1K3-OS-17a-6

ロボットへの話しかけやすさモデルの多数の被験者実験による評価

Empirical Evaluation of Model that Predicts When People will Speak 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 have tackled a novel problem to predict how likely a humanoid robot is to be talked by a user. We previously used common parts to which three participants gave the same labels as the training and evaluation data of our model. In this paper, we collect annotated data from 25 participants recruited from general public and use them for training our new model. We then evaluate the model and compare it with the previous one. We show that our new model performs almost equivalently to the previous one, which shows that our proposed method does not depend on a specific set of participants.

1. はじめに

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している [3]．モデルの全体像を表 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で，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．

本研究の手法により，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ため，より高精度な誤動作回避ができると考える．つまり，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．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 [1, 5]．例えば，Lee らは，Gaussian Mixture Model によって，ユーザ発話と雑音の音響的特徴に基づき，これらを判別する手法を提案している [2]．

本論文の貢献は次の 2 点である．

1. 多数の一般ユーザを対象としても，本手法が有効であることを確認した．
2. 話しかけやすさの付与で生じる 3 つの変動が，本モデルで説明できるかどうかを検証した．

まず，以前の我々の研究で対象とした被験者は，本研究室の学生であり，彼らは我々の研究に関して既知だったことが，話しかけやすさの付与に影響している可能性があった．さらに，対象人数は 3 名と少なく，一般的な話しかけやすさを予測できていたかどうかを検証する必要がある．そこで，新たに 25 名の一般ユーザに対して実験を行った．この実験により収集したデータから話しかけやすさモデルを再構築し，評価を行う．これにより，多人数かつ，一般ユーザを対象としても，[3] で提案した手法が有効であることを確かめる．

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

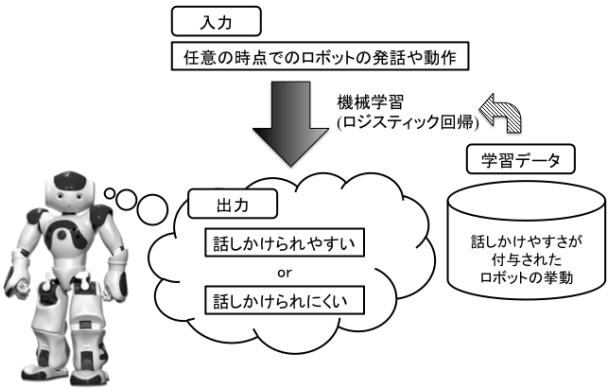

図 1: 本手法の全体像

次に，ユーザが話しかけやすさを付与する際に本質的に生じる 3 つの変動について取り組む．つまりこれらが，我々のモデルで説明可能かどうかを検証する．

(1) 個人差
(2) 同一個人内での試行毎の揺れ
(3) 実験前に与えた教示による影響

話しかけやすいと感じるタイミングはユーザによって異なる．また，同じ状況で付与を行なっても施行毎に揺れが生じ得る．さらに，実験前に被験者に教示した内容によって，付与するタイミングは異なる．そのため，ユーザや試行，教示の内容に対応して話しかけやすさを予測できる必要がある．

2. 多数の被験者からのデータ収集

本研究では，特定のユーザに依存しない，話しかけやすさモデルの構築を目指している．そのため，25 名の被験者から話しかけやすさを付与したデータを複数回収集する．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．

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否かである．ここでは，次の 3 つの場合を仮定して

いる [3]. (i) 話しかけたい内容は，緊急度が非常に高いものではない．緊急度が非常に高い内容の場合，ロボットの状態にかかわらずユーザは話しかける可能性がある．(ii) ユーザはロボットを人間同様に扱う．つまり，ユーザはロボットの状態を考慮して発話することとする．(iii) ユーザが 1 名だと仮定する．ユーザが複数存在する場合，ユーザとロボットの関係によって話しかけやすさは異なる．そのため，簡単のために一人の場合を考える．

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る．被験者には，話しかけやすいと感じる時にクリックさせ，話しかけにくいと感じるまで押し続けさせる．これを挙動が終了するまで行い，話しかけやすいと感じた区間を記録する．被験者がクリックし続けた区間を，ユーザが話しかけやすいと感じた区間とした．

ロボットの挙動は，[3] と同じ 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．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．ユーザの位置はロボットの正面であるとしている．

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，年代が均等になるように考慮した．

実験の手順を図 2 を用いて説明する．実験の準備では，事前に用意した実験の設定や手順に関する文書を被験者に渡した．被験者にそれを読み進めさせながら，以降の実験を行わせた．次に，実験で使用するロボットの挙動を鑑賞させながら，GUI の練習をさせた．これは，一般人がロボットの挙動を初めて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，被験者の話しかけやすさの付与し忘れを防ぐ．

各実験では，被験者にある状況を想定させて，ロボットの挙動に対して話しかけやすさの付与を行わせた．具体的には，図 3 の内容を想定させた．これらの緊急度合の順序も事前に伝えた．これは，被験者全員に同じ状況を想定させるためである．実験 1，2 では，被験者に教示 a を想定させ，挙動のみを変えて実験を 3 回続けて行う．これは，後に話しかけやすさモデルの学習データとテストデータを作成するためである．実験 3 は，実験 2 から教示のみを変えて各 1 回ずつ実験を行う．これは，教示による話しかけやすさへの影響を調べるためである．また，同じ実験を連続して行ったことによる，被験者の疲労を防ぐために，実験毎に 5 分間休憩をとった．

全ての実験の終了後，被験者にアンケートを記入させた．その一部を図 4 に示す．ここでは，教示 b，c で挙動 Y に対して話しかけやすさを付与した際に，教示 a と比べた場合の話しかけやすいと感じた区間の増減を 7 段階で記入させた．

3. 多数のユーザによる付与結果を用いた話しかけやすさモデル

データ収集により，被験者 25 名分のデータを得た．具体的には，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と，教示により話しかけやすさがどの程度変化したかを被験者が記入した結果である．本章では，これらのデータを用いて，新たに話しかけやすさのモデルを構築・評価する．次節では，

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

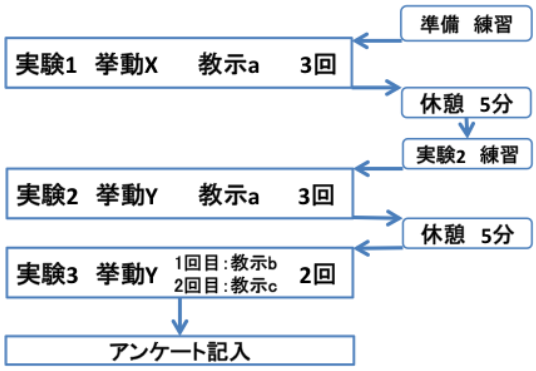

図 2: 実験の手順

【被験者に想定させた状況】
教示a: もう少し大きい声でしゃべってください
教示b: たった今お腹が痛くなったので，席を外したい
教示c: 少し肌寒いので，部屋の温度を上げてください

《緊急度合》
［低い］ 教示c < 教示a < 教示b ［高い］

図 3: 実験で用いた教示内容

学習データの作成方法について説明する．3.2 節では，多数の一般ユーザを対象にしても，本手法を適用できることを確認する．3.3 節では，ユーザの試行毎にロジスティック回帰の閾値を設定することで，個人や試行回数に依らず，話しかけやすさを予測できることを示す．

3.1 学習データの作成

25 名の被験者から得られた話しかけやすさの付与データから，学習データを作成する．学習データには，実験 2 の 2 回目の収集データを用いる．実験 2 を選択したのは，2 つの挙動を比較すると，挙動 Y の方が，挙動の時間が長く，動作や発話のバリエーションが多いため，より様々なデータが得られると考えたためである．3 回の試行のうち，2 回目を選択したのは，[3] と同様に，1 回目は被験者がロボットの挙動に慣れていない，3 回目は被験者が同じ挙動を続けて見たことによる疲労が生じる可能性があるためである．

学習データとして，大多数の被験者の話しかけやすさが一致した区間を用いる．以前の論文 [3] では，3 名の被験者全員が一致した区間のみを学習データとして利用していた．本研究では，25 名の被験者を対象としており，ユーザ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非常に少ない．そこで，話しかけやすさが一致する人数と話しかけやすさの関係から，ある一定数以上の被験者の話しかけやすさが一致した区間を，学習データとして利用する．話しかけやすさの一致した被験者の人数と，その時に一致した区間長の関係を，図 5 に示す．横軸は，話しかけやすさが一致した人数である．縦軸は，0.1 秒を 1 フレームとして，話しかけやすい，話しかけにくいが一致したとみなす区間のフレーム数である．図を見ると，一致している人数が 25 名付近は，被験者毎の話しかけやすいとした区間で一致している区間が非常に少ない．本研究では，3 名の被験者の話しかけやすさが偶然一致する割合は $(1/2^3)$ であることから，25 名の 7/8 に相当する 21 名以

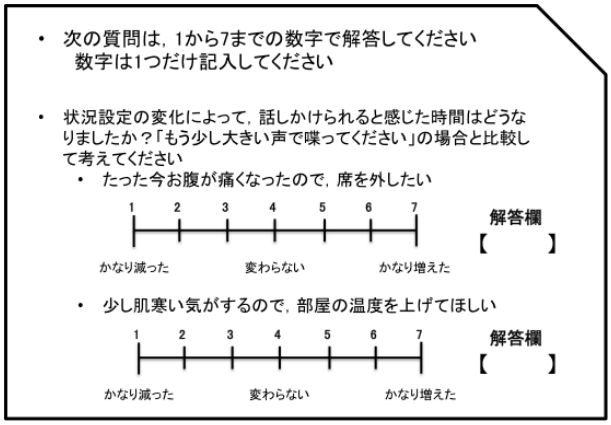

図 4: 実験後に実施したアンケートの一部

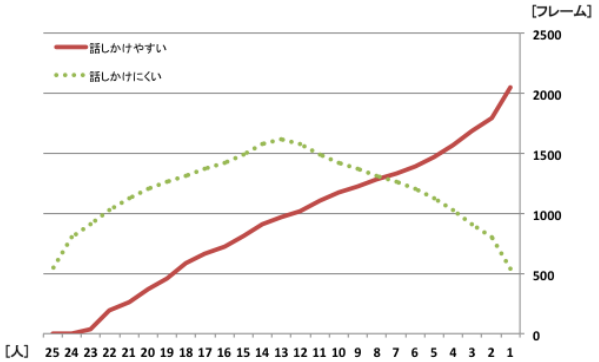

図 5: 話しかけやすさが一致した人数とフレーム数の関係

上の話しかけやすさが一致した区間を学習データとして利用する．このとき，話しかけやすいとした区間は 259 フレーム，話しかけにくいとした区間は 1123 フレームだった．

3.2 多数の被験者による話しかけやすさモデルの評価

多数の被験者かつ，一般のユーザを対象として，論文 [3] の手法を適用できるか調査する．具体的には，新たに構築したモデルと論文 [3] で構築したモデルの性能を比較し，同等の性能が得られることを確かめる．話しかけやすさの識別モデルには，ロジスティック回帰を用いる．

ロジスティック回帰の入力特徴は，表 1 に示す 9 つの特徴を用いる [3]．主に，ロボットの発話に関する特徴，動作に関する特徴，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られやすさの判定に用いる．

モデルの性能を評価するための指標として，「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計算できる MacroF1 [4] を利用する．MacroF1 は，「話しかけやすい」の F 値と「話しかけにくい」の F 値の平均値である．対象データは，前節で説明した，21 名の被験者による話しかけやすさが一致した 1382 フレームのデータである．これを，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，1382 個の対象データとする．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 4.34 を重みとして与え，学習及び評価を行う．

MacroF1 を次の 2 つの場合で計算し，性能を検証する．

(1) 対象データに対する 10 分割交差検定

表 1: 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（脚）	0.1 秒前の角度との差の両脚の和
(7)	動作（右腕）	0.1 秒前の角度との差
(8)	視線（水平方向）	正面を基準とした位置
(9)	視線（垂直方向）	正面を基準とした位置

表 2: 学習データを変えた場合のモデルの性能（MacroF1）

	10 分割交差検定	オープンテスト
新たに構築したモデル	91.9	69.8
論文 [3] のモデル	92.0	69.6

(2) オープンなデータに対するテスト

オープンなデータには，今回収集した挙動 X に対して被験者 25 名が話しかけやすさを付与した 2 回目のデータ 1500 フレームを利用した．

表 2 にモデルの性能比較を示す．比較対象として，論文 [3] のモデルで同様の評価を行った結果を用いる．10 分割交差検定では，今回のモデルと論文 [3] のモデルとの差は 0.1 ポイントであった．オープンなデータに対するテストでは，今回のモデルの方が 0.2 ポイント高かった．オープンテストの値が 10 分割交差検定の値に比べて低い．これは，10 分割交差検定では話しかけやすさが共通した区間をテストデータに利用しているのに対し，オープンテストでは被験者が付与した全てのデータをテストデータに利用しているためである．この結果より，新たに構築したモデルは，論文 [3] のモデルに比べて，ほぼ同等の性能であることがわかった．つまり，多数の一般ユーザを対象とした場合でも，論文 [3] の手法を適用できることを示した．

3.3 個人や試行に対応した話しかけやすさの予測

今回構築したモデルにより，個人や試行に対応して，話しかけやすさを予測できることを確かめる．ここで試行とは，同一個人内での施行毎の揺れを指す．まず，今回構築したモデルで 25 名から収集したデータをどの程度予測できるかを調べる．対象データとして，実験 1 で挙動 X に対して，被験者 25 名に話しかけやすさを 3 回付与させたデータを全て用いる．さらにこの時，ロジスティック回帰の閾値を変化させ，最も性能が良い時の閾値も調べる．これを調べる理由を以下で説明する．ある被験者の話しかけやすさを予測するのに，最も良い性能だった閾値が 0.3 で，その被験者が話しかけやすいとした区間のフレーム数が 300 個だった．また別の被験者は，最も良い性能だった閾値が 0.8 で，話しかけやすいとした区間のフレーム数が 83 個だった．このように，話しかけやすいとする区間長には個人差があるものの，それが同一のモデルのパラメータ（閾値）の変化で表現できると考えた．具体的には，話しかけやすいと感じやすい人は閾値が低く，話しかけにくいと感じる人は閾値が高くなっていた．これらから，このフレーム数と最も良い性能が得られる閾値を調査する．

まず表 3 に，閾値を変化させた場合，固定した場合の性能の比較を示す．閾値を固定した場合とは，ロジスティック回帰の閾値を 0.5 に固定した場合を指す．ベースラインとして，全てを 1 とした時の F 値と，全てを 0 とした時の F 値の平均を用いる．この結果，この値は 47.2 であった．表に示した値は，

表 3: 閾値の変化の有無による性能比較

	MacroF1 の平均		
試行	1 回目	2 回目	3 回目
閾値 変化あり	76.1 ± 6.1	74.3 ± 8.2	74.7 ± 7.7
閾値 固定	70.0 ± 7.2	69.8 ± 8.4	69.1 ± 7.8
ベースライン	47.2		

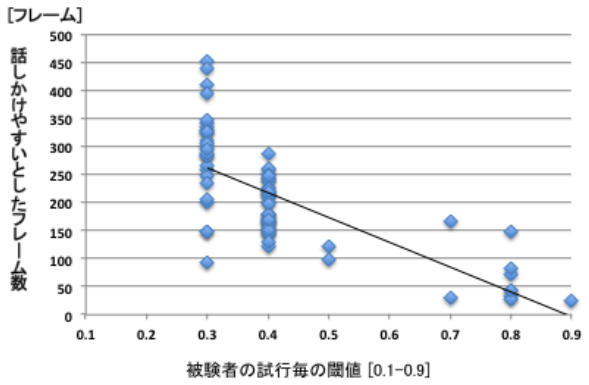

図 6: 最適な閾値とフレーム数

試行毎に被験者 25 名の MacroF1 の値の平均と標準偏差である．閾値変化ありの MacroF1 は，ロジスティック回帰の閾値を 0.1 刻みで 0.1 から 0.9 まで変化させ，最も MacroF1 が高い値を選択した．表 3 より，閾値変化ありの方が固定の場合に比べて全体的に性能が高いことがわかる．これにより，個人毎に閾値を変化させると良い性能が得られることを確認した．

次に，最も性能が良い時の閾値と，被験者が話しかけやすいとした区間のフレーム数の相関を調べる．これらの関係を図 6 に示す．図を見ると，負の相関を示していることがわかる．相関係数は −0.72 であり，回帰直線の方程式は，$y = -443x + 395$ であった．y は話しかけやすいとしたフレーム数，x は被験者の施行毎の閾値である．これにより，ユーザが話しかけやすいとした区間のフレーム数がわかれば，そのユーザがどの程度話しかけやすいと感じやすいかがわかる．

ユーザや試行に対応して話しかけやすさを予測するには，このフレーム数を知る必要がある．この一端として，話しかけやすさに関する被験者の主観的評価とフレーム数の関係を利用する手法を次節で説明する．

4. 教示内容と話しかけやすさの関係

教示内容と被験者の主観的評価の関係を調査する．具体的には，被験者に記入させた図 4 のアンケート結果と，被験者が話しかけやすいとした区間のフレーム数の関係を調べる．これらに相関があれば，アンケート結果から，被験者が話しかける内容により，話しかけやすさがどの程度変化したかがわかる．

実験 2 で使用した教示 a の場合と比較して，教示 b，c で話しかけやすいとした区間のフレーム数がそれぞれどの程度増減したかを調べた．これら 2 つの関係を図 7 に示す．図を見ると，正の相関を表している．相関係数は 0.828 であり，回帰直線の方程式は $y = 204x - 834$ であった．y が教示 a からの増減，x が被験者のアンケート結果である．つまり，被験者のアンケート結果と教示によるフレーム数の増減には相関があることがわかった．これにより，被験者が緊急度によってどのくらい話しかけやすさが変わるかがわかれば，その被験者がどの程度話しかけやすいと感じる区間が増えるかがわかる．つまりこ

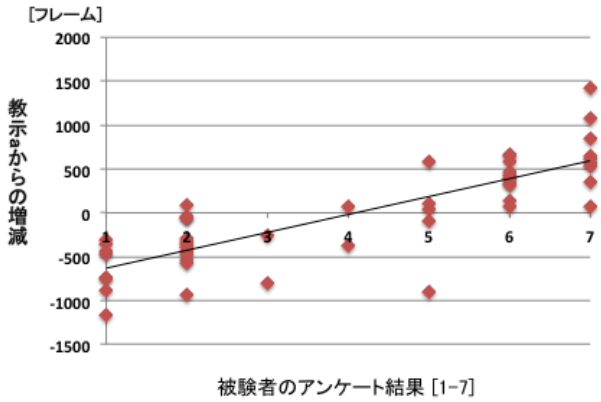

図 7: 教示の変化によるフレーム数の増減とアンケート結果

れらの関係を見ることで，話しかける内容が変化しても話しかけやすさを予測できる可能性が示されている．

5. おわりに

本研究では，25 名の被験者に話しかけやすさを付与したデータから，新しい話しかけやすさモデルを構築した．これにより，論文 [3] で提案したモデルの構築手法が個人や試行に依存しないことを示した．さらに，話しかけやすさを付与する際に生じ得る 3 つの変動について取り組んだ．

(1) 個人差
(2) 同一個人内での試行毎の揺れ
(3) 実験前に与えた教示による影響

(1)，(2) に対して，ロジスティック回帰の閾値を，ユーザが話しかけやすいと付与したフレーム数との関係から設定することで，個人や試行に応じて話しかけやすさを予測できる可能性を示した．(3) に対して，話しかける内容の変化によって，ユーザの話しかけやすさの変化に関する主観的評価と，教示の変化によるフレーム数の増減との関係から，ユーザが話しかける内容が変化しても話しかけやすさを予測できることを示した．

本研究では，被験者の予測が一致した区間を対象にした．今後は，被験者間で話しかけやすさが異なる区間の予測の検証を考えている．そこで，話しかけやすさが異なる区間を学習データに利用して話しかけやすさモデルを構築し，これを予測する．これが出来れば，被験者間で話しかけやすさが共通していた人数によって，話しかけやすさモデルを変えることで，さらに多くの状況を予測できると考える．

参考文献

[1] 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.

[2] 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. of Interspeech*, pp. 173–176, 2004.

[3] T. Sugiyama, K. Komatani, and S. Sato. Predicting when people will speak to a humanoid robot. *International Workshop on Spoken Dialog Systems*, 2012.

[4] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, Vol. 1, pp. 69–90, 1999.

[5] 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.