E-013

データセット拡張を用いた音声感情認識

Speech emotion recognition using dataset augmentation

生形 優也‡　杉山 颯‡　田村 仁‡
Yuya Ubukata　Hayato Sugiyama　Hitosi Tamura
日本工業大学院工学研究科　機械システム工学専攻

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．AI が感情認識を行うことができればより自然に人間とのコミュニケーションを行うことができ，昨今での高齢化や福祉・介護分野での人手不足解消につながると考えられる．人間と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，精度の向上を検討している．しかしながら，音声感情認識の問題として感情ラベル付きのデータセットの不足という課題がある．先行研究では[1]学習を行うと特定個人の声質やその他の特徴量が過学習されて認識精度が低下してしまう問題がある．そこで本研究ではデータセットにボイスチェンジャなどを使用し，感情ラベル付きデータセットの声質やその他の特徴量を変更することでデータセット拡張手法を検討する．これらの手法を用いて感情分類の精度に影響を与えずにデータセットを拡張できるかを実験で検証する．

2. 関連研究

関連研究として，RAVDESS データセット[1]を用いたデータセット拡張実験[2]がある．この実験では拡張方法としてノイズ追加、タイムシフト，タイムストレッチ，ピッチシフトを使用している．これらの手法では従来の音声データセット拡張方法であり，音声信号の振幅と位置の変更や音の高さを変更せずに持続時間やピッチ変えることで元のサンプルに似てはいるが，ある程度のサイズと多様性を持つデータを生成することでモデルのパフォーマンスと一般化の強化を目的としている．拡張なしのモデルでは感情の精度は 62%だがデータセット拡張により 92%という高精度を実現している．しかしながら声質に着目したデータセット拡張が見当たらないため，本実験では声質が音声感情認識にどのような影響があるのかを検討し，従来の手法との比較を行う．

3. 提案手法

本研究では，RAVDESS データセットをボイスチェンジャを利用して拡張を行う．RAVDESS とは北米英語を話す 24 人(男性 12 人，女性 12 人)で構成されており 1440 の音声発話で構成されそれぞれが上記の 8 つの感情（中立，穏やか，幸せ，悲しい，怒り，恐怖，嫌悪，驚き）を乗せた語彙が一致する 2 つの文の発音を wav 形式で収録されているデータセットである．本実験の音声感情認識では拡張された RAVDESS データに収録されている感情を乗せた音声データ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．ボイスチェンジャには RVC[3]（Retrival-based-voice-conversion）を使用する．RVC とは中国で開発された AI を使用した高性能なボイスチェンジャであり，変換したい声を学習データとして与えることでより他のボイスチェンジャよりも人間的で自然な声を再現することができる．データ拡張方法については，RAVDESS データセットに RVC を使用し声質変換を行い元々の話者 24 名を拡張する．拡張に使用する RVC の学習データは音声認識の研究開発を目的として Mozilla が立ち上げた大規模多言語音声コーパスの Common Voice[4]を使用する．Common Voice は世界中のボランティアによるデータセットの作成が日々行われており，多量の RVC の話者学習データを入手することが可能になる．今回は Common Voice の英語音声 Common Voice Delta Segment14.0 を使用する．収録人数は 1212 人，サイズは 1.43GB である．

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係数の画像を音響特徴抽出に使用する．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

Speech emotion recognition using dataset augmentation
† Yuya Ubukata
† Hitosi Tamura
† Hayato Sugiyama
‡ Nippon Institute of Technology, Graduate School, Mechanical System Enginrrting Major

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[5]とよばれる深層学習モデルを使用して学習を行う．ResNet50 とは深さが 50 層の畳み込みニューラルネットワークであり本実験では 100 万個を超えるイメージで学習させた事前学習済みモデルを使用し感情の判定を行う．

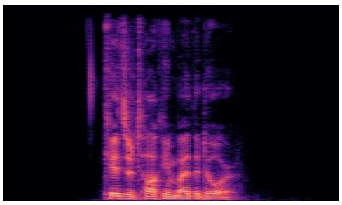

図 1 STFT スペクトログラム

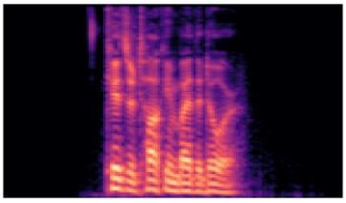

図 2 メル周波数スペクトログラム

図 3 メル周波数ケプストラム係数

4. 実験

4.1 データセット拡張なし

初めにデータセットの拡張を行わない場合での 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での実験を行った．データセットの 24 名のうち 23 名を学習用データとし残り 1 名をテストデータとした．1 名につき 60 音声ファイル収録されており，それぞれ中立 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．評価方法はテストデータの感情ラベルをどれだけ正しく分類できたかの正解率(accuracy)を使用する．分類結果を表 1，2、3 に示す．

表 1 STFT ペクトログラム

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

正解率は 76.67%となった．

表 2　メル周波数スペクトログラム

実際＼予想	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	0	0	0	0	0	0
幸せ	4	1	2	0	1	0	0	0
悲しみ	0	4	1	2	1	0	0	0
怒り	0	0	0	0	8	0	0	0
恐怖	0	1	1	0	0	6	0	0
嫌悪	0	3	0	0	1	0	4	0
驚き	0	0	1	0	0	0	0	7

正解率は 68.8%となった．

表 3　メル周波数ケプストラム係数

実際＼予想	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	2	0	0	2	0	0	0	0
穏やか	1	7	0	0	0	0	0	0
幸せ	0	0	5	1	2	0	0	0
悲しみ	1	2	1	2	1	0	1	0
怒り	0	0	0	0	8	0	0	0
恐怖	0	0	1	0	0	5	2	0
嫌悪	0	0	1	0	1	0	6	0
驚き	0	0	0	0	0	0	0	8

精度は 71.7%となった．

３つの実験の結果では３つの音響特徴量とも精度はほとんど同等であることがわかり，この特徴量から感情分類を行えることがわかる．関連研究でもデータセット拡張なしの場合の精度は 62%となっており、やや上の精度であることがわかった．

4.2 データセット拡張あり

次に RVC でのデータセットの拡張を行った場合での 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での実験を行った．話者を 24 名から 58 名へと増加させて男性話者には男性の声で，女性話者には女性の声で RVC によって学習させたモデルで声質変換を行った．58 名のうち 54 名を学習用データとし残りの 4 名をテストデータとした．テストデータはそれぞれ中立 16，穏やか 32，幸せ 32，悲しい 32，怒り 32，恐怖 32，嫌悪 32，驚き 32 とする．分類の結果を表 4，5，6 に示す．

表 4　STFT スペクトログラム

実際＼予想	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	12	3	0	1	0	0	0	0
穏やか	5	22	2	3	0	0	0	0
幸せ	2	6	5	3	1	0	0	15
悲しみ	2	16	0	4	2	4	4	0
怒り	0	5	0	3	12	5	2	5
恐怖	3	1	5	6	1	14	0	2
嫌悪	0	4	0	1	1	1	25	0
驚き	1	0	1	0	0	2	0	28

精度は 50.1%となった．

表 5　メル周波数スペクトログラム，

実際＼予想	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	13	0	2	1	0	0	0	0
穏やか	3	20	3	6	0	0	0	0
幸せ	7	4	3	0	3	1	0	14
悲しみ	1	12	3	9	2	2	3	0
怒り	2	2	2	0	20	1	1	4
恐怖	0	1	4	9	4	14	0	0
嫌悪	0	4	0	1	1	1	25	0
驚き	0	0	2	0	1	0	0	29

精度は 55.4%となった．

表 6　メル周波数ケプストラム係数

実際＼予想	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	9	6	0	1	0	0	0	0
穏やか	8	14	11	1	1	3	0	2
幸せ	3	6	5	2	3	0	0	13
悲しみ	0	13	2	10	0	4	3	0
怒り	1	2	3	3	19	0	1	3
恐怖	0	1	2	9	0	20	0	0
嫌悪	0	2	1	2	3	1	23	0
驚き	0	0	0	0	0	0	0	32

精度は 55%となった．

関連研究と RVC でのデータセット拡張では精度が下がる結果となった．

５．考察

データセット拡張なしとありとのそれぞれの感情ごとの精度を音響特徴量ごとに図 4,図 5，図 6 に示す．

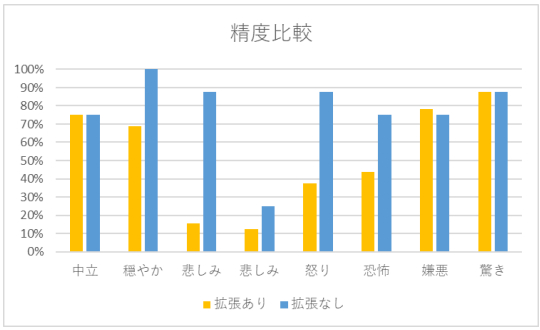

図 4STFT での精度比較

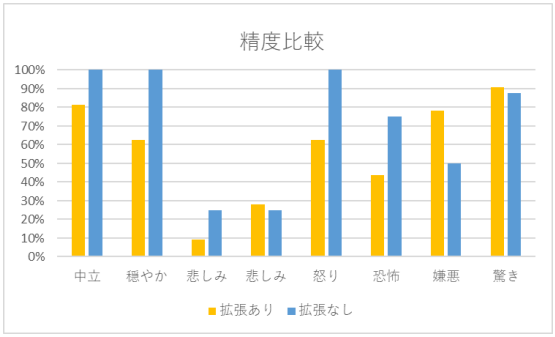

図 5 メル周波数スペクトログラムでの精度比較

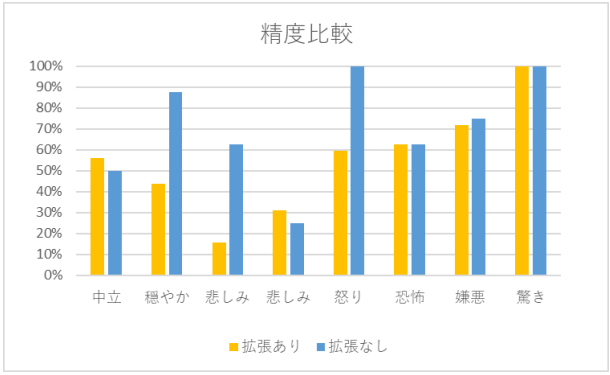

図 6　メル周波数ケプストラム係数での精度比較

データセット拡張で精度が下がってしまった要因として RVC での感情変換がうまくいかなかったことが挙げられる．変換された音声を聞くと感情が乗っていない中立での声ならば声質変換が人間的で自然な声であるが、怒りや悲しみなどの音の高低差が激しく変化しやすくなる感情ではノイズが発生したり，声が急激に高くなり不自然に感じる音声が多くなり感情をうまく変換できないことがわかった．変換前の怒りのメルスペクトログラム図 7，変換後の図 8 をみると高周波数帯ほど情報が失われていることが確認できる．このことから RVC での声質変換では高周波数

帯での変化に鈍く声の高低差が激しい怒りの感情が大幅に低下したと考えられる．

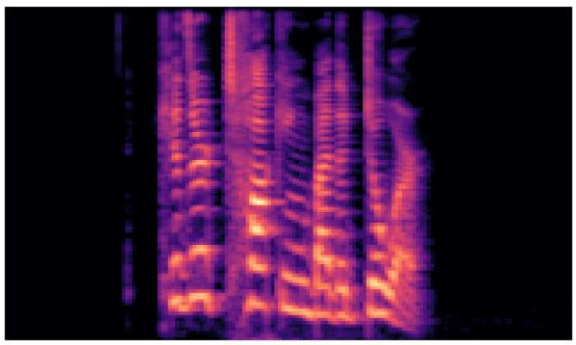

図 7 変換前のメル周波数スペクトログラム

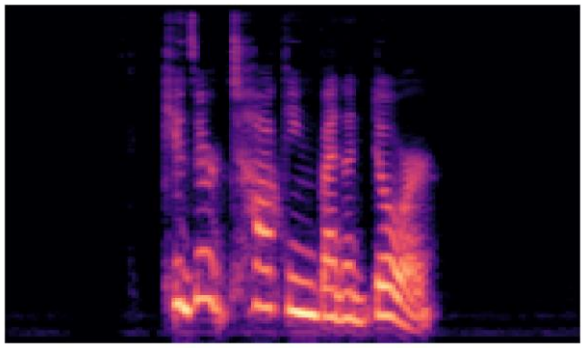

図 8　変換後のメル周波数スペクトログラム

RVC で拡張されたデータセットの結果では幸せが驚き，悲しみが穏やかになるといった誤判定が多かった．拡張していないデータセットでは悲しみから穏やかでの誤判定もあり，データセットによる特徴だと考えられるが，幸せから驚きへの誤判定がないため，RVC による特徴だと思われる．さらに RVC に使用される学習データについても個人による作成が可能であることからその音質にばらつきが生じることがある．また，マウスのクリック音や無音区間などのノイズが含まれている場合があり音声の前処理が適切に行われていないことも一因として考えられる．

6. おわりに

ボイスチェンジャを用いたデータセット拡張実験を行った結果、RVC では感情を保持したまま自然な声に変換することが困難であることが判明した。今後の研究では、自然な声の再現性に限界があるとしても、感情を保持したまま変換できるボイスチェンジャを用いてデータセットの拡張を進めていく．また、RVC の性能向上を目指し、高周波数成分を多く含む学習データを用意することで、変換後の音声品質の改善が行えるかを検討する．

参考文献

[1] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.PLoSONE13(5):e0196391. https://doi.org/10.1371/journal.pone.0196391

[2] V. Singh and S. Prasad, "Speech Emotion Recognition using Fully Convolutional Network and Augmented RAVDESS Dataset," 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA), Mumbai, India, 2023, pp. 1-7, doi: 10.1109/ICACTA58201.2023.10392486.

[3] Retrival-based-voice-conversion, https://github.com/RVC-Project , (最終閲覧日　2024/6/6)

[4] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M. and Weber, G. (2020) "Common Voice: A Massively-Multilingual Speech Corpus". Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). pp. 4211—4215

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778). DOI: 10.1109/CVPR.2016.90