音声言語情報処理 45−18
（２００３．２．８）

ロボットとの音声対話におけるユーザの心的状態の分析

伊藤 亮介　駒谷 和範　河原 達也　奥乃 博

京都大学 情報学研究科 知能情報学専攻
〒 606-8501 京都市 左京区 吉田本町
e-mail: rito@kuis.kyoto-u.ac.jp

あらまし　ロボットとの音声対話をより円滑にするためには，言語的情報だけでなく話者の心的状態 (=感情) を取り扱う必要がある．本研究では，親近感，喜び，困惑の感情を対象として，WOZ 方式によって収集された子供とロボットとのリアルな対話データを用いて，韻律的特徴に基づく分析・判別を行う．特に，対話であるという状況を考慮して，それらの特徴量の発話ごとの変化量や，発話間の時間間隔を利用する．これにより，事前学習を必要としないリアルタイムな判別を可能にする．判別には SVM 及び C5.0 により学習した決定木を用い，困惑で 79%，喜びで 74%，親近感で 87%の判別精度を得た．この感情判別を導入した音声対話機能を実ロボット Robovie に実装し，動作の確認を行った．

Analysis and Detection of Emotional States in Spoken Dialogue with Robot

Ryosuke Ito　Kazunori Komatani　Tatsuya Kawahara　Hiroshi G. Okuno

School of Informatics, Kyoto University, Kyoto 606-8501, Japan
e-mail: rito@kuis.kyoto-u.ac.jp

Abstract　We address analysis of emotional states in order to improve quality of spoken dialogue with a robot. We define a sense of intimacy, joy, perplexity as target states to be detected. Realistic dialogue data between children and a robot are collected using WOZ method. Effective prosodic features are investigated. Especially, we introduce change of those features and time interval between utterances to realize real-time on-line detection without prior training. We used SVM and decision tree trained by C5.0 for classification and obtained accuracy of 79% for perplexity, 74% for joy and 87% for a sense of intimacy. The spoken dialogue system is implemented to a real robot Robovie.

1 緒論

近年，音声認識技術を導入した様々なアプリケーションが開発され，情報案内を対象とした音声対話システムの実用化や愛玩ロボットへの音声認識の導入が行われている．しかし，それらの音声対話の多くは音声に含まれる言語的情報のみを扱うにとどまっている．そのため，どのような相手に対しても画一的な応答を示す．人間どうしの対話においては，視覚や触覚等から得られる音声以外の情報や，音声に含まれる非言語的な情報をともに用いることにより，深いインタラクションが行われる．柔軟な音声対話の実現には，そのような非言語的な情報を統合した対話戦略が必要であると考えられる．

本稿では，従来音声対話においてあまり扱われてこなかった情緒性情報に重点をおいて，その自動判別及びそれに基づいた対話の実現について検討する．

音声における感情の分析や認識に関する従来研究では，言語的な意味から感情を推定する研究 [1] もあるが，多くは韻律的特徴から感情を推定 [2][3][4][5] するものである．これらの研究において対象としている感情は通常，怒り，悲しみ，喜び，平静である．その他，驚き，嫌悪等を対象とした研究もある．また，判別に用いる韻律的特徴もおおよそ同じである．韻律的特徴量については Kiebling ら [6] によって詳細に報告されている．

これらの研究の多くはどの感情にも分類できる文章を，それぞれの感情をこめて役者に発話してもらうことによってデータを収集している．例外として WOZ 方式によって，データを収集した研究も報告されている [7]．また判別の方法に関しても基本的に，各感情音声の特徴量を平静の場合で正規化して比較するため，実際の状況に適用するには事前学習が必要となる．判別率は 5 クラス (怒り・悲しみ・喜び・嫌々・平静) の判別で 64%ほどである [3]．また 2 クラス (怒りの有無) の判別で 86%の精度を得ているものもある [5]．対話において発話間の間隔による意図の違いを検証した研究も行われている [8]．

これに対して本研究では，人間と機械，特にロボットとの音声対話において重要と考えられる心的状態 (=感情) に着目し，事前学習を行うことなく当該対話から得られる特徴のみを利用して，判別する方法を検討する．さらに，実ロボット Robovie においてこの判別を利用した対話機能を実装する．

2 ロボットとの対話におけるユーザの心的状態の分類

本研究では，ロボットと人間 (特に子供を対象) とのインタラクションをよりスムーズに行うために有用な心的状態として．以下の 4 つを扱う．またそれぞれの感情を判別する意義，システムの対応について述べる．

- 怒り
  音声認識の誤りによりユーザが感情を害し，怒りの感情を持った発話はさらに認識を困難にする．このような訂正発話を検出する研究も行われている [9]．このような場合には，ユーザの気持ちを落ち着かせるような対話を展開する．
- 親近感 (緊張感，なれなれしさ)
  機械, ロボットと話すことになれていないユーザも多いと考えられる．そこで緊張しているユーザには緊張をほぐすようにインタラクションを進める．
- 喜び
  興味のある話題については，掘り下げて聞く．
- 困惑
  ユーザが返答に困るような話題の場合，話題の転換をする．

これらの感情は，その性質により以下の 2 つに大別できる．

- 一時的感情
  怒り，喜び，困惑のように，発話単位での変化がみられ，その後数発話のシステムの対応に影響を与える．
- 持続的感情
  親近感等のように，その人個人の性格にも依存し，対話を通して大きく変化しにくい．システムの全体的な対話及びインタラクションの戦略に対して影響を与える．

そこで，一時的感情は発話単位で，持続的感情は話者単位で判別し，評価を行う．

3　対象データとラベル付け

本研究では，神戸市科学館の会話ロボット開発のために，WOZ 方式により収集した一般の来客者との音声対話データを使用する．ユーザは，その多くが 5-15 才ぐらいの子供である．対話は，いくつかの項目についてロボットから質問を行うことにより進められている．対話数は 47，ユーザの総発話数は 498 である．一対話における平均的なユーザの発話数は 10-15 発話で，時間は 2, 3 分程度である．実際の対話例を以下に示す．

・(対話例)
R:「こんにちは」
U:「こんにちは」
R:「僕ロボビー。きみは？」
U:「○○です。」
R:「○○ちゃんかぁ。君いくつ？」
U:「7 才。」
R:「7 才。小学校 2 年生ですか？」
U:「はい。はは、なんでわかるん？」
(**喜んでいるのがわかる**)
R:「小学校は楽しいですか？」
(**もう少し深く聞いている**)
U:「はい。」
R:「今日は科学館で何が楽しかったですか？」
U:「うーん … 、なんやろう … 」
(**困惑しているのがわかる**)
R:「じゃあ質問かえるね。ロボビーは好きですか？」
(**話題の転換をしている**)
U:「好きです。」
R:「やったぁーー!!」
. . . . . .

このデータに対して，人手で親近感 (緊張，普通，馴れ馴れしい)，喜び (有無)，困惑 (有無) のラベルをつけた．WOZ 方式による収集では音声認識誤りがなく，怒りの感情を含む発話はみられなかったため扱っていない．喜び，困惑について発話ごと (計 498 発話) に対してラベル付けを行った．親近感は，対話ごと (計 46 名) に対してラベル付けを行った．同一データに対して複数人 (3 人) によるラベル付けを行った (ただし親近感は一人によるラベル付けのみ)．ラベル付けの結果を表 1(A, B, C, はラベル付けを行った人物)，表 2 に表す．このように人間による感情の認識においても個人差が大きいので，困惑，喜びに関しては，3 者 (A，B，C) のラベルの一致した部分のみで評価・実験を行う．

表 1: 感情のラベル付け結果 (困惑，喜び)(発話単位)

	困惑		喜び	
	あり	なし	あり	なし
A	81	417	193	305
B	67	431	72	426
C	76	422	168	330
3 者で一致したラベル	36	405	47	251

表 2: ラベル付け結果 (親近感)(話者単位)

親近感			
馴れ馴れしい	普通	緊張	合計
17	21	8	46

4　感情判別システムの構成

本研究では、対話中の発話からその時点での心的状態を推定し，ユーザとの間で状況に応じた柔軟なインタラクションを実現することをめざす．また，先行研究にあるような事前学習を必要としない方式を考える．

図 1 にシステムの構成を示す．ユーザの音声から音声認識と同時に心的状態を推定し，その 2 つの結果から次の応答を決定する．また，判別した感情によりシステムの全体的な行動戦略の変更を行う．その際に，音声から得られる特徴量に加えて，判別結果のそれぞれの履歴も考慮する．

心的状態の推定には，先行研究にもあるように韻律的特徴量が大きく関与していると考えられる．そこで，以下の 7 つの特徴量を本研究でも用いた．

- F0 の最大値
- F0 の初期値 (オンセット)
- F0 の平均値
- F0 の最大値と最小値の差
- パワーの最大値
- パワーの平均値
- 発話時間

ただし，事前学習なしで心的状態の推定を行うには，従来のように平静時の韻律的特徴量で正規化する方

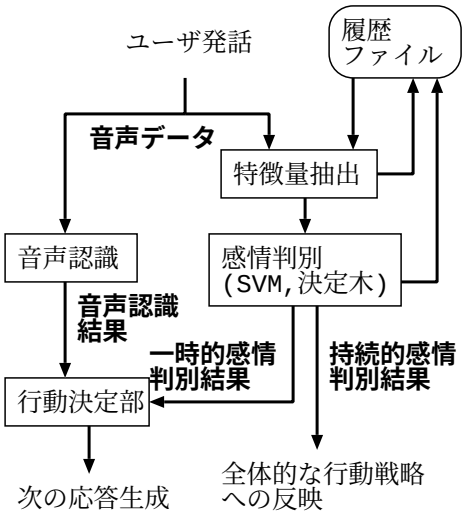

図 1: システムの構成

法を用いることができない．そこで，対話という前後の発話の関係が存在する状況に依存した特徴量を用いる．具体的には，発話間間隔や韻律的特徴量の変化として前発話との差分値や第一発話で正規化した値をともに用いる．

- 前発話との差分値，及び現発話での正規化
  特徴量の絶対値は個人差があるため，比較の対象としにくい．一方，感情が変化する際には，特徴量も変化し，この特徴量の変化は特徴量の絶対値に比べて個人差が少ないと考えられる．さらに，各特徴量の差分値を現在の値で正規化した値も特徴量としてともに用いる．

- 第一発話を基準とする正規化
  対話において，もっとも平静の感情に近いのは対話の開始時であると仮定し，第一発話の特徴量で正規化を行う．

7 つの特徴量の値そのものに，以上で挙げた 3 つの操作 (差分値，差分値の現在の値による正規化，第一発話での正規化) をそれぞれに行って得られる特徴量を加えた 28 個の特徴量と，発話間間隔を用いて判別を行う．発話間間隔は，直前のシステムの発話の終了からユーザの発話が発声されるまでの時間である．

判別には，決定木学習アルゴリズム C5.0 と SVM(Support Vector Machine) を用いる．

5 判別実験と考察

5.1 実験条件

3 章で述べたデータを用いて判別実験を行った．評価は 10 クロスバリデーション (データを 10 分割し，9/10 で学習し，1/10 で評価する過程を 10 回くり返す) で，その分割方法をランダムに 10 回入れ替え，合計 100 回の判別実験による判別率の平均によって比較を行っている．ただし親近感ではデータ量が少ないため 5 クロスバリデーションで行った．感情クラスによりサンプル数に差があるため，コストをつけて学習を行った．コスト比はサンプル数に反比例するように付与した．判別率は，各クラスのサンプル数が等しいと仮定した場合に相当し，各クラスの再現率の平均によって求める．

5.2 一時的感情 (困惑・喜び)

まず，一時的感情である困惑・喜びについて評価実験を行った．先行研究では、平静の発話から抽出した特徴量で，各感情における発話の特徴量を正規化している．そこで，話者毎に各感情がない場合の特徴量の平均値を計算し，その平均値で各発話の特徴量を正規化した場合を，ベースラインとした．

提案手法では，現発話の特徴量の値そのものに加えて，前発話との差分値および現発話の値による正規化を行う手法，第一発話で正規化する手法，発話間間隔を導入した手法とを比較する．そして，上記の 3 つの手法を混合して 29 の特徴量全てを用いた場合を評価した．さらに全ての特徴量を用いると過学習が起きてしまうため，テストセットにおける判別率が向上するように，いくつかの特徴量を取り除くことで最適な特徴量を選択した手法に関しても参考のため試みた．

C5.0 により学習した決定木による結果を表 3 に示す．これより，いずれも 70%に近い判別率が得られた．事前学習を必要とするベースラインの手法と比較しても同程度以上の判別が可能である．また学習した決定木を分析すると，困惑では学習したサンプルによらず発話間間隔が木構造の上位にくることが多く，判別に有効であることがわかる。喜びでは，特に判別に影響を与えている特徴はみられなかった．

次に SVM による判別実験結果を表 4 に示す．これより，困惑で 79%，喜びで 74%程度の判別率が得

表 3: 感情の判別結果 (困惑，喜び)(C5.0)

判別率 (%)	困惑	喜び
平静発話で正規化 (ベースライン)	63.1	66.9
差分値および現発話で正規化	59.3	66.3
第一発話で正規化	66.3	68.0
発話間間隔	66.4	68.8
手法の混合	69.0	66.8

表 4: 感情の判別結果 (困惑, 喜び)(SVM)

判別率 (%)	困惑	喜び
平静発話で正規化 (ベースライン)	73.5	71.8
差分値および現発話で正規化	78.3	73.6
第一発話で正規化	75.4	72.9
発話間間隔	76.6	72.5
手法の混合	79.0	71.9

られており，ベースラインよりも高い性能が得られた．1 つずつ特徴量を除いた場合の判別率を調べることにより，判別に影響を与えている特徴量が，困惑の場合では，パワーの最大値，F0 の平均値，発話間間隔で，喜びに関しては，パワーの最大値と平均値であることもわかった．

C5.0 による決定木を用いた場合と比較して，SVM を用いた場合の方が，高い精度が得られた．SVM 及び C5.0 の両方の場合において，事前学習を行うベースライン手法以上の判別が行えたことから，提案する特徴量が実時間での判別において妥当であると考えられる．

5.3 持続的感情 (親近感)

次に，持続的感情である親近感に関する判別実験を行った．持続的感情は，発話ごとに大きく変化するとは考えにくいため，抽出した特徴量の値そのものを用いる．用いる特徴量は，4 章で述べた 7 つの特徴量と発話間間隔である．持続的感情は，システムの全体的な行動戦略に影響を与えるため，対話の初期段階で判別する必要がある．そこでこれらの特徴量を対話の最初の 1 発話，2 発話，3 発話でそれぞれ平均をとり，それを特徴量として判別に用いる．

表 5: 親近感の判別結果 (C5.0)

	3 クラスの判別率 (%)	緊張のみの判別率 (%)
最初の 1 発話の平均	44	66
最初の 2 発話の平均	57	87
最初の 3 発話の平均	56	79

判別は C5.0(決定木) を用いて行う．学習に用いるデータが少ないため．SVM による判別は行わなかった．親近感については，特に緊張を判別することの意義が大きいと考えられるため，3 クラス (緊張，普通，なれなれしい) での判別と，緊張 (有無) のみの判別について評価を行う．

判別結果を表 5 に示す．緊張のみの判別では，最大 87%の判別精度が得られた．判別に大きく影響を与える特徴量は，パワーの最大値であった．しかし 3 クラスの判別では，57%と大きく性能が低下した．これは,「なれなれしい」と「普通」のラベル付けが曖昧であることによると考えられる．親近感に関してはデータが少ないため複数人によるラベル付けを行わなかった．

6 ロボットへの実装

本システムを ATR 知能ロボティクス研究所で開発されているロボット Robovie[10] に実装した．Robovie は，人間同様の身体表現や移動能力，コミュニケーション能力を持たせることを目標とした日常型ロボットである．多数のセンサ，人間同様の上半身，移動機構を持ち，音声やジェスチャーによって人とのコミュニケーションを可能にしている．

音声認識は，本研究室で開発された記述文法ベースの認識エンジン Julian[11] を用いている．感情の判別は，一時的感情については SVM で発話毎に行い，持続的感情については C5.0 で学習した決定木により，ユーザとの対話の最初の 2 発話の平均を用いて行う．Robovie との会話の様子を図 2 に示す．以下に感情判別部分を導入したロボットの実際の対話例を示す．

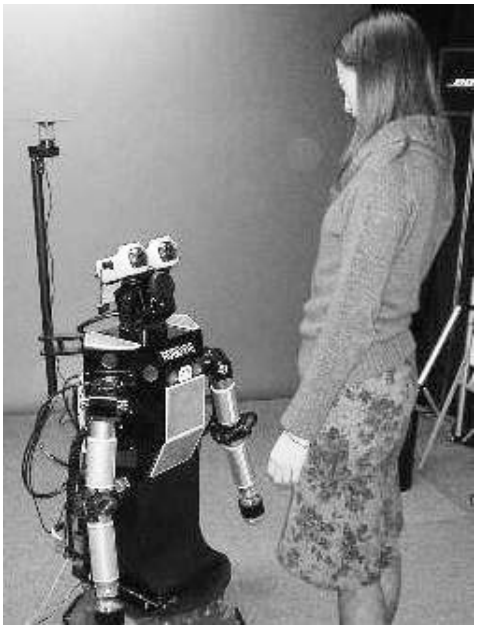

図 2: Robovie との会話

・(対話例)
R:「こんにちは」
U:「こんにちは」
R:「どこからきましたか？」
U:「名古屋からきました」(**感情判別**)
R:「わからない」
判別結果が喜びの場合
R:「そこは遠くですか近くですか？」(**深く追求**)
U:「遠くです」
R:「大変だね」
R:「ロボビーは ATR からきたよ」
判別結果が困惑の場合
R:「ロボビーは ATR からきたよ」
(**話題の追求を避ける**)
· · · · · · · · ·
R:「ロボビーってかわいい？」
U:「かわいいよ」(**感情判別**)
R:「わーい、うれしいうれしい」
喜びの場合
R:「どこがかわいいですか？」(**深く追求**)
U:「全部」
R:「全部。わーい」
困惑の場合
R:「握手しよう」(**話題の転換**)
U:「うん」

7 結論

本研究では，ロボットと人間との音声インタラクションをより円滑にするために，困惑，喜び，親近感の 3 つの感情を扱った．それらの感情を，発話単位で変化がみられる一時的感情と，対話単位 (話者単位) で大きく変化しにくい持続的感情に分類し，韻律的特徴量を用いて判別を行った．従来研究においては，それらの値を平静時の値で正規化するため事前学習が必要であったが，本研究では特に対話という状況を考慮した前発話との差分値の計算や第一発話での正規化を行ったり，特徴量に発話間間隔を加えることで，実時間での判別を可能にした．

またこれらのシステムを実際のロボットに実装し，動作を確認した．実ロボットを用いた評価については種々の要因を考慮しながら検討する必要がある．

謝辞 ロボットへの実装に際して御協力頂いた ATR 知能ロボティクス研究所の石黒浩先生と神田崇行氏に深く感謝いたします．

参考文献

[1] 目良和也, 市村匠, 相澤輝昭, 山下利之. 語の好感度に基づく自然言語発話からの情緒生起手法. 人工知能学会誌, Vol. 17, No. 3, pp. 186–195, 2002.

[2] 北原義典, 東倉洋一. 音声の韻律情報と感情表現. 電子情報通信学会技術研究報告, SP88-158, 1988.

[3] 重永実. 感情の判別分析からみた感情音声の特性(VII) — open な判別について. 電子情報通信学会技術研究報告, SP99-134, 2000.

[4] 森山剛, 斎藤英雄, 小沢慎治. 音声における感情表現語と感情表現パラメータの対応付け. 電子情報通信学会技術研究報告, SP95-67, 1995.

[5] R. Huber, E. Noth, A. Batliner, V. Warnke, and H. Niemann. You BEEP Machine - Emotion in Automatic Speech Understanding System. In *Proceedings of the First workshop on Text, Speech, Dialogue (TSD'98)*, pp. 223–228, 1998.

[6] A. Kiebling, R. Kompe, A. Batliner, H. Niemann, and E. Noth. Classification of Boudaries and Accents in Spontaneous Speech. In *Proceedings of the CRIM/FORWISS Workshop*, pp. 104–113, 1996.

[7] R. Huber, A. Batliner, J. Buckow, E. Noth, V. Warnke, and H. Niemann. Recognition of emotion in a realistic dialogue scenario. In *Proc. ICSLP*, Vol. 1, 2000.

[8] 木村大生, 橋彌和秀. 発話間間隔が発話意図解釈におよぼす影響. 人工知能学会研究会資料, SIG-SLUD-A201-10, 2002.

[9] 山肩洋子, 河原達也. 音声対話システムにおける訂正発話の韻律的特徴の分析. 人工知能学会研究会資料, SIG-SLUD-A101-3, 2001.

[10] ATR. robovie. http://www.irc.atr.co.jp/~m-shiomi/Robovie/index-ja.html.

[11] 音声認識エンジン julius/julian. http://julius.sourceforge.jp/.