サロゲートデータに基づく動作予測器を用いた傾聴対話ロボットの開発

Development of a Attentive Listening Robot Using The Motion Prediction Model Based on Surrogate Data

野口 翔平 [1]　中村 泰 [2]　岡留 有哉 [1,2] *
Shohei Noguchi[1]　Yutaka Nakamura　Yuya Okadome[1]

[1] 東京理科大学 工学部 情報工学科
[1] Tokyo University of Science, Faculty of Engineering, Department of Information and Computer Technology
[2] 理化学研究所 情報統合本部
[2] RIKEN Information R&D and Strategy Headquarters

Abstract: The development of nonverbal behavior functions such as nodding is also important for natural dialogue robots. However, since natural nonverbal dialogue functions is not developed, it is not possible to gather natural human-robot dialogue data. In this study, we developed the attentive listening robot system that uses human-human dialogue data as surrogate data. The nodding prediction model substitute human-human dialogue data for human-robot dialogue data to predict the behavior of the dialogue robot. The proposed system makes a judgment on whether to nod based on the output of the prediction model, audio and image information are input to the model. The results of the attentive listening experiment suggested that the proposed system that generates noddings that make it easier to speak was developed.

1　はじめに

近年、介護や接客の分野における人手不足を解消する手段として、コミュニケーションロボットの需要が高まっている [1]。このようなロボットは、主に会話や動きを用いて人間とのインタラクションを行う [2]。実体を持ちながら周囲の環境に応じた行動をとるため、人間に対して人のようにインタラクションを行うことが期待できる。

コミュニケーションは言語的要素と非言語的要素の 2 つの要素からなる。対話ロボットの言語的要素、特に発話要素は ChatGPT[3] などの大規模言語モデルの発展により、性能が飛躍的に向上している。一方の非言語的要素は、顔の表情やジェスチャーといった身体の動き、発話のテンポなど多岐にわたる [4]。

非言語的動作を行うロボットの研究として、Joanna ら [5] は声量を指標として頷き、目線移動、まばたきの 3 つを行うロボットシステムを開発した。東条ら [6] は発話者の顔や方向、発話内容などを認識して頷きなどを行うロボットシステムを開発し、表情や動作がある場合の方が印象が良いことを示した。Chaoran ら [7] はルールベースのシステムを用いて、発話中のロボットに適切なタイミングで頷きなどを行わせることを目的とする研究を行った。Chidchanok ら [8] はロボットに用いる非言語的コミュニケーションシステムの違いによる人間からの好感度の差異を調査した。自然な対話ロボットの実現のためにはこのような非言語的動作機能の開発が重要となる。しかしながら、人-ロボット対話で生じる全ての状況を考慮し、動作を設計することは容易でない。

そこで本研究では、人-人の対話データをサロゲート動作データ [9] として用いる傾聴対話ロボットシステムを開発する。サロゲートは代理を意味しており、本システムにおいては人-ロボット対話データの代わりに人-人対話データを用いて対話ロボットの動作予測を行う。提案システムでは取得した話者の音声や、首と目線の角度、表情を事前に用意したサロゲートデータと組み合わせる。組み合わせたデータを頷き予測モデルに入力し、ロボットの動作予測を行う。

提案システムを傾聴対話実験に適応し、サロゲートデータを用いた動作予測について検証した。本実験では実験参加者は対話ロボットに 1 分の発話を行い、その間対話システムは頷き動作を表出する。実験結果から、提案システムの頷きは話者に話しやすくなる印象を与えるものであることが示唆された。

*連絡先：東京理科大学 工学部 情報工学科
東京都葛飾区新宿 6 − 3 − 1
E-mail: okadome@rs.tus.ac.jp

2　提案手法

本研究では非言語的動作として、頷きを行う対話ロボットシステムを開発する。図 1 に提案システムの概要を示す。取得した音声データと画像データから抽出した特徴量に対して事前処理を施す。その特徴量を頷き予測モデルに入力し、頷き動作を予測する。

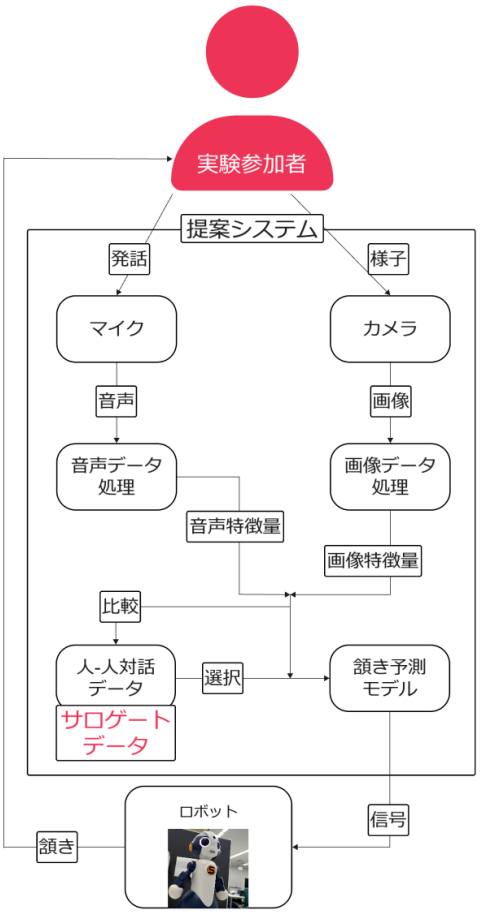

図 1: 今回のシステムの概要図

2.1　取得データへの事前処理

音声データにはラウドネス正規化を事前処理として行う。ラウドネス正規化を行うことで、マイクからの口位置の変化に起因する声量の増減の影響を低減できる。

画像データに OpenFace[10] を適応することで、顔に関する特徴量を計測する。画像データから取得した特徴量には、首と目線の角度及び Facial Action Unit(FAU) が含まれる。取得した特徴量に対して、首と目線の角度データの差分変換及び 0 次ホールドを事前処理として行う。首と目線の角度データの差分変換は、実験参加者の初期配置の影響を減らすために行い、前フレームのデータからの相対的な変化量である時系列差分を用いる。0 次ホールドは、顔認識の際に首の角度や目線の情報が欠落したデータを 1 フレーム前のデータで補完する方法である。

2.2　頷き予測モデル

頷き予測モデルでは 0.5 秒先で頷くかどうかを判断する。アノテーションのコストを削減するために、岡留ら [9] と同様の 2 者間対話を対象とした自己教師あり学習を用いた事前学習を行った。自己教師あり学習の際はラグ操作 [9] と呼ばれるデータ拡張を行う。ラグ操作により片方の話者の時刻をずらすことで、対話データに含まれる時間構造を崩す。このような自己教師あり学習を行うことで事前学習重みを獲得する。その後、少量の頷きラベル付きデータを用いた教師あり学習を行うことで頷き予測モデルを得る。なお、本頷き予測モデルは人-人コミュニケーションデータを用いて訓練する。

頷き予測モデルを f とすると、頷き予測モデルの推定結果は $\hat{l} = f(X_L(t), X_R(t); \omega)$ と表現できる。$\hat{l}$ は頷き予測タスクにおける 0.5 秒後の頷きラベルを表す。なお、$X_L(t)$ と $X_R(t)$ はある時刻 t から過去 T ステップ分の 2 人の特徴量を表す。ω は予測モデルのパラメータを表す。頷き予測モデルの構造は図 2 に示す 5 層の CNN モデルとした。なお、経験的に画像データと音声データをどちらも 10fps のデータにダウンサンプリングする。

2.3　頷き表出判断

頷き予測モデルに入力するためのサロゲート動作データの選択方法を述べる。システム実行時には頷き予測モデルの訓練時と同様に、入力データとして 2 人分の特徴量が必要となる。しかしながら、セッション中は発話者の特徴量のみ取得可能であるため、聞き手の特徴量をサロゲート特徴量で代用する。リアルタイムで取得した発話者の特徴量 $X_R^T(t)$ と、サロゲート特徴量 X_L^S を組み合わせて入力データとする。

組み合わせの概要は図 3 に示す。人間の動きは 0.1 秒間隔では連続した動きであると期待されるため、$\hat{l}(t)$ は $\hat{l}(t-0.1)$ と同じラベルである確率が高いと期待できる。そのため、直前の頷き予測結果を条件として、2 つのサロゲートデータ群 $\bar{X}_0^S, \bar{X}_1^S$ のどちらを用いるのかを選択する。$\bar{X}_0^S, \bar{X}_1^S$ は、それぞれ頷き予測モデルの訓練データを頷いていない場面と頷いている場面に分別した部分集合を表す。

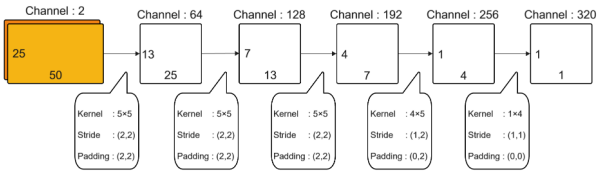

図 2: CNN のネットワーク構造

頷き予測モデルへの入力データ $X(t)$ は、リアルタイムで取得した発話者の特徴量 $X_R^T(t)$ とサロゲート特徴量 X_L^S を組み合わせたデータであり、組み合わせるサロゲートデータのインデックス k は

$$p = \hat{l}(t - 0.1) \tag{1}$$

$$c_i = \frac{\langle X_R^T(t), \bar{X}_{p,R_i}^S \rangle}{||X_R^T(t)||\ ||\bar{X}_{p,R_i}^S||} \tag{2}$$

$$k = \text{argmax}\{c|c = c_0, c_1, ...\} \tag{3}$$

と計算する。ここで、$\bar{X}_{p,i}^S = (\bar{X}_{p,L_i}^S, \bar{X}_{p,R_i}^S)$ は i 番目のサロゲートデータの 2 者の特徴量を表す。リアルタイムの発話者の特徴量 $X_R^T(t)$ とサロゲートデータにおける発話者の特徴量 $\bar{X}_{p,R_i}^S$ の cos 類似度を c_i と表す。式 3 で得られた k を用いると入力データは $X(t) = [\bar{X}_{p,L_k}^S, X_R^T(t)]$ と表現できる。

頷き予測モデルから出力された確率に基づき、ロボットが頷くかどうか判断する。なお、本研究では経験的に設定した閾値 (0.6) を超えた場合にのみロボットが頷くとした。

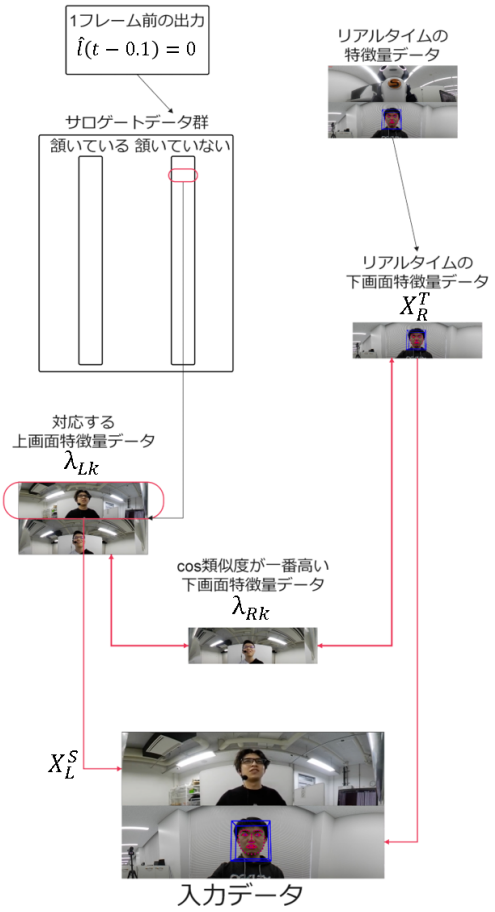

図 3: データの組み合わせの概要

話し手が交代する際の無言区間の時間が約 0.2 秒である [11] ことを考慮して、ルールベースシステムを設計した。

3 実験内容

3.1 実験方法

実験参加者は 1 分間の対話セッションを行い、ロボットに関する印象評価を行う。本実験の実験参加者は東京理科大学所属の学生 12 人 (男子 12 名、年齢 22.0 ± 1.0 歳) であり、有効回答数は 10 であった。

実験参加者はロボットに対面して着席した状態で話しかける。実験参加者は以下の 2 つのシステムに対して発話を行い、それぞれのシステムを用いたロボットへの印象評価を行った。

- 提案システム：
  サロゲートデータを用いた頷き予測を行う
- ルールベースシステム：
  発話が 0.2 秒途切れた場合に 90%の確率で頷く

3.2 アンケート内容

本実験で使用したアンケートは 5 段階のリッカート尺度を用いた 14 問の質問で構成した。アンケート内容は、井上ら [12] の傾聴対話ロボットへの主観評価アンケートをもとにして、質問数は実験参加者の負担を考慮して作成した。頷き予測モデルとルールベースの頷きモデルに関する質問ごとの差異を Mann-Whitney の U 検定を用いて求めた。

3.3 実験結果

表 1 に質問項目と検定結果を示す。U1 は提案システムの検定統計量を表し、U2 は比較手法の検定統計量を表す。表 1 における 14 問の項目のうち、特徴的な

表 1: 検定結果

質問の内容	U1	U2	p 値
(Q1) しゃべりやすかった	67.5	32.5	0.087^{+}
(Q2) 気恥ずかしさを感じた	48.5	51.5	0.565
(Q3) 嬉しい気分になった	52.0	48.0	0.448
(Q4) 落ち着かない部分があった	50.0	50.0	0.516
(Q5) ロボットの振る舞いは自然だった	56.0	44.0	0.327
(Q6) ロボットはタイミングよく反応していた	46.5	53.5	0.623
(Q7) ロボットの反応は人間らしかった	65.0	35.0	0.125
(Q8) ロボットの反応はあなたの話を適切に促してた	43.0	57.0	0.725
(Q9) このロボットとまた話したい	45.5	54.5	0.657
(Q10) ロボットは真面目に話を聞いていた	62.5	37.5	0.173
(Q11) ロボットは話を理解していた	39.5	60.5	0.814
(Q12) ロボットは話に対する関心を示していた	49.0	51.0	0.548
(Q13) ロボットはあなたに対して共感を示していた	47.0	53.0	0.610
(Q14) ロボットはあまり集中して話を聞いていなかった	56.0	44.0	0.330

$+ : p < 0.1$

結果が得られた Q1, Q7, Q10 について述べる。それぞれ”(Q1) しゃべりやすかった”で $p = 0.087$、”(Q7) ロボットの反応は人間らしかった”で $p = 0.125$、”(Q10) ロボットは真面目に話を聞いていた”で $p = 0.173$ となっていた。”(Q1) しゃべりやすかった”の質問項目において $p < 0.1$ であるため、有意傾向がみられた。

4 議論

(Q1) の質問項目で有意傾向が見られた理由について考える。頷き予測モデルでは、実験参加者が発話し続けているタイミングでも頷く場合がある。実験参加者はこのような発話中の頷きに発話を促す意味合いがあると認識していた可能性が考えられる。しかし、「ロボットの反応はあなたの話を適切に促してた」(Q8) では有意性が見られないことや、聞き取りの際に「急かされているように感じた」と答えた人がいたことから、ロボットの頷きの量やタイミングが適切でない可能性が考えられる。その一方で、頷き動作のみであっても人の発話に影響を与えることが示唆された。

対話システムの頷き判断は、取得した表情の大きさを表す Facial Action Unit(FAU) の影響を強く受けると考えられる。特に OpenFace における FAU の推定は個人差があることから、頑健な FAU 推定がシステムのふるまいの安定性につながると考えられる。

5 まとめ

本研究ではサロゲート動作を用いた頷き生成システムを作成した。本システムではリアルタイムに得られる特徴量に対し、サロゲート動作としてデータベースから取得した特徴量を組み合わせる。さらに、音声データや首と目線の角度、顔の表情といったマルチモーダルなデータを扱う。実験より、本システムが話し手が話しやすくなるような頷きを生成できることが示唆された。

本研究で作成した頷き予測モデルの挙動は FAU の影響を強く受ける。そのため、より正確に顔の表情を取得できるようにシステムを改善することが今後の課題である。また、「うん」や「あー」といった音声の表出も今後の課題として挙げられる。

参考文献

[1] Alessandro Vercelli, Innocenzo Rainero, Ludovico Ciferri, Marina Boido, and Fabrizio Pirri. Robots in elderly care. *ICT & Neurodegenerative Diseases [special issue] Vol 2 No 2*, pp. 37–50, Mar,2018.

[2] Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems,Volume 63, Part 1*, pp. 22–35, Jan, 2015.

[3] ChatGPT OpenAI. (https://openai.com/chatgpt), 閲覧日 (2024/2/11).

[4] Shane Saunderson and Goldie Nejat. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics Volume 11*, pp. 575–608, Jan,2019.

[5] Joanna Hall, Terry Tritton, Angela Rowe, Anthony Pipe, Chris Melhuish, and Ute Leonards. Perception of own and

robot engagement in human–robot interactions and their dependence on robotics knowledge. *Robotics and Autonomous Systems Volume 62*, pp. 392–399, Mar,2014.

[6] T. Tojo, Y. Matsusaka, T. Ishii, and T. Kobayashi. A conversational robot utilizing facial and body expressions. *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0*, pp. 858–863, Oct,2000.

[7] Chaoran Liu, Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. *2012 7th ACMIEEE International Conference on Human-Robot Interaction (HRI)*, pp. 285–292, Mar,2012.

[8] Chidchanok Thepsoonthorn, Ken ichiro Ogawa, and Yoshihiro Miyake. The relationship between robot's nonverbal behaviour and human ' s likability based on human ' s personality. *Sci Rep 8, 8435 (2018).*, May, 2018.

[9] 岡留有哉, 阿多健史郎, 石黒浩, 中村泰. 対話中の振る舞い予測のための時間的整合性に注目した自己教師あり学習. 人工知能学会論文誌 37 巻 6 号, pp. B–M43_1–13, Nov,2022.

[10] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. pp. 59–66, May,2018.

[11] Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language Volume 67 101178*, May,2021.

[12] 井上昂治, ラーラーディベッシュ, 山本賢太, 中村静, 高梨克也, 河原達也. アンドロイド erica の傾聴対話システム–人間による傾聴との比較評価–. 人工知能学会論文誌 36 巻 5 号, pp. H–L51_1–12, Sep, 2021.