適切なタイミングで挨拶を行う 対話エージェントシステムの実現可能性

○小林弘幸　杵鞭健太　山本知仁（金沢工業大学）三宅美博（東京工業大学）

概要　本研究では, 挨拶のタイミングを考慮した簡易的な音声対話エージェントシステムの構築を行った. 具体的には, まずこれまでわれわれが行ってきた挨拶動作の同調に関する実験結果に基づいて, 挨拶のタイミング生成モデルを構築した. つづいて, このモデルを簡易的な音声対話エージェントに導入し, そのタイミングの制御可能性について検証を行った.

キーワード: 対話エージェントシステム, 同調, タイミング制御, 挨拶動作

1　はじめに

近年, コミュニケーションロボットがエンターテイメントの観点だけでなく, 医療や介護, 生活支援といった観点からも注目されてきている. しかし, 現在開発されているコミュニケーションロボットは, ときに人が実際に行う身体動作と異なる動作をすることがあり, 違和感を与えることがある. 今後, 一般的なユーザがこのようなロボットをさらに利用していくことを考えると, まずは人がコミュニケーションを行う際の身体動作とその内容の関係を分析し, 得られた結果を体系化していくことが重要であると考えられる.

これまでコミュニケーションにおける身体動作に関する研究は, リズムの同調という観点から解析されてきていることが多い. われわれの研究グループもこれまで, 人の対話における発話リズムの同調を解析してきた. 具体的には, 指示発話と応答発話からなる対話において, 発話速度の意図的な変化が発話に関わる時間特徴量の同調にどのような影響を与えるかを明らかにしてきた[1]. また, 挨拶動作に注目した対話の実験を行い, 動作の内容によって異なる同調が現れることを明らかにした[2].

本研究では, われわれが行った挨拶に関する先行研究から得られた解析結果に基づき, 挨拶のタイミングを生成するモデルを構築する. また, そのモデルを簡易的な対話システムに導入することで, 挨拶におけるタイミング制御の実現可能性について検証を行う.

2　挨拶タイミングを考慮する対話システム

本研究では, ユーザの挨拶タイミングを考慮した簡易的な音声対話エージェントシステムを構築する. 本システムは主に, 身体動作検出, 音声の認識, 応答タイミングの制御, 応答の生成の4つの処理から構成される. 本システムにおいて, ユーザの挨拶動作開始タイミングの検出はKinectを用いて行う. このタイミングを検出した後, システムはユーザの発話を認識し, 得られた情報に基づいて, 対話エージェントの挨拶動作と発話を生成する. このようなシステムの処理において使用する時間特徴量をFig.1に示す.

これまでのわれわれの知見[2]では, 人の挨拶において後行動作開始タイミングは先行動作終了タイミングにオーバーラップするか, ほぼ同時になることがわかっている. このような動作を対話システムにおいて実現するためには, ユーザの動作が終了するまでに, ユーザの身体動作長を推定する必要がある. このことを実現するために, 本システムでは先の研究で得られた知見である, 先行する発話者の発話長 (UU:Duration of User Utterance) と動作長 (UM:Duration of User Motion)

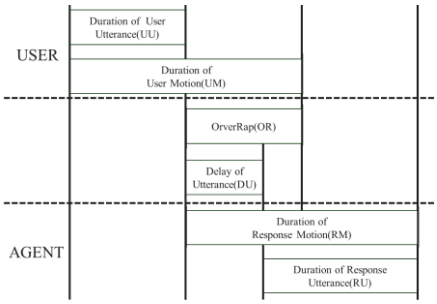

Fig.1: Durations of greeting

にある強い正の相関関係を利用した. 具体的には, 身体動作よりも早く終了する発話長を検出した後, この時間長から身体動作長を推定する式 (1) を, 両時間長の正の相関関係を基に構築した (本研究では, 以下の式 (1)-(5) における定数と係数を, 典型的な挨拶動作のデータから算出して適用した.).

$$UM = UU * a + b \tag{1}$$

挨拶の音声認識に関しては, これまでわれわれの先行研究[3]でも用いてきた, 音声認識エンジン「Julius [4]」を用いた. このシステムにおいて発話長の算出は, 発話区間の開始と終了点で割り込み処理を行い, その差分を求めることで行われた. ただし, 認識システムは発話の信号レベルの減衰部分を正確に検出するために, 発話区間の前後にマージンを設けているため, 実際より長く発話長を算出する. 本研究では, このマージン分を補正した値を発話長とした.

発話長検出後, 推定したユーザの動作長 (UM) のオーバーラップ (OR) 分前から応答を行うことで, 応答タイミングの制御を行った. オーバーラップの時間長は先行発話者の発話長と負の相関関係にある[2]ことがわかっているため, 以下の式 (2) を用いて算出した.

$$OR = UM * c + d \tag{2}$$

応答の生成では, これまでの実験結果[2]に基づき, ユーザの動きに同調するようエージェントに挨拶を行わせた. 推定したユーザの動作長 (UM) とエージェントの動作長 (RM) には正の相関関係があり, エージェントの動作長 (UM) と発声遅延 (DU), 発話長間 (RM) にも正の相関関係があることから, 以下のような式 (3), (4) を用いて, 各時間長を算出した.

$$RM = UM * e + f \tag{3}$$
$$DU = RM * g + h \tag{4}$$

エージェントの動作生成に関しては, 予めモーションキャプチャで取得した人による挨拶動作のデータを用い, 再生速度を変更することで時間長の制御を行った. また, エージェントの動作長 (RM) とエージェントの発話長 (RU) にも正の相関があるため, 以下の式 (5) を用いて発話長を制御した. ただし本システムでは, 合成音声にアクエスト社の「Aques Talk2」によって作成されたデータを用いている. そのため, 発話長を50msec間隔で調整した挨拶発話を予め用意しておき, 式から算出された発話長に最も近い音声ファイルを再生することで擬似的に発話長の制御を行った.

$$RU = RM * i + j \quad (5)$$

3 対話システムの動作検証実験

3.1 検証手法

本研究では, 構築した音声対話エージェントシステムにおいて, 各開始タイミングと時間長の制御が行えているかについて検証実験を行った. 実験は, 被験者1 名がディスプレイに表示される対話エージェント(本研究ではボーンモデルとして構成した) に向かって「こんにちは」と発話をしながらお辞儀をすることで行った. このとき被験者には, 挨拶速度を「はやく」, 「自然に」, 「ゆっくり」の 3 つで変化させるように指示した. 被験者とエージェントの動作は, モーションキャプチャとビデオカメラを用いて記録した. また, ユーザとエージェントの発話は PC 上の音声解析ソフト (Sound Engine Free) を用いて録音した. 実験では速度 3 条件でそれぞれ 9 試行行い, これらのデータを基に検証を行った.

3.2 実験結果

本システムはユーザの動作長を発話長から推定しているため, 実際の動作長に対してユーザの動作長をどの程度正確に推測できているかが重要となる. 検証の結果, 誤差の平均値は 170.34msec であり, ユーザの身体動作長の平均値が 1283.33msec であることを考えると, 誤差がその 10%程度になることがわかった. また, このような誤差により, 期待するオーバーラップ長が確保できない場合があり, 平均で動作開始タイミングが 197.36msec 遅延することがわかった.

次に, ユーザとエージェントの動作長間の相関関係を Fig.2 に示す. 図中, ユーザの挨拶動作長はモーションキャプチャから得られた値, エージェントの動作長はビデオカメラから得られた値を示している. 図より, 両時間長に強い正の相関があるのがわかる (相関係数:0.818) . この結果は, ユーザの動作長が長くなれば, エージェントの動作長も長くなっていることを意味し, ほぼ期待通りにシステムが動作していることを示している.

最後に, ユーザとエージェントの発話長間の相関関係を Fig.3 に示す. 図中, エージェントの発話長は再生されたファイルより求め, ユーザの発話長は録音された音声データから求めた. 図より, 両時間長に強い正の相関 (相関係数:0.917) があることがわかる. この結果は, ユーザの発話長が長くなるにつれてエージェントの発話長も長くなっていることを意味し, システムがほぼ期待通りに動作していることを示している.

Fig.2: Relation between UM and RM

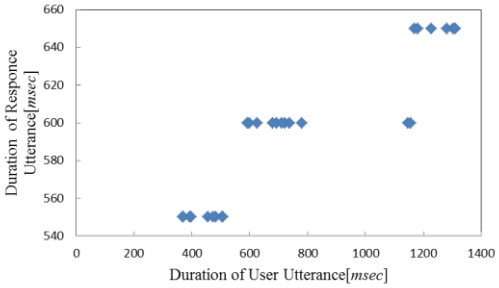

Fig.3: Relation between UU and RU

4 考察と今後の展開

本研究では先行研究を基に, ユーザの発話長によって身体動作長を推定して, それを用いることで挨拶タイミングを制御し, ユーザの身体-発話動作に同調する音声対話エージェントシステムを構築した. 結果として, やや誤差があるものの挨拶タイミングを制御し, ユーザの身体-発話動作長にシステムの身体-発話動作長を, ほぼ同調させることができた.

今回構築したシステムでは, 身体動作長の推定にやや誤差が残った. この原因の 1 つとして, 各時間長を計算する式における係数, 定数を決めるために用いたデータが, 実験を行った被験者のデータと乖離していたことが考えられる. 今後より汎用性のあるモデルを構築するために, データ数を増やすことを考えている.

また本システムでは, エージェントをユーザの身体動作にオーバーラップする形で動作させているが, まだ十分にその動作を実現できていない. この理由の 1 つとして, 発話終了を識別するために存在する時間的マージンが, システム全体を遅延させていることが考えられる. 今後, この時間をうまくキャンセルすることが必要であると考えている.

本研究では以上のことを考慮し, より精度の高い対話エージェントシステムを構築したいと考えている. また, ユーザの印象評価実験を行うことで, 本システムの有用性についても示していきたいと考えている.

参考文献

1) 山本, 武藤, 阿部, 三宅 : 対話コミュニケーションにおける 2 種類の発話タイミング構造, 計測自動制御学会論文集, Vol.45, No.10, pp.522-529 (2009)
2) 杵鞭, 山本 : 挨拶動作における身体リズムと韻律情報の同調, 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, Vol.114, No.67, pp.247-252 (2014)
3) 小林, 大村, 山本 :音声対話システムにおける適切な発話タイミング生成に関する考察, ヒューマンインターフェース学会研究報告集, Vol.15, No.9, pp.23-28 (2013)
4) Julius: http://Julius.sourceforge.jp/