2M5-OS-20b-3

話しかけやすさの予測に基づく雑音に頑健なロボット音声対話

Noise-Robust Speech Interaction based on Online Prediction of How Likely the 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

A human speaker considers her interlocutor's situation when she determines to begin speaking in human-human interaction. We assume this tendency is also applicable to human-robot interaction when a human treats a humanoid robot as a social being and behaves as a cooperative user. As a part of this social norm, we have built the model of predicting when a user is likely to begin speaking to a humanoid robot. In this paper, we construct a spoken dialogue system with this model to verify whether it is valid for actual spoken dialogues. We use Robot Operating System (ROS) that can synchronize modules in time. We implement the module of predicting when a user is likely to begin speaking as a ROS package.

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，人間は相手の状態を考慮して話しかけることや，相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研究では社会的規範と呼ぶ．私は，人間に類似したロボットとユーザとの対話でも，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る．人間が人工物を無意識に擬人化するという傾向は，心理学実験により確かめられている [2]．

これまで我々は社会的規範の一部として，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行ってきた [3]．この研究では，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を構築した．話しかけやすさを予測する枠組みを図 1 に示す．入力は，任意の時点でロボット自身から得られる情報であり，例えば，ロボットの姿勢や動作，発話中か否かなどである．これら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．

本稿では，ロボット用音声対話システムの構築と，話しかけやすさをオンラインで予測するモジュールの実装について報告する．対話中に，ユーザの話しかけやすさをロボットが予測できれば，ロボットは自身の状態から，話しかけられやすいか否かを認識できる．これにより，ロボットは話しかけられにくいタイミングでの入力音を雑音の可能性が高いと判断できる．これまでの研究では，ユーザの話しかけやすさをロボットが高精度に予測できることを目指していた [3, 4]．そこで，この話しかけやすさの予測モデルを実際の音声対話システムに導入し，雑音に頑健な音声対話の実現に有用であるか否かを検証する．

この音声対話システムを構築する際に課題となるのは，話しかけやすさをオンラインで予測する際に生じる処理の遅延である．複数のモジュールを音声対話システムに組込む場合，これらを時間的に同期させる必要がある．これらを個々に管理する場合，処理の遅延を防ぐのは難しい．本システムでは，これら

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

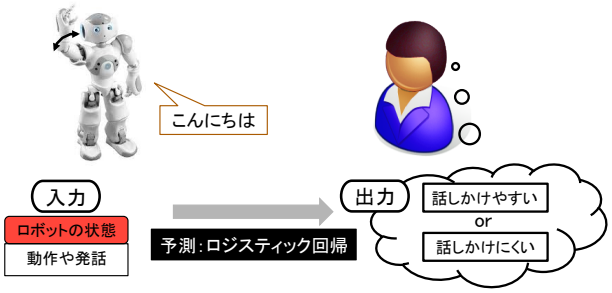

図 1: 話しかけやすさを予測する枠組み

の入出力管理に Robot Operating System（ROS）[*1]を利用する．ROS は，各モジュールの入出力を時系列毎に管理できる．ロボットの応答生成に必要な情報を ROS で管理することで，処理の遅延を未然に防げる．さらに，我々は話しかけやすさの予測モジュールも ROS のパッケージとして実装する．これにより，ロボットの応答生成時に，話しかけやすさの予測結果を考慮できる．

2 章では，話しかけやすさの予測を用いた音声対話システムで実現したデモの例を挙げる．3 章では，本研究で構築したロボット用音声対話システムについて説明する．ここでは，特に ROS のパッケージによる対話システムの実装方法と，話しかけやすさの予測モジュールの実装方法について述べる．4 章では，まとめと，話しかけやすさモデルの利用方法を今後の展望として述べる．

2. 実現したデモの例

話しかけやすさモデルは，例えば，ロボットとユーザとの対話中における，ロボットへの入力音による誤動作回避に利用できる．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．例えば，李らは，入力音の音響的特徴から GMM（Gaussian Mixture Models）を作成し，これに基づきユーザ発話と周辺雑音を判別する手法を提案している [1]．こ

*1　http://www.ros.org/wiki/

図 2: ロボットがスライドの方を見ながら，ユーザに説明している風景

の手法を用いれば，携帯電話の電子音や，咳ばらいなど，明らかにユーザ発話とは異なる音は高精度に判別できる．

一方，入力音の判別に基づく手法では，自身以外に向けたユーザ発話や，テレビから流れる音声などを適切に判別することは困難である．例えば，図 2 の状況で図 3 のような対話を行う場合を想定する．図 2 は，ロボットが正面に座っているユーザに，スライドの方を向きながら，研究室の紹介をしている風景である．図 3 は，ロボットとユーザの対話中に救急車が近くを走行した時の対話の失敗例と成功例である．この例では，救急車は歩行者に対して注意喚起を行いながら，走行している．失敗例では，ロボットは救急車の注意喚起に対して誤って応答してしまっている．救急車の注意喚起は人間の音声であるため，注意喚起に対する入力音判別の結果は「ユーザ」と判別される．本来ならば，成功例のように，ロボットは，入力音判別の結果を棄却し，注意喚起には反応せず，次のユーザの発話を待つことが望ましい．

本モデルを利用すれば，入力音の判別時に，その時点での話しかけられやすさをロボットが考慮できる．ユーザが話しかけにくいと感じるタイミングでは，ロボットはユーザに話しかけられる可能性が低いと考えられる．例えば，図 2 のように，ロボットの視線がスライドの方を向いており，かつ，何か説明している時は，ユーザはロボットに対して話しかけにくいと感じるだろう．このように，協調的な対話において，ユーザがロボットに話しかけられるか否かの事前確率を与えるモデルとして，話しかけやすさモデルが利用できる．これにより，ロボットは話しかけにくいタイミングでの入力音を雑音の可能性が高いと判断できるため，より頑健な音声対話が実現できる．

3. 話しかけやすさの予測機能を備えたロボット用音声対話システム

ここでは，2 章で述べたデモを実現するシステムの実装方法について説明する．システムの全体像を図 4 に示す．なお，この図には実装した ROS のパッケージのうち話しかけやすさの予測モジュールのみを記載しているが，実際には，3.1 節で述べる 6 つのモジュールを実装している．現状では，Julius の入力音判別のモジュール以外の実装は完了している．本システムのタスクは，研究室紹介である．図 3 の対話例のように，ロボットとユーザは一問一答形式で対話を行う．

以降では，まず，ROS を用いた入出力管理の実装方法について説明する．次に，ROS のパッケージとして話しかけやす

	発話	ロボットの視線
R:	何か聞きたいことはありますか？	U方向
U:	研究について教えてください.	
R:	自然言語処理と音声対話システムの研究が. . .	スライド方向
周辺雑音:	ピーポーピーポー	
	左へ曲がります. ご注意ください.	
【失敗例】		
R:	私はNAOです. フランスで開発. . . （救急車の注意喚起に誤って応答）	
【成功例】		
R	（救急車の音には反応せず，待機）	
U:	音声対話について教えて	
R:	音声対話グループの研究には，. . .	U方向

図 3: ロボットとユーザの対話例

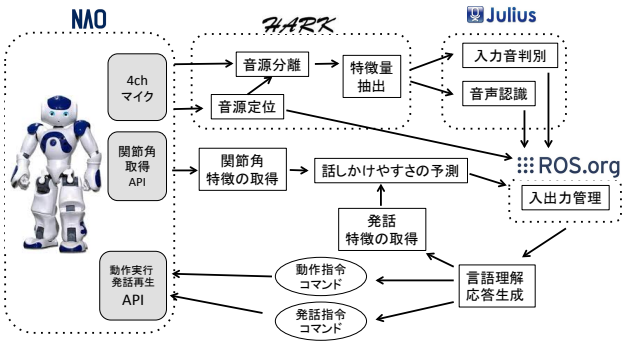

図 4: ロボット用音声対話システムの全体像

さの予測モデルを実装する方法について説明する．最後に，実装に用いたロボットとソフトウェアについて簡単に述べる．

3.1 ROS を利用した入出力管理

本システムは，各時刻でセンサから得た情報を，ROS を利用して管理する．ROS を用いることで，各モジュールの入出力を時系列毎に管理し，これらのモジュールを分散・並列処理ができるため，時間的同期が容易になる．さらに，システムを複数のモジュールに分散しているため，機能を追加したい場合も，その機能をモジュールとして実装すればよい．

本システムに実装した ROS モジュールとその情報の流れを図 5 に示す．これらのモジュールは，それぞれ下記のような役割がある．

1. SUBSCRIBER：他のモジュールから得られる情報を管理
2. FACEDETECT：ロボットの API からユーザの顔認識結果を取得
3. HARK：HARK から音源定位結果，パワー等を取得
4. ACTIVITY：ユーザが対話に積極的に参加しているか否かを出力 [5]

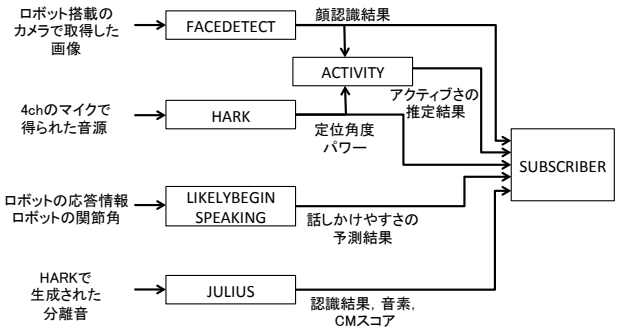

図 5: 実装した ROS モジュールとその情報の流れ

5. LIKELYBEGINSPEAKING：ユーザが話しかけやすいと感じているか否かを出力

6. JULIUS：音声認識結果等を取得

それぞれのモジュールの出力は，全て SUBSCRIBER で管理を行う．SUBSCRIBER のみが言語理解・応答生成部に情報を出力するように実装することで，各モジュールの時間的同期が可能になる．その他のモジュールでは，例えば，HARK モジュールは，HARK の音源定位結果から，定位角度やパワー，同時に発生した音源数などを取得する．また，JULIUS モジュールでは，Julius_mft の出力から，音声認識結果や音素，CM スコアなどを取得し，これを SUBSCRIBER に出力する．

3.2 「話しかけやすさの予測」モジュール

3.2.1 ROS のパッケージによる実装

本研究では，話しかけやすさを予測するモジュールを，ROS のパッケージとして実装する．まず，ロボットの関節角に関する特徴はロボットの API から，発話に関する特徴は言語理解・応答生成部から 0.1 秒毎に取得する．次に，得られた情報からロジスティック回帰式を計算する．その後，話しかけやすさの予測結果を SUBSCRIBER に送る．

「話しかけやすさの予測」モジュールでは，主にロジスティック回帰式の計算及び，閾値による判別を行う．取得した全ての特徴をロジスティック回帰式（式 1）の $x_1, x_2, ..., x_n$ に代入し，これを求める．

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

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

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．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．なお，ロジスティック回帰の特徴を 0.1 秒毎に取得しているため，判別結果も 0.1 秒毎に出力される．この判別結果は SUBSCRIBER を介して，言語理解・応答生成部に出力される．ロジスティック回帰の係数は，論文 [4] で作成した学習データで機械学習し，そこで得られた係数を利用する．

図 4 の入力音判別モジュールが実装できれば，話しかけやすさの予測と入力音の判別結果を統合できる．例えば，話しかけやすいタイミングでは，入力音判別の結果に基づきロボットの挙動を実行し，話しかけにくいタイミングでは，その間の入力音を棄却する，といったことが可能になる．入力音判別モジュールの具体的な実装方法は，まず ROS の JULIUS モジュールで取得する特徴に，GMM による入力音判別結果を追加する．そして，この結果を SUBSCRIBER に出力するという方法である．今回は，次のように実装している．判別結果が「話しかけやすい」の場合，その間の入力音の音声認識結果に基づき，応答を生成する．一方，出力結果が「話しかけにくい」の場合，その間の入力音は棄却される．

3.2.2 入力特徴

ここでは，話しかけやすさの予測に用いる入力特徴と，それらをオンラインで得る方法について述べる．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 [3]．これらは主に，ロボットの発話，動作，視線に関する特徴である．以降では，これらの特徴の概要とその収集方法について説明する．

発話間間隔　発話間間隔は，ロボットの発話終了時から次のロボットの発話開始時点までの無音区間である．そのため，ロボットの発話開始・終了タイミングを得る必要がある．そこで，応答生成部からロボットの API に発話指令コマンドを送るタイミングを発話開始，再生した音声ファイルの終了時を発話終了とし，これらのタイミングを「話しかけやすさの予測」モジュールに送る．ここで，人間は，ロボットの発話終了後から，話しかけやすいと感じるまでにある程度時間がかかると考える．そこで，予備実験によりこの時間長を $t_0 = 1.1[s]$ と設定し，発話間間隔 t から t_0 を引いた値を特徴 x_1 とする．つまり，発話中，または発話終了後から 1.1 秒間は $x_1 = 0$ とし，それ以降の区間では $t - t_0$ を特徴 x_1 とする．

発話の文末表現・韻律　発話の文末表現は，ロボットの発話末が発話交替を表す表現で終わったか否かである．例えば，「〜ですか？」，「〜ですよね？」など疑問形で終わった場合や，「教えてください」のようにユーザの発話を促す表現をした場合が該当する．また，発話末の韻律は，ロボットの発話末の韻律が上昇したか否かである．そこで，これらの特徴を 2 値（0 または，1）で表現する．この特徴の値を応答生成部からロボットの API に転送する際に，その発話の文末表現と韻律に応じて，2 値をこのモジュールに送る．例えば，文末表現が発話交替を表す表現ならば，その後の無音区間は 1，発話交替を表す表現でない場合は 0 とする．また，発話中はどちらも 0 とする．

ロボットの動作　ロボットの動作に関する特徴は，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 を用いて取得する．次に，これらそれぞれに対し，0.1 秒前に取得した角度と比較し，差の絶対値を求める．さらに，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，頭，左腕，右腕，脚の 4 つの部位ごとに足しあわせ，これを特徴とする．なお，ロボットの関節角度は 26 箇所あり，その内訳は，頭部で 2 箇所，右腕・左腕でそれぞれ 6 箇所ずつ，右脚・左脚でそれぞれ 6 箇所ずつである．

ロボットの視線　ロボットの視線は，ロボットの頭の水平方向の角度の向きと垂直方向の角度の向きを特徴とする．そこで，ロボットの動作と同様にこれらの関節角度をロボットの API で取得する．ロボットが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視線方向との差の絶対値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仮定しているため，ロボットが正面を向いた状態を 0 度とした．

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]

3.3 実装に用いたロボットとソフトウェア

ロボットは，アルデバランロボティクス社で開発されたヒューマノイドロボット NAO[*2]を用いる．音声の入力には，ロボットの頭部に搭載されている 4ch のマイクを利用する．NAO には CPU が搭載されているが，音源分離などを実行するための十分な計算能力がない．そのため，個々の入力情報は外部のコンピュータに転送し，そこで音声認識や応答生成，また話しかけやすさの予測を行う．本システムでは，ROS により，音声認識結果と音源定位結果，話しかけやすさの予測結果を言語理解・応答生成部に入力する．言語理解・応答生成部がこれらに基づき，ロボットの API である NAOqi[*3]に発話や動作の指令を送る．そして，応答発話は，NAO 内に配置された音声ファイルをロボットのスピーカーから再生する．この発話は，HOYA 社製の VoiceText[*4]を利用している．同様に，動作も，言語理解・応答生成部から NAO の API に指令を出すことで，実行される．

4. まとめと今後の展望

本研究では，話しかけやすさの予測モデルが頑健な音声対話の実現に有用であることを示すために，話しかけやすさの予測機能を備えたロボット用音声対話システムを構築した．本システムは，入出力管理に ROS を利用することで，言語理解・応答生成に必要な情報の時間的同期を行った．話しかけやすさの予測も，ROS のパッケージとして実装した．「話しかけやすさの予測」モジュールは，外部センサ等から取得した情報からロジスティック回帰式を解き，得られた値を閾値によって判別するモジュールとして実装した．ここでは，動作・視線に関する特徴をロボットの API から，発話に関する特徴を言語理解・応答生成部から取得した．

最後に，今後の展望として，話しかけやすさの予測モデルの応用例を述べる．本稿の成果により，ユーザがロボットに対して話しかけやすいと感じているかを否かを，ロボットがオンラインで認識できるようになった．そこで今後は，これが実際の音声インタラクションで有用であることを示す必要がある．具体的には，例えば，以下の 3 つへの展開が考えられる．

1. 雑音による誤動作回避システム
2. ユーザの属性の推定
3. ユーザの挙動の抑制

まず，今回構築した音声対話システムを利用し，話しかけやすさの予測が雑音に頑健な音声対話の実現に有用であることを，実験的に確かめる必要がある．具体的には，実際に GMM のみでは判別しにくいような区間で適切に誤動作回避できるかどうかを調査する．これは，社会的規範を信号レベルの問題に適用していることに相当する．これにより，社会的規範がロボット用音声対話システムに利用可能であることを示す．

次に，ユーザの属性を推定するシステムの構築が考えられる．ユーザがシステムの発話中に話し始める，バージインという現象からは，急いでいる，システムについてよく知っている，といったユーザの属性が推定できることが知られている．話しかけやすさの予測ができれば，例えば，バージインをさらに 2 つの場合に分類できる．ユーザが，「話しかけにくい」区間から「話しかけやすい」区間に変わってすぐにロボットに話しかけた場合，そのユーザはとても急いでいることがわかる．一方，「話しかけやすい」区間が十分に続いた後に，話しかけてきたユーザは，あまり急いでいないことがわかる．そこで，対話中にユーザ側から得られる情報を特徴として，ユーザの属性を推定する手法 [6] と統合し，上記のような推定ができるかどうかを調べる．これにより，ロボット自身の状態をユーザの特性を推定するための特徴として利用できることを示す．

最後に，ロボットにとって不都合なユーザの挙動を抑制するロボットの挙動を，ロボット自身の判断により生成できる．これは，先に述べた 2 つは社会的規範を受動的に利用していたのに対し，このシステムは能動的視点へと展開している．ここで，能動的とは，以降のインタラクションでのロボットの挙動生成に用いることである．例えば，ロボットが話しかけられたくない状態（例えば，周囲の雑音がうるさい時）でユーザに話しかけさせないように，ロボットが話しかけにくい動作を生成することである．このような，社会的規範の能動的視点での利用方法を確立も検討する．

謝辞

本研究の一部は，JST 戦略的創造研究推進事業さきがけの支援を受けた．

参考文献

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 中島大一, 駒谷和範, 佐藤理史. 複数人会話におけるロボットによる視聴覚情報に基づくアクティブユーザの推定. 情報処理学会研究報告, Vol. 2013-SLP-095, No. 20, 2013.

[6] 駒谷和範, 上野晋一, 河原達也, 奥乃博. 音声対話システムにおける適応的な応答生成を行うためのユーザモデル. 電子情報通信学会論文誌, Vol. 87, No. 10, pp. 1921–1928, 2004.

∗2 http://www.aldebaran-robotics.com/
∗3 https://community.aldebaran-robotics.com/doc/1-14/dev/naoqi/index.html
∗4 http://voicetext.jp/