine を用いて高精細なグラフィックスで表現され、人としてのリアルな質感と CG らしい豊かな表現力を備えている。2D-CG モデルは子どもや若年層に親しまれるアニメ調の表現で、アニメ調の豊かな表情および動作表現を持つ。これら２体のモデルの制作が本年度内で完了し、現在はシステムと組み合わせてプロジェクト内へ公開するための準備が進行している。特に 2D モデルは研究開発項目２および７のグループ内へ先行公開し、社会実験への準備と試用を進めている。なお、ジェネリック型以外の２つの類型「キャラクター型」「着ぐるみ型」についても進行中で、来年度の前半に全ての CG-CA が完成する見込みである。

図 1：ジェネリック型 3D-CG アバター「Rubina」

図 2：ジェネリック型 2D-CG アバター「ジェネ」

（２）CG エージェントの対話生成の実施

（C） ＣＧ会話における特有の誇張や強調を伴う言葉・声・動きのモデルの研究について、VTuber を対象としてＣＧ会話の特有性の分析を行った。近年 VTuber と呼ばれる人間がＣＧキャラクターを操演する形の動画配信や会話コンテンツが広

がりつつあるが、人間に比べリアリティが低く動作の制約が大きい CG キャラクターとの会話が広く受け入れられることから、本年度は VTuber を対象とした「in the wild なＣＧ会話」の事例分析を行った。まずツールとしてはほとんどの VTuber が Live2D と呼ばれる平面の絵を変形させて動かす軽量な規格を用いており、キャラクターの上下左右移動（XY 平面）と表情変化しかできない中で独特の CG 会話ふるまいが用いられている傾向が見られた。次に、日本の VTuber の人気上位 30 名（合計チャンネル登録者（購読登録者）数 3430 万人、日本の総合計の 21%）の雑談配信計 15 時間を対象に Dialogue Act のタグ付けとふるまいの分析を行った。その結果、人間対人間の雑談に比べて共感と感謝が若干多いものの全体ではほぼ同じ出現傾向であるが、会話時のしぐさの大きさが約 2.5 倍大きく、乏しい実在感をカバーするために誇張や強調を伴う会話様式が好まれる傾向が確認できた。

（３）CG-CA 対話システムの研究開発の実施

(D)自律動作と遠隔操作の両方が行える CG-CA システムを、MMDAgent をベースに構築した。システムは Windows で動作する CG-CA 表示・操作フロントエンドであり、外部から表情情報や動作情報をネットワーク越しに受信して CG-CA をリアルタイムに動かすことができる。また、音声認識エンジンＪｕｌｉｕｓによるリアルタイム音素認識を内蔵しており、受信した音声波形をもとにリップシンクを生成しながら音声再生できる。また、自律動作システムを想定して、「挨拶」「驚き」といったコマンドを受け取ることで、対応するプリセットアクションを再生する機能を備える。操作者側のフロントエンドとして OpenFace を想定しているが拡張は容易である。本システムは 2D-CG 用であるが現在 Unreal Engine をベースにした 3D-CG 用のシステムも並行して構築中である。1 月にシステムのベータ版と最低限のマニュアル、CG-CA モデルをまとめて GitHub で限定公開しており、研究開発項目２および７の各グループのメンバーと個別に共有を行っている。

（E）また、対話システムにおいては会話の流れに合わせてキャラクターが相槌や感情表現等のアクションを行うが、これまではタスクごとに必要なアクションが個別に定義され、共通の規格や標準的な実装は存在しなかった。そこで、本研究課題では CG-CA システムの構築に先立って、案内対話や接客対話、傾聴対話など想定される多くの対話タスクをカバーするアクションセットを定義した。策定にあたって調査・参考にした情報は以下のとおり：京大の傾聴対話研究で用いられた相槌セット、Ekman の基本表現セット（６基本表情＋４追加表情）、名工大正門に設置されている案内システム「メイ＆タクミ」で長年運用されている応答会話セット、CG アバターの既存プラットフォームの標準仕様（MMD / VRoid / VR Chat）、代表的なフェイシャルキャプチャーAPI の仕様（FACS/Apple ARKit）。この調査をもとに多様な対話タスクをカバーできる 34 種類のアクションセットを考案した。今年度制作に着手したすべての CG-CA はこのアクションセットに準拠するものとなる。

課題推進者：李晃伸（名古屋工業大学）