

多人数対話システムにおける ロボットの挙動に対するユーザ反応の分類

水野 壮†

駒谷 和範†

佐藤 理史†

†名古屋大学 大学院工学研究科 電子情報システム専攻

1. はじめに

人間は対話時に、相手の反応から対話状況を理解し、これに応じて自身の発話や動作を変更することができる。ロボットはユーザの反応を検出できない場合、対話状況に適した発話や動作を生成できない。ロボットが社会的パートナーとして受け入れられるには、ユーザの質問などの意図的な情報だけでなく、ユーザの反応などの意図的でない情報も理解する必要がある。

本研究では、対話状況に適した発話や動作を生成させるため、ユーザの反応に着目する。ユーザの反応を検出できれば、図1のような対話が可能になる。ユーザが頷いていれば、ユーザが興味を示しているという状況をロボットが理解し、それに応じた発話を行うことで対話の活性化が図れる。また、ロボットの挙動に対してユーザが笑ったなら、何かがおかしいという状況をロボットが理解し、それに応じた発話を行うことで、誤動作から復帰できる。

本稿では、インタラクション中のロボットの挙動に対するユーザの反応を調査する。服部ら [1] は、ロボットの誤動作時におけるユーザの反応を調査した。これに対して、誤動作時以外も含めてユーザの反応を調査する。

研究の流れは、インタラクションデータの収集、反応の分類、反応の検出の順である。まず、ロボットとのインタラクション中にユーザのどんな反応が見られるかを調査するため、ロボットとユーザのインタラクションデータを収集する。次に、収集したデータからユーザの反応の分類を行う。本研究では、ユーザの反応の発生回数が多く、ロボットの挙動の変更にも有用な反応に着目し、その検出を目指す。

2. インタラクションデータの収集

データの収集は、以下に紹介するシステムを用いて行った。本研究室で開発された、2体の Aldebaran Robotics 社製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]。多人数でデータ収集を行うことで多くのユーザの反応を観察できると考えたため、多人数対話を行うこのシステムを用いた。音源定位と音源分離には、ロボット聴覚ソフトウェア HARK[§]を用い、音声認識器には Julius[¶]を用いている。各モジュールの統合にはロボット用ミドルウェア ROS^{||}を用いている。インタラクション中の音声認識結果や音源定位結果、顔検出結果を含むロボットの挙動の生成結果は、システムのログに記録される。

このシステムにおいて、ユーザが発話をする、次の3つの手順でロボットが応答する。



図1: 実現できる対話例

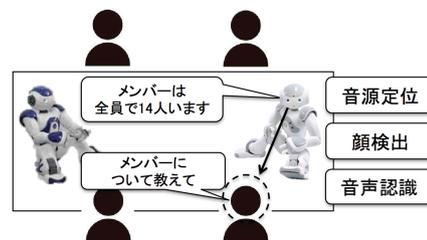


図2: インタラクション時の配置図

1. ユーザの発話に対し、ロボットは入力音の音源定位と音源分離を行う。
2. ロボットは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。確認できない場合は「あれ?」と発話し、顔を正面に戻す。
3. 分離音に対する音声認識結果に応答する。

データ収集は、以下のように行った。ロボットを机の上に配置し、その周りにユーザを座らせた。その例を図2に示す。ユーザには、積極的に対話に参加してもらうことを意図して、ロボットに最低でも1回は発話させるべく自然体を心がけ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。参加者は本研究室の学生、計14名である。

3. ユーザ反応の分類

ユーザの発話とそれに対するロボットの挙動を1組とし、組ごとにユーザの反応を分類する。このユーザの反応は、ロボットの挙動開始後で、かつ次のユーザの発話より前に発生するものを対象とした。この例を図3に示す。組を単位として、観察区間内に発生したユーザの反応の回数を数えた。

ユーザの反応はロボットの挙動(正動作・誤動作)ごとに分類する。正動作・誤動作を、ユーザの発話に対してロボットの挙動が適切であったか否かとする。正動作・誤動作の例を図4に示す。つまり正動作では、ユーザの発話に対して、ロボットが発話者の方向を向き、正しい

Classification of User Reactions for Robot Behaviors in Multi-Party Dialogue System: Takeshi Mizuno, Kazunori Komatani, and Satoshi Sato (Nagoya Univ.)

[‡]<http://www.aldebaran-robotics.com/ja/>

[§]<http://winnie.kuis.kyoto-u.ac.jp/HARK/>

[¶]<http://julius.sourceforge.jp/>

^{||}<http://wiki.ros.org/ja>

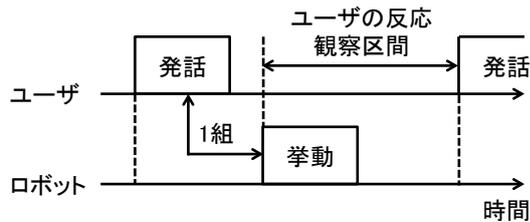


図3: ユーザの反応の分類する単位とその範囲

1. ロボット: 「それでは、研究室紹介を始めます」
- 1組 [2. ユーザA: 「研究室のメンバーについて教えて」
3. ロボット: 「研究室のメンバーは全員で14人います」(正動作)
- 1組 [4. ユーザB: 「音声対話システムについて教えて」
5. ロボット: 「あれ？」(誤動作)
- 1組 [6. ユーザC: 「音声対話システムについて教えて」
7. ロボット: 「研究室のメンバーは全員で14人います」(誤動作)

図4: 対話例

内容を応答した場合である。誤動作は大きく分けて3つある。

1. 音源定位が失敗し、発話したユーザがいる方向とは違う方向を向く。
2. 顔検出が失敗し、ユーザの発話に対して応答できない。
3. 音声認識が失敗し、発話したユーザの意図とは異なる応答をする。

ユーザの反応の分類にあたり、ロボットのセンサから検出できそうなものにまず着目した。ここではセンサとして、ロボット自身のマイクロフォン (4ch) とカメラ (ステレオ) を想定している。

収集したデータから、ロボットの挙動 (正動作・誤動作) に対するユーザの反応を分類する。分類に用いたデータは、7回分のインタラクションデータで、合計時間は21分であった。このうち、ユーザの反応は合計44回発生し、そのときのロボットの挙動は正動作が10回、誤動作が34回であった。ユーザの反応ごとに、その発生回数と、その直前のロボットの挙動 (正動作・誤動作) の発生回数を計数した。この結果を表1に示す。

分類結果に基づき、着目する反応を決定する。ここでは、発生頻度が高く、かつ対話状況の理解に有用な反応に着目する。ロボットの正動作時と誤動作時でユーザの反応が異なるものは、状況の理解に有用、つまり検出結果に基づくロボットの挙動の変更に使え。本研究では、表1の結果より、頷きと笑いに着目する。次章ではこのうち、まずは笑いの検出について検討する。

4. 笑いの検出

ここでは、入力音の音響的特徴に基づく笑いの検出を行う。具体的には、分離音に対して Gaussian Mixture Model (GMM) を適用することで、検出を試みる。

この音響的特徴に基づく手法は、我々が昨年度に提案した手法 [1] と、用いる情報が異なるため、相補的である。このためこの2つの手法の併用により、高精度な笑いの検出を狙う。昨年度の手法 [1] は、音源定位結果の検出時刻を利用している。具体的には、多人数対話にお

表1: ユーザの反応の分類結果

ユーザの反応	発生回数	ロボットの挙動回数 (正動作/誤動作)
頷く	9	9/0
質問し直す	11	1/10
笑う	21	0/21
視線を逸らす	1	0/1
首を傾げる	2	0/2
計	44	10/34

表2: 実験結果

GMM \ 人手	笑いを含む	笑いを含まない	合計
laugh	26	4	30
adult	0	26	26
child	11	5	16
cough	0	0	0
noise	0	1	1
合計	37	36	73

いて、複数のユーザが同時に笑う傾向があることに着目した。これにより、ロボットの挙動開始後、一定時間内に複数方向から音源定位結果が観察された場合、笑いが起こったと判別するものである。

4.1 実験条件

判別に用いる GMM は、生駒市北コミュニティセンターの公共音声情報案内システム「たけまるくん」[3] のデータで学習されたものをそのまま使用した。この GMM では、子供 (child)、成人 (adult)、笑い声 (laugh)、咳 (cough)、その他雑音 (noise) の5つのクラスが定義されている。

テストセットとして、2章で説明した収録データ中で HARK により分離された音声ファイル 73 個を用いる。これらの音声ファイルの長さは 1.5 秒から 4.0 秒である。これらを人手で聴取し、笑いを含むものと笑いを含まないものの2クラスとして、正解ラベルを付与した。その数は、前者が37個、後者が36個であった。笑いを含まない音声ファイルは、全てユーザの発話音声である。

4.2 実験結果

笑いの検出結果を表2に示す。表2は、人手で2つのクラスに分類したテストセットが、GMM によって5つのどのクラスに判別されたかを示す。まず、笑い (laugh) であると判別された30回のうち、26個が笑いを含む音声ファイルであり、笑いの検出の適合率は 87% であった。一方で、笑いを含むにも関わらず、誤って child と判別された場合が11回存在した。これは、他のデータで学習された GMM をそのまま用いているため、音響環境にミスマッチだったと考えられる。

今後の課題として、以下の3点が挙げられる。まず、本研究の実験環境での録音データで学習した GMM を新たに作成する。次に、音響的特徴に基づく笑いの検出手法と、昨年度開発した手法 [1] の統合を試みる。さらに、画像処理により頷きを検出する手法も検討する。

参考文献

- [1] 服部真之, 駒谷和範, 佐藤理史: “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”, 情報処理学会全国大会講演論文集, Vol.75, No.2, 6T-4, pp.517-518, 2013.
- [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.
- [3] 西村竜一: “10年間の長期運用を支えた音声情報案内システム「たけまるくん」の技術”, 人工知能学会誌 28(1) pp.52-59, 2013.