ヒューマノイドロボットへの話しかけやすさの予測

Predicting How Likely User is to Talk to Humanoid Robot

杉山　貴昭[1*]　駒谷　和範[1]　佐藤　理史[1]
Sugiyama Takaaki[1], Komatani Kazunori[1], Sato Satoshi[1]

[1] 名古屋大学大学院　工学研究科　電子情報システム専攻
[1] Graduate School of Engineering, Nagoya University

Abstract: We tackle a novel problem to predict how likely a humanoid robot is to be talked by a user. A human speaker usually takes his/her addressee's state into consideration and chooses when to talk to the addressee; this convention can be used when a system interprets its audio input. We formulate this problem by using machine learning whose input features are a humanoid's behaviors such as its posture, motion, and utterrance. A possible application of the model is to reject environmental noises that occur at timing when a cooperative user hardly talks to a robot.

1　はじめに

人間同士の対話には，対話者同士が無意識のうちに守っているルールが存在する．例として，聞き手は話し手の状態を考慮して話しかけることや，話し手は聞き手の方向を向いて発話することが挙げられる．このようなルールを本稿では社会的規範と呼ぶ．我々は，対話相手が人間ではなく，ヒューマノイドロボットである場合も同様の傾向があると考える [1]．つまり，人間に類似したロボットとユーザとの対話では，ユーザは社会的規範を守りながら，ロボットと対話すると考えられる．

本研究では，この社会的規範のうち，特に聞き手が話し手の状態を考慮して話しかけることのモデル化を試みる．具体的には，ロボットの一連の発話や挙動に対して，ユーザが話しかけられると感じるタイミングを，ロボットが予測するモデルの構築を目指す．協調的なユーザは，社会的規範を守って話しかけることから，話しかけやすいタイミングと話しかけにくいタイミングが存在することを利用している．

話しかけやすさを予測する枠組みを図 1 に示す．入力は，任意の時点でのロボットの状態であり，これから話しかけやすさに寄与する特徴を設計する．この特徴を用いてロジスティック回帰を行い，話しかけやすい，話しかけにくいの 2 値を出力する [2]．

本研究では，まずユーザスタディを行い，ロボットの挙動に対し，実際に話しかけやすいかどうかを付与したデータを収集する．次に，収集したデータから機械学習に用いる学習データを作成し，話しかけやすさを予測するモデルを構築する．そして，評価実験により，本モデルが話しかけやすさを予測できることを確認する．本稿の最後に，今後の展望について述べる．

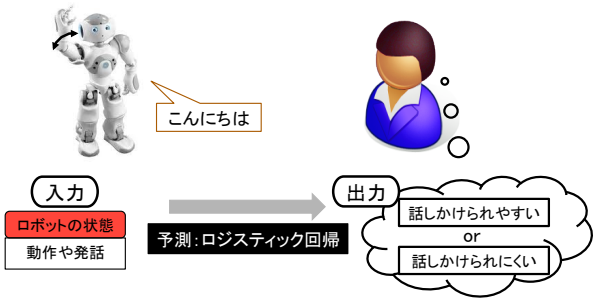

図 1: 話しかけやすさを予測する枠組み

2　実現可能な対話例

本モデルにより，人とロボットとの音声対話において，例えば以下のようなことが実現できる．

まず，その時点での話しかけられやすさを考慮することで，ロボットは自身への入力音をより高精度に解釈できる．ヒューマノイドロボットを用いる場合，ロボットの動作音など様々な雑音が入力されうるため，近接マイクを利用する場合より解釈が難しい．そこで，本研究では入力音ではなく，その受け手であるロボットの挙動に着目する．つまり，ロボットが自身の発話や動作に基づき，ユーザにとって話しかけやすい状況にあるかどうかを予測する．ユーザが話しかけにくいと感じるタイミングでは，ユーザからロボットへの入力

*連絡先：名古屋大学 工学研究科 電子情報システム専攻
愛知県名古屋市千種区不老町 C3-1(631)
E-mail:takaak_s@nuee.nagoya-u.ac.jp

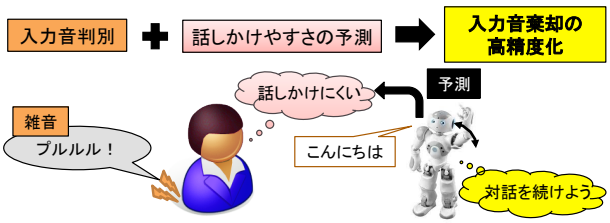

図 2: 実現可能な対話例

がある可能性が低い．そのため，このタイミングでの入力音は，雑音等である可能性が高いとみなせる．このように，協調的な対話において，雑音とユーザ発話が生じる事前確率を与えるモデルとして，本モデルが利用できる．これにより，雑音等をより高精度に棄却できる．例えば，図 2 に示すような，ロボットとユーザとの対話中に，ユーザ方向で雑音が発生した場合を想定する．従来このような誤動作回避は，入力音の判別に基づき行われることが多い [3, 4, 5]．本モデルを利用すれば，ロボットは入力音判別に加えて，話しかけられやすさの予測ができるため，図のような対話が実現できる．

この他にも，ロボットに対してユーザが話しかけにくい状況を，ロボット自身で作り出すことが可能になる．実環境では，入力音を受理しにくい状況が存在する．例えば，周辺雑音やロボットのファン雑音が発生している場合である．この時に，ロボットがユーザにとって話しかけにくい状況を作れれば，ユーザに話しかけられないように振る舞うことができる．逆に，その後入力音を受理できる状況になれば，話しかけやすいように振る舞うこともできる．このように，ユーザの発話タイミングをコントロールできる可能性がある．

3　話しかけやすさの定義と予測方法

本研究で議論する話しかけやすさは，ロボットがユーザに説明しているときに，ユーザがロボットに話しかけやすいと感じるか話しかけにくいと感じるかであるとする．本研究では，以下の 3 点を仮定している．まず，ユーザがロボットに話しかけたい内容は深刻性ではないとする．これは，ユーザが深刻な内容をロボットに話しかけたい場合，ロボットの状態に関わらず，ユーザはその内容を伝えようとするためである．次に，ユーザはロボットを擬人化された存在として感じていると仮定する．この仮定が満たされなければ，ユーザはロボットの状態を考慮しないで発話し得る．本研究ではヒューマノイドロボットを用いることで，この仮定は満たされているとする．最後に，本稿では簡単のため，ユーザは 1 名であるとする．

本稿で扱う話しかけやすさの予測は，人間同士の会話における transition relevance place（TRP）の予測に相当する．TRP とは，発話交替が起こり得る場所である．Sacks らは，聞き手がこの TRP を予測しながら対話を行っていることを示している [6]．さらに，非言語行動（例えば，視線）が発話交替に寄与することも知られている [7, 8]．これらは人間同士の会話を想定した研究であるのに対し，我々は人間とロボットのインタラクションを想定している．ロボットのマルチモーダルな動作から，聞き手であるユーザがターンを取る可能性があるタイミング (TRP) を，ロボットが予測する．これにより，例えば，ロボットがその時点でユーザに話しかけられる事前確率の取得に，提案モデルの出力を用いることができる．

話しかけやすさの予測にロジスティック回帰式（(1) 式）を用いる．

$$P(y|x_1, x_2, ..., x_n) = \frac{1}{1 + \exp(-f(\boldsymbol{x}))} \qquad (1)$$

$$f(\boldsymbol{x}) = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n$$

ここで，$y \in \{0, 1\}$ は目的変数，$P(y|x_1, x_2, ..., x_n)$ は，入力特徴 $x_1, x_2, ..., x_n$ に対して，y が 1 の値をとる条件付き確率であり，a_n は係数である．本研究では，0.1 秒ごとのロボットの挙動に対して，話しかけやすいか，話しかけにくいかをロジスティック回帰により判別する．ここでは目的変数としてユーザがロボットに「話しかけやすい場合」は 1,「話しかけにくい場合」は 0 を割り当てる．判別は，確率 P に対する閾値処理（閾値 0.5）として行われる．つまり，$P \geq 0.5$ で話しかけやすい，$0.5 > P$ で話しかけにくいと判別される．

ロボットに対する話しかけやすさに寄与する要素として，ロボットの動作や姿勢，発話を扱う．これらは，後に機械学習の特徴として用いるため，ロボット内部で自動的に取得可能なもののみを考える．ロジスティック回帰の入力特徴として，表 1 に示す 9 つを用いる [2]．主に，ロボットの発話，動作，視線に関する特徴である．これらの特徴は 0.1 秒ごとに取得し，その時点での話しかけやすさの判定に用いる．なお，本研究で用いたロボットの眼球は動かないため，首の角度と視線の方向は完全に対応する．

4　予測モデルの構築

4.1　話しかけやすさを付与したデータの収集と学習データの作成

データの収集方法として，被験者にロボットの一連の挙動を通して見せ，話しかけやすさを付与させる方法を採用した．これは，話しかけやすさは前の発話や

表 1: ロボットの挙動を表す入力特徴

特徴	取得方法
発話間間隔	ロボット発話終了からの経過時間 [秒]
発話の文末表現	発話交替表現を用いたか (0 または 1)
発話の文末の韻律	韻律が上昇する表現を用いたか (0 または 1)
動作（頭）	0.1 秒前の角度との差 [度]
動作（左腕）	0.1 秒前の角度との差 [度]
動作（脚）	0.1 秒前の角度との差の両脚の和 [度]
動作（右腕）	0.1 秒前の角度との差 [度]
視線（水平方向）	首の関節角の，正面からの角度差 (水平方向)[度]
視線（垂直方向）	首の関節角の，正面からの角度差 (垂直方向)[度]

表 2: 学習データの作成方法

	被験者	挙動	利用区間
データ α	一般ユーザ	挙動 Y	21 名以上一致
データ β	研究室の学生	挙動 Y	3 名全員一致

挙動に関係すると考えたためである．話しかけやすさを記録する方法として，計算機のディスプレイに表示された GUI を用い，被験者にマウスをクリックさせた [2]．

我々は，被験者や人数の違いによる性能の違いを調べるため，2 種類の被験者集合を対象に実験を行った．1 回目は，被験者として本研究室の学生 3 名を対象とした．2 回目は，被験者として 20 代〜50 代までの 25 名（男性 13 名，女性 12 名）を一般から募集した．年齢の平均は 37.9 才であり，各年代の人数が均等になるようにした．

ロボットの一連の挙動は，[2] で作成した 2 種類の挙動（以降，挙動 X，挙動 Y とする）を用いる．内容は，どちらもロボットの自己紹介である．長さは，1 つ目の挙動 X は 150.0 秒，2 つ目の挙動 Y は 259.3 秒である．挙動 Y は，挙動 X に比べて，発話や動作のバリエーションや組み合わせが多い．これらは，後に，学習・テストデータに用いる．ヒューマノイドロボットには Aldebaran Robotics 社製の NAO[1]を使用し，音声合成には VoiceText[2]を使用した．

話しかけやすさは，ユーザがロボットにどの程度すぐに話しかけたいと感じているか（緊急度合）によって異なる．そのため，本実験では，全被験者に下記のような状況を想定させて，ロボットの挙動に対して話しかけやすさを付与させた．

> あなたは「もう少し大きい声で喋ってほしい」とロボットに伝えたい

この内容は，被験者に緊急度合をわかりやすく教示するためのものであり，話しかける内容を限定するものではない．なお，論文 [9] では，本モデルのパラメータ（ロジスティック回帰の閾値）を教示内容の変化によって変更することも検討している．

[1]http://www.aldebaran-robotics.com/
[2]http://voicetext.jp/

表 3: 学習データに用いる話しかけやすさのフレーム数

	話しかけやすい	話しかけにくい	合計
データ α	239	1123	1382
データ β	161	1269	1430

実験で収集したデータから，ロジスティック回帰の学習に用いるデータを作成する．被験者や人数の違いによる性能の違いを調べるため，条件が異なる 2 種類の学習データ α，β を準備した．表 2 に学習データ α，β の作成方法を示す．被験者として，学習データ α では一般ユーザ，学習データ β では本研究室の学生を用いた．学習データに利用した区間は，学習データ α では，21 名以上の被験者の話しかけやすさが一致した区間であり，学習データ β では 3 名全員が一致した区間とした．それぞれの学習データに用いた話しかけやすさのフレーム数は，表 3 に示す通りである．21 名以上が一致した区間を利用した理由は，本実験では 25 名の一般ユーザを対象としており，個人によって話しかけやすさは異なるため，話しかけやすさが全員一致する区間は少ないからである．以降では，各被験者の話しかけやすさが一致した区間を共通区間とする．この共通区間内の，各時点でのロボットの挙動から特徴の値を得て，対象データとする．被験者が話しかけやすいとしたサンプル数は，話しかけにくいとした場合よりも少ない．このため，被験者が話しかけやすいとしたサンプルに対して，Over Sampling [10] を行い，サンプル数の比を重みとして与え，学習を行った．

4.2 モデルの性能評価

挙動 Y を用いて収集したデータを学習データ，挙動 X を用いて収集したデータをテストデータとして，評価を行う．評価指標として,「話しかけやすい」「話しかけにくい」の正解ラベルと，ロジスティック回帰の出力が一致した数から，MacroF1[3] を計算する．

MacroF1 を次の条件で計算し，性能を検証する．

条件 (1) 被験者の共通区間に対して予測
条件 (2) 被験者の全区間に対して予測

条件 (1) では，一般ユーザ 25 名のうち，多くのユーザの話しかけやすさが一致した区間（共通区間）に対して予測を行う．ここでは被験者 25 名全区間のデータのう

[3]「話しかけやすい」の F 値と「話しかけにくい」の F 値の平均値

表 4: 学習データによるモデル性能の差（MacroF1）

モデルの学習データ（挙動 Y）	挙動 X に対するオープンテスト		
	(1)	(2) 全区間	
	共通区間	一般ユーザ 25 名	研究室学生 3 名
データ α	84.9	69.8 ± 8.4	69.0 ± 2.5
データ β	84.3	69.6 ± 9.1	69.4 ± 2.5

ち，本モデルで予測した時に最も MacroF1 が高くなった，18 名以上の話しかけやすさが一致した区間をテストデータとした．条件 (2) では，一般ユーザ 25 名の全区間のデータ，または，本研究室の学生 3 名の全区間のデータに対して予測を行う．これらのデータは，ロボットの挙動 X の 1500 フレーム全てに対して，各被験者が話しかけやすさを付与したものである．

表 4 に各モデルの性能を示す．なお，条件 (2) の結果は，被験者それぞれに対して MacroF1 を計算し，これらの平均をとった値である．データ α のモデルの性能を見ると，共通区間に対しては 85%程度，それを含めた全区間に対しても 70%程度予測できることがわかった．また，データ α のモデルとデータ β のモデルを比較すると，各条件でほぼ同等の性能が得られた．これにより，モデルの構築方法とその性能は，ある特定の被験者集合に依存しないことがわかった．

さらに，表 4 の各条件間の予測性能の差について考察する．まず，条件 (1) の値に比べ，条件 (2) の値が低い．これは，本モデルによって共通区間の予測はできているが，被験者によって話しかけやすさが異なる区間の予測は難しいことが示されている．この理由は，本モデルが共通区間を学習データとして利用しているためである．また，条件 (2) では，一部の被験者の「話しかけやすい」の F 値が，他の被験者と比較して，低かった．これは，話しかけやすさの感じ方には個人差があり，特に話しかけやすいと感じる区間の方がこの差は顕著なためである．そのため，この個人差を考慮して話しかけやすさを予測する必要がある．

5 今後の展望

本モデルをオンラインで利用するうえでの今後の展望を述べる．

まず，データの収集方法が適切であったかどうかを検証する必要がある．本研究で実施したデータ収集の方法は，ユーザに話しかけやすいと感じたタイミングでマウスをクリックさせるという方法であった．この方法はロボットとユーザが実際に対話するようなインタラクティブな方法ではないため，得られたデータはユーザが本来感じる話しかけやすさとは異なる可能性がある．そこで下記の 2 点について確認する必要がある．1 点目は，インタラクティブな方法で新たにデータを収集し，これまでに得たデータの妥当性を確認することである．具体的には，話しかけやすいタイミングでユーザに発話させる方法でデータを収集し，ユーザが発話したタイミングとマウスをクリックしたタイミングがほぼ一致しているかどうかを調べる．これらが一致していれば，これまでに収集したデータが本モデルの学習データとして妥当であることが示せる．2 点目は，ロボットの挙動の変化がユーザの話しかけやすさに影響を与えるかどうかを確認する必要がある．よりインタラクティブなデータ収集の方法として，ユーザがロボットに話しかけた時にロボットの挙動を変化させることが考えられる．この方法では，ユーザが話しかけやすさを付与するタイミングは，ロボットの挙動の変化によって異なる可能性がある．そこで，よりインタラクティブな方法で実験を行い，これまでに得たデータとの比較を行う．ロボットの挙動の変化や被験者への教示内容などの実験条件については，今後検討する．

次に，被験者間で話しかけやすさが異なる区間の予測精度の向上である．4.2 節で示した通り，本モデルは被験者間で話しかけやすさが異なる区間の予測が難しい．そこで，この区間を予測するために，被験者間で話しかけやすさが異なる区間を学習データに利用してモデルを作成し，これを予測することを検討している．これができれば，学習データの異なる複数のモデルの出力を統合することで，全体的な性能をさらに向上させられると考えている．また，本モデルのパラメータを個人毎に変更することも検討している．話しかけやすさの感じ方（ユーザの属性）は個人毎に異なる．例えば，ロボットがある動作をしている時に，ユーザ A は「話しかけやすい」，ユーザ B は「話しかけにくい」と感じる可能性がある．我々はユーザの属性をロジスティック回帰の閾値の変化で表現できるかどうかを検討している [9]．これにより，ユーザの属性が推定できれば，この閾値の変化で個人差に対応できることを示した．将来的にはこのユーザの属性は，対話中の情報（例えば，当該ユーザの発話頻度）から推定できると考える．

6 おわりに

本研究では，ユーザがロボットに話しかけやすいと感じるかどうかを予測するモデルを構築した．具体的には，ロボットの発話表現や動作，姿勢を入力としたロジスティック回帰により，ロボットが話しかけられやすい状況にあるかどうかを予測した．評価実験により，被験者が話しかけやすさを付与した全てのデータに対

し，70%程度予測できることを示した．また，本手法がある特定の被験者集合に依存しないことがわかった．さらに，本モデルをオンラインで利用するために必要な今後の課題について述べた．

参考文献

[1] B. Reeves and C. Nass. The media equation: How people treat computers, televisions, and new media as real people and places. *Cambridge University Press*, 1996.

[2] 杉山貴昭, 駒谷和範, 佐藤理史. ヒューマノイドロボットが話しかけやすさを予測するモデルの構築. 人工知能学会論文誌, Vol. 28, No. 3, pp. 255–266, 2013.

[3] W. Kim and H. Ko. Noise variance estimation for Kalman filtering of noisy speech. *IEICE Transactions on Information and Systems*, Vol. E84-D, No. 1, pp. 155–160, 2001.

[4] 野村行弘, 呂建明, 関屋大雄, 谷萩隆嗣. 雑音量に依存しない音声領域と雑音領域との判別を用いた音声強調. 電子情報通信学会技術研究報告. SP, 音声, Vol. 104, No. 30, pp. 29–34, 2004.

[5] A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari, and K. Shikano. Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. *Proc. Interspeech*, pp. 173–176, 2004.

[6] H. Sacks, A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, Vol. 50, No. 4, pp. 696–735, 1974.

[7] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, Vol. 23, pp. 283–292, 1972.

[8] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychol*, Vol. 26, pp. 22–63, 1967.

[9] 杉山貴昭, 駒谷和範, 佐藤理史. ロボットへの話しかけやすさモデルの評価と個人差や教示による変動への対応. 人工知能学会論文誌, Vol. 29, No. 1, pp. 32–40, 2014.

[10] N. E. Breslow and K. C. Cain. Logistic regression for two-stage case-control data. *Biometrika*, Vol. 75, No. 1, pp. 11–20, 1988.