5R-03

深層学習を用いた音声感情認識

生形優也† 田村仁†

日本工業大学先進工学部 ロボティクス学科†

1. はじめに

感情は人と人でのコミュニケーションにおいて重要な要素となっている．去年登場したChatGPTは，人での自然な対話が可能であり人と AI とのコミュニケーションの実現に近づくことができた一方で，相手の声や表情では感情を推定しながら話すことはできない．人は目や耳で相手の表情や声をとらえそれらを元に相手の感情を推し量り言葉を選びながらコミュニケーションを行うことが一般的であるが，表情や音声感情認識技術はまだ精度が低く確実に感情を推定することが難しい状況である．表情の推定は多くの研究があるが音声感情認識においては表情に比べて数が少なく言語によっても精度が変わる可能性があるため精度の向上を検討している

2. 関連研究

音声感情認識では、データセットを用いて音声データから特徴量抽出をして学習させる．
音声感情認識の学習によく使われる音響特徴量は STFT スペクトログラム、メル周波数スペクトログラム[2],メル周波数ケプストラム係数[3]が多く，今回はそれらの音響特徴量を使用して感情分類の正答率が同等に出せるかを再現する．今回使用するデータセットである RAVDESS データセット[4]を使った研究での精度はほとんどの場合で 70～85%であった．

3. 提案手法

本研究では，人の感情の音声から音響特徴を抽出して，中立，穏やか，幸せ，悲しい，怒り，恐怖，嫌悪，驚きの 8 つの感情を人の音声から分類していくことを目的としている．使用する RAVDESS データセットは男性 12 人，女性 12 人で構成されており，それぞれが上記の 8 つの感情で発音している，wav 形式のみで録音されている．
今回の実験では STFT（短時間フーリエ変換）スペクトログラムとメル周波数スペクトグラム、メル周波数ケプストラムの画像を音響特徴抽出に用いる．
STFT スペクトログラム(図 1)とは，音声データの周波数成分を時間ごとに分解したもので，色がオレンジ色になるほど，音が強くなることを表している．縦軸が周波数で，横軸は時間である．時間ごとに周波数成分がどのくらいの変化があるのかを示している．メル周波数スペクトログラム(図 2)とは先ほどの STFT スペクトログラムを人間の音の知覚に近いメル尺度に変換されたものである．メル尺度とは人間の聴覚の聞こえ方に基づいた尺度である．人間の聴覚には周波数の低い音に対して敏感であり，周波数の高い音に対して鈍感であるという性質がある．つまりメル周波数スペクトログラムは音声信号の周波数情報をより人間の聴覚に適した形で表現したスペクトログラムである．メル周波数ケプストラム係数（図３）はメル周波数スペクトログラムの値の対数をとり、離散コサイン変換をすることによって求められる。この特徴量は音の声質を表しており、感情・音声認識に用いられている。この 2 つのスペクトログラム画像を 8 つの感情ごとに分け ResNet50 とよばれる深層学習モデルを使用して学習させる．エポック数は 10, PyTorch を使用する．最後にテストデータを用意して，8 つの感情に分類できるかを判定する．

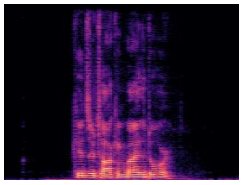

図 1 STFT スペクトログラム

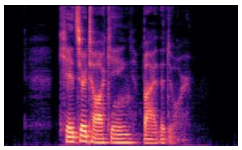

図 2 メル周波数スペクトログラム

図 3 メル周波数ケプストラム係数

4. 実験

学習させたモデルでテストデータから 8 つの感情を分類する実験を行った．テストデータは wav 形式から STFT スペクトログラムとメル周波数スペクトログラム、メル周波数ケプストラム係数に変換し、

Speech Emotion Recognition using deep learning
†Yuya Ubukata，Hitoshi Tamura
Department of Robotics，Faculty of Advanced Engineering, Nippon Institute of Technology

ファイル数は全部で 60 である．中立 4，穏やか 8，幸せ 8，悲しい 8，怒り 8，恐怖 8，嫌悪 8，驚き 8 となっている．評価指数は正解率を用いる。

5．結果

表 1 STFT スペクトログラムでの分類結果（赤い文字が正しく分類された数）

	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	4	0	0	0	0	0	0	0
穏やか	0	8	1	3	0	0	0	0
幸せ	0	1	4	0	0	0	0	0
悲しみ	0	3	1	2	1	0	0	0
怒り	0	0	1	0	7	0	1	0
恐怖	0	0	0	0	1	4	0	0
嫌悪	0	0	1	3	0	2	7	0
驚き	0	0	0	0	0	2	0	8

表 2 メル周波数スペクトログラムでの分類の実験結果（赤い文字が正しく分類された数）

	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	0	0	0	0	0	0	0	0
穏やか	0	6	1	4	0	0	2	0
幸せ	0	0	6	0	0	2	3	1
悲しみ	4	2	1	3	0	1	0	0
怒り	0	0	0	0	7	0	0	0
恐怖	0	0	0	0	0	5	0	1
嫌悪	0	0	0	1	1	0	3	0
驚き	0	0	0	0	0	0	0	6

表 3 メル周波数ケプストラム係数での分類結果（赤い文字が正しく分類された数）

	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	0	0	0	0	0	0	0	0
穏やか	3	7	0	2	0	0	0	0
幸せ	0	0	6	2	0	4	2	2
悲しみ	4	1	0	2	0	0	0	0
怒り	0	0	2	0	8	0	4	0
恐怖	0	0	0	0	0	4	0	0
嫌悪	1	0	0	0	0	0	2	0
驚き	0	0	0	2	0	0	0	6

STFT スペクトログラム、メル周波数スペクトログラム、メル周波数ケプストラム係数の精度はそれぞれ 73.3%、60%、58.3%となった．
音響特徴量を画像にした際に STFT スペクトログラムが一番精度が高いと考えられる。

6．考察

5 の結果では STFT スペクトログラム、メル周波数スペクトログラム、メル周波数ケプストラム係数は関連研究よりも精度が低くなった。今回使用したデータセットは無音の部分があり、特徴量抽出に問題があった．さらに画像にすることで元々の音声信号の小数点以下の情報が失われたことも考えられる．さらに STFT・メル周波数スペクトログラム、メル周波数ケプストラム係数や画像に変換する際に、パラメータの調節が適切でなく、精度が低下したことが考えられる．しかしながら、３つの実験の結果では怒りと穏やか、驚きの精度は高く、この３つの感情は画像からでも特徴量抽出できることがわかり判別しやすい感情であることがわかる。怒りと驚きは周波数の変化が他の感情と比べて激しく、音圧や周波数成分も高いものが多いためだと考えられる。逆に穏やかでは周波数の変化がなく一定であり、音圧と周波数成分も低いものが多いためだと考えられる。

6．おわりに

本研究では，人の音声から音響特徴量を抽出して感情を分類することを目的としている．今回は STFT 今回の実験で画像を使った特徴量抽出では精度が悪かったため、スペクトグラムの numpy 配列データなどほかの特徴量を使用していきたいと考える．中立、幸せ、悲しみ、恐怖、嫌悪の感情が精度が低いため、どの特徴量が適切なのかをしらべていきたいと思う。さらに今回使ったデータセット以外にも，ほかのデータセットを使用したり，自分や他の人の感情音声を録音するなど，データ数をさらに追加を予定している．

参考文献

[1] K.Tarunika , R.B Pradeeba , P, Aruna
“Applying Machine Learning Techniques for Speech Emotion Recognition” ,, in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp.1–5, IEEE,2018.

[2] R Vijaya Saraswathi ,Dheeraj Nandigama, R Vasavi ,B Ganesh Babu, D sai Shivani
“Voice Based Emotion Detection using Deep Neural Networks” , 2021 International Conference on Smart Generation Computing, Communication and Networking(SMART GENCON), pp.1-6,IEEE, 2021

[3] “深層学習を使って楽曲のアーティスト分類をやってみた！” . Platinum Data Blog . 2018-04-17
https://blog.brainpad.co.jp/entry/2018/04/17/143000
(2024-01-10 閲覧)

[4] “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)” .zenodo .2018-04-5
https://zenodo.org/records/1188976