1O2-OS-18-7

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

Building Model to Predict How Likely User is to Talk to Humanoid Robot

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

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

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. The proposed method predicts it by machine learning that uses a humanoid robot's behaviors as input features, such as its posture, motion, and utterance. 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

ヒューマノイドロボット（以下，ロボット）との音声インタラクションの実現には，周辺雑音による誤動作が問題になる．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 [2, 3]．我々は，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，ロボットが話しかけられやすい状況にあるかどうかを，ロボット自身がその発話や動作に基づき予測するモデルの構築を目指す．モデルの全体像を図 1 に示す．入力は，その時点でのロボットの動作や発話であり，これら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うことにより，話しかけられやすい，話しかけられにくいの 2 値を出力する．

ロボットが任意の時点で，話しかけられやすさを予測できれば，協調的なユーザが話しかけるであろうタイミングを，ロボットが知ることができる．逆に，ユーザにとって話しかけにくいと思われる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し，これを棄却できる．つまり入力音の解釈時に，その時点での話しかけられやすさの状態を考慮できるようになる．さらにこのモデルを用いて，ロボットにとって都合の悪い状況（内部ファンの動作中や周囲のキャリブレーション中など）に，話しかけられにくい挙動を生成・選択することも可能である．

本研究では，まずロボットの様々な挙動に対し，ユーザが実際に話しかけやすいかどうかを付与した学習データを作成する．その後，ロジスティック回帰により，任意の時点での話しかけられやすさを予測するモデルを構築する．

2. 話しかけやすさに寄与するロボットの動作や発話

2.1 話しかけやすさの定義

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．この話しかけやすさは，単にロボットが発話しているか否かのみでは予測できない．例えば，ロボットが発話せずに，ユーザに対して後ろを向いて動作している場合，ユーザは話しかけにくいと感じる可能性が高い．そのため，動作やロボットの目線などを特徴として用いて，これを予測する．本研究では，以下の 3 つの場合を想定している．

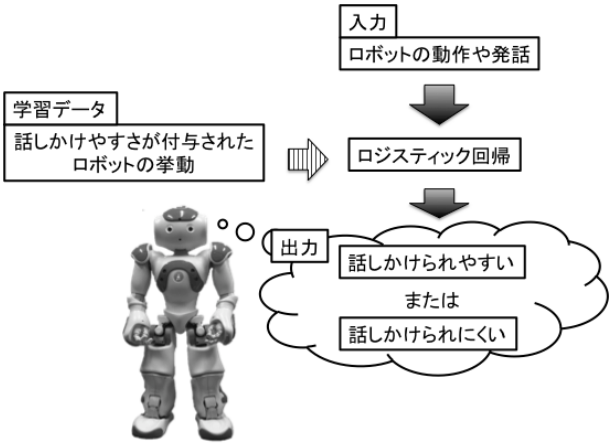

図 1: 話しかけられやすさの予測の全体像

1. ユーザは緊急性のない内容を話しかけようとしている
2. ユーザがロボットに対して社会性を感じている
3. ユーザが 1 名である

まず 1 点目として，話しかけやすさは，聞き手が話しかけようとしている内容に依存する．例えばユーザが「救急車を呼んでほしい」など緊急の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るので，話しかけやすさを定義するにはそぐわない．

次に 2 点目として，ユーザはロボットに対して社会性を感じていると仮定する．例えば，ユーザがロボットに命令しないと動作しないロボットである場合，ユーザはロボットに話しかけやすさを考慮して話しかけないだろう．本研究では，人間の外観に類似したロボットを用いているため，この仮定は満たされているとする．

最後に 3 点目として，本稿ではユーザは 1 名であると仮定して議論を進める．複数のユーザが異なる場所に位置する場合，ユーザによって話しかけやすさは異なる．例えば，ロボット

連絡先: 杉山貴昭，名古屋大学大学院 工学研究科 電子情報システム専攻，〒 464-8603 愛知県名古屋市千種区不老町 C3-1(631) IB電子情報館南棟159，052-789-4435，takaak_s@nuee.nagoya-u.ac.jp

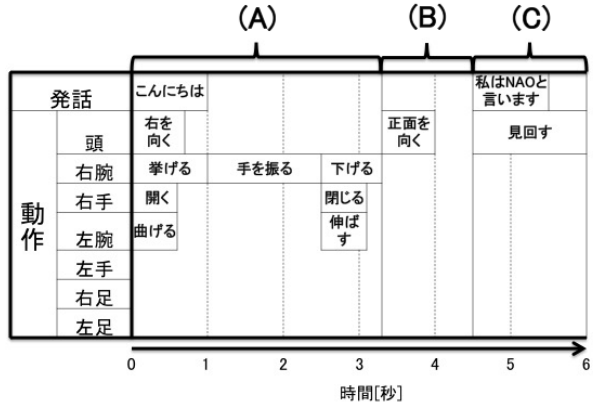

図 2: 作成した挙動の一部

がユーザ A の方を向いて無言で静止しており，ユーザ B はロボットの左側に存在している場合を考える．このとき，ユーザ A はロボットが自分のほうを向いているため話しかけやすいと感じ，ユーザ B はロボットがユーザ B の方を向いていないため話しかけにくいと感じると予想される．このように，ユーザが複数存在する場合，ユーザの位置によって話しかけやすさが異なるため，本稿ではまず簡単のために 1 人の場合を考える．

2.2 ロボットの挙動の作成

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．これらの要素と話しかけやすさの関係を，個々に議論することはできる．しかし実際のロボットの挙動では，これらが連続的かつ複合的に起こる．このため，実際のロボットの挙動において話しかけやすさを予測するには，個々の要素ごとの話しかけやすさを考えるだけでは不十分である．ここでまず連続的とは，ロボットがある挙動の後に続けて別の挙動を行うことである．連続的に挙動を行うと，ユーザはロボットの前の挙動との関係も考慮して話しかけやすさを考える．例えば，後ろを向いて話していたロボットが，次にユーザ方向を向いて話し始めるような場合，ユーザは話しかけやすいと感じる可能性が高い．

次に，複合的とは，ロボットが一度に複数の要素を含んだ挙動を行うことである．複数の要素を含んだ挙動は，要素の組み合わせ方によって話しかけやすさが異なる．例えば，発話していない時に，ロボットが同時にお辞儀をした場合，ユーザは話しかけにくいと感じる可能性がある．

本研究では，話しかけやすさに寄与する要素を連続的・複合的に含んだロボットの挙動を作成した．この挙動に対してまず人手で話しかけやすさを付与する．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．作成した挙動の内容はロボットの自己紹介であり，長さは 150 秒である．ユーザの位置はロボットの正面であるとしている．

作成した挙動の一部を図 2 に示す．区間（A）ではロボットが手を振りながら,「こんにちは」と発話している．その間，ロボットの頭・右腕・右手・左手が動いている．また顔は右方向を向き，正面に位置するユーザの方を向いていない．区間（B）では，ロボットは正面のユーザ方向を向き，発話をせず静止している．区間（C）では，1 秒程度の発話があり，周りを見回している．これら 3 つの区間のうち，区間（A）と区間（C）では要素を複合的に用いている．区間（A）では要素を連続的に用いている．

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

3. 学習データの収集と分析

3.1 データの収集

本研究では，特定のユーザに依存しない，話しかけやすさのモデルの構築を目指している．そのため，以下の 2 点をデータ収集の方法として採用する．

- 複数の被験者から話しかけやすさを付与したデータを収集する．
- 1 名の被験者から数回データを収集する．

作成したロボットの挙動に対して，実際に話しかけやすさの付与を行う．具体的には，被験者がロボットの挙動に対して話しかけやすいと感じた区間を記録する．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．何も操作していない時は「話しかけにくい」，マウスをクリックすると「話しかけやすい」と表示され，t 秒後に「話しかけにくい」に戻る．クリックから t 秒以内にもう一度クリックした場合，その時点から再度 t 秒間「話しかけやすい」が表示される．ここでは $t = 0.5$ 秒とした．この「話しかけやすい」が表示されていた区間を，ユーザが話しかけやすいと感じた区間とした．被験者には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た．これは，話しかけやすさは前の発話や挙動に関係すると考えたためである．

被験者は本研究室の学生 3 名とした．被験者は以下の手順で実験を行う．

1. 被験者に実験の事前教示をする．
2. GUI の使用方法を説明し，被験者に実際に練習させる．
3. 被験者をロボットの真正面に位置するように座らせる．
4. 被験者に実験で使用するロボットの挙動を数回鑑賞させる．
5. 同じ挙動に対して，3 回続けて，GUI により話しかけやすさを入力させる．

事前教示として，実験の設定や手順に関する文書を用意し，被験者にそれを読ませた．これは被験者ごとに説明が異なることや不足することを防ぐためである．被験者には，わからないことがある場合のみ質問させた．被験者には，話しかけやすさの判断に際して以下の状況を想定させた．

> あなたはロボットに「もう少し大きい声でしゃべってください」と伝えるために,「ねえ」と呼びかけるタイミングをうかがっている．

さらに，話しかけやすいと感じる区間が続く間は，続けてマウスをクリックするように指示した．被験者にロボットの挙動を数回見させたのは，初めてロボットの挙動を見る場合には見入ってしまうことが多いためである．これにより，被験者が話しかけやすさを付与し忘れることを防ぐ．

3.2 被験者間での一致度合の分析

3.1 節で述べたデータ収集の結果，被験者が話しかけやすいと感じた区間を 3 回ずつ記録したデータが，3 名分得られた．これを用いて，本節では被験者間の比較を行う．具体的には，被験者間で共通して話しかけやすいと感じる区間と，話しかけにくいと感じる区間が存在するかどうかを調査する．これによ

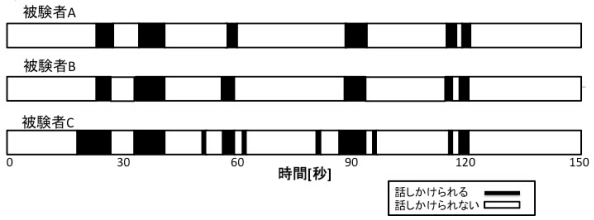

図 3: 話しかけやすさの付与結果

り，収集したデータが機械学習のための学習データとして有用であるかを調べる．

被験者間の比較には，各被験者の 3 回のデータのうち，以下で述べる理由により，2 回目のデータを用いた．1 回目のデータは，被験者が GUI の使用に慣れていないため，操作ミスをしている可能性がある．さらに，ロボットの挙動に見入ってしまい，話しかけやすさの付与し忘れが起こる可能性が 3 回の実験で最も高い．3 回目のデータでは，慣れや疲労が生じる可能性がある．

図 3 は，各被験者がクリックしていた時と何も操作していない時を示している．つまりこれは，ロボットの挙動に対して各被験者が話しかけやすいと感じた区間と，話しかけにくいと感じた区間を表す．横軸はロボットの挙動の再生時間 [秒] である．話しかけやすい区間は黒，話しかけにくい区間は白で表されている．3 名の被験者のデータを比較すると，再生時間 150 秒のうち，25 秒，35 秒，58 秒，90 秒，119 秒，121 秒の付近でそれぞれ，3 人とも話しかけやすいとしている．一方，話しかけにくい区間も 0 秒付近，150 秒付近など，ほとんどの区間で一致している．これより，各被験者とも共通して話しかけやすいと感じた区間や，話しかけにくいと感じた区間が存在することがわかる．

どの程度話しかけやすいとした区間が一致するかを，定量的に示す．各被験者が話しかけやすいとした区間長と，被験者間でのその重なりを表 1 にまとめている．表中の x と y は被験者を表し，例えば，A B は被験者 A と被験者 B の比較である．$t(x)$ と $t(y)$ はそれぞれの被験者が話しかけやすいと感じた区間の長さ，$t(x \cap y)$ はこの 2 名の被験者が共通して話しかけやすいと感じた区間の長さ [秒] である．さらにこれらの結果に対して重複度を以下の式で定義する．

$$\text{重複度} = \frac{t(x \cap y)}{min(t(x), t(y))} \quad (1)$$

これは話しかけやすいと感じた区間が短い被験者の結果が，どの程度, 共通して話しかけやすいと感じた区間を包含しているかを表す．話しかけやすいと感じた区間の共通部分の割合により，話しかけやすいとした区間の一致度を示す．

表 1 より，A と B，B と C，C と A の重複度はそれぞれ 0.889，0.927，0.963 である．これは，話しかけやすいと感じた区間が短い被験者が付与した区間の約 9 割で，他の被験者も同様に話しかけやすいと感じたことを表す．つまり，話しかけやすいとした区間の長さに個人差はあるものの，その共通部分の割合は多いことが示されている．同様に，ABC3 者間の重複度を計算すると，0.877 であった．これらより，共通して話しかけやすいと判定した区間が存在することが確認された．

話しかけにくいと感じた区間についても同様の比較を行い，重複度は A B，B C，C A の順に，0.986，0.989，0.995 であった．3 者間では 0.986 であった．これより，ユーザによらず，共通して話しかけやすいと感じる区間と，話しかけにくいと感じる区間が存在することを確認した．

表 1: 話しかけやすいとした区間長とその重なり

x	y	$t(x)$	$t(y)$	$t(x \cap y)$	重複度
A	B	16.2	17.8	14.4	0.889
B	C	17.8	27.5	16.5	0.927
C	A	27.5	16.2	15.6	0.963

被験者 C は他の被験者と比べて，話しかけやすいと感じた区間が長かった．しかし被験者 C が話しかけやすいとした区間で，他の被験者が話しかけにくいとした区間はほとんど無いことから，この違いは話しかけやすいと感じる程度の個人差によるものと考えられる．3 者で判定が一致した区間は，150 秒中に 135 秒存在した．この判定が一致した区間のデータを，次章での学習及び評価用データとして用いる．

4. ロジスティック回帰による話しかけられやすさの予測

ロジスティック回帰による機械学習を行い，話しかけられやすいか話しかけられにくいかをロボット自身が予測するモデルを構築する．

4.1 入力特徴

話しかけやすさの予測に用いるロジスティック回帰の入力特徴について順に述べる．本研究で用いる入力特徴の一覧を表 2 に示す．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られやすさの判定に用いる．

(1) は直前のロボット発話の終了からの経過時間 (秒) である．これを仮想的に，発話間間隔とする．ここでユーザがロボット発話の終了を知覚するには一定時間 t_0 がかかると考え，ロボット発話の終了から t_0 秒経過した後に，この特徴が有効になるとした．つまり，この特徴 x_1 は，直前のロボット発話の終了時刻を t_i，現在時刻を t で表すと，$x_1 = t - (t_i + t_0)$ とする．予備実験の結果，$t_0 = 1.1$ とした．$x_1 < 0$ の場合は $x_1 = 0$ とした．

(2) , (3) は発話に関する特徴である．これらは，直前のロボットの発話末で，発話交替を促す表現や韻律を用いているかどうかを特徴とする．具体的には，発話交替を促す表現や韻律が含まれた場合，ロボット発話終了時から，次の発話開始までの区間を 1 とし，それ以外の区間を 0 とした．

(4) 〜 (8) は動作に関する特徴である．動作は，ロボットの関節角度の一定時間内における変化量 [度] を特徴とする．まず，ロボットの関節角度をロボットの API を通じて取得する．取得できる関節角度は 26 箇所あるが，ロボットの動作を大まかに表現するため，同じ部位の関節角度の差の絶対値を足しあわせて，特徴とする．

(9) , (10) はロボットの視線に関する特徴である．これらは，ロボットの首関節の角度を API から取得して利用する．ユーザ方向を向いているかどうかを表現するため，ユーザのいる方向とロボットの目線方向との差の絶対値 [度] を特徴として用いる．本研究では，ユーザはロボットの正面に位置するとしているため，ロボットが正面を向いた状態を基準とする．

4.2 評価実験

ロボットの挙動に対して，話しかけられやすいか，話しかけられ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，ロボットがユーザに「話しかけられやすい場合」に 1,「話しかけられにくい場合」に 0 を割り当てる．正

表 2: ロボットの挙動を表す入力特徴

	特徴	取得方法
(1)	発話間間隔	ロボット発話終了からの経過時間
(2)	発話の文末表現	発話交替表現を用いたか
(3)	発話の文末の韻律	韻律が上昇する表現を用いたか
(4)	動作（頭）	0.1 秒前の角度との差
(5)	動作（左腕）	0.1 秒前の角度との差
(6)	動作（左脚）	0.1 秒前の角度との差
(7)	動作（右脚）	0.1 秒前の角度との差
(8)	動作（右腕）	0.1 秒前の角度との差
(9)	視線（水平方向）	正面を基準とした位置
(10)	視線（垂直方向）	正面を基準とした位置

表 3: 提案手法による正解率

	正解率 (%)
ベースライン	72.5
全ての特徴	86.2
特徴選択後	87.4

解率は「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割合である．

正解率は，対象データに対する 10 分割交差検定により求める．対象データは，前章で 3 名の被験者による話しかけやすさが一致した 135 秒分のデータである．これを 0.1 秒単位でサンプリングし，各時点でのロボットの挙動から得た特徴値と組み合わせることで，1350 個の対象データとする．1350 個のデータのうち，話しかけやすいとされたのは 142 個，話しかけにくいとされたのは 1208 個であった．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かなり少ない．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，サンプル数の比 8.51 を重みとして与え，学習及び評価を行う．

4.3 全ての特徴を用いた話しかけやすさの予測

まず，ベースラインを設定した．ここでは，ロボットが発話しているときを 1，発話していないときを 0 とし，これのみを特徴としてロジスティック回帰を行った．これをベースラインとした理由は，相手が発話しているか否かが，話しかけやすさの大きな要因であると考えたためである．この結果，ベースライン手法の正解率は 72.5%であった．

次に，表 2 に示した特徴 10 個全てを用いて，ロジスティック回帰を行った．この場合の正解率は 86.2%であった．これらの結果を表 3 にまとめる．発話しているかどうかのみを特徴としたベースライン手法に比べて 13.7%高い．ロボットの動作や視線，発話表現から得た特徴が，話しかけやすさの予測に有効であることが示されている．

4.4 特徴選択による有効な特徴の分析

さらに特徴選択を行い，判別に有効な特徴について調査した．特徴選択は backward stepwise feature selection[1] により行う．これは，特徴を 1 つだけ除いてロジスティック回帰を行った時に，正解率が下がらなければその特徴を除くという操作を，どの特徴を除いても正解率が下がるようになるまで繰り返す方法である．この結果，正解率は 87.4%であった．これはすべての特徴を用いた場合と比較して 1.2%向上している．

特徴選択により選択された特徴と除外された特徴を表 4 に示す．表に示されているように，(1) 発話間間隔，(2) 文末表現，(3) 文末の韻律，(4) 動作（頭），(9) 視線（水平方向）の 5 つが，判別に有効な特徴であることがわかった．さらに，この 5 つの特徴のうち，最も有効だった特徴を調べた．具体的には，この 5 つから特徴をさらに 1 つ取り除いて学習及び判別を行い，最も正解率が低下する場合を調べた．この結果，(1) 発話間間隔が最も有効な特徴であることを確認した．一方で，これら 5 つ以外の，体の動作（(5) から (8)）や，視線の垂直方向（(10)）に関する特徴は，話しかけやすさの予測には寄与していないという結果になった．直観的には，例えばロボットの手脚が動いている場合や，ロボットが大きく上や下を向いている場合は，ユーザは話しかけにくいと感じると予想されるため，この結果は直観に反する．この原因として，今回作成し，被験者が話しかけやすさを付与したロボットの動作の中に，これらの特徴が単独で現れる箇所が極めて少なかった点が考えられる．この点を含め，別の挙動を作成し，提案したモデルの一般性を今後検証する．

表 4: 特徴選択により選択された特徴と除外された特徴

選択された特徴	(1),(2),(3),(4),(9)
除外された特徴	(5),(6),(7),(8),(10)

5. おわりに

人間は任意のタイミングで対話相手に話しかけるのではなく，相手の状況を考慮して話しかける．本研究では，人間がヒューマノイドロボットに話しかける際にも同様の傾向があると考え，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．

今後の展開として，まず構築したモデルの一般性の検証が必要である．今回の評価実験は交差検定により行ったため，オープンなデータセットに対する性能の検証が必要である．この検証のためにも，新たな挙動の作成とそれに対するデータ収集を現在進めている．次に，学習データにほぼ現れなかった挙動に対しては，直観に反する予測結果となる可能性が示唆された．あらゆる挙動を作成するのは不可能ではあるが，学習データ収集時に用いる挙動の満たすべき条件（作成指針）についても今後検討する．さらに，本来話しかけやすいと感じる度合はユーザごとに異なる．このユーザごとの違いを，本モデルで表現可能かどうかの検証も考えている．

参考文献

[1] Isabelle Guyon. An introduction to variable and feature selection. *Journal of Machine Learning Research*, Vol. 3, pp. 1157–1182, 2003.

[2] Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura, Hiroshi Saruwatari, and Kiyohiro Shikano. Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. *Proc. of INTERSPEECH*, pp. 173–176, 2004.

[3] 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.