口部筋電位認識に対する深層学習パラメータの影響に関する研究

飯尾和司*　呉詩源*　朝倉義裕**

Study on the Influence of Deep Learning Parameters on Oral EMG Recognition

Kazushi IIO*　Shiyuan WU*　Yoshihiro ASAKURA**

ABSTRACT

The purpose of this study is to realize silent speech recognition using surface myoelectric potential. In the experiment, we measured the surface electromyogram on three points of the body; two of the orbicularis oris muscles and one of the zygomatic muscles. Learning was performed using a 5-layer convolutional neural network. We investigated the effect on validation accuracy by changing Filter, Kernel size and Pooling size. As a result, the validation accuracy was 84.6% when Filter was 32, the validation accuracy was 81.9% when Kernel size was 6, and the validation accuracy was 88.6% when Pooling size was 4.

Keywords : EMG, Deep learning, CNN, Hyperparameters

1. はじめに

自発的な発音が困難になった人のコミュニケーション方法として，手話や筆談，人口喉頭などの手法が存在する．しかし，現代の情報社会では，通話やビデオ通信などでのやり取りを通信ネットワークを介してリアルタイムで行うことが主流となってきており，対面を前提とした従来のコミュニケーション方法には限界がある．また，騒音が大きい，静寂を保つ必要があるなど，音声を介したコミュニケーションが取りにくい場面も想定される．

本研究では，発声による通話と同等のスムーズさを持つコミュニケーション手法として，表面筋電位を利用した無発声での音声認識の実現を目的として，筋電位データを入力としたニューラルネットワークで構築を行い，畳み込みニューラルネットワークを用いた学習のパラメータが識別精度に与える影響を「あ」から「こ」に限定し調査を行った．

2. 実験方法

2.1 表面筋電位の測定方法　表面筋電位計測には，筋電アンプ(OpenBCI 製 Cyton Biosensing Board (8-channels))を使用した．筋電位の導出には，塩化銀製皿電極と導電ペーストを介して皮膚表面に貼り付けることで行った．アンプで計測された筋電位データは，Bluetooth を介して PC へ転送し，数値データとして記録した．その時のサンプリング周波数は 250Hz である．

測定対象筋は図 1 に示すように，口輪筋上部，口輪筋下部，頬骨筋の 3 か所とした．口輪筋上部は口の突き出しの検出，頬骨筋は，口角の外方への引き上げの検出，口輪筋下部は口を開ける動作を検出するためこの 3 つの筋を測定対象とした．

測定する筋電位は動作学的筋電位であるため，双極導出法を用いて測定を行った．動作学的筋電位の場合は，2 つの電極を一定の間隔をあけて位置する．筋活動は筋の部位で変化するために，2 つの電極の電位差を記録する[1]．本研究での電極間隔は 1.5cm とした．

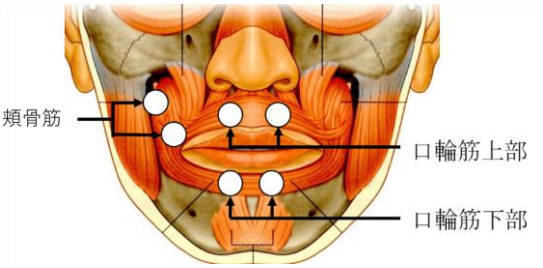


図 1　取得した筋．

*　機械工学科 5 年
**　機械工学科　准教授

被験者(20 代男性 3 名)に対して「あ」から「こ」までの 10 音を発声してもらった．電極を貼り付けた状態での発声に慣れてもらうため，実験前に 10 分の練習を行った[2]．発音時の筋電位データは単音ごとにとり，発声時間は 0.5 秒程度の短時間とし，前後が安定状態になるように 3 秒間 750 のデータを計 2400 個取得した．発声する音の順番は発生順による影響を減らすため，乱数によって決定した．取得したデータは平均を0，分散を1に正規化し，偏差の値が2を超えるものは外れ値とし，取り除いて学習データとした．その後，1660個の学習用データと，720個の検証用データに取り分けた．これら 10 音の筋電位を用いて，「あ」から「こ」までの 10 音の識別精度を調査し，母音の区別と子音の区別が可能であるか検証を行った．

2.2 ネットワーク構造および検証内容　本実験では図 2 に示す畳み込みニューラルネットワークを使用した．入力層には，±1 に正規化した筋電位の時系列データ×3 筋とし，出力層は「あ」～「こ」の 10 カテゴリの One-Hot encoding とした．このネットワーク構造[3]を基準とし，各パラメータを変更することによる認識精度の差の検証を行った．畳み込み層の Filter 数，Filter の Kernel size と，Pooling 層の Pooling size の 3 種類をそれぞれ変化させ学習を行った．

学習には，発声時の 1660 個の筋電位時系列データにラベル付けを行った教師あり学習とした．認識精度は，学習用とは別に用意した 720 個の検証用データに対する学習後のモデルの正答率を表す．学習後の認識精度と中間層の出力から，学習パラメータを評価した．ここにおける学習パラメータとは，層の数，層のサイズなどを表すハイパーパラメータを指す．パラメータの評価には，Filters 数，kernel size，Pooling size を変化させたときの認識精度で行った．

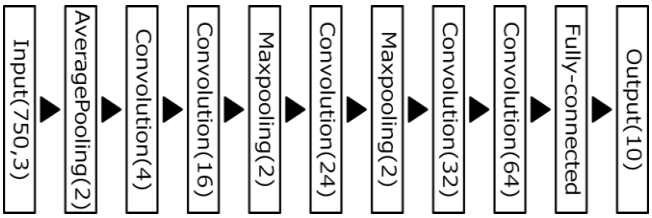


図 2　ネットワーク構造．

Filter 数は 1 つ目の畳み込み層の値を変化させ，その値を基準に 2 つ目以降の畳み込み層の値を増加させた．図 2 は，畳み込み層の Filter 数の基準の値を 4 とし，2 層目は 4 倍の 16，3 層目は 6 倍の 24，4 層目は 8 倍の 32，5 層目は 16 倍の 64 としたときのネットワーク構造を示す．実験では Filter の Kernel size を 6，Pooling size を 4 で固定し，この基準の Filter 数を 4，6，8，16，32，64 と変化させた．Kernel size では，Filter 数を 16，Pooling size を 4 で固定し，すべての畳み込み層で，2，4，6，8，10 と同じ値で変化させた．Pooling size では，Filter 数を 16，Kernel size を 6 で固定し，図中の Max pooling 層の値だけを 2，3，4，5，6 と変化させ検証を行った．それぞれ学習 Epoch 数 400 回まで計算を行った．この値は事前に何度か学習を行い，学習精度が一定になるまでの回数を調べ決定した．

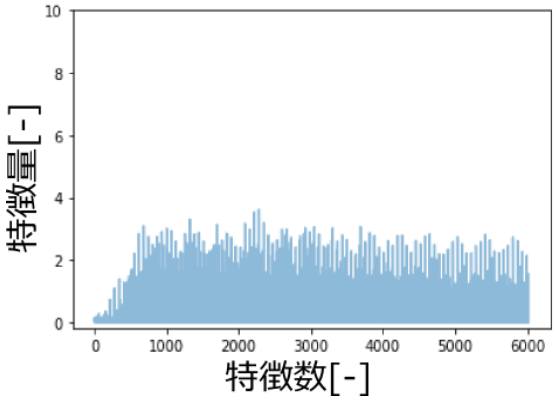
3. 結果及び考察

3.1 Filter 数の変更による影響　Filter 数を 4～64 まで変化させた場合の検証精度を表 1 に示す．Filter 数を増やすほどニューラルネットの表現力が増加することから，実験結果の精度もよくなる結果となったが，Filter 数が 16 よりも多くした場合，総合的な認識精度に大きな差異は見られなかった．この実験では最適な Filter 数が 64 であるという結果を得た．どの Filter 数においても「こ」と「く」の間での誤認識が多く，母音である「う」，「お」と誤認識する場合もあるため正答率が低くなる結果となった．

表 1　Filter 数と検証精度の関係．

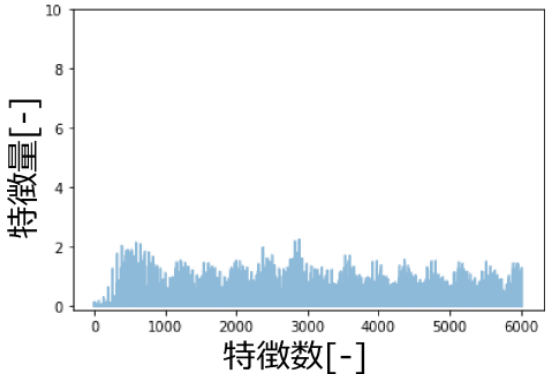
Filter数	あ	い	う	え	お	か	き	く	け	こ	計
4	74	53.8	71.9	50	68.3	61.9	94.4	54.1	64.9	55.9	64.6
6	79.2	76.9	67.2	61	83.3	55.6	98.6	64.9	77.9	58.8	72.4
8	75.3	76.9	89.1	62.2	73.3	71.4	98.6	63.5	75.3	70.6	75.4
16	85.7	75.6	92.2	67.1	80	61.9	98.6	66.2	76.6	67.6	77
32	85.7	78.2	90.6	68.3	81.7	63.5	98.6	66.2	79.2	61.8	77.3
64	79	79.2	78.5	77.4	83.3	77.5	95	62.8	83.1	70.2	78.2

Filter 数 4，32 のときの学習後の「あ」と「い」の特徴量の強さを表したグラフを図 3 に示す．ここでの特徴量とは，図 2 中の出力直前の全結合層の各ニューロンへ入力される値をすべて直線状に並べたものであり，前段のニューロンから全結合層の各ニューロンへ入力される数が特徴数である．Filter 数を多くすると，前段のニューロン数が多くなり，全結合層で取得する特徴数が多くなる．このため，ニューラルネットで表現できる形が多くなる．図 3 で「あ」と「い」の場合を比較すると，Filter 数が 4 のときは表現力が低く「あ」と「い」がほとんど同じ形となっている．対して，Filter 数を増やし 32 にした場合は「あ」は特徴が強く，「い」は特徴となる部分が弱いということがわかる．

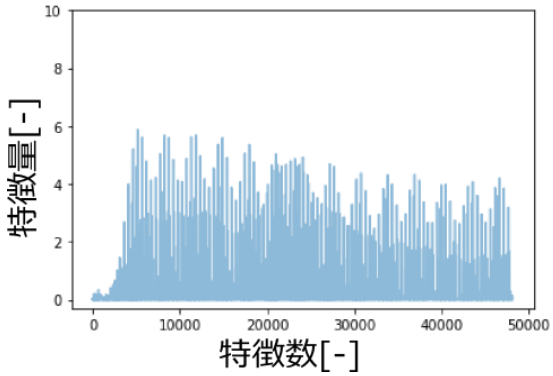


(a)「あ」の Filter 数 4 のとき．

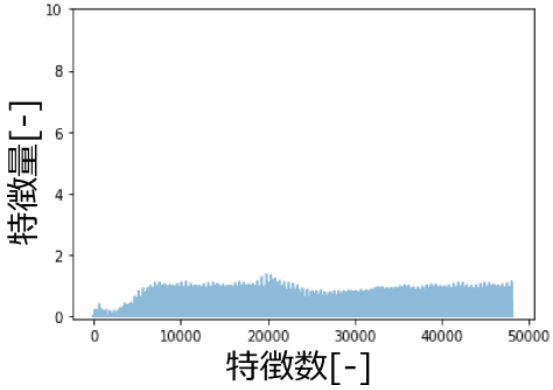
図 3　Filter 数と特徴量の関係



(b) 「い」の Filter 数 4 のとき.



(c) 「あ」の Filter 数 32 のとき.



(d) 「い」の Filter 数 32 のとき.

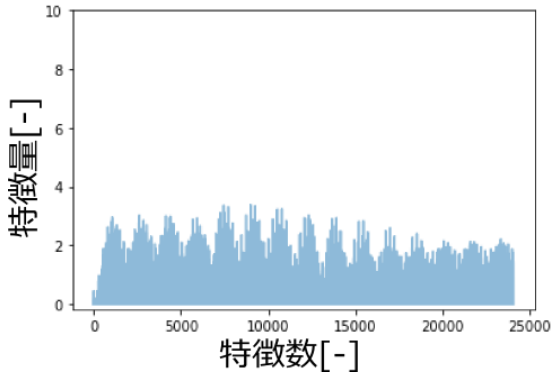
図 3 Filter 数と特徴量の関係(つづき)
Kernel size = 6, Pooling = 2.

3.2 Kernel size の変更による影響 Kernel size を 2～10 まで変化させた場合の検証精度を表 2 に示す. Kernel size を増やすことで取得することのできる周波数帯が大きくなるため, 本実験においては精度がよくなると考えられるが, Kernel size が 4 以上のときの検証精度に大きな差は出なかった. 音ごとに見ると Filter 数を変えた時に比べ子音の正答率はあまり変わらなかったが母音の正答率が少し高くなる結果となった.

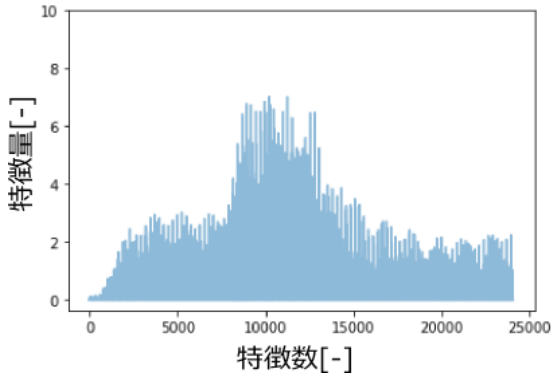
Kernel size が 2, 6, 10 のときの「あ」の特徴量の強さを表したものを図 4 に示す. 図 4 より Kernel size を増やすことで計算に使用するデータ幅が大きくなり, 低周波数まで表現できるようになる. そのため, 開口時に入力される低周波数の波形が特徴として現れ, 2 に比べ 4 以上は検証精度が上がったと考えられる.

表 2 Kernel size と検証精度の関係.

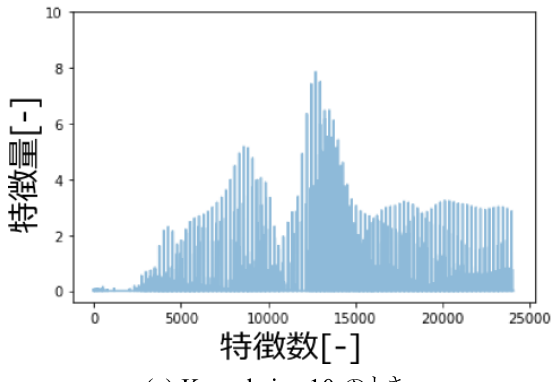
Kernel size	あ	い	う	え	お	か	き	く	け	こ	計
2	84.4	78.2	84.4	65.9	76.7	66.7	98.6	71.6	74	61.8	76.2
4	85.7	80.8	90.6	80.5	83.3	73	100	74.3	87	67.6	82.4
6	89.6	78.2	93.8	68.3	90	79.4	100	85.1	81.8	63.2	82.6
8	80.5	78.2	89.1	63.4	86.7	76.2	100	74.3	71.4	64.7	78
10	92.2	80.8	87.5	73.2	88.3	69.8	100	71.6	83.1	61.8	80.8



(a) Kernel size 2 のとき.



(b) Kernel size 6 のとき.



(c) Kernel size 10 のとき.

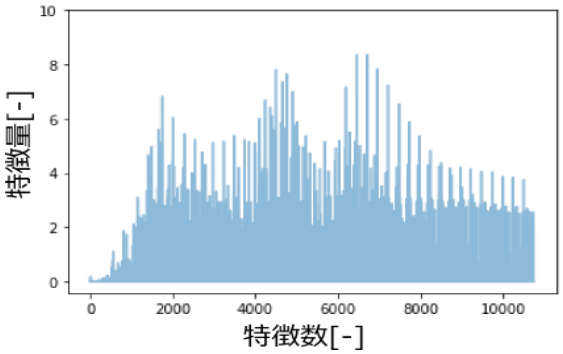
図 4 Kernel size と特徴量の関係
Filter 数 = 16, Pooling = 2.

3.3 Pooling size の変更による影響　Pooling size2～6 まで変化させた場合の検証精度を表 3 に示す. Pooling size を大きくすると, 特徴が不鮮明になり汎化性能が上がる. 結果として, Pooling size が 4 のときに最適値となり, 5 以上では下がる傾向になった. Pooling size が 4 のとき子音が母音と誤認識することが減り, 正答率が上がる結果となった.

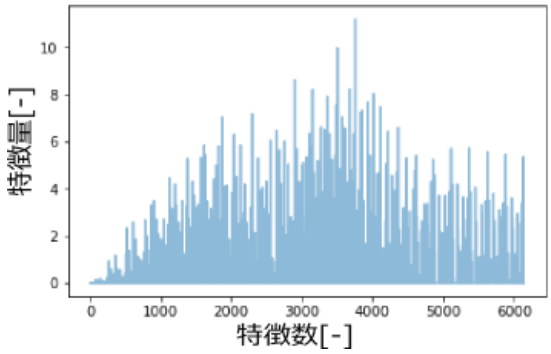
表 3　Pooling size と検証精度の関係.

Pooling size	あ	い	う	え	お	か	き	く	け	こ	計
2	89.6	78.2	93.8	68.3	90	79.4	100	85.1	81.8	63.2	82.6
3	84.4	76.9	90.6	70.7	78.3	79.4	100	77	81.8	63.2	80.1
4	93.5	92.3	95.3	75.6	88.3	85.7	100	79.7	85.7	77.9	87.3
5	87	84.6	84.4	74.4	81.7	71.4	98.6	78.4	80.5	60.3	80.3
6	92.2	83.3	89.1	74.4	86.7	87.3	98.6	81.1	80.5	64.7	83.6

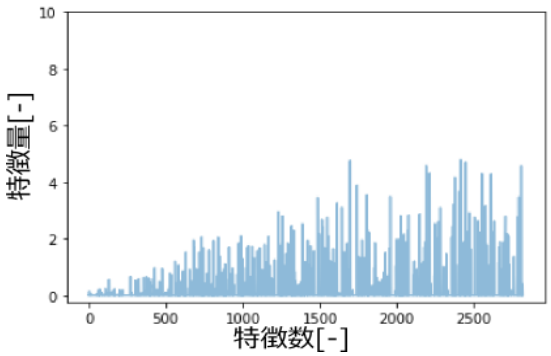
Pooling size 3, 4, 6 での学習後の「あ」の特徴量の強さを表したグラフを図 5 に示す. 図 5 より Pooling size を大きくすることでグラフの形がぼやけたものとなっていることがわかる. Pooling size が 3 のとき特徴数が多く Pooling size を 4, 6 と大きくすると特徴数は少なくなっている. 特徴数が多いと波形の特徴が鮮明に残ってしまい, 特徴数が少なすぎると波形の特徴が不鮮明になり判断に影響が出て精度が落ちたと考えられる.



(a) Pooling size 3 のとき.



(b) Pooling size 4 のとき.



(c) Pooling size 6 のとき.

図 5　Pooling size と検証精度の関係
Filter 数 = 16, Kernel size = 6.

4. まとめ

本実験では, 顔の表面筋電位を用いた無発声での音声認識の実現のための, 前実験としてネットワークのハイパーパラメータの変化によるデータへの影響の確認と適正値の調査を行った. 3 つの実験を通して, 子音の識別率は母音に比べて少し低い結果となった. また, 誤認識した対象は似た口の形となるものが多く今回取得した 3 つの筋では, 同じ母音の音の区別は難しいと考えられる. そのため, 音数を増やした場合, 精度が低くなると考えられる. 母音の特徴づけた一人の発声における母音の識別率は約 95%[4] であるため, 高い識別率だと考えられる. しかし, 今回の実験では同じ条件でとった筋電位データを学習用と検証用に分けて実験したため, 張り直しや発声中の電極のずれによる誤差が乗った, 未学習のデータに対しては有効性を示すことができなかった. そのため, 誤差の影響を減らすために, より多くの学習データの取得を進めるとともに, 誤差の影響を抑えるためのデータの処理方法の検討が課題となる.

参考文献

(1) 鈴木　俊明, 谷　万喜子:「筋電図からわかること 臨床で筋電図をどう生かすか」, 関西理学　17: 1–2, 2017.
(2) 福本　尚生, 倉富　勲, 古川　達也, 相知　政司, 伊藤　秀昭, 和久屋　寛:「無発声母音認識のための訓練システム」, 計測自動制御学会論文集　Vol.49, No.12, 1106/1112 (2013)
(3) 岡谷　貴之 :「機械学習プロフェッショナルシリーズ/深層学習」, 講談社, 2015.
(4) 福田　修, 藤田　真治, 辻　敏夫 :「EMG 信号を利用した代用発声システム」, 電子情報通信学会論文誌 Vol. J88-D-II No.1, 2005/1