2Z-8

ニューラルネットワークを用いた 音声と顔画像のバイモーダル情報による感情認識

根本 直樹　槫松 明　金子 正秀　長井 隆行

電気通信大学

1　はじめに

我々の日常生活においては感情や感性などの伝達が重要な役割をしている [1]. 人間同士の会話の場合, 相手の話す速さ, 声の高低, 強弱および顔の表情などにより, 相手話者の会話時における感情が知覚でき, それがその場における会話をスムーズに行えるようにするのに一役を担っていると考えられる. したがって, 人間とコンピュータシステムの円滑なコミュニケーションの実現においても, 人間の感情認識は重要な研究テーマの一つと考えられる. マルチモーダルによる感情認識としては文献 [2] があげられる. 本稿では, 感情カテゴリーとして「怒り」「驚き」「悲しみ」「喜び」「嫌悪」の 5 種類と「無感情」を合わせた計 6 種類を扱い, 音声における感情に関する特徴量と顔画像からの感情に関する特徴量のバイモーダル情報による感情認識をニューラルネットワークを用いて行う手法について検討する.

2　音声と顔画像の特徴量

2.1　音声の特徴量

音声の発声で, 音韻的特徴と韻律的特徴は分かち難く結び付いていると考えれば, 音韻的特徴の制御のみによって感情を表現するのは困難ではないかと考えられる [3]. 本研究では、この考えにのっとり, 音韻的特徴を表現する特徴量と韻律的特徴を表現する特徴量の両方を用いた.

2.2　顔画像の特徴量

顔表情の記述においては, Ekman らによって開発された Facial Action Coding System(FACS) が知られているが, 実際の顔画像から動作単位 (Action Unit:AU) を抽出するのが複雑である. そこで肖ら [4] は表情と無表情の顔画像全体に対する 2-D DCT 係数の低周波成分の差分のみをそのままニューラルネットワークに学習させ, 表情空間のマッピングを実現している. 2-D DCT は, 画像圧縮分野において最も広く用いられており, 表情変化の多くが直接周波数成分に反映し, エネルギーの多くが低周波数成分に集中する. 本研究でもこの手法を用いて顔画像の特徴量の抽出を行った.

3　実験

3.1　データベース

本研究で扱うデータベースは音声は 16[kHz] 標本化,16[bit] 量子化, モノラル, 画像は 10[f/s],320x240[pixel],RGB 各 8[bit], 照明は蛍光灯のみである. 各感情ともに発声文は「え、そうですか」の一文のみ, 1 人の被験者 (男性) に対して収録日を変えて計 2 セットを収録した.

3.2　特徴抽出

3.2.1　音声からの特徴抽出

音韻的特徴量としては, フレーム毎の線形予測係数を用い, 韻律的特徴量としては, フレーム毎の 2 乗平均エネルギーとゼロ交叉数, 線形予測分析における予測誤差の自己相関関数から得られるピッチ, そしてフレーム毎のピッチの差分の平均を用いた. フレーム長を 25[msec], フレーム周期を 10[msec] とした. また, 画像との時間同期をとるために本研究での音声のフレームの切り出しは, 画像 1 フレームに対して音声 10 フレームが対応するようにした. 表 1 に音声の 10 フレームあたりの特徴量とその次元数を示す. また, 音声処理の結果の例として「怒り」を図 1に示す.

3.2.2　顔画像からの特徴抽出

原画像から色情報を用いて頭髪を除いた顔領域を抽出し, 各フレーム毎に顔の位置あわせを行い, 8[bit] グレイスケールの画像に変換した上で 128x128[pixel] の顔領域を切り出した. この切り出された画像全体に対して 2-D DCT を施し, 2-D DCT 係数の低周波成分 16x16 領域を用いた. したがって, 画像 1 フレームあたりの特徴ベクトルの次元数は 256 次元となる.

3.3　バイモーダルの特徴量

バイモーダル情報で感情認識を行う場合の特徴量は, 音声 10 フレーム分と顔画像 1 フレーム分の特徴量を合わせたものを 1 つの特徴ベクトルとして扱うため, 次元数は 407 となる.

Emotion Recognition using Neural Network with Bimodal Information

Naoki NEMOTO, Akira KUREMATSU, Masahide KANEKO, Takayuki NAGAI

University of Electro-Communications

表 1: 音声 10 フレームあたりの特徴量

特徴量	次元数
線形予測係数	120
2 乗平均エネルギー	10
ゼロ交叉数	10
ピッチ	10
フレーム毎のピッチの差分の平均	1
合計	151

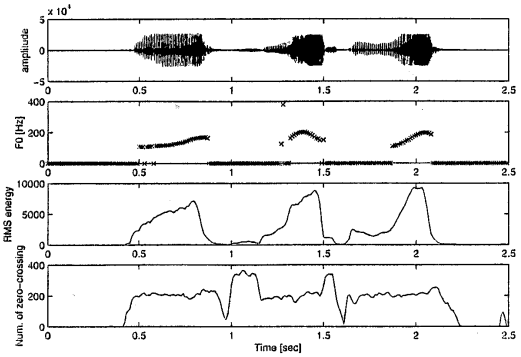

図 1: 「怒り」の音声処理の結果

3.4 ニューラルネットワーク

3.4.1 本研究で扱うニューラルネットワーク

本研究では時系列データを扱うため, 時間を考慮した Elman 型のニューラルネットワークを用いた. 学習アルゴリズムは結合荷重の変化項に慣性項をつけた誤差逆伝搬法を用いた. 慣性項を付け加えることにより, 結合荷重の変化に一種の慣性が生じるので, 誤差局面の細かな凹凸を無視するという効果が期待できる.

3.4.2 ニューラルネットワークの構成

各ネットワーク中の中間層におけるユニット数や学習比, 学習回数は, 予備実験を行った中で最適なものをそれぞれ採用した. なお, モーメンタム定数はいずれのネットワークでも 0.95 とした.

音声のみで感情認識を行う場合とバイモーダル情報で感情認識を行う場合は, 感情毎にネットワークを構築し, 最も出力値の高いネットワークに対応する感情を認識結果とした.

顔画像のみを用いた感情認識は,1 つのネットワークで行う. 出力層のユニット数は 6 個で, 各ユニットに 1 つの感情を割り当て, 最も出力値の高いユニットに対応する感情を認識結果とした.

3.5 実験結果

それぞれ構築されたネットワークにおいて学習に用いたデータを用いてクローズドな以下の 3 種類の認識実験を行った.

- 音声のみによる感情認識実験
- 顔画像のみによる感情認識実験
- バイモーダル情報による感情認識実験

各実験結果を表 2に示す.

表 2: 認識結果 [%]

情報/感情	無感情	怒り	驚き	悲しみ	喜び	嫌悪
音声	61	65	55	63	53	56
顔画像	91	94	97	93	91	90
バイモーダル	79	81	83	81	79	78

3.6 考察

表 2の実験結果では, バイモーダルにすると顔画像のみの場合よりも認識率が下がっている. これは顔画像のみによる認識と音声のみによる認識結果の組合わせ方が, まだ適切でないためと考えられ, 両者の結果を相補的に組み合わせる方法を検討していきたい.

また, 表 2では, 顔画像による認識率がいずれの感情の場合にも 90[%] 以上と高くなっているが, 顔画像による認識率が低くなってしまうケースにおいて, 音声からの情報が役に立つと考えられる. 現時点では評価用のデータも限られたものであるので, データを増やすことも含めて, 検討を進めたい.

4 むすび

本稿では, 顔画像と音声のバイモーダル情報を用いて, 感情認識を行う方法について検討した. 顔画像, 音声それぞれでの感情認識手法の組合わせ方に工夫を加え, 相補的な効果により安定した認識率を目指していきたい.

また, 本研究で作成したデータベースを実際に人間に対して提示したとき, どのような判断をするのかを確かめ, その結果と本研究での計算機による感情認識の結果とを比較したときにどのような相関があるのかを確かめる必要がある.

参考文献

[1] 黒川 隆夫, “ヒューマンコミュニケーション工学シリーズ - ノンバーバルインターフェース”, オーム社, 1996.

[2] Liyanage C. De Silva, Tsutomu Miyasato and Ryohei Natkasu, “Use of Multimodal Information in Facial Emotion Recognition”, IEICE Trans. Inf & Syst., vol.E81-D, no.1, Jan. 1998.

[3] 土佐 尚子, 中津 良平, “芸術とテクノロジー Lifelike, Autonomous Character “MIC” & Feeling Session Character “MUSE””, 第 4 回日立中研究所研究会予稿集, pp.75-82, 1996.

[4] 肖 業貴, N.P. チャンドラシリ, 田所 嘉昭, 尾田 政臣, “2-D DCT とニューラルネットワークを用いた顔画像の表情認識”, 信学論 A ,vol.J81-A, no.7,pp.1077-1086,1998