# 機械学習を用いた人の状態認識システムの構築

電子デバイス技術課　佐々木克浩*1　　機械情報システム課　釣谷浩之　　企画管理部　金森直希*2

## 1. 緒言

音声データと機械学習を用いて人の感情や健康状態を推定する技術があり、人と機械との対話や疾病の簡易自動検知への応用が期待される。感情認識では、正解感情が定められた音声データベースの作成にコストがかかるため、学習用の音声データが少量となる。これが感情認識を困難とする一要因となっており、人の状態認識システムを開発するうえで課題である。

感情認識を行う機械学習の方法は、大別して古典的手法と深層学習に基づく手法がある。深層学習に基づく手法は、近年主流であり、一般的に古典的手法より高い認識精度が得られるが、通常は大量の学習データを必要とする。少量の学習データへの対応策のひとつとして、大量のデータを用いて事前に学習された深層学習モデルの一部を感情認識モデルに転用する手法がある。

本報告では、少量の音声学習データによる人の状態認識の基礎として、事前学習済モデルを用いた音声感情認識システムに関して検討した。

## 2. システム

本研究で検討する感情認識システムを図 1 に示す。同図より、音声波形について短時間ごとに周波数変換した 2 次元データを画像認識用の事前学習済モデルに入力し、そのモデルより抽出した特徴量を分類器に入力して感情分類を行う。この方法は、計算負荷が比較的小さく簡便である特徴がある。学習済モデルは、画像認識で良く知られており、音声による感情認識への適用例[1]がある VGG16[2] を選定した。このモデルは図 2 に示すように複数の層で構成されている。最初に、複数の畳み込み層等により構成される 5 つのブロック (block1 から block5) がある。続いて、2 つの全結合層 (fc1、fc2) があり、最後に画像の予測分類を行うための全結合層を有する。学習済モデルからの特徴量の抽出は、最終層付近の利用[1]がベースと成り得る。一方で、モデルの最初の方の層は汎用性の高い特徴を抽出するため、分類対象と事前学習のデータが大きく異なる場合は、モデルの最初の層から特徴量を抽出するほうがよい可能性が示唆されている[3]。このため本研究では、上記ブロックの出力および全結合層の出力を対象に、特徴量を抽出する適切な部位を検討した。その際に、各ブロックからの出力は次元数が大きいため、特徴量マップ内で平均化するグローバルアベレージプーリング (GAP) 処理を施した。

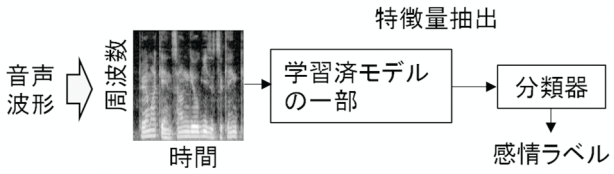


図 1　システムの概要

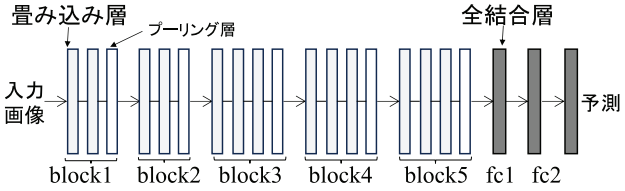


図 2 学習済モデル VGG16 の概要

## 3. 実験条件

実験に用いる音声データセットは、オンラインゲーム音声チャットコーパス[4]とした。そのコーパスには日本語の自発音声と演技音声が収録されており、今回は比較的評価が容易と考えられる演技音声を対象にした。話者 4 名 (男性 2 名、女性 2 名) の収録音のうち、表 1 に示す 5 種の感情を対象とした。音声データについて、VGG16 の入力サイズである 224×224×3 とするため、固定時間長の音声波形を抽出して対数メルスペクトログラムに変換し、補間処理および 3 チャンネル同一値の設定を行った。分類器はサポートベクトルマシン (SVM) を用い、話者 3 名分の音声データを用いて SVM を学習させ、残りの 1 名の音声データに対してシステムの性能を評価した。この学習・評価の過程について、すべての話者が評価対象となるように繰り返した際の平均値から評価値を算出した。開発言語は Python とし、信号処理やモデル等の実装には librosa、keras および scikit-learn ライブラリを用いた。

特徴量の抽出部を検討する実験では、2.23 秒間の 1 フレームの音声波形を対象にした。さらに、発話単位の感情分類への拡張として、1 フレームより長い発話では複数フレームの音声波形を用いることも試みた。この場合、映像分野[5]等の方法[6]を参考に、複数の音声波形において上記と同様の手順で特徴量を抽出し、それらのフレーム間における最大値を分類器に入力した。

表 1 音声データセット

| 感情 | 発話数 |
|---|---|
| 怒り | 240 |
| 喜び | 252 |
| 平静 | 344 |
| 悲しみ | 252 |
| 驚き | 288 |
| 総発話数 | 1 376 |

*1 現 機械情報システム課、*2 現 ものづくり研究開発センター

## 4. 実験結果

学習済モデルの最終層付近である全結合層の出力を特徴量として用いて、各感情の再現率の平均値を評価した結果を表 2 に示す。この値は、SVM の正則化係数 $C$ を 1～$10^4$、パラメータγを $10^{-8}$～$10^{-3}$ の範囲において一桁間隔で変更した際の最良値である。同表より、fc2 より手前の層である fc1 の出力を用いた方が高い平均再現率が得られている。更に手前の層を検討するため、ブロック出力に GAP 処理を施した特徴量を用いて、同様に評価した結果を表 3 に示す。同表と表 2 より、本実験の条件では、block4 の出力を用いた場合に最も平均再現率が高いことがわかった。この時の感情ごとの再現率を表 4 に示す。同表より、再現率は各感情で 5 割から 7 割の範囲であり、平均は 65%であった。

表 2 全結合層の出力部に対する評価結果

| 特徴量抽出部 | 平均再現率 [%] |
|---|---|
| fc1 出力 | 59 |
| fc2 出力 | 56 |

表 3 ブロック出力部に対する評価結果

| 特徴量抽出部 | 平均再現率 [%] |
|---|---|
| block3 出力 | 63 |
| block4 出力 | 65 |
| block5 出力 | 56 |

表 4 block4 の出力利用時の評価結果

| 感情 | 再現率 [%] |
|---|---|
| 怒り | 56 |
| 喜び | 63 |
| 平静 | 72 |
| 悲しみ | 69 |
| 驚き | 64 |

次に、1 フレームより長い発話では、時間的に 75%オーバーラップさせて複数フレームを取得するようにし、block4 の出力を利用して同様に評価を行った結果、平均再現率 68%が得られた。これにより、表 4 の場合の平均再現率と同等以上の性能で発話単位の感情を分類できる可能性が得られた。

本実験では 1 つの演技音声データセットを対象に評価したが、今後は、複数の音声データセットを対象にした評価が必要であり、実用上は自発音声に対する有効性の検証が課題となる。その際には、既存手法による感情分類結果との比較・検証も必要である。

## 5. 結言

事前学習済モデル VGG16 と SVM 分類器を用いた音声感情認識システムを構築し、そのモデルにおける特徴量抽出部を検討した。演技音声における 5 種の感情を対象に評価実験を行った。その結果、モデルの中間の層付近である block4 の出力に GAP 処理を施した方が、最終層付近の出力を用いるよりも高い平均再現率が得られた。更なる再現率向上のために、他の学習済モデルの適用や学習済モデルのパラメータを微調整する手法に関して検討することが必要である。また、本システムを基礎的な枠組みとして、他の状態認識への展開も模索したい。

## 参考文献


1) Mehmet Bilal Er: *IEEE Access*, **8** (2020) 221640
2) K. Simonyan *et. al.*: *arXiv*:1409.1556v6 (2015) 1
3) F. Chollet, マイナビ出版 (2022) 225
4) 有本 他, 日本音響学会 2013 年秋季研究発表会講演論文集, 1-P-46a (2013) 385
5) 植木, 日本神経回路学会誌, **24** (2017) 13
6) S.Zhang *et. al.*: *IEEE Access*, **8** (2020) 23496


## 謝　辞


本研究では、国立情報学研究所 音声資源コンソーシアムから提供を受けた「感情評定値付きオンラインゲーム音声チャットコーパス (OGVC)」を利用した。



キーワード：音声、感情認識、事前学習済モデル、特徴量抽出、分類器


# Construction of Recognition System for Human Condition Using Machine Learning


Electronics and Device Technology Section; Katsuhiro SASAKI*[1]
Mechanics and Digital Engineering Section; Hiroyuki TSURITANI and
Planning and Management Department; Naoki KANAMORI*[2]



A speech emotion recognition system was constructed using a support vector machine classifier and features extracted from the VGG16 pre-trained model. Appropriate layer depth for the feature extraction were examined, and the unweighted average recall (UAR) was evaluated for the five categories of emotions in acted Japanese speech. As a results, the highest UAR was obtained when the feature after applying a global averaging pooling to the block4 (near middle layer) output was used.