2ZB-05

音声対話ロボットにおける音響情報と話者位置情報を用いた笑いの検出

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 ‡ 大阪大学 産業科学研究所

1. はじめに

笑いは，対話中のユーザや場の状況を推定するために有用な情報である．これを検出できれば，状況に応じてロボットの応答を変更することができる．例えば，ロボットとユーザとの音声インタラクションにおいて，笑いはロボットの誤動作時に多く発生する [1]．このため，笑いを検出することで，ロボットが自身の誤動作を検知することを狙う．

本研究では，音響情報と話者位置情報を用いた 2 つの手法を併用し，対話中に発生する笑いを検出する．1 つ目は，音響情報を特徴量とする Gaussian Mixture Model (GMM) を用いた手法である．2 つ目は，多人数対話における話者の位置情報を用いた手法 [2] である．

本稿の流れは以下である．まず，実際にロボットとユーザのインタラクションデータの収集方法と，テストデータの作成について述べる．次に，笑い声を含む複数の発話データを学習データとした GMM の作成について述べる．最後に，音響情報による手法と話者位置情報による手法の併用に基づく，笑いの検出について述べる．

2. テストデータの作成

データの収集には，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 [3] 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 を用いた．インタラクション中の音声認識結果や音源定位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．

このシステムを用いて収集したインタラクションデータを，テストデータとして使用する．テストデータは，1 回当たり約 3 分，計 6 回分のインタラクションである．参加者は本研究室の男子学生，計 9 名である．テストデータ中には，ロボットの動作は計 77 回，それに対するユーザの笑いは計 47 回存在した．

3. GMM の作成と評価

3.1 学習データの収集

まず，GMM の学習に用いる音声データを収集した．本研究では，ユーザの笑い声 (laugh)，ユーザ発話 (user)，ロボット発話 (robot) の 3 クラスの GMM を作成する．この 3 つのクラスに対応する音を再生し，これらを実際のインタラクションと同じ環境となるように，ロボットのマイクを通じて録音した．

各クラスの学習データは以下のとおりである．laugh の学習データには，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」[4] で収録された音声データのうち，成人男性の笑い声のみを用いた．user には，国立国語研究所で作成された日本語話し言葉コーパス‡から，成人男性 40 話者が読み上げた音声を 2 分ずつ使用した．これら 2 種類は，ロボットの前に置いたスピーカから音を再生し録音した．robot には，実際のインタラクションでロボットの声として使用している合成音声を使用した．これをロボットのスピーカから再生し，収録した．収録した音声データの合計時間を表 1 に示す．

表 1: 収録したクラスの音声と削除後の音の合計時間 [秒]

クラス	収録した時間	削除後の時間
laugh	246.28	194.64
user	5785.03	4897.64
robot	975.48	745.59

3.2 判別手順

まず，収録した音声データに対して，パワーの小さい区間を削除する．学習データとして用いた音声データには，無音区間や雑音区間が含まれるが，これらは判別対象の音声データではないため削除する．これには SoX を用い，最大音量の 2 %以下の音量が 0.1 秒以上続く区間を削除した．ただし笑い声の音声ファイルは，笑っている間の無音区間を削除しないように，同様の音量が 0.2 秒以上続く区間とした．これらのパラメータは実験的に決定した．削除後の音声ファイルの合計時間を，表 1 右側に示す．

さらに，GMM により得られる尤度を補正する．これは，学習データとテストデータの収録環境が必ずしも同一ではなく，一部のクラスの尤度が相対的に小さくなっていたためである．具体的には，あるクラスの尤度を定数倍してから尤度を比較し，尤度最大であるクラスを出力することとする．

3.3 性能評価

GMM の作成には，HTK (version 3.4.1) を利用した．特徴量は，MFCC (12 次元)・Δ MFCC・Δパワーの 25 次元を用いた．混合数は 4 とした．

テストデータとして，2 章で説明したインタラクション中に HARK により分離された音声ファイル計 167 個を用いた．これらを人手で聴取し分類した結果，笑いの音声ファイルは計 37 個，ユーザの発話音声ファイルは計 95 個，ロボットの発話音声ファイルは計 35 個であった．音声ファイルの長さは 1.5 秒から 4.0 秒であった．

以下の 3 つの条件で判別性能を比較した．(A) パワーの小さい区間を削除する前と (B) 削除した後の 2 つの場合と，(C) 削除した後に尤度の補正を行った場合である．尤度の補正は，laugh の尤度を 1.0093 倍した．この値は笑いの正解率が上がるように実験的に決定した．これは本システムでは全体の精度とともに，笑いの正解率が重要であるからである．

テストデータに対する GMM の判別結果を表 2 に示す．まず，パワーの小さい区間を削除することで，全体の精度 (3 クラスの正解数の合計 / 全体の数) が 65% (108/167) から 74% (123/167) に向上した．これは，学習データからパワーの小さい区間を削除したことによる結果である．

Detecting Laughter by Acoustic Information and Speaker Localization Information for Dialogue Robot: *Takeshi Mizuno (Nagoya Univ.), Kazunori Komatani (Osaka Univ.), and Satoshi Sato (Nagoya Univ.)*

‡http://www.ninjal.ac.jp/corpus_center/csj/

表 2: GMM による判別結果の混同行列

			Input			
			laugh	user	robot	計
Output	A	laugh	4	1	0	5
		user	32	92	23	147
		robot	1	2	12	15
	B	laugh	11	4	0	15
		user	23	89	12	124
		robot	3	2	23	28
	C	laugh	27	18	0	45
		user	8	76	12	96
		robot	2	1	23	26

A = 元データ, B = パワーの小さい区間削除後, C = 尤度補正後

次に，尤度を補正したことで，全体の精度は 75%とほぼ変わらないものの，笑いの判別数が 16 個増えた．以降では，(C) の GMM を用いる．

4. 音響情報と話者位置情報を併用した検出

4.1 検出する笑いの定義

本研究では，ロボットの動作開始時点から，次の動作開始までに笑いが起きたのかどうかを判定する．状況の具体例を図 1 に示す．これは，ユーザが「メンバーは何人？」と質問し，それに対してロボットが誤った応答をした結果，ユーザが笑ったという場面である．この図において，ロボットとユーザは入力を表しており，GMM から下の 4 つはシステムのログに記録された出力を表す．観察時間は，説明のために記載した．

笑いの検出に用いる分離音とユーザの定位結果について，詳細に述べる．本研究では 2 台のロボットを使ってインタラクションを行っている．そのため，2 台のロボットからそれぞれ定位結果が得られる．定位結果 1 は主にユーザとインタラクションを行うロボットから，定位結果 2 はもう 1 台のロボットから，それぞれ出力される．また，なるべくノイズの少ない音声データを対象とするため，ユーザから距離が近いロボットが出力する定位結果 1 から得られる分離音のみを用いた．

4.2 検出手法

まず，音響情報による検出手法を述べる．インタラクション中に生成される分離音に対して，前章で作成した GMM を用いてクラス判別を行う．ロボットの動作開始時点から次の動作開始までに，複数の分離音が存在することがある．このとき，分離音に対する GMM による判別結果の 1 つ以上が laugh であった場合，笑いが起こっていたと判定する．図 1 の例の場合，ロボットが「さようなら」と発話した後，3 つの分離音が得られており，1 つは user，2 つは laugh と判別されている．このとき laugh が 1 つ以上得られたため，笑いが起こったと判定する．

次に，話者位置情報による検出手法を述べる．これは 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う現象に着目する手法である [2]．そのため，笑いの発生区間中には，システムのログに同時に複数方向の話者の定位結果が観測される．本稿では，ロボットの動作開始時点から 7 秒以内に，定位結果 1 または定位結果 2 のどちらか一方から 3 方向の定位結果が得られたとき，笑いと判定する．ただし，複数の定位結果の角度の差が 20 度以下のものは同一方向とみなす．これは，音源定位の角度分解能が 10 度であり，定位結果に揺れが生じるためである．図 1 の例の場合，観察時間中に定位結果 1 では 3 つ，定位結果 2 でも 3 つの定位結果がほぼ同じタイミングで得られた．2 つのロボットの定位結果からそれぞれ 3 つの定位結果が得られたため，この手法では笑いが起こったと判定する．

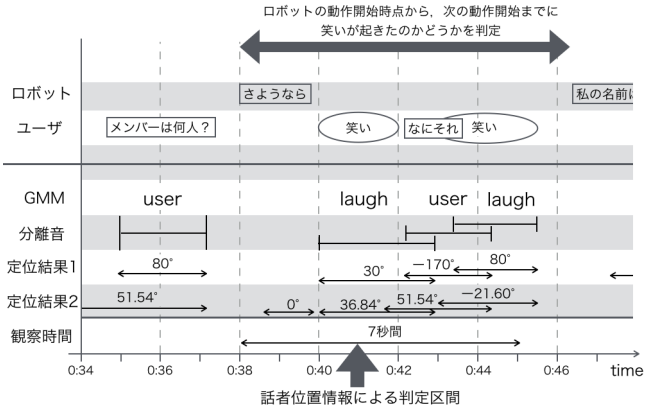

図 1: 笑いを検出する状況の具体例

表 3: 各手法の正解率

	正解率		
	笑いあり	笑いなし	全体
音響情報のみ	0.33 (16/48)	0.86 (25/29)	0.53 (41/77)
話者位置情報のみ	0.33 (16/48)	0.93 (27/29)	0.56 (43/77)
併用 (or)	0.63 (30/48)	0.79 (23/29)	0.69 (53/77)

最後に，2 つの手法を併用し，どちらか片方の結果が笑いであれば，笑いと判定する．話者位置情報を用いた手法では，笑い声以外の音声が混ざっても検出可能だが，複数人が笑わないと検出ができない．一方，音響情報による手法は，音が発生すれば検出可能だが，笑い声以外の発話が混ざると検出できない場合がある．この 2 つの手法は相補的なため，併用することは効果的である．

4.3 実験結果

音響情報，話者位置情報，併用の 3 つの手法での笑いの正解率を比較した．テストセットとして，2 章で作成したインタラクションデータ計 6 回分を使用する．

併用手法を用いた結果，どちらか一方を使った手法より全体の正解率が向上した．結果を表 3 に示す．表 3 は各手法に対して，左から順に笑いが発生したときの正解率，笑いが発生していないときの正解率，両者を合わせたときの正解率を表す．どちらか片方を用いた手法では，全体の正解率がそれぞれ 53%，56%であるが，併用することで全体の正解率が 69%まで向上した．また，どの手法においても，笑いでないときの正解率は高い．

併用手法で検出できなかった笑い 18 回を分析した．このうち，鼻笑いなどの短い笑いが 7 回，ユーザやロボットの発話と重なって発生した笑いが 6 回であった．これらの笑いを検出するためには，音声情報だけでなく，画像情報を合わせて使う必要があると考える．

参考文献

[1] 水野壮, 駒谷和範, 佐藤理史: “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”, 情報処理学会全国大会 講演論文集, Vol.76, No.1, 4S-7, pp.453-454, 2014.
[2] 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.
[3] 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.
[4] 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.