たたみ込みニューラルネットワークを用いた病的音声検知の検討*

☆石原一樹，荒井隆行（上智大）

1　はじめに

病的音声の検知は，様々な方法で行われてきており，近年，機械学習を用いたコンピュータによる自動検出が試みられている．本研究では，ドイツのザールラント大学がウェブ上で公開しているSaarbruecken Voice Database (SVD) の音声データを使用した分類を試みた．SVDは，健常音声が 687 人 (男性 259 人，女性 428 人)，病的音声が 1356 人 (男性 629 人，女性 727 人) 収録されており，合計 2000 人を超える音声データが記録されている．声の病気の症状は71種類あり，すべての音声データは，サンプリング周波数 50 kHz，量子化ビット数 16 bit である．各参加者の音声は,ピッチが 3 種類あり，それぞれ/a/，/i/，/u/の持続母音とドイツ語で“Guten Morgen, wie geht es Ihnen?” (おはようございます．お元気ですか？) の発話が記録されている[1].

SVDの音声データを使用し，機械学習による病的音声と健常音声を分類する研究が数多く行われている．これまで，MFCCやLPCCなどの音声特徴量を用いた研究が行われてきており，機械学習の手法として，サポートベクターマシンやニューラルネットワークが用いられていた[2][3][4][5][6]．また，近年，畳み込みニューラルネットワーク (CNN: convolutional neural network) が画像認識の分野において注目されるようになり，SVDに関する研究でも，音声データからスペクトログラムを生成し，それを画像認識により分類する研究が行われている[7][8][9][10]．表 1 は，先行研究の概要をまとめたものである．

本研究では，SVDの各音声サンプルをスペクトログラムに変換し，CNNのモデルであるEfficientNet によりスペクトログラムの画像データから健常音声と病的音声を分類した．EfficientNetは，Googleが2019年に発表したモデルであり，画像の解像度，モデルの深さ，モデルの幅をスケーリングし，効率的かつ高性能な分類を行うために開発されたモデルである．EfficientNetのモデルは，B0からB7までの 8 種類のモデルが提案されており，画像サイズや計算コストのバランスを考慮しながら使用できる[11]．本研究では，B0からB4までのモデルの分類性能の比較を行った．また，スペクトログラムの画像データを生成する際のパラメータの違いにより，分類結果に違いが生じるか検証した．図 1 は，本実験で使用した健常音声と病的音声のスペクトログラムの一例である．

表 1 先行研究の概要

論文	特徴量	機械学習の手法	サンプル数		正解率 (%)
			健常音声	病的音声	
[7]	スペクトログラム	CNN (VGG16, CaffeNet)	686	1616	93.90
[2]	MFCC	ANN，SVM	50	70	87.82
[3]	ComParEの特徴量, eGeMAPSの特徴量, MFCCなど	SVM	記載なし	679	82.80
[4]	MFCC, CPP	SVM	482	482	81.03
[5]	MFCC, LPCC, HOS	FNN，CNN	482	482	75.18
[8]	スペクトログラム	CNN (RESNET)	869	597	69.27
[6]	音声信号	DNN (LSTM)	686	1353	68.08
[9]	スペクトログラム	CNN，CDBN	482	482	68.00
[10]	スペクトログラム	CNN	482	482	66.20

* Detection of pathological speech using convolutional neural network, by Kazuki ISHIHARA, Takayuki ARAI (Sophia University).

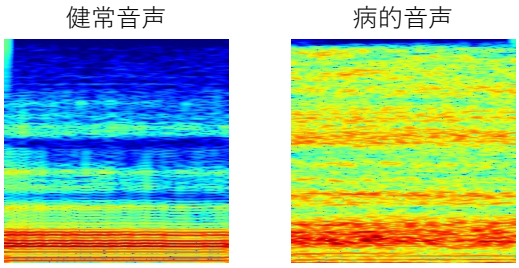

図 1 健常音声と病的音声の
スペクトログラム
(サンプリング周波数 16000 Hz, 窓の長さ 40 ms, フレームシフト幅 1.25 ms)

2 提案手法

2.1 音声データ

本研究では，病的音声として，先行研究[4][5][9][10]で使用されていた 6 つの症状に注目した. SVD には，喉頭炎 140 サンプル，白板症 41 サンプル，ラインケ浮腫 68 サンプル，反回神経麻痺 213 サンプル，声帯癌 22 サンプル，声帯ポリープ 45 サンプルがあり，1 つの音声ファイルで複数の症状を持つものもあった. 今回，病的音声を 482 サンプル，健常音声も同数の 482 サンプル使用した. すべての音声サンプルは，持続母音の/a/のみを使用し，サンプリング周波数を 50 kHz から 25 kHz と 16 kHz にそれぞれダウンサンプリングしたものを用意した.

2.2 スペクトログラムの生成

Python のライブラリである librosa を用いて，スペクトログラムを作成した. SVD の音声データは，サンプルごとに音声の長さが異なるため，各音声サンプルの真ん中部分の 400 ms を使用した. 窓の長さは 40 ms で固定し，フレームシフト幅を 20 ms と 1.25 ms の 2 種類を試した. また，スペクトログラムのカラーマップもカラー画像 (jet) と白黒画像 (gray) の 2 種類を作成し，カラーマップの違いにより分類性能に影響があるか検証した.

画像の拡大縮小を行う際，OpenCV を用いて，最近傍補間法とバイリニア補間法を試した. 画像のサイズは，それぞれの EfficientNet のモデルに合わせてリサイズした. EfficientNet のモデルごとの解像度を表 2 にまとめた.

表 2 EfficientNet のモデルと解像度

モデル	解像度
EfficienNetB0	224×224
EfficienNetB1	240×240
EfficienNetB2	260×260
EfficienNetB3	300×300
EfficienNetB4	380×380
EfficienNetB5	456×456
EfficienNetB6	528×528
EfficienNetB7	600×600

2.3 CNN のモデル

本研究では，ImageNet による事前学習を行った EfficientNet を使用し，EfficientNet のモデルの重みは事前学習のまま値を更新しないようにした. さらに，EfficientNet の構造に加えて，新たに 2 層の全結合層を追加した. 1 つは，ドロップアウト率 25%，ノード数 1024，活性化関数を ReLU とした. 最終層は，活性化関数を sigmoid 関数とし，出力が 2 種類となるようにした. CNN のモデル構築は，オープンソースで公開されている Keras および Tensorflow を用いて実装した. また，使用する際は，Google Colaboratory 上で，ランタイムのタイプを GPU に設定して使用した.

2.4 機械学習

本実験では，全体のデータのうちの 30%をテスト用データ，70%を訓練用データに分割した. また，訓練用の中で 20%を検証用データとし，ホールドアウト検証を行った. 学習はバッチサイズ 32 のミニバッチで行い，エポック数を 15 とした. モデルの最適化手法は Adam を使用し，学習率は 0.0001 とした.

3 実験方法と結果

3.1 実験 1

EfficientNet の B0 から B4 までのテスト用データに対する正解率をもとに，分類性能の比較を行った. スペクトログラムを生成する際，サンプリング周波数 16 kHz，窓の長さ 40 ms，フレームシフト幅 20 ms，カラーマップは jet のカラーのスペクトログラムのみを使用した. また，最近傍（nearest neighbor）補間法を用いて，画像のリサイズを行った.

表 3 は，これらの実験結果をまとめたものである．本研究のデータに対し，EfficientNetB3 の正解率が一番高い結果となった．

3.2 実験 2

実験 1 より，本研究では EfficientNetB3 が一番適していると推測された．そのため，実験 2 では，EfficientNetB3 を用いて，スペクトログラムを作成する際のパラメータを変化させながら，正解率を評価した．フレームシフト幅が, 20 ms と 1.25 ms の 2 種類の比較を行った．フレームシフト幅を 20 ms にした理由は，窓の長さ 40 ms の半分の値であるためである．また，フレームシフト幅を 1.25 ms にした理由は，時間軸の次元数を周波数軸の次元数と合わせるため設定した．さらに，スペクトログラムのカラーマップの違いによる分類性能の差も検証するため，カラー画像 (jet) と白黒画像 (gray)を用意した．画像をリサイズする際の補間方法は，最近傍補間法を用いた．

以上のパラメータを組み合わせて実験を行い，結果をまとめたものが表 4 である．テスト用データに対する正解率は，全体的にカラーマップがカラー (jet) のものよりも白黒 (gray) の方が高い結果となった．また，フレームシフト幅が短いほど正解率が高い傾向となった．一方で，今回の実験では，サンプリング周波数の違いによる分類性能の差はあまり見られなかった．

3.3 実験 3

実験 3 でも，実験 2 で使用した EfficientNetB3 を使用し，画像の補間法による差を検証した．今回の実験では，画像の補間方法として，バイリニア（bilinear）補間法を使用し，実験 2 で使用した最近傍補間法と比べて，分類性能に変化があるか検証した．その他のパラメータに関しては，実験 2 と同様にした．

表 3 EfficientNet のモデルごとのテストデータに対する正解率

モデル	サンプリング周波数 (Hz)	窓の長さ (ms)	フレームシフト幅 (ms)	補間方法	正解率 (%)
EfficienNetB0	16000	40	20	nearest	75.17
EfficienNetB1	16000	40	20	nearest	72.07
EfficienNetB2	16000	40	20	nearest	75.86
EfficienNetB3	16000	40	20	nearest	**76.55**
EfficienNetB4	16000	40	20	nearest	74.83

表 4 最近傍補間法を用いた際のテストデータに対する正解率

モデル	サンプリング周波数 (Hz)	窓の長さ (ms)	フレームシフト幅 (ms)	補間方法	スペクトログラムのカラーマップ	
					jet (カラー)	gray (白黒)
					正解率 (%)	正解率 (%)
EfficienNetB3	16000	40	1.25	nearest	74.83	**79.31**
EfficienNetB3	16000	40	20	nearest	76.55	78.62
EfficienNetB3	25000	40	1.25	nearest	**76.90**	78.62
EfficienNetB3	25000	40	20	nearest	74.48	73.45

表 5 バイリニア法を用いた際のテストデータに対する正解率

モデル	サンプリング周波数 (Hz)	窓の長さ (ms)	フレームシフト幅 (ms)	補間方法	スペクトログラムのカラーマップ	
					jet (カラー)	gray (白黒)
					正解率 (%)	正解率 (%)
EfficienNetB3	16000	40	1.25	bilinear	**77.93**	77.59
EfficienNetB3	16000	40	20	bilinear	75.86	78.28
EfficienNetB3	25000	40	1.25	bilinear	74.83	**78.97**
EfficienNetB3	25000	40	20	bilinear	74.48	77.24

表 5 は, 実験 3 の結果をまとめたものである. 実験2と実験3を比較した際, 画像の補間方法による大きな違いは見られなかった. また, 実験2と同様に, 実験3もスペクトログラムのカラーマップがカラー (jet) のものよりも白黒 (gray) の方が, 全体的に正解率が高かった. また, フレームシフト幅が短いものの方が, 比較的高い正解率であった.

4 考察

本実験では, 白黒画像のスペクトログラムを用いた際に, 最大で79.31%の正解率で病的音声と健常音声を判別できた. また, スペクトログラムのカラーマップが白黒画像のものの方が, 比較的高い分類が可能であることが示唆された. 逆に,サンプリング周波数の違いによる分類性能の差があまり見られなかった. また, 画像の拡大縮小における画像の補間方法の違いによる大きな差異も見られなかった.

5 おわりに

本研究では, カラーマップとしてjetとgrayのみしか使用しなかったが, 他のカラーマップを使用した際に, 分類性能に変化が生じるか検証したい. また, 本研究では病気の有無を 2 値分類したが, 多クラス分類の研究においてもスペクトログラムのカラーマップによる影響があるのか検証したい. 今後は, 病気の重症度を分類する研究も試みたい.

謝辞

本研究は, 上智大学重点領域研究の一部として助成を得た.

参考文献

[1] B. William, and P. Manfred, “Saarbrucken voice database,” Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.uni-saarland.de/, 2007.

[2] N. Souissi and A. Cherif, “Artificial neural networks and support vector machine for voice disorders identification,” Proc. International Journal of Advanced Computer Science and Applications (IJACSA), 7(5), 339–344, 2016.

[3] P. Barche, *et al.*, “Towards automatic assessment of voice disorders: A clinical approach,” Proc. INTERSPEECH, 2537–2541, 2020.

[4] 石原, 荒井, “ケプストラム法を用いた音声障害の有無の判定の試み”, 日本音響学会音声コミュニケーション研究会資料, 1(1), 41–45, SC-2021-8, 2021.

[5] J. Lee, and H.-J. Choi, “Deep learning approaches for pathological voice detection using heterogeneous parameters,” Proc. IEICE Trans. Inf. Syst., e103-d, 1920–1923, 2020.

[6] P. Harar *et al.* , “Voice pathology detection using deep learning: A preliminary study,” Proc. Int. Conf. Workshop Bioinspired Intell. (IWOBI), 1–4, 2017.

[7] M. Alhussein and G. Muhammad, “Voice pathology detection using deep learning on mobile healthcare framework,” IEEE Access, 6, 41034–41041, 2018.

[8] M. Huckvale and C. Buciuleac, “Automated detection of voice disorder in the Saarbrucken voice database: effects of pathology subset and audio materials,” Proc. INTERSPEECH, 1399–1402, 2021.

[9] H. Wu *et al.*, “A deep learning method for pathological voice detection using convolutional deep belief networks,” Proc. INTERSPEECH, 446–450, 2018.

[10] H. Wu *et al.,* “Convolutional neural networks for pathological voice detection,” Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), 1–4, 2018.

[11] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,’ Proc. International Conference on Machine Learning (ICML), 2019.