

2024 年度 卒業論文

顔画像認識を活用したキャラクターとの音声対話システム

龍谷大学 先端理工学部 数理・情報科学課程

Y210016 大畑 舞奈

指導教員 佐野 彰

概要

本研究では、音声対話システムの必要性に着目し、特定のシチュエーションやユーザー層において従来のテキスト対話システムでは満たすことが難しい課題に対応することを目指した。テキスト対話システムは、プログラムコードのやり取りや詳細な情報交換には優れているが、文字入力に不慣れなユーザーには十分な利便性を提供できない場合が多い。例えば、日常の中で音声で雑談したい時や、複数人の子供の世話をしている手が離せない状況や、あるいは幼い子供のように文字入力にハードルを感じるユーザーにとって、音声対話システムは直感的に操作でき、誰でも簡単に利用できる仕組みとなり得る。また、音声による対話は、より人間らしいコミュニケーションを提供し、親しみやすさを強化する効果も期待される。そこで、本研究では、音声対話システムに親しむための手段としてアニメキャラクターをアバターに採用した。その一例として、子どもから大人まで幅広い年齢層に知られており、親しみやすさやユーモラスな性格が特徴的である「スポンジ・ボブ」を用いた。

本システムでは、対話ユーザーがPCの前に座ると、接続されたカメラからOpenCVを用いて顔および目の検出を行い、OpenCVのFaceDetectionを用いてユーザーの顔を検出する。本システムが顔を検出すると、Azure Speech ServiceとAzure OpenAI Serviceを活用することで、リアルタイムでの会話生成と音声合成を実現し、ユーザーに会話開始を促すような対話音声を出力する。ユーザーからの発話が始まると、システムはマイクから取得した音声データをAzure speech to text Serviceを用いてテキストデータに変換する。変換されたテキストデータはクラウド上の自然言語処理サービスによって解析され、ユーザーの発言内容に応じた応答文が生成される。この応答文はさらにAzure Speech Serviceの音声合成技術を用いて音声データとして変換され、スピーカーを通じてユーザーに返答される。これらのプロセスはすべてリアルタイムで実行されるため、シームレスな対話ができる。

また、システムは対話中にユーザーの顔や視線の動きを継続的に監視する機能を持つ。ユーザーの視線が大きく逸れた場合や、一定時間顔が検出されない場合には、「こっち向いて」といった注意喚起の音声を自動で生成し、対話を円滑に進めることを目指している。このように、顔認識と音声認識、さらに自然言語処理を統合することで、単なる音声応答に留まらず、ユーザーの行動や状況に応じたインタラクションが可能となっている。

さらに、本システムではユーザーが視線をPCカメラの正面に戻すと即座に注意喚起を停止し、通常の対話モードに復帰することで、自然な会話体験を維持している。また、Azure OpenAI Serviceを活用した自然言語処理により、対話の文脈を理解し、ユーザーの発言に応じた柔軟で多様な応答が可能である。加えて、本システムはリアルタイム処理に対応するため、顔認識と音声認識の処理で効率的なデータのやり取りを行うために結果は非同期でし、それぞれが並列的に動作することで、全体の処理時間を最小限に抑えている。これにより、スポンジ・ボブとユーザーが直接対話を楽しめる環境を構築し、より親しみやすいものとした。

一方で、本システムにはいくつかの課題が残されている。特に、音声合成においてスポンジ・ボブのキャラクター性を十分に再現することができていない点が挙げられる。よりリアルな対話を提供するためには、スポンジ・ボブの声を学習した音声モデルを構築する必要がある。この実現には、キャラクター特化型の音声データの準備と、それに基づくモデルトレーニングが求められる。また、会話内容についても、スポンジ・ボブ特有のユーモアやキャラクター性を反映した応答を生成するため、より高度な自然言語処理技術が必要である。さらに、キャラクターとの対話の臨場感を向上させるために、音声合成と連動した口パクの動きを実装することが重要である。これにより、ユーザーにキャラクターが実際に話しているような印象を与えることが可能となり、システムの完成度を大きく向上させることが期待される。