6T-4

音声インタラクションでの参加者の反応に基づくロボットの誤動作の自動検出

服部 真之 † 駒谷 和範 ‡ 佐藤 理史 ‡

† 名古屋大学 工学部 電気電子・情報工学科 ‡ 名古屋大学 大学院工学研究科

1. はじめに

ロボットとの音声インタラクションでは，音声認識誤りなどにより，ロボットがユーザの意図とは異なる動作をしてしまうという問題が生じる．本研究では，ユーザの意図しないロボットの動作を誤動作と呼ぶ．ロボットが，自身の誤動作を検出できれば，それに対応した動作，発話を生成できる．例えば，図1に，ロボットの誤動作の例を示す．ロボットは音声認識を誤り，ユーザの質問に対して適切でない返答をしている．このとき，この誤動作を自動で検出できれば,「すみません．間違えました．もう一度言って下さい」といった発話を生成できる．

本研究では，ロボットの誤動作が生じた際のユーザの反応を用いて，誤動作を検出する．インタラクションデータを記録する手法として，ユーザの体や実験室内に特別なセンサを装着する研究 [1] も行われているが，本研究ではロボットに搭載されたセンサから得られる情報のみを使用する．これは，特別な環境の中だけでなく，様々な状況における誤動作検出を目指すためである．

具体的にはまず，ロボットとユーザのインタラクションデータを収集し，誤動作時のユーザの反応を典型的なものに分類する．次に，データより，発生数の多いユーザの反応の発生区間を設定する．この発生区間中のシステムのログに現れる，ユーザの反応を示す特徴を用いて，ロボットの誤動作の自動検出を行う．

2. ユーザの反応の分類

まず，ロボットとユーザとのインタラクションデータを収集した．データの収集には，本研究室で開発している，2体のヒューマノイドロボット NAO[§]による研究室紹介システムを用いた [2]．ユーザはシステムに，本研究室に関する質問ができる（例えば,「研究室のメンバーについて教えて」）．ユーザが発話すると，まずロボットは入力音の音源定位と音源分離，分離音の音声認識を行う．次に，定位方向に顔を向け，顔検出によりユーザの存在を確認する．ユーザの存在が確認できる場合は，音声認識結果に対応した返答をし，できない場合は「あれ？」と返答する．顔検出中と返答中以外は，ロボットは正面を向いており，この状態の時のみシステムはユーザの発話を受理する．インタラクション中の音声認識結果や音源定位結果，顔検出結果は，システムのログに記録される．

図1のように，机の上にロボットを配置し，その周りにユーザに座ってもらった．ユーザには，ロボットと対話をしていることを意識し，積極的にインタラクションに参加するように，教示を行った．インタラクションは全てビデオで記録した．

次に，収集したデータ中の，誤動作中のユーザの反応を分類した．分類するデータは，4回分のインタラクションで，合計時間は 12 分 9 秒であった．このなかで，ロボットの誤動作は 47 回発生し，それに対するユーザの反応は 32 回存在した．

[§]http://www.aldebaran-robotics.com/ja/

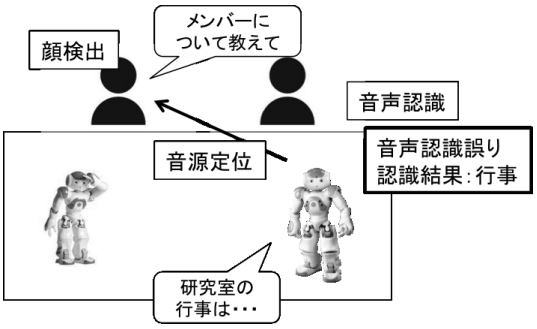

図 1: インタラクション時の配置と返答例

発生タイミング (ロボットの動作)		ユーザの反応	発生回数
定位方向に顔を向けた直後		笑う (a)	6
		発話する	1
返答した直後	顔検出失敗時の返答「あれ?」	笑う (b)	4
		同じユーザが質問を繰り返す	1
	顔検出成功時の返答 例:「研究室の」行事は …」	笑う (c)	7
		発話する	2
		顔を背ける	1
		首をかしげる	1
反応がないとき		同じユーザが質問を繰り返す (d)	9

表 1: 誤動作時のユーザの反応の分類結果

これらのユーザの反応を，ロボットの動作やそのタイミングに応じて，9種類の典型的な反応に分類した．この結果を表1に示す．表はロボットの動作ごとのユーザの反応と，その発生回数である．本研究では，発生回数の多かった反応 (a)〜(d) について詳細な分析を行う．

3. 誤動作の検出方法

前章の分類で得られたユーザの反応が生じている場合は，ロボットに誤動作が発生したと仮定する．ユーザの反応を検出することで，ロボットの誤動作を検出する．

3.1 ユーザの反応の発生区間の分析

ユーザの反応の発生区間を設定する．反応 (a)〜(d) では，ユーザが発話をするため，その特徴は音源定位結果のログに現れると考えられる．そのため，それぞれの反応に対する音源定位結果を調査した．例として，反応 (c) とそれ以外による音源定位結果の開始時刻の分布を図2に示す．ここで，それ以外には，雑音等による音源定位結果が含まれる．グラフの縦軸は音源定位結果の数で，横軸は，ロボットの誤動作発生からの時間であり，1秒ごとにプロットしている．このグラフの横軸の起点は，本来，ロボットの誤動作を，ユーザが認識した時刻とするべきである．これを正確に検出するのは不可能であるため，反応ごとに，システムがログに記録可能な時刻を

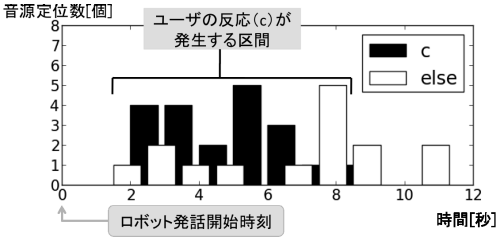

図 2: 反応 (c) による音源定位結果の分布

	N	T 1.0	1.5	2.0	2.5	3.0	3.5
適合率	1	.63	.58	.51	.50	.47	.39
	2	.85	.81	.77	.70	.70	.61
	3	1.00	.89	.85	.81	.80	.78
再現率	1	.81	.92	.95	.97	1.00	1.00
	2	.30	.59	.65	.70	.81	.81
	3	.16	.22	.30	.35	.43	.46
F 値	1	.71	.71	.67	.65	.63	.56
	2	.44	.69	.71	.70	.75	.70
	3	.24	.35	.44	.49	.56	.58

表 2: 反応 (a)(b)(c) による誤動作検出の結果

T_d	1	2	3	4	5	6	7
正解数	0	1	2	2	2	5	5
不正解数	0	1	1	1	1	1	2
検出できなかった数	11	10	9	9	9	6	6

表 3: 反応 (d) による誤動作検出の結果

近似的に設定した．具体的には，反応 (a) では，ロボットが首を発話者の方向へ向ける時刻，反応 (b) と反応 (c) では，ロボットの発話の開始時刻，反応 (d) では，直前のユーザの発話終了時刻とした．

図 2 より，反応 (c) は，誤動作開始から一定時間後の区間に集中して発生していることが分かる．そこで，反応 (a)〜(c) では，ユーザの各反応による音源定位結果の定位開始時刻の中央値を求め，この時刻の $\pm T$ 秒を反応の発生区間とした．反応 (d) では，直前のユーザ発話の定位終了時刻に続く T_d 秒の間を発生区間とした．T と T_d は実験的に設定する．

3.2 反応ごとの検出条件

反応 (a)〜(c) は，ユーザの笑いである．収集したデータでは，1 人のユーザのみでなく，複数のユーザが同時に笑うといった現象が見られた．そのため，これらの反応の発生区間中には，複数方向からの音源定位結果が観測される．よって，区間内に発生する音源定位の方向数が N 以上であることを，反応 (a)〜(c) の検出条件とする．ここでは，複数の音源定位結果の角度の差が 20 度以下のものは同方向と見なす．これは，音源定位の角度分解能が 10 度であり，定位結果には揺れが生じるため妥当と考える．N は実験的に決定する．

反応 (d) は，同じユーザによる質問の繰り返しである．そのため，反応の発生区間において，音源定位結果が同方向から観測される．よって，反応 (d) の検出条件を以下のようにした．

$$|\theta_1 - \theta_2| \leq 20^\circ \quad \text{かつ} \quad t_2 - t_1 \leq T_d$$

ここで，θ_1 は直前，θ_2 は現在の音源定位結果の定位角度である．t_1 は直前の定位終了，t_2 は現在の定位開始時刻である．ここでは、システムがユーザの発話を受理できるときに入力された音源定位結果を対象とする．

4. 評価実験

ユーザの反応 (a)〜(d) を用いて誤動作が検出できることを確認する．システムが出力する誤動作の発生箇所と，人手で判断した誤動作の発生箇所が一致した場合に，正しく誤動作を検出できたとする．

4.1 評価データと評価指標

評価実験に使用するデータは，分析に使用した 4 回分のデータに新しく 3 回分のデータを加えた，合計 7 回分のデータである．合計時間は 21 分 59 秒で，参加者は合計 11 人である．ロボットの誤動作は合計 85 回発生しており，そのうち，ユーザの反応 (a) (d) が生じる誤動作は，反応 (a) が 13 回，反応 (b) が 6 回，反応 (c) が 18 回，反応 (d) が 11 回発生していた．反応 (a)〜(c) による音源定位開結果の定位開始時刻の中央値は，反応 (a) が 4.2 秒，反応 (b) が 3.4 秒，反応 (c) が 4.8 秒であった．評価の指標には適合率 (P)，再現率 (R)，F 値 (F) を用いる．これらは以下のように算出される．F は，P と R の調和平均である．

$$P = \frac{\text{誤動作を正しく検出した数}}{\text{システムの誤動作検出数}}$$

$$R = \frac{\text{誤動作を正しく検出した数}}{\text{誤動作発生数}}$$

4.2 実験結果

反応 (a)〜(c) が生じる誤動作の検出結果を表 2 に示す．この表は，T, N の各値ごとの適合率，再現率，F 値を示している．T を大きくすると再現率が高くなっている．これは，区間が大きくなるとより多くのユーザの反応 (a)〜(c) による音源定位結果が検出されるようになり，誤動作の検出数が増えるからである．N を大きくすると適合率が高くなっている．N が 2 以上のときは，区間内に複数人のユーザからの発話があったことを検出している．複数のユーザの笑いによる音源定位結果と，その他の音源定位結果を区別できるため，誤検出の数が減っている．F 値より，$T = 3.0, N = 2$ のときに最も高い性能が得られた．これは，反応 (a)〜(c) の定位開始時刻の中央値の平均が約 4 秒なので，ユーザはロボットの誤動作が発生してから，おおよそ 3 秒後から 7 秒後までの時間に笑うことを意味する．

反応 (d) が生じる誤動作の検出結果を表 3 に示す．表は T の各値ごとの，ロボットの誤動作を正しく検出した数 (正解数)，間違って検出した数 (不正解数)，検出できなかった数を示している．$T_d = 6$ で高い性能が得られている．これは，ユーザの質問にロボットが反応しないとき，同じユーザがおおよそ 6 秒以内に繰り返し質問をすることを意味する．

参考文献

[1] 角康之: ”マルチモーダルデータを用いた会話的インタラクションの構造理解”, 人工知能学会誌 27(4) pp.405–410, 2012.
[2] 中島大一, 駒谷和範, 佐藤理史: ”複数人会話システムにおける複数の音源定位結果の統合による発話者の特定”, 情報処理学会全国大会 講演論文集, Vol.74, No.2, 4U-3, pp.579-580, 2012.