4X-5

音声情報と顔画像情報の統合によるパラ言語レベルの感情認識

松本 祥平 †　　駒谷和範 †　　尾形哲也 †　　奥乃博 †
† 京都大学大学院情報学研究科

1　はじめに

近年、単純な命令通りの動作を行うだけでなく、人間とインタラクションを行いながらプランニングを行うマシンや、人間とのインタラクションを目的とするマシンの社会への普及が進んでいる。それに伴ない、人間とマシンの高度な協調のために、正確かつ簡単に人間の意思がマシンに伝わる高度なマンマシンインターフェースが求められている。

そのようなマンマシンインターフェースには、コミュニケーションにおいて重要な情報である話者の感情の認識は欠かせない。例えば、コンピュータが話者の感情を認識することができれば、人間が明示的に命令を与えなくても、相手が喜んでいたら現在の戦略を続け、相手が怒っているようなら戦略を変えるといったことが可能となる。

しかし、感情は曖昧なものであること、感情表現は多種多様であることから、感情認識というのは非常に困難な課題である。

本研究では意識的に表出される快・不快情報に着目すること、音声情報と顔画像情報という特性の異なるモダリティを用いることで曖昧性を解消し、話者の感情情報を抽出することでより高度なインタラクションを行うことを目的とする。

2　パラ言語レベルの感情認識

感情表現は、そこに含まれる話者の意図の多寡によって大きく二つに分けることができる。一つは、驚いたときに目を見開いて悲鳴を上げてしまうような、話者の内部状態が自然と表出したもの。もう一つは、感謝していることを伝えるために大袈裟に喜ぶといったように話者が相手に情報を伝えるために意図的に表出したものである。

本研究では後者のように表現される感情を「パラ言語レベルの感情」として認識対象とする。パラ言語とは「話者が意図的に操作する非言語情報」のことで、近年、コミュニケーションにおけるその重要性が注目されている。例えば、藤江ら [8] はパラ言語情報として韻律情報やうなづきなどに着目し、より自然な対話を実現している。

さらにパラ言語レベルの感情の中でも快 · 不快情報に着目する。多数ある感情の中から快・不快情報を取り上げる理由を以下に挙げる。

- 表情の研究において、Schlosberg による二次元モデルを始めとする多くの感情的意味次元の研究において快次元が基本的な軸として存在することが繰り返し確認されている [6]
- 音声に含まれる感情の研究においても、感情が含まれる音声の人間による主観評価に対して因子分析を行なった結果、快 · 不快軸が確認されている [7]
- 話者の快 · 不快情報はマシンが今までの動作が適当だったか、これからどうするべきかを決定するための重要な情報である

感情を認識するためのセンサとしてはマイクとカメラを用いる。マイクとカメラはロボットを始めとした人間とのインタラクションを意識して設計されたマシンの多くに搭載されている。また、音声情報と顔画像情報は感情認識においては相補的な情報であることが報告されており [5]、AV(Audio-Visual) 統合による曖昧性の解消が期待できる。さらに、インタラクション中の感情認識において音声情報だけでは相手が発話しているときしか認識できず、顔画像情報だけでは相手が発話しているときに表情表出のための顔の変化なのか発話のための顔の変化なのかの判別が非常に困難という問題があるが、この問題も AV 統合によって解決することが期待できる。

3　音声情報による感情認識

音声に含まれる感情については韻律的特徴を用いて認識を行った研究が数多く報告されている。本研究においても伊藤ら [9] の手法を参考にして韻律的特徴量から快 · 不快を認識する。

特徴量として以下のものを用いる。

- 基本周波数の最大値
- 基本周波数の初期値
- 基本周波数の平均値
- 基本周波数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

さらに以上の特徴量を第一発話、前発話によって正規化したものも特徴量とする。第一発話、前発話により正規化された値は当該話者に対する事前学習を行わずに快 · 不快の認識を行うために加えた特徴量である。基本周波数は自己相関法による相関値に基づいて抽出した。

学習データには神戸市科学博物館の会話ロボット開発のために WOZ (Wizard of Oz) 方式により収集された一般の来客者との音声対話データを使用した。話者数は 46、ユーザの総発話数は 498 である。このデータの対して一発話毎に快、やや快、平静、やや不快、不快のラベル付けを行い、そのラベルを正解とした。快、不快に対してそれぞれ 2 種類のラベルを用意したのは、音声に含まれる感情認識の認識率が低い原因の一つとして、激しい喜びと穏やかな喜びでは韻律的特徴が異なるにも関わらず、同じ「喜び」のクラスに認識しようとしていることが指摘されているからである [4]。

認識器には SVM (Support Vector Machine) を用いた。ラベル付けを行なった 5 クラス全てを認識対象とした場合や、不快とやや不快を 1 つのクラス、快とやや快を 1 つのクラスにしてそこに平静を加えた 3 クラスを認識した場合などの 10

Para-Linguistic Emotion Recogntion based on Audio-Visual Integration *Shohei Matsumoto, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno (Kyoto University)*

フォールドクロスバリデーションにより得た認識率を表 1 に示す。

5 クラスの認識を行うことは難しいが、快なのか不快なのか、平静なのか不快なのかといった情報は抽出できることがわかる。5 クラス全てを認識対象とした場合の認識率が低いことの原因として、今回用いた学習データが WOZ 方式で収集されたものなので音声に含まれる感情が単純な快や不快ではなく複雑なものであることなどが考えられる。

表 1: 音声に含まれる感情認識結果

認識クラス	認識率
不快 vs やや不快 vs 平静 vs やや快 vs 快	47.4%
不快、やや不快 s 平静 vs やや快、快	54.1%
不快、やや不快 vs やや快、快	78.8%
不快、やや不快 vs 平静	75.2%
平静 vs やや快、快	65.7%

4 顔画像情報による感情認識

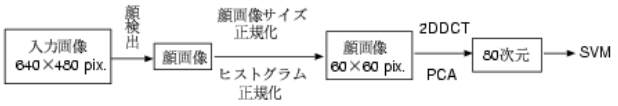

図 1: 表情認識の処理の流れ

表情認識においては、FACS (Facial Action Coding System[3]) に基づいた Action Unit の認識などによる基本 6 感情 (喜び、悲しみ、恐怖、嫌悪、怒り、驚き) の認識を行う研究が盛んに行われている [2]。しかし、本研究では快 · 不快の認識を目的としているので基本 6 感情の表情認識を行なうのは目的に沿わないと考える。

そこで、意図的に表出された快、平静、不快の 3 種類の表情を学習データとして空間周波数を特徴量に用いた表情認識を行う。具体的には、2DDCT(2 次元離散コサイン変換) によって算出された係数を主成分分析で圧縮したものを特徴量として SVM で認識を行う。

表情認識の処理の流れを図 1 に示す。まず、MPIsearch[1](http://kolmogorov.sourceforge.net で入手可能) を用いて顔領域の抽出を行う。顔の大きさの個人差やマンマシン間の距離の変動の影響を軽減するために顔画像の大きさを 60×60 ピクセルに正規化する。次に、照明の影響を軽減するためにヒストグラムの正規化を行う。そして 2DDCT を行い、得られた係数を主成分分析で 80 次元 (累積寄与率 94.8%) に圧縮したものを用いて SVM による認識を行った。

2DDCT の結果には、表情による顔の変化を示す情報だけでなく、顔のつくりを示す情報も含まれている。話者に対してオープンな表情認識を行なうためにはこの 2 つの情報を切り分ける必要がある。そこで、個人差を吸収するためにそれぞれの係数の平静時との差も特徴量に加えた場合、表情の認識には目、口の顔器官の変化が重要であると考え、顔の下半分 (口)、上半分 (目、眉) に対して 2DDCT を行った場合の SVM による認識も行った。

学習データは男子大学生 2 名、男子大学院生 2 名がカメラの前で意図的に平静、快、不快の表情を表出したものを用いた。データ数は 640×480 ピクセルの画像が一人につき約 600 枚 (1 つのクラスにつき約 200 枚) である。

認識対象人物のデータが学習データにも含まれるように 10 フォールドクロスバリデーションを行った場合の認識率は 99.4%であり、2DDCT が表情を認識するための特徴量として有効であることを示せた。

認識対象人物のデータが学習データに含まれない場合の表情の認識率を表 2 に示す。単純に 2DDCT を行なって認識した場合より、平静時との差分を特徴量とした場合、顔器官に着目して 2DDCT を行った場合の方が認識率が約 15%向上しており、今回検討した特徴量が個人差を吸収する効果が有ったといえる。

認識誤りの傾向としては、不快の表情を快と誤認識してしまうことが圧倒的に多い。理由としては、FACS の基本 6 感情において快に含まれる感情は幸福のみなのに対して、不快に含まれる感情は怒り、嫌悪、恐れ、悲しみの 4 種類が存在するので一つのクラスとして認識することが難しいということが考えられる。したがって、不快の表情の多様性を吸収する枠組みが必要である。また、実際のインタラクションでは典型的な平静、快、不快以外の様々な表情が現れるので、その中からどのように快、不快情報を抽出するかを検討する必要がある。

表 2: 不快、平静、快の表情認識結果

	2DDCT	2DDCT(差分)	2DDCT(顔器官)
認識率	52.7%	66.6%	68.4%

5 まとめ

マンマシンインターフェース高度化のためにパラ言語レベルの感情認識システムを提案し、音声情報、顔画像情報からの感情認識モジュールの開発を行なった。

今後は、ヒューマノイドロボット Robovie に今回開発したシステムを実装し、Robovie と人間のインタラクションで得られたデータから顔画像情報と音声情報の統合手法の検討と洗練を行う。そして、感情認識システムによって高度なインタラクションを取れたかどうかを確かめる評価実験を行う予定である。

本研究の一部は科学研究費補助金、21 世紀 COE プログラム、SCAT 研究助成の支援を受けた。

参考文献

[1] Ian Fasel, Bret Fortenberry, and Javier R.Movellan. A generative framework for boosting with applications to real-time eye coding. In *INC MPLab TR*, 2003.

[2] M.S.Bartlett, G.Littlewort, I.Fasel, and J.R.Movellan. Realtime face detection and facial expression recognition: development and applications to human computer interaction. In *SVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003.

[3] P.Ekman and W.V.Friesen. *The Facial Action Coding System*. Consulting Psychologists Press, 1978.

[4] Klaus R.Scherer. Vocal communication of emotion: a review of research paradigms. In *Speech Comunication*, Vol. 40, pp. 227–256, 2003.

[5] Thomas S.Huang, Lawrence S.Chen, and Hai Tao. Bimodal emotion recognition by man and machine. In *ATR Workshop on Virtual Communication Environments*, 1998.

[6] 乾敏郎, 安西祐一郎（編）. コミュニケーションと思考, 第 4 章, pp. 115–138. 岩波書店, 2001.

[7] 森山剛, 斉藤英雄, 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. 電子情報通信学会論文誌, Vol.J82-D-II,No.4, pp. 703–711, 1999.

[8] 藤江真也, 江尻康, 菊地英明, 小林哲則. パラ言語の理解能力を有する対話ロボット. 情報処理学会研究技術報告, SLP-48, pp. 13–20, 2003.

[9] 伊藤亮介, 駒谷和範, 河原達也, 奥乃博. 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. 情報処理学会研究報告, 2003-SLP-45-18, 2003.