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　　機械情報システム課　釣谷浩之　　企画管理部　金森直希

1. 緒言

機械学習を用いて音声から人の感情を認識する技術があり、健康状態の推定や、人と機械との対話の要素技術としての応用が期待される。一方、機械学習に必要な正解感情が定められた音声データベースの作成には非常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とが、感情認識を困難とする一要因となっている。また、そのデータベースには自発音声と演技音声があり、自発音声の感情認識は難易度が高い[1-2]。

本研究では、機械学習を用いた音声による感情認識システムを構築し、日本語の自発音声を対象とした認識精度を高める方法に関して検討する。

2. システムおよび実験

2.1 検討指針

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[3]がある。前者は、人が設計した特徴量を分類器に入力することで感情を分類する。一方、近年主流である深層学習では、特徴量抽出と分類がひとつのネットワークになっており、特徴量が自動で抽出される。この方法は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、画像分類等の大量のデータを用いて学習させ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。一方で、日本語の自発音声の少量の学習データに対しては、古典的手法のほうが適している例[2]がある。以上の状況から、本研究では、古典的手法をベースにして、後述の特徴量抽出部に深層学習(事前学習モデル)を適用する方法に関して検討する。本年度は、ベースとなるモデルの性能確認の目的で、従来の古典的手法に基づいたシステムを構築した。

2.2 音声データ

日本語の自発音声において感情評定ラベルが付与されている音声データセットとしては、自発音声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)[4]がある。このコーパスは、感情が音声に反映されるように、オンラインゲームの音声チャットによるコミュニケーションを行わせている。収録音声のうち、話者 11 名（男性 8 名、女性 3 名）の感情 10 種を含む合計 6578 発話については、3 名の評価者により評定されている。本研究では、怒り、恐れ、喜び、平静、悲しみの 5 種の感情を対象とし、評価者 3 名中 2 名以上が一致した感情を正解感情ラベルとして使用した。各感情の発話数を表 1 に示す。

表 1 音声データセット

感情	発話数
怒り	237
恐れ	142
喜び	595
平静	798
悲しみ	243
総発話数	2015

2.3 モデルおよび実験方法

構築した感情認識システムを図 1 に示す。音声波形データの短時間区間から抽出する特徴量は、音声認識等において代表的な、メル周波数ケプストラム係数(Mel-Frequency Cepstrum Coefficient : MFCC)と二乗平均平方根(Root Mean Square : RMS)を用いた。MFCC は、音声の周波数スペクトルの概形に関する特徴量であり、その次数は 1～12 次を選択した。また、MFCC と RMS の変化量として、時間的に隣接する値の一階差分も特徴量とした。これら短時間の特徴量(26 次元)の時系列データから、表 2 に示す 7 種の統計量を求め、合計 182 次元の特徴量とした。

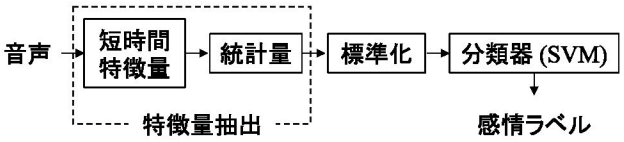

図 1 感情認識システム

表 2 特徴量

短時間の特徴量	MFCC（1～12 次）、RMS これらの一階差分 D
統計量	平均、標準偏差、最大、最小、四分位数
合計次元数	182 次元

特徴量は標準化した後、分類器に入力した。以上のシステムの開発には、Python を用いて行った。短時間の特徴量は、LibROSA ツールキットを用いて、音声データを 16kHz でリサンプリングし、窓関数(幅 25ms)を 10ms 間隔でシフ

トさせるように設定して生成した。標準化と分類は scikit-learn ライブラリにより実装し、分類器はサポートベクトルマシン(Support Vector Machine : SVM)を採用した。実験は、話者 10 名分の音声データを用いて分類器を学習させ、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[2,5]。

3. 実験結果および考察

感情認識システムを評価した結果例を表 3 に示す。同表は、分類器のパラメータ設定を、kernel は rbf、class_weight は balanced、他は既定値とした場合の結果である。同表では、例えば正解感情ラベルが怒りについては、19%が正しく認識できているが、恐れに 6%誤認識しており、同行に他の感情に対する誤認識の率が示されている。他の正解感情ラベルの場合も同様に示されており、同表の太字が各感情ラベルの正解率を表している。各感情ラベルの正解率を平均した結果は、39%であった。

表 2 に示す発話数が多い感情ラベルである喜び、平静では、表 3 より 60%前後の精度が得られている。一方、発話数が少ない怒りや恐れでは 20%以下の低い精度となっており、平静への誤認識率が高い。この一対策として、データ拡張等により発話数を均一化させる検討が考えられる。また、精度向上のための一方法として、分類器におけるパラメータの最適化が挙げられるが、未知データへの汎化性の観点からは学習用のデータを分割して検証することが望ましく、これは今後の課題とする。

表 3　混同行列

	認識感情(%)				
正解感情	怒り	恐れ	喜び	平静	悲しみ
怒り	**19**	6	30	43	3
恐れ	13	**16**	17	33	20
喜び	13	5	**57**	20	5
平静	7	5	14	**65**	9
悲しみ	4	12	8	35	**40**

4. 結言

深層学習モデルを適用する前段のベースシステムとして、代表的な音響特徴量と分類器を用いた感情認識システムを構築した。本システムを用い、日本語の自発音声の 5 感情を認識する実験を行い、システムの基礎的な動作とともに、認識精度が概して低い課題を確認した。今後は、精度向上のため、深層学習(事前学習モデル)の特徴量抽出部への導入を検討する予定である。また本研究では、長期的な展開のひとつとして、本研究のシステムや知見を基盤に、健康推定に繋げていきたいと考えている。

参考文献

1) 森 他, 電子情報通信学会誌, 101, 9 (2018) 902.
2) T. Iizuka et. al., Acoust. Sci. & Tech. 43, 4 (2022) 228.
3) 安藤 他, 日本音響学会誌, 79, 1 (2023) 72.
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) 阿部 他, 情報処理学会東北支部研究報告, 2016-7-A3-3 (2016) 1.

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、機械学習、特徴量、分類器

Construction of Recognition System for Human Condition Using Machine Learning

Electronics and Device Technology Section; Katsuhiro SASAKI
Mechanics and Digital Engineering Section; Hiroyuki TSURITANI
Planning and Management Department; Naoki KANAMORI

An emotion recognition system for Japanese spontaneous speech was constructed to confirm the performance of the base model using well-known hand-crafted features and a Support Vector Machine (SVM) classifier. In the system, the five categories of emotion were recognized using statistical values of short-time features such as Mel Frequency Cepstrum Coefficient (MFCC), Root Mean Square (RMS) and those delta coefficients. The recognition accuracy (unweighted average recall) of the system was confirmed.