1ZL-05

音声対話システムの行動表出と沈黙時の相互理解

前土佐 勇仁† 三枝 亮†
†神奈川工科大学 創造工学部

1 はじめに

介護施設の日常会話では介助者が被介助者の発話を十分に待ち，発話がない場合は介助者が続けて発話することで会話を継続させる場面が見られる．一方，従来の音声対話システムではユーザの発話がないと待機を続けるか会話を終了させる場合が多い．発話行動は動作といった非言語情報からでも取得できることから，音声対話システムがユーザの無言を認識して発話を開始するためには，その動作が言語化可能な情報かを判断することが有効であると考えられる．本研究では，ユーザの沈黙や身体動作を認識し，CG キャラクタの身体動作で発話の継続や話者の交代を円滑化する音声対話システムを提案する．ユーザとシステムの行動表出により沈黙時の相互理解が高まり，自然な日常会話が実現される．

2 音声対話システム

本研究で提案する音声対話システム [1] の構成を図1に示す．音声認識には Whisper を用い，応答文の生成には GPT-2 を用いた．システムの応答には rinna 社の学習済みモデル「japanese-gpt-1b」を使用した．会話のセッションでは,「バイバイ」,「さよなら」,「さようなら」のいずれかを認識することで応答を終了する．

本システムの沈黙の定義を図 2に示す．User が発話者で System が傾聴者のターンと仮定する．本研究では文前に発生した「間」を『文前「間」』，文の中に発生した「間」を『文中「間」』と定義する．User の発話終了時刻を起点とすると，無言は User が再び発話し始めるまで，同様に文前「間」は User の発話開始までと定義する．これをシステムに組み込んだ場合，システムの待機時間は使用者の発話開始時刻から機体の音声再生終了時刻までの時間となる．一方，文中「間」は User の発話中に起きる現象であり，User の発話行動停止時刻から User の発話行動再開時刻までと定義する．

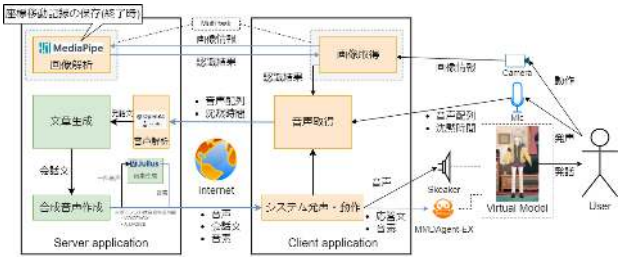

図 1: 音声対話システム

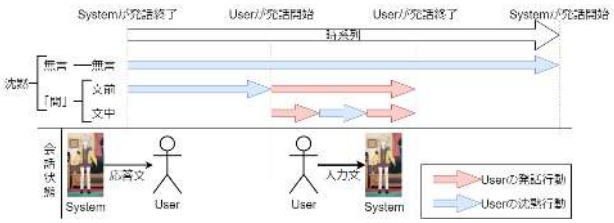

図 2: 沈黙の定義

本システムはユーザの発話時とシステムの無言認識時にユーザが何らかの動作を行った場合，直前に行った動作の記録を行う．x 座標における微分値導出例を図 3に示す．MediaPipe により顔・両手・骨格の x, y 座標を取得する．画像を 20fps で取得している場合，1 会話ターンにつき最大 2.5 秒間に取得した座標を記録する．対象部位の座標取得回数が 50 を超えると，動作の記録と各値間の微分値導出を開始する．x を座標値，t を配列番号とすると，データに格納された一区間の微分値 S_t は以下の式で導出される．微分値は最大 9 個生成される．

$$S_t = \frac{x_{t+5} - x_t}{5} \quad (1)$$

なお，取得データの時間長は対応する配列番号により正規化されている．

微分値が -0.5 未満，0.5 以上の状態が連続して 3 区間続いた場合，システムは動作の変化があったと判断し，該当する部位を別途記録する．動作記録は応答終了時に S_t をサーバ側に保存する．

Embodied Speech Interaction System for Mutual Understanding in Silence
Yuto MAETOSA†, Ryo SAEGUSA†
†Faculty of Creative Engineering, Kanagawa Institute of Technology, 243-0203, Shimoogino 1030, Atsugi, Japan
{yuto.maetosa, ryo.saegusa}@syblab.org

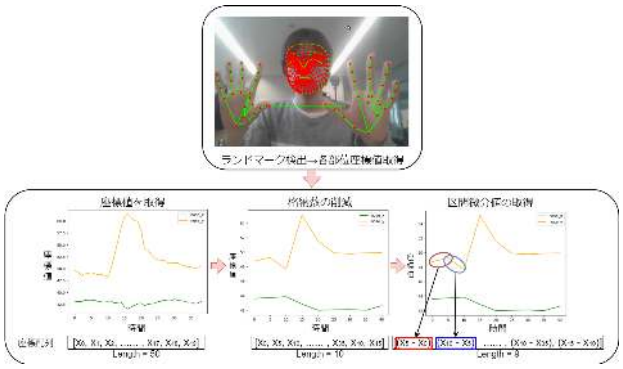

図 3: 各区間の微分値の導出例

図 4: 実験風景

表 1: サンプル構成

	動作あり	動作なし
発話	105	68
無言	13	13

表 2: F1-Score 解析結果

	精度	再現率	F1-Score
発話	0.79	1.00	0.88
無言	0.00	0.00	0.00

3 実験

本システムと被験者の二者間で自由に対話させる実験と並行してユーザの発話行動の記録を行った．実験風景を図 4に示す．会話中に違和感を生じさせないように，被験者には録音状態と文章内容を表示しなかった．代わりに，システムが動作表出することでターン・テイキングを明示的に表現する．

実験後には頭部の x, y 移動に対して沈黙時の動作とユーザ発話時の動作ごとにラベリングを行い，サポートベクターマシン（SVM）を用いて分類処理を行った．今回は頭の動作情報を分析した．また，沈黙時，ユーザ発話時の共に直前に x, y 値のいずれかが大きく移動したサンプルを使用した．大きく移動したかどうかの判定は 2項の該当部位導出手法を用いた．

分析には被験者 9 人 18 会話の動作記録を使用した．サンプルの構成を表 1に示す．このうち，録音切り上げ時に直前の動作が確認されたサンプル 118 個に対して機械学習を行った．訓練データとテストデータの比率は 8:2 である．SVM の機械学習で得られたデータは F1-Score を用いて精度を導出した．解析結果を表 2に示す．通常文の F1-Score は高い精度であることが示唆されたが，無言に対してはサンプル数が少なかったために F1-Score 値を得られなかった．

4 まとめ

本研究では画像認識により動作の有無を判定する音声対話システムを提案した．また，ユーザの動作記録をもとに機械学習を利用してユーザが発話に入るか否かの推定を行った．これにより，発話の兆候を認識することが可能になり，ユーザとシステムの行動表出による沈黙時の相互理解に近づいた．

しかしながら，既知のデータ以外の動作情報が出現した場合に誤った判断を行う恐れがある．また，被験者の主観評価から発話行動に移る場合には必ず頭部をマイクに近づける行動をしていた．今後の展望として，自然にジェスチャを誘発できる会話環境を構築する必要がある．実験では無言時にユーザが動作したサンプル数が少なく，分類処理では精度を導出することができなかったため，対話実験を通じて動作記録を引き続き収集する．

謝辞

本研究は神奈川工科大学先進技術研究所の助成を受けた．

参考文献

[1] 前土佐 勇仁, 三枝 亮, CG キャラクターの行動表出によるユーザ無言時の話者交替の明確化, 2022-AAC-20(4), pp.1-7, 12 月 9-10 日, 2022.