表情・音声・言語からのリアルタイム感情認識システム

Real-time emotion recognition system from facial expression, voice, and language

岡田 敦志
Atsushi Okada
広島市立大学大学院 言語音声メディア工学研究室
Language and Speech Research Laboratory, Graduate School of Information Sciences, Hiroshima City University

概要 言語音声メディア工学研究室では，これまで発話時の表情，発話音声の音響的特徴，発話文字列それぞれから感情を推定する手法について研究してきた．しかし，各推定感情を総合的に捉えて話者の心理状態を推定するには，これらの感情推定処理を同時に実行できるシステムが必要となる．そこで本研究では，マイク付き Web カメラで撮影している動画像に対して，表情，音響的特徴，文字列を抽出し，並列に感情推定処理を実行するシステムを構築する．

1　はじめに

近年，我々の生活にロボットが浸透するにつれて，人間とロボットがコミュニケーションをする機会が多くなった．人間とロボットがより円滑なコミュニケーションを行うためにはロボットが人間の感情を推定する必要がある．感情推定を行わなければ，ロボットは人間の発話の字面しか考慮しないため，ロボットは誤った返答をしてしまう．例えば，人間が元気なく「元気だよ」と発話したとき，感情推定を行わなければロボットは「それは良かったね」と不自然な返答をしてしまう．

そこでこれまで人間の感情を推定するためのさまざまな手法が提案されている．その情報源としては主に，表情，発話音声の音響的特徴，発話文字列などが挙げられる．ほとんどの手法は単独の情報源だけから感情推定を行っているが，実際の人間の心理状態としては，皮肉やツンデレのように，各情報源から推定される感情に食い違いがあることによって，複雑な心理状態が表現されていることもある．

そこで本研究では，話者の表情，音響的特徴，文字列からそれぞれの感情をリアルタイムで推定し，出力するシステムを構築することを目指す．

2　リアルタイム感情認識システムの処理の流れ

本論文で提案するリアルタイム感情認識システムの処理の流れを図 1 に示す．まず，話者が web カメラの前でマイクに向かって話すと，発話シーンの動画が web カメラにより取得される．この動画像から約 100msec ごとに静止画像を抽出し，それぞれの画像から話者の表情を推定する．そしてそれらの表情推定結果を統合し，動画全体の感情を推定する．顔画像処理と並行して，音声信号に対する処理も行う．まずマイクから取得した音声の波形から音響的特徴を抽出し，その特徴を用いて感情を推定する．また，音声から音声認識ソフトを用いて発話文字列を取得する．抽出された文字列に形態素解析を行い，解析された字句の意味から感情推定を行う．このようにして表情，音声，言語から推定された感情について，それらの一致あるいは食い違いに着目することで，話者の複雑な心理状態の推定を行う．

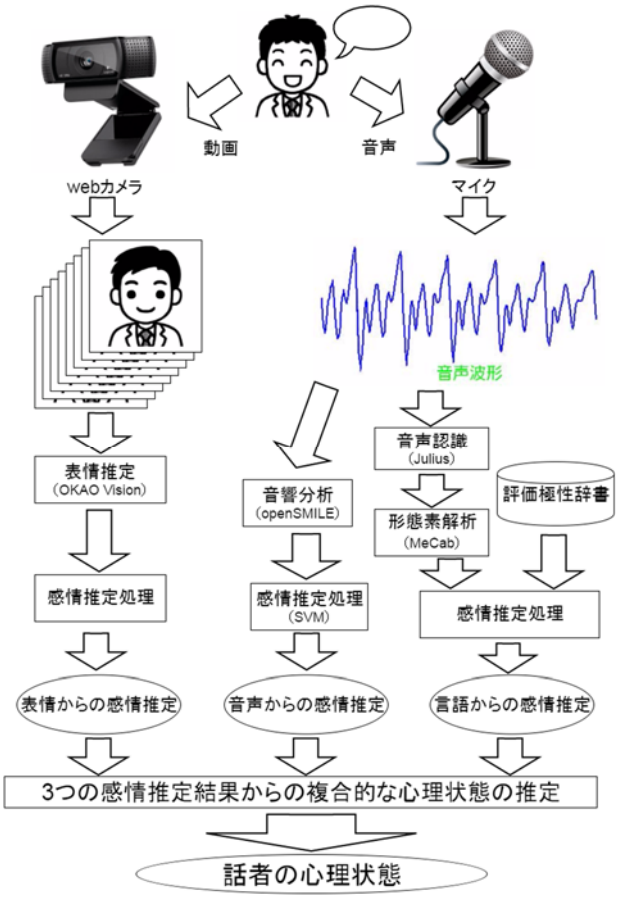

図 1　システムの構成図

3　感情の推定手法

3.1　表情からの感情推定手法

表情からの感情推定は，web カメラからの映像をフレーム単位で切り出し，顔の静止画像を取得する．それらの画像からオムロンの OKAO Vision[1]を用いて，表情を推定（無，喜，驚，怒，悲）する．その推定表情を用いて動画全体の表出感情を Positive，Negative，Neutral に分類する．現段階では，推定されたフレームごとの感情を全て出力する方法と，推定された画像の中でもっとも頻度が高い感情 1 つを出力する方法を想定している．

表情から人間が推定した結果と提案手法を比較すると，感情を全て出力する方法では一致率が 0.38 となっており，頻度が高い感情 1 つを出力する方法では一致率が 0.55 となっている．また，Positive では精度が 0.76 となっているが，Negative，Neutral ではそれぞれ精度が 0.43，0.13 となっている．

3.2 音声からの感情推定手法

音響情報からの感情推定は，マイクを通して取得した音声から openSMILE[2]によって音響的特徴を抽出し，その特徴を用いて識別器である Support Vector Machine（SVM）で音声を Positive，Negative，Neutral の 3 クラスの感情に分類する．openSMILE は，音声の波形からさまざまな特徴量を計算するツールである．算出する特徴量は，音の大きさ，メル周波数ケプストラム（12 段階），ゼロ交叉率，声である確率，基本周波数と，各波の一次導関数の合計 32 種類の波形である．そしてその各波形から最大値や平均など 12 種類の静的特徴量を算出するため，1 つの音声から取得できる特徴量の数は 16×2×12=384 種類となる．

音声から人間が推定した結果と提案手法を比較すると，一致率は 0.56 となっている．また，精度は Positive が 0.70，Negative が 0.62，Neutral が 0.51 となっている．

3.3 言語からの感情推定手法

文字列からの感情推定は，まず，発話音声を音声認識エンジン Julius[3]によって文字列に変換する．次に，分割した単語列を形態素解析システム MeCab[4]によって形態素解析する．最後に，発話文字列に含まれる語に対して日本語評価極性辞書[5]を用いて，発話全体の持つ感情を Positive，Negative，Neutral の 3 クラスの感情に分類する．

提案手法では品詞・係り受け情報，深層格情報を基に文字列の極性推定を行う．文字列に対して，日本語評価極性辞書を用いて文字列に含まれている名詞，用言 (動詞，形容詞，形容動詞) の評価極性を抽出し，係り受けの関係を考慮しつつ文全体の感情極性を推定する．

4　3 つの感情推定結果からの複合的な心理状態の推定

表情，音声，言語それぞれにより推定された感情より，複合的な話者の心理状態を推定する．

提案手法では，3 つの推定結果にそれぞれ重みを付けて統合し，最終的な話者の感情を出力する．重み付けは上村らの研究[6]を参考にする．出力としては，Positive，Negative，Neutral の度合を取得し，その度合より感情を決定し出力する． 本研究では，Positive，Negative，Neutral の度合の出力を目指す．

5　まとめ

本研究では，表情，言語，音声のそれぞれの感情を Positive，Negative，Neutral に分類し，出力するシステムについて述べた．

今後は，3 つの出力結果を基に話者の感情を“喜び”や“悲しみ”のように出力できるようにしていきたい．そのためには，どのような感情を出力として持ってくるかを決める必要がある．また，それぞれの感情の出力も Positive，Negative，Neutral ではなく，より多様な感情を出力できるようにすることで，複合感情などを推定できるのではないかと考えられる．またそれには，表情，音声，言語についてどのような感情を出力するか，それらを合わせたときに，どのような感情を出力するかを検討する必要がある．

参考文献

[1] OKAO Vision｜オムロン人画像センシングサイト，http://plus-sensing.omron.co.jp/technology/，(2016/07/04 アクセス).

[2] openSMILE | audEERING | Intelligent Audio Engineering – openSMILE ， http://audeering.com/research/opensmile/ ，(2016/07/07 アクセス).

[3] Julius，汎用大語彙連続音声認識エンジン Julius – OSDN，http://julius.osdn.jp/，(2016/07/07 アクセス).

[4] MeCab ｜ MeCab: Yet Another Part-of-Speech and Morphological Analyzer，http://taku910.github.io/mecab/，(2016 /07/07 アクセス).

[5] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一. 意見抽出のための評価表現の収集. 自然言語処理，Vol.12, No.3, pp.203-222, 2005.

[6] 上村譲史，目良和也，黒澤義明，竹澤寿幸，字句情報，音響情報，表情から推定した話者の感情の食い違い状況の分析と食い違い自動検出手法の提案，情報処理学会（2016）

[7] 小林峻也，萩原将文，ユーザの嗜好や人間関係を考慮する非タスク指向型対話システム，第 29 回人工知能学会全国大会論文集（2015）