E-011

音声感情認識のための複数の手法を用いた多様なデータ拡張方法の検討
Examination of various data augmentation methods using multiple techniques for audio emotion recognition.

生形優也† 田村　仁† 大久保　友幸†
Yuya Ubukata Hitoshi Tamura Tomoyuki Okubo

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． LLM は，人との自然な対話を可能としたが，依然として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，相手の感情を推し量りながらコミュニケーションを行う．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，福祉・介護分野などでの人手不足解消にもつながると考えられる．

人間の表情や音声感情認識の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は表情に比べて数が少なく,精度の向上が期待されている．

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が挙げられる．感情表現には多種多様なものがあり，感情を高い精度で認識するために膨大なデータを収集するのは非常に困難である．そこで少数のデータから多様なデータを生成するデータ拡張が注目されている．本研究ではデータ拡張手法の検討と，データ拡張による感情推定への影響を確かめる．

2. 先行研究

先行研究[1]では，RAVDESS データセット[3] を 20 倍程度まで拡張し，時間伸縮，時間マスキング，ピッチシフト，そして CopyPaste といったデータ拡張手法を用いて，音声感情認識の精度向上を検証した．結果としては，拡張前が 68.3%，拡張後が 69.6%となったが，正解率の顕著な上昇は見られなかった．そこで本研究では音声データの特徴量をメル周波数ケプストラム係数からメル周波数スペクトログラムへと変更を行い，データ拡張の大規模化を行うことができる２種類の拡張手法を用いる．

3. 関連研究

音声感情認識におけるデータ拡張は，時間マスキング，ピッチシフト，時間収縮などさまざまな手法があるが，特にデータの規模を大容量になおかつ感情認識に特化させた学習データ拡張手法である，感情音声を結合する CP(CopyPaste)[2]，感情音声を混合する EMix[3]手法を用いた研究がある．この２つの手法では IEMOCAP[4]と Crema-D[5]データセットでは他の拡張手法より高い精度を実現している．（表１）

†日本工業大学　Nippon Institute of Technology

一般に機械学習においては学習するデータ量に応じて正解率の向上が見られる．そこでこれら２つの拡張手法でデータ数の規模をさらに増やすことでデータセットの正解率がさらにどのように変化するのかをデータセット拡張の有効性を検証する．

表１　手法による正解率([5]より引用)

拡張手法	IEMOCAP	Crema-D
拡張なし	72.59	74.28
CopyPate	71.36	70.27
EMix	77.32	77.25

3.提案手法

本研究では，RAVDESS[3]データセットをデータ拡張して CopyPaste で約 57 倍，EMix で約 56 倍まで規模を拡大し，それぞれのモデルへ学習を行い，結果の精度の比較を行う．

RAVDESS データセット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 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．これらの拡張された 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である log メルスペクトログラムを学習モデルへの入力として 8 つの感情の分類を行う．

4. 実験

4.1 データ拡張手法

本実験では以下のデータ拡張手法を使用した．

4.1.1 CopyPaste (CP)

Pappagari[2]らによって提案された CopyPaste は，人間が感情をどのように知覚するかという点に着目した，知覚に基づいたデータ拡張手法である．この手法は，「中立発話とそれ以外の感情発話が連続している場合，後半の感情をより強くその発話全体の感情として強く認識する」という人間の心理特性に基づいたものである．中立的な発話と感情的な発話（感情 E）とを連結しても，その連結された発話が引き続き感情 E としてラベル付けできるというアイデアが提案されている．CopyPaste には以下の３つの拡張手法がある．

①Neutral CopyPaste (N-CP):中立的な発話と任意の感情発話を連結し，感情 E としてラベル付けを行う．(図１)

②Same Emotion CopyPaste (SE-CP): 同じ感情 E を持つ 2 つの感情的な発話を連結し，生成された発話に感情 E のラベルを行う．(図 1)

③ N+SE-CP: 上記の N-CP と SE-CP の両方を組み合わせて一つの学習データとする．本実験では③の拡張方法を用いた学習データを作成した．

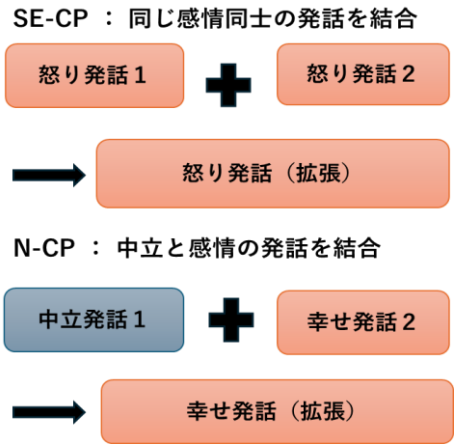

図 1　CopyPaste 拡張方法

4. 1 .2 EMix

EMix[3]では中立発話や任意の感情発和を混合させることによりデータペアの特徴ベクトルを線形補間して新しいサンプルの生成を行う．EMix には以下に 3 つの拡張手法がある．

①EMix-N:任意の感情を持つサンプルと，中立感情のサンプルを組み合わせ混合する方法である．(図 2)xが log メルスペクトログラム，yが感情ラベルとしたとき，感情 e を持つサンプル (x^e, y^e) と，中立nを持つサンプル (x^n, y^n) に対して，新たなサンプル$(\tilde{x}, \tilde{y})$は式(1)のように定義される．

$$\tilde{x} = \lambda x^e + (1-\lambda)x^n \quad (1)$$
$$\tilde{y} = y^e$$

ここで，$\lambda \sim \cup(0.5, 1)$は，感情を持つサンプルからより多くの情報を保持するためにランダムに選ばれる．EMix-N は，中立サンプルが背景ノイズのように機能するノイズ拡張の一種とみなすことができる．

②EMix-S: 同じ感情状態を持つ音声のみを混合する方法である．(図 2)混合されたサンプルのラベルは，それらを構成する元の音声と同じラベルになる．感情 e を持つサンプル $(x_i^e, y^e), (x_j^e, y^e)$に対して，新たなサンプル$(\tilde{x}, \tilde{y})$は式(2)のように定義される．

$$\tilde{x} = \lambda x_i^e + (1-\lambda)x_j^e \quad (2)$$
$$\tilde{y} = y^e$$

ここで，λ は一様分布$\cup(0, 1)$はランダムに選ばれる．同じ感情ラベルを持つ音声がサンプル中に多く含まれることで，混合後のサンプルにも一貫した感情がはっきりと表れるため，元のサンプルよりも感情がより明確で信頼性の高いデータとして扱うことができる．そのため，EMix-S は，ノイズが多く曖昧な音声感情認識データに対して効果的に働くと期待されている．

③EMix-NS: EMix-N と EMix-S を組み合わせて一つの学習データとする．　本実験では③の拡張方法を用いた学習データを作成した．

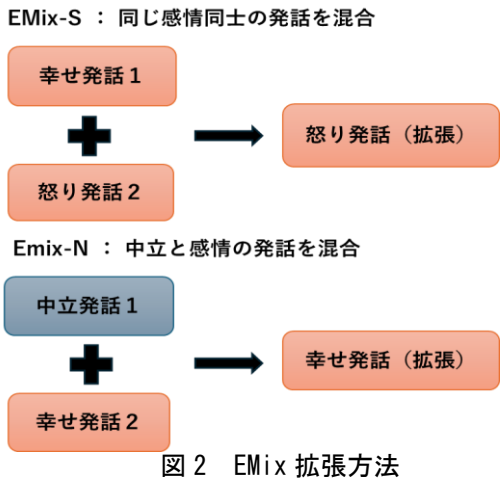

図 2　EMix 拡張方法

4. 2 実験準備

Ravdess データセットの 24 名話者のうち 22 名分をデータ拡張に使用し，残り男女各 1 名をテストデータとする．拡張方法については CopyPaste と EMix 手法を上記の①と②のデータ拡張手法を用いて音声データをそれぞれ約 57 倍と 56 倍拡張を行う．つぎに CopyPaste から拡張を行った音声データから logmel スペクトログラムの画像を生成をし，学習データとした．本研究における音声感情認識の学習モデルとして ResNet-50[6]を使用した．ResNet-50 は，深層残差学習に基づく畳み込みニューラルネットワーク（CNN）であり，50 層の深い構造を持ちながらも，残差接続（residual connections）により勾配消失問題を効果的に回避し，高い表現能力と学習安定性を両立している．学習条件は表 2 に示す．

表 2　学習条件

データ数	E-Mix: 73828　CP:75000
元データ数	1320
テストデータ	各男女一組で合計120
モデル入力	Logメル周波数スペクトログラム画像
エポック数	50
バッチサイズ	32
学習率	10^{-4}
評価	正解率

5. 結果

拡張前と CopyPaste，EMix でそれぞれ拡張されたデータセットの学習を行い，その結果の感情ごとの正解率を混合行列にそれぞれ表 1，表 2，表 3 に示す．

5. 1 拡張前での分類結果

先行研究と比較したところ，log メル周波数スペクトログラムでの正解率が 58. 3%となり特徴量を変えたことで分類精度が低下してしまった．また特に「悲しみ」の感情が他の感情と比較して特に誤分類されやすい傾向が確認された．

表 3　拡張前

実際＼予測	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	1	0	0	7	0	0	0	0
穏やか	0	6	0	10	0	0	0	0
幸せ	0	0	2	6	1	0	0	7
悲しみ	1	3	0	11	1	0	0	0
怒り	0	0	0	0	14	0	0	2
恐怖	0	0	1	4	1	10	0	0
嫌悪	0	0	1	1	1	1	12	0
驚き	0	0	1	1	0	0	0	14

5. 2 CopyPaste での分類結果

拡張前の正解率と比べて 39.2%となり，大幅に低下した．中立や悲しみでは正しく分類された音声がなく，怒りや驚き悲しみなどこれまで比較的良好に分類できていた感情においても，正解率が軒並み低下した．嫌悪では唯一認識率が高いが，誤判定も多く全体的に嫌悪に偏っていることによりデータセット拡張による正解率の向上ができなかった．

表 4　CopyPate での拡張

実際＼予測	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	0	4	0	1	0	0	3	0
穏やか	0	8	0	3	0	0	5	0
幸せ	0	0	7	1	0	1	4	3
悲しみ	0	7	0	0	0	1	8	0
怒り	0	0	4	0	2	0	10	0
恐怖	0	2	5	0	0	6	3	0
嫌悪	0	1	0	0	0	0	15	0
驚き	0	0	4	0	0	0	3	9

5. 3　EMix での分類結果

拡張前と比べて，正解率が 65.83%となり，データセット拡張の有効性が示された．拡張前と比べて正解率が低下した感情があったが，中立と穏やかは正解率が 2 倍以上の上昇がみられた．しかし先行研究とくらべて拡張前のメル周波数ケプストラム係数を使用した場合の正解率よりは低くなることがわかった．

表 5 EMix での拡張

実際＼予測	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	7	0	0	0	0	1	0	0
穏やか	3	12	0	0	0	1	0	0
幸せ	1	0	6	2	0	1	0	6
悲しみ	0	4	2	4	0	0	6	0
怒り	0	0	0	0	12	0	1	3
恐怖	1	0	1	2	0	10	1	1
嫌悪	0	0	2	0	0	2	12	0
驚き	0	0	0	0	0	0	0	16

6. 考察

本実験で，先行研究より拡張前の正解率が低下してしまったのは特徴量を log メル周波数スペクトログラムに変更した影響と考えられる．これは，関連研究では log メル周波数スペクトログラムが多く使用されているが，データセット（RAVDESS）の感情特性や，利用するモデルとの相性によっては感情分析には必ずしも適しているとは言えないと考えられる．今後は特徴量をメル周波数ケプストラム係数に変更をして正解率の比較を行う．

また CopyPaste 手法において正解率が大幅に低下したのは，テストデータの構造が要因であると考える．この問題に対処するため学習データとテストデータの入力データ形式の統一を行った．学習データにおいて中立音声や同じ感情の音声が連結されることにより，音声データの長さが伸長する一方でテストデータは連結を行っておらず，元のデータ長のままで分類を行っていた．そこでテストデータにも CopyPste でのデータ生成を行なった．120 個あるテストデータの中，８つの感情音声データに N-CP，SE-CP を適用し，中立や同一の感情同士を連結させて新しいテストデータを作成した．分類の結果は表 6 に示す．

表 6 CP テストデータの結果

実際＼予測	中立	穏やか	幸せ	悲しみ	怒り	恐怖	嫌悪	驚き
中立	3	0	0	1	4	0	0	0
穏やか	1	7	0	5	0	3	0	0
幸せ	1	1	2	2	3	0	0	7
悲しみ	0	4	0	5	0	1	6	0
怒り	0	0	0	0	15	0	0	1
恐怖	0	0	1	4	0	10	1	0
嫌悪	0	4	0	0	0	0	12	0
驚き	0	0	0	0	0	0	0	16

分類の正解率は 58.3%となった．拡張前の分類結果と同等ではあるが，テストデータを適した形式にすることで Copypaste 手法での分類ができることがわかった．

正解率が大幅に低下した要因は ResNet50 が画像認識モデルであるため，音声のデータサイズや構造の一貫性に対して敏感である可能性が指摘される．図 3 のように無音区間やデータ長の違いで，スペクトログラムの視覚的な構造が根本的に異なるためテストデータに対して適切に特徴を分類するが困難であったため、結果として正解率の低下につながったと考えられる．Copypaste 手法での分類の正解率に、時間領域ではなく周波数領域での特徴量の使用や無音区間の削除，結合するデータとテストデータをパディングやリサンプリングを行いデータの形式の統一などをすることでどのような影響があるのかを検討する．また，画像認識モデルではなく LSTM や Transformer など時系列データに適したモデルを使用することで正解率が向上するかを検証する．

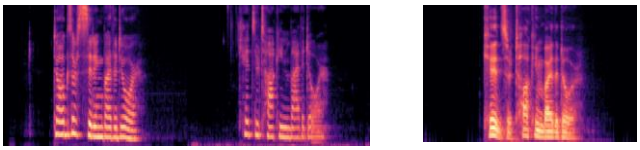

図 3 CP 手法での学習データ(左)とテストデータ(右)の構造

10. まとめ

本研究では，音声感情認識におけるデータセット拡張の有効性を検証するため，RAVDESS データセットを対象に CopyPaste (CP) と EMix の 2 種類のデータ拡張手法を適用し特徴量に log メルスペクトログラム，学習モデルに ResNet50 を使用して大規模なデータセットでの認識精度を検証した拡張前の感情の分類の正解率は 58.3%と，先行研究より低下し，モデルとの相性や特徴量などが適していなかった．また CopyPaste で約 57 倍に拡張したデータの正解率は 39.2%と大幅に低下した．中立や悲しみなどの感情で分類が著しく悪化し，全体的に嫌悪に偏る傾向が見られました．これは，学習データとテストデータ間の入力形式（データ長）の不整合，および ResNet50 モデルがデータの一貫性に敏感であることが原因と考えられ，追加実験でテストデータも CP で拡張し入力形式を統一した結果，正解率

は 58.3%と同等に回復し，適切な形式であれば CP での分類が可能であることが示された．今後は，特徴量の再検討や，ど時系列データに適したモデルの導入を行い，データセット拡張による音声感情認識の正解率の向上に目指していきたい．

参考文献

[1] 生形　優也，田村仁，　データセットの規模による音声感情認識の検討，情報処理学会　第 87 回全国大会講演論文 2 分冊，pp. 551-552(2025)

[2] Raghavendra Pappagari, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez and Najim Dehak: CopyPaste: An Augmentation Method for Speech Emotion Recognition, Proc. ICASSP 2021, pp.6324-6328, 2021.

[3] A. Dang, T. H. Vu, L. Dinh Nguyen and J. -C. Wang, "EMIX: A Data Augmentation Method for Speech Emotion Recognition," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096789

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” Language resources and evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[5] Houwei Cao, David G Cooper, Michael K Keut mann, Ruben C Gur, Ani Nenkova, and Ragini Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” IEEE transactions on affective computing, vol. 5, no. 4, pp. 377–390, 2014.

[6] 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