ユーザーの態度推定に基づき適応的なインタビューを行うロボット対話システムの開発

Developing Adaptive Interview Strategy Based on User's Attitude Recognition for a Interview Robot

長澤 史記[1*] 石原 卓弥[1] 岡田 将吾[2] 新田 克己[1]
Fuminori Nagasawa[1] Takuya Ishihara[1] Shogo Okada[2] Katsumi Nitta[1]

[1] 東京工業大学 情報理工学院 情報工学系
[1] Dept. of Computer Science, School of Computing, Tokyo Institute of Technology
[2] 北陸先端科学技術大学院大学
[2] Japan Advanced Institute of Science and Technology (JAIST)

Abstract: The goal of this research is to develop a robot dialogue system that elicits story of users by choosing appropriate questions which user would like to answer. The key technique is the inference of users' motivation from multimodal information: gesture, posture, speaking status and prosodic status.
In this paper, for this purpose, we developed a robot dialogue system that collects multimodal information of interviewees for examination of useful feature quantity for state estimation. Based on the collected data, state estimation was performed by machine learning, and utterance motivation was estimated from multimodal information.

1 はじめに

広報や案内のような, 特定の事柄について他人に伝えるべき情報を収集する手段として, インタビューが広く用いられている. インタビューでは, 事前に収集するべき情報についての質問事項を作成して質問するだけではなく, 相手の反応によって, 話題を掘り下げる, 話題を転換するなどといった質問戦略の選択を行う必要がある. 相手に語らせることによりより, 多くの情報を引き出すため, 情報源であるインタビュー対象者の発話意欲を推定し, より意欲的な発話が得られるような話題選択を行うことも重要な技術となる. また, インタビュー結果をもとに情報を発信する際には第三者にとっても理解しやすいように具体的な情報を多く含む事が好ましい. そのためインタビュー対象者の回答内容を考慮して, より具体的な回答を得られるような戦略判断も重要な要素となる. 本研究では, インタビュー対象者であるユーザーのマルチモーダル情報から発話意欲の推定を行い, 推定結果に基づいた適切な質問戦略を選択することで, インタビュー対話によってユーザーからより深く掘り下げた内容を聞き出すインタビュアーロボットシステムの構築を目的とする.

この目的に向けて, インタビュー対話を通してユーザーのマルチモーダル情報を収集するシステムを構築し, インタビュー実験を実施してマルチモーダル情報の収集を行った. 実験で得られたマルチモーダル情報をもとに機械学習を用いて発話意欲の推定を行い, 意欲推定の精度評価を行った.

2 関連研究

Saito ら [Saito2015] は，高齢者との問診対話を目的とした対話システムにおいて，ユーザの発話意欲推定に関して検討した．視線の動き，フィラーなどの特徴量や，書く非言語イベントの生起タイミングが有効な特徴量であることを示したが，多くの特徴量は手動でアノテーションされており，発話意欲の自動推定は課題として残っている．本研究はユーザの態度を自動推定することを目指し，特徴量は全て自動的に抽出する．

岡田ら [岡田 16] は, 対話に参加する人間のコミュニケーション能力を評価する方法として, マルチモーダル情報を用いることが有用であることを示した. この研究では，対話の開始から終了までに観測されたマルチモーダル情報を対象としてモデルを構築しているのに

*東京工業大学大学院情報理工学院 情報工学系 情報工学コース
神奈川県横浜市緑区長津田町 4259 J2-53
E-mail: nagasawa.f.aa@m.titech.ac.jp

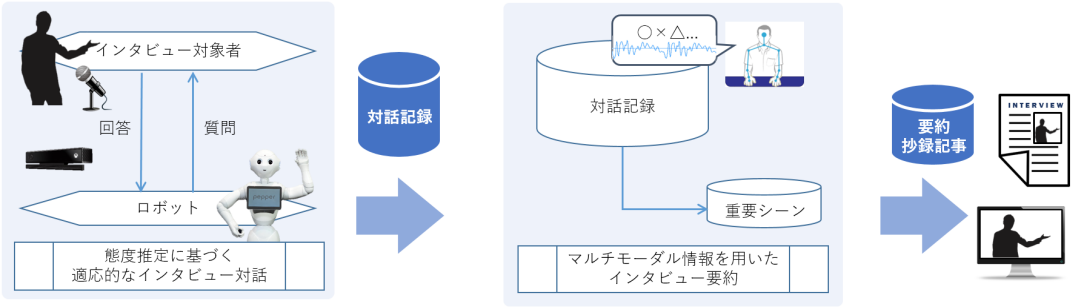

図 1: インタビューシステム全体概要

対し，本研究では，マルチモーダル情報を用いて発話ターン毎にオンラインで発話態度を推定するモデルの構築・評価を目的としている．

千葉ら [千葉 14] は，ユーザーの対話意欲を考慮した話題選択を行うインタビュー対話システムのために，人-人間でのインタビュー対話について分析を行った．インタビュー対話を通じてユーザーから多くの情報を引き出せるように相手の意欲の高い話題毎の対話意欲を推定するためには，音声や顔や全身の動きなどのマルチモーダル情報を用いる事が有効であることを示した．

本研究では，対話ロボットの質問戦略を逐次更新するために，対話の応答毎に意欲の推定を行うほか，骨格を考慮した姿勢に関わる非言語特徴量も考慮する．このため本研究では人-ロボット間でのインタビュー対話におけるマルチモーダル情報を取得し，発話意欲との関連性を分析する．また，本研究では発話意欲の尺度として，対象が意欲的な発話を行っているかという積極性の観点からアノテーションを行う．

小堀ら [Kobori16] は，テキストベースでのインタビューシステムにおいて，ユーザーの対話への意欲を高める方法として，質問に関係のない小発話をシステムが行う事が有効であることを示した．本研究では，質問発話を適応的に選択することによって，ユーザーの対話への意欲を高める事によってより多くの情報を引き出すインタビューシステムを目指す．

井上ら [井上 16] は自律型アンドロイドと人との対話における，聞き手の振る舞い（笑い，うなづき，相槌など）からエンゲージメントの推定を試みている．一方で，本研究は主に発話時の振る舞いより意欲を推定しようとしている点が異なっている．

本研究の初期検討 [長澤 17] においては，インタビュー対話中のインタビュー対象者のマルチモーダル情報から，発話意欲に有用な特徴量について，T 検定を用いて被験者毎に検定を行い，得られた検定結果の平均を求める事により有意な特徴量を調べた．その結果，音声についてはピッチ，姿勢については特徴量の発話区間平均が

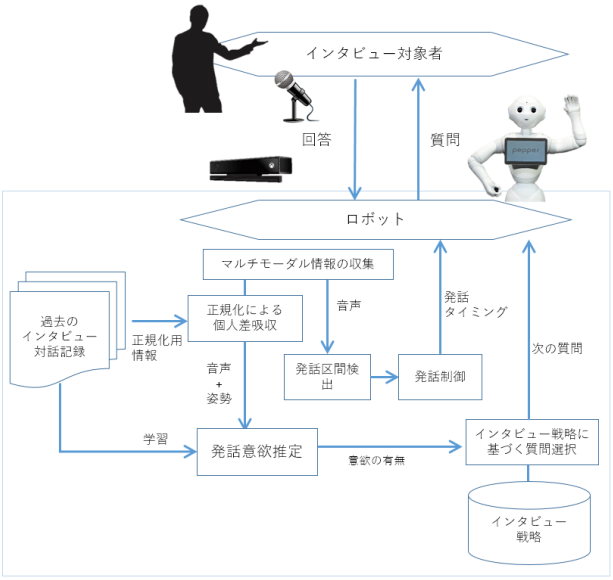

図 2: インタビューロボットシステム

有用であることがわかった．本論文では，分析結果をもとに機械学習を用いて意欲推定を行う．

3 システム概要

本研究で目標とするインタビュアーロボットシステムについて説明する．本ロボットシステムは，インタビューで話を引き出し，詳細なインタビューログデータを得て，そのデータから記事やハイライト動画を作成することを目的とする． システムの構成を図 1 に示す．ロボットがインタビュー対象者と対面し，一対一のインタビュー対話を行う．ロボットは質問を行い，ユーザーの回答時の状態を音声・画像・姿勢の情報から，リアルタイムで解析し，これらの情報を基に次に行う質問の選択を行う．ユーザと直接対話するロボットには

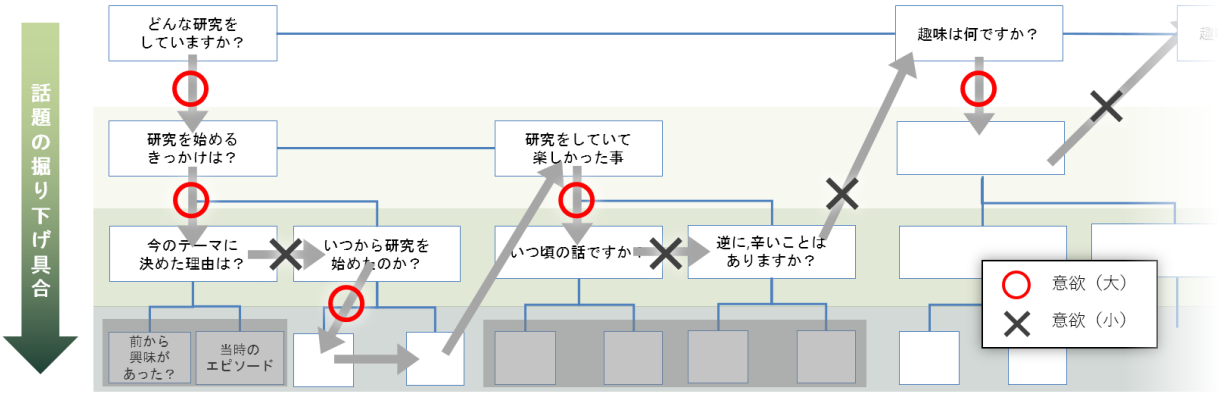

図 3: 推定発話意欲に基づく質問戦略の例

pepper を使用し, 各種センサを用いて機械学習および戦略判断を行うコンピュータからその挙動を制御する.

本論文では, インタビュー対象者からより多くの情報を引き出すために, 発話ターンごとに態度推定を行い, 推定結果に基づき質問戦略決定を行う方法について検討を行う.

インタビューログデータから記事やハイライト動画を作成するための技術・検討に関しては, [石原 17] で述べられている.

3.1 マルチモーダル情報の取得

インタビュー対話を行っている間, ロボットの後方に設置した Logicool 社製 Web カメラを用いて対象者の顔画像を 30fps で撮影し, Kinect v2 から対象者の姿勢の情報を取得する. Web カメラと Kinect は近くに設置し, インタビュー対象者を同じ方向から撮影する. 取得した顔画像に対して seeingmachine 社製顔認識ライブラリである faceAPI を用いて顔認識を行い, 顔の向きや視線方向の情報を取得する予定である.

音声については, Web カメラ内臓のマイクから対話全体の音声を取得するほか, Shure 社製の指向性マイクを対象者が装着し, 回答発話を行う際の音声を取得する.

3.2 発話区間検出と発話制御

システムは音声から Julius による発話区間検出を行い, 発話終了を検出すると, インタビュー戦略に則った次の質問を行う. この際, 発話終了とみなすタイミングは, 音が途切れるまでの連続区間だけではなく, 息継ぎやいいよどみなどを含めた一つながりの発話全体の終了を考える [中野 15]. これにより検出された終了タイミングから, 次の質問に移ってよいかを判断する発話制御を行う.

3.3 推定発話意欲に基づくインタビュー対話戦略

測定した姿勢や音声などのマルチモーダル情報を基に機械学習を用いて, 対象者の回答発話が意欲的であるか, 多く語っていたかなどといった発話態度の推定を行うことを想定する.

推定された発話態度ラベルに応じて対象者からより多くの情報を引き出すために, 質問に対する応答における発話意欲が大きいと推定された場合には, より掘り下げた内容の質問を行う. 具体的には, 以下のような行動を選択する

- 意欲的かつさらに情報を引き出すことが可能なら, 直前に質問した話題についてより深く掘り下げた質問を行う
- 意欲的であるが回答が少ない場合には, より具体的な質問をして回答を促す.
- 意欲が小さい場合には, 質問する話題の変更を行う.

これらの行動を行うために, 本研究では二分木探索による対話戦略の選択を行う. 深さ優先探索をベースとして, 意欲が大きい場合には現在の質問ノードの子ノードを選び, 意欲が小さい場合には子ノードを無視した探索を行う.

3.4 個人差吸収のための正規化

マルチモーダル情報についてはすべて被験者毎に正規化を行う. しかし, 本研究で目指すインタビューシステムの想定する使用状況では, 学習用データの正規化を行う事は可能であるが, インタビュー対象者については全てのデータが集まる前に発話意欲の推定を行う必要があるため, 対象者のマルチモーダル情報について正規化を行うことが出来ず, 上述する手法をそのまま適用することができない. そこで, 本研究では, 教師データを正規化しつつ, インタビュー対象者の測定値を正規化できない状況で予測を行う手法について検討する.

4 インタビュー対話実験

態度推定による適応的なインタビュー対話のための初期検討として, インタビュー対話を行いながらインタビュー対象者のマルチモーダル情報を収集するロボットシステムを開発し, 8 人にそれぞれ 1 セッションずつ, 計 8 セッションのインタビュー実験を行った.

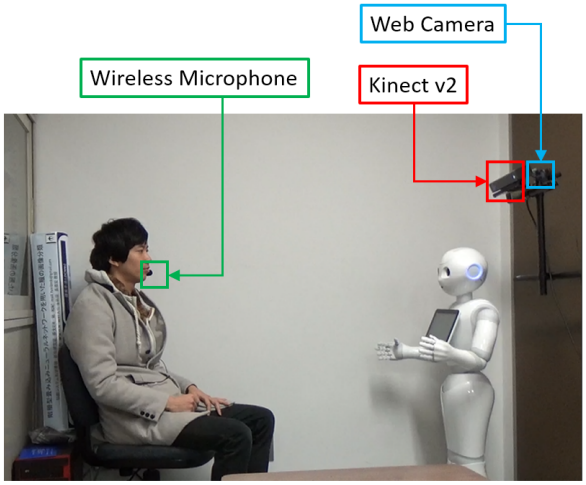

図 4: インタビュー実験のようす

対象者は図 4 のように椅子に座ってロボットと対面した状態で, 自身の研究活動に関する 15 項目についての質疑応答を行った. 全ての実験で, 同じ質問内容, 同じ順番で質問を行った. 今回の実験に使用した質問内容のリストを表 1 に示す. これによって得られたマルチモーダル情報を用いて, 機械学習による発話意欲の推定を行った.

4.1 アノテーション

実験により得られた対話データから, 各応答発話について, インタビュー対象者が積極的に発言を行っているか, という観点で 5 段階評価のアノテーションを行った.

表 1: 実験に使用した質問リスト

順	内容
1.	現在の研究テーマ
2.	研究を始めた時期
3.	研究テーマ選択の動機
4.	研究をしていて 楽しかったこと
5.	研究をしていて つらかったこと
6.	研究の魅力
7.	その研究にはどのような応用があるか
8.	研究のほかに興味のあること
9.	以前の研究
10.	以前の研究ではどんな成果を得たか
11.	今の研究とどちらが楽しいか
12.	それはなぜ？
13.	どんな趣味を持っているか
14.	私生活と研究との両立
15.	対話の感想

各発話の積極性 A_n について 1 から 5 の五段階でのアノテーション結果について, 意欲が高い ($A_n > 3$) 発話を A 群, 積極性が低い ($A_n < 3$) 発話を B 群とした T 検定を行い, 有効な特徴量の検討を行った.

5 意欲推定モデルの構築

5.1 検討した特徴量

本実験では, マイクを用いて録音した音声について得られた音声特徴量と, Kinect によって得られた姿勢特徴量について検討した.

発話区間

インタビュー実験の終了後, 質問に対応する 1 応答毎の発話区間について人手によるアノテーションを行った. 1 質問に対する回答発話を 1 発話区間とし, 発話区間の時間を発話長として特徴量に用いた.

姿勢

本実験では, インタビュー対象者は椅子に座った状態であるため, Kinect を用いて得られた骨格情報のうち上半身の情報である頭, 肩, 肘, 手の位置について検討した. 各部位について測定した (x,y,z) 座標のそれぞれの測定値を使用した. さらに, 原点からのノルム $\sqrt{x^2+y^2+z^2}$ を使用したほか, 肩, 肘, 手の左右がある部位については左右の値を加算したものも用いた. これらの特徴量について, 発話区間毎に平均と分散を計算した.

音声

ユーザーに装着したマイクから録音した音声について, 各応答の発話時間長の他, Speech feature extraction code [1] を用いてピッチ, エネルギー, MFCC を求めた. これらの特徴量のそれぞれについて, 1 発話区間当たりの最大, 最小, 平均を計算した.

5.2 機械学習を用いた発話意欲の推定

マルチモーダル情報から発話意欲を推定することを目的とした機械学習を行い, 予測精度を検証した. 各発話毎に, 積極性についてのアノテーション結果を教師データとした 機械学習を行い, One-Person-Out 交差検定により発話意欲を正しく推定できているかを確かめた. 以下に挙げる条件の組み合わせ毎にそれぞれの場合で交差検定を行った.

学習器の種類

本研究では, 学習器として Random Forest と Liner SVM の二種類を使用した.

特徴量の組み合わせ

音声と姿勢の二つの特徴量カテゴリのついて, それぞれ単独で使用した場合と, 両方を使用した場合について検証を行った.

特徴量選択の有無

T 検定による特徴量の分析により, 有意差が見られた特徴量のみを使用した場合と, すべての特徴量を使用した場合の比較を行った.

5.3 データの正規化による影響の分析

対話システムを運用する際，既学習済みのモデルを用いて，未知の話者の発話意欲を推定する必要がある. その場合，その話者のデータを正規化するための情報が得られない. 本節では，以下に挙げる条件の組み合わせ毎にそれぞれの場合で交差検定を行った.

正規化

学習データ・検定データの両方とも正規化した.

正規化展張

学習データの正規化を行った後, 検定データについて, 各特徴量の値域を, それぞれの平均値域と等しいとみなして正規化と同様の変換を行った.

無加工

学習データ・検定データの両方について正規化などの加工を一切行わない.

[1]Speech feature extraction code, http://groupmedia.mit.edu/data.php

表 2: 交差検定の結果 (Random Forest)

		音声		
		全使用	特徴量選択	不使用
姿勢	全使用	71.32%	65.44%	45.59%
	特徴量選択	66.18%	61.77%	60.29%
	不使用	66.91%	61.03%	

表 3: 交差検定の結果 (SVM)

		音声		
		全使用	特徴量選択	不使用
姿勢	全使用	71.32%	61.77%	46.32%
	特徴量選択	69.85%	62.50%	61.77%
	不使用	69.85%	61.77%	

6 意欲推定モデルの評価

6.1 機械学習を用いた発話意欲の推定

Random Forest と SVM の両者での交差検定の結果を表 2 と表 3 に示す.

結果から, すべての特徴量を使用した場合に SVM で 72.79% の割合で発話意欲を正しく推定できた. 本実験は意欲の有無についての二値分類タスクであるため, 正答率が 50% 以上の場合にランダムよりも意欲推定が正確に行われたとみなせる.

姿勢特徴量については, 特徴量選択を行った場合に全使用の場合よりも高い正答率を得た. 結果として，今回のタスクでは RF が安定であった.

6.2 正規化による影響の調査

学習データの違いによる正答率の変化を図 5 に示す. 結果から, 学習データ・検定データの全てを正規化した場合に最も正答率が大きくなっていた. また, 正規化データを平均値域で展張した場合では, 学習器に RandomForest を使用した場合で, 生の値をそのまま使用した場合と比較してより高い正答率を得た.

表 4: 交差検定の結果 (SVM)

学習データ	Random Forest	SVM
すべて正規化 (参考)	71.32%	72.79%
提案手法	68.59%	53.57%
無加工データ	50.74%	52.21%

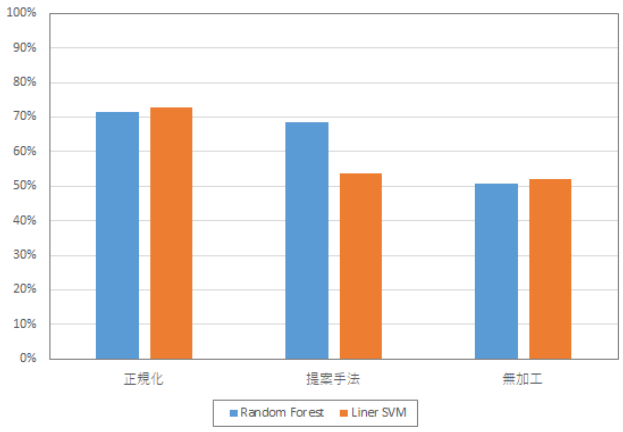

図 5: 正規化手法による交差検定結果

7 結論

本論文では, インタビュー対象者の態度を推定することにより, 対象者の対話への意欲を高めつつ, より多くの情報を引き出すインタビュー対話ロボットシステムの開発に向け, マルチモーダル情報を用いた態度推定を行った.

この検討を行うために, 対象者のマルチモーダル情報をインタビュー対話中に取得するロボットシステムを構築した. インタビュー実験を行い, 得られたマルチモーダル情報に対して, 発話の積極性との関係について T 検定を用いた分析を行った結果, 音声特徴量については, 発話長とピッチの最大と最小と平均, および MFCC とエネルギーについて有意差が存在することを確認した. 姿勢特徴量についてはインタビュー対象者によって左右の違いが見られたが, 肩や肘の座標に共通して有意差が見られた.

特徴量の分析結果を用いて, 実際に RandomForest と LinerSVM の二つの学習器を用いて機械学習による発話意欲の推定を行い, 交差検定により推定精度を検証した結果, マルチモーダル情報を用いて, 最大で 72.29% の割合で発話意欲を正しく推定できた.

また, 対話中にインタビュー対象者のマルチモーダル情報を正規化できない課題に対して, 正規化学習データを平均値域で展張することにより生の値で態度推定を行う手法について検討した結果, 観測データをそのまま使用する場合よりも高い正答率が得られることを確認し, 同手法が有用であることを確認した.

参考文献

[Saito2015] Naoko Saito, Shogo Okada , Katsumi Nitta, Yukiko Nakano, and Yuki Hayashi: Estimating User's Attitude in Multimodal Conversational System for Elderly People with Dementia, 2015 AAAI Spring Symposium Series(2015)

[岡田 16] 岡田将吾, 松儀良広, 中野有紀子, 林佑樹, 黄宏軒, 高瀬裕, 新田克己: マルチモーダル情報に基づくグループ会話におけるコミュニケーション能力の推定, 人工知能学会論文誌 (2016 年 08 月)

[石原 17] 石原卓弥, 長澤史記, 岡田将吾, 新田克己: インタビュー対話における重要シーン抽出のための言語・非言語特徴量の分析, 人工知能学会全国大会 (2017)

[千葉 14] 千葉祐弥, 伊藤彰則: ユーザの対話意欲を考慮したユーザプロファイリング対話システムのためのインタビュー対話の分析, 電子情報通信学会技術研究報告 (2014)

[Kobori16] Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura : Small Talk Improves User Impressions of Interview Dialogue Systems, 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue

[中野 15] 中野幹生, 駒谷和範, 船越孝太郎, 中野有紀子 : 対話システム (自然言語処理シリーズ), コロナ社 (2015)

[井上 16] 井上 昂治, Divesh Lala, 高梨 克也, 河原 達也: 階層ベイズモデルを用いた聞き手の多様なふるまいに基づく対話エンゲージメントの推定, SIG-SLUD 102-107(2016)

[長澤 17] 長澤史記, 石原卓弥, 岡田将吾, 新田克己: ユーザーの態度推定に基づき適応的なインタビューを行うロボット 対話システム構築への一検討, 人工知能学会全国大会 (2017)