Vol.2021-HCI-191 No.11
2021/1/29

筋電位によるCNNを用いた黙声の音素セグメント情報の推定

宮坂 清貴[1,a)]　坂本 雄児[1,b)]

概要：騒音環境など声を出せない環境で，あるいは、声をうまく出せない人が音声コミュニケーションをとるための手法として，音以外の情報から発話動作を認識する方法（黙声認識）がある．黙声時の顔の筋電位図 (Electromygram:EMG) から黙声認識する手法は微小な動きから認識が可能である他，発話する単語の認識が高い精度で可能である．しかし，発話する単語の長さなどの情報が失われてしまうため表現が乏しくなる．そこで，本研究では，推定した単語情報と EMG から黙声する単語の音素セグメントを CNN を用いて推定し，既存のシステムによって生成した音声に長さの情報を加えることにより，EMG から長さの情報の加わった音声を生成する手法を提案する．EMG から音素セグメントを推定する実験の結果，提案手法を用いることにより，音声に長さの情報を加えることが可能であることがわかった．

キーワード：EMG，黙声認識，CNN，筋電位

1. はじめに

会話などの音声を使ったコミュニケーションは誰でも，直感的に情報を伝えることができるため，非常に便利なコミュニケーション方法である．一方で，音声を自由に使用できない状況も存在する．秘匿性が求められる内容や，電車の中などの騒音環境，公共の場などの静かさが求められる環境など，声を出せない環境では音声を使用することはできない．また，喉頭がんなどが原因で声をうまく出せない人も音声を使用できない．音以外の情報から発話動作を認識する方法（黙声認識）がある．黙声認識の方法には口の映像から発話動作を認識する方法や，口内の超音波画像から，認識する方法，脳波や顔の筋電位図（Electromygram:EMG）などの生体情報から認識する方法などがある．なかでも EMG から認識する手法は，EMG が実際に筋肉が動くよりも少し前に発生するため，微小な動きから認識が可能である他，リアルタイム認識の可能性があるため，黙声認識によるコミュニケーションに向いている．現在まで，EMG による黙声認識の研究では，90%以上の精度で単語の分類を行うことができている．しかし，現在の方法でコミュニケーションを行うと，長さなどの発話の表現の情報が失われてしまう問題があり，EMG と発話の長さとの関係については研究が行われていない．本研究では．EMG から長さの情報を取り出すことを目的として．CNN を用いた長さの推定方法と，学習データの測定，処理方法を提案する．この研究が進めば，実際には声をださずに好きな声で表現豊かなコミュニケーションができることが期待される．

[1] 北海道大学
Hokkaido University
[a)] miyasaka_meijin@ist.hokudai.ac.jp
[b)] yuji@ist.hokudai.ac.jp

2. 提案システムの概要

EMG から音声を生成するために推定を 2 段階にわけて行う．１段階目は単語の推定を行う部分であり，十分な量の学習データを用意すれば高い精度で分類が可能であることがわかっている．2 段階目は音節の長さを推定する部分である．本研究では 2 段階目について研究を行う．これによって推定された単語情報を音節アライメント情報から既存の音声生成の手法を用いて様々な声の音声を出力することができる．

3. 学習に使用するデータの測定

音節の長さを学習するために EMG と対となる音声の長さの情報が必要となる．そのために，学習に使用する単語を声をだし発話し，その際の顔の EMG の測定と音声の録音を同時に行うことにより，EMG と音声の対のデータを得た．本研究では以下の 11 の単語と一部の音節の長さが異なる単語の EMG と音声の測定を行った．

測定は読み上げるテキストと音声が再生される動画を実

ⓒ 2021 Information Processing Society of Japan

験者がみながら行った．再生されるテキストと音声連続で 2 回ずつ再生され，1 回目は発声せず，2 回目の音声にあわせて実験者が発声を行う．EMG の測定にはミユキ技研の PolymateⅡ 使用し，図の 6 箇所の EMG の測定をサンプリング周波数 1kHz で行った．また，再生する音声には Google の Text-to-Speech を使用した．

4. 音節アライメントの推定

4.1 前処理

発話する時間は最大 1500ms であるが，EMG は発話の前にも口の動きに伴い発声するため，発声前 500ms を含めた 2000ms ぶんの EMG データを切り出す．さらに， EMG はノイズ除去のため 30Hz のローパスフィルタをかける．音声データは教師データとなる形式に変換する．教師データには音節アライメントの位置を使用する．音節アライメントの推定には julius を使用した．例えば/warui/の場合/wa/と/ru/と/i/のそれぞれの始まりと終わりの時間が教師データとなる．教師データの大きさは/wa/の終わりと/ru/の始まりが同じになるため (音節の数+1) 個の数のデータの大きさとなる．

4.2 データオーグメンテーション

学習データに起こりうる変化を加え，新たな学習データとし，データ量を増やして学習に使用する方法がデータオーグメンテーションである．どのような変化が音節の長さの情報の推定に有効かを調べるため，2 種類のオーグメンテーションを行った．1 つ目は **magnitude-warping(WM)** で，EMG に以下のような変化を加え，新たなデータ EMG' を得る．

$$EMG'[t] = EMG[t] * (0.1 \sin(2\pi f t) + 0.9)$$

$$f = \frac{1}{f_s + rand(0, f_s/20)}$$

ここで，t:時間，f_s:サンプリング周波数，rand(a,b):a から b-1 の範囲の乱数，である．

2 つ目は **time-shift(TS)** で，EMG に時間方向に以下のような変化を加え，新たなデータ EMG" を得る．

$$EMG'[t] = EMG[t + rand(-f_s/40, f_s/40)]$$

4.3 学習

音節アライメントの EMG による推定に使用する CNN は 3 層の畳み込み層と，2 層の全結合層からなる．畳み込み層は活性化関数にランプ関数と 20%の dropout を適用している．全結合層では活性化関数に線形関数を適用している．損失関数は平均二乗誤差を用いて adam で最適化を行った．1 種類の単語のデータのうち 20 個は評価用に学習には使用せず，残りのデータのみを学習に使用し，評価を行った．オーグメンテーションしない学習では 100epoch，オーグメンテーションでは 20epoch 学習を行った．

4.4 複数の種類の音節アライメントの推定

あまりできていないが誤差は少ない MW がよい

5. 音の長さによる分類

音節セグメンテーションの推測は音節セグメンテーションの位置に着目すると誤差が非常に小さい単語もあったが，音節を伸ばすかどうかの分類問題としてみた場合，正答率が低くなったしまった．そこで，次に分類問題の正答率に着目し，実験を行った．分類を行うために音節の長さの異なる単語を，別の単語として考え，EMG から，どの音節の伸ばし方なのかという分類を学習させた．このとき，前処理やオーグメンテーションなどは同じであるが，CNN の構造が異なり，出力されるのは分類する種類の大きさのラベルデータとなる．

5.1 学習

長さの異なる単語の分類の EMG による推定に使用する CNN は 3 層の畳み込み層と，2 層の全結合層からなる．最終層以外は活性化関数にランプ関数と 20%の dropout を適用している．全結合層では活性化関数に softmax 関数を適用している．損失関数はクロスエントロピー関数を用いて adam で最適化を行った．

5.2 複数の種類の音節アライメントの推定

分類による方法のほうが複数の分類の正答率は良くなった

6. 2 種類の長さの異なる単語の分類

一度に複数の種類の音節の伸ばし方をする，場合を扱ったが，伸ばす音節によって分類しやすさが，大きく異なることもわかった．また，日本語として不自然な伸ばし方のものも多く扱っていた．実際には，自然な音節の伸ばし方が 2 種類の単語は多い．そこで，2 種類の単語の分類をそれぞれ行い，どのような音節を伸ばした場合，分類しやすいのかを調べた．

7. 考察

長音同士の分類は難しい．短音と長音符 2 個分の長音は精度高く分類できた．るい，かい，ない，のような自然と口が大きく動く単語は精度高く分類できた． EMG のオーグメンテーションでは，オーグメンテーションをしない場合よりもほとんどの場合，magnitude-warping を行った場合の方が良い結果となった．一方で，時間方向に変化を加えた場合はオーグメンテーションをしない場合よりも悪い結果となった．音節の長さのように

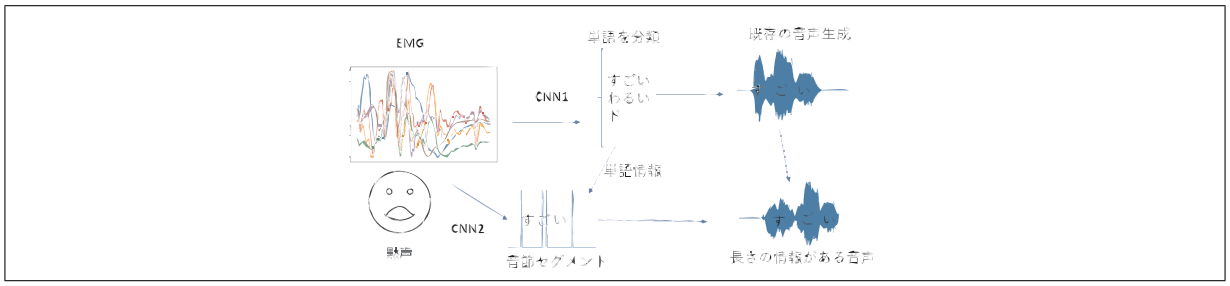


図 1 システムの概要

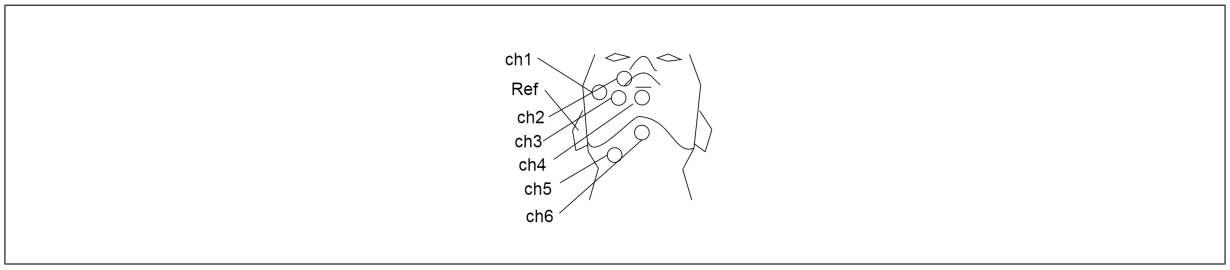


図 2 電極位置

8. まとめ

実験の結果三角関数を用いた magnitude-warping はすくない周波数や変化量で同じ位相の三角関数しか用いなかったが，どのような与える変化の量を増やすことによって，より，学習効率が良くなる可能性がある．一部の音節の組み合わせは音節アライメントと EMG との関係性が高いことがわかった．より，

参考文献

[1] Yannis,M.,Brendan ,S. ,, Shimon ,W. and Nando ,F.: *LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING*, ICLR (2017).
[2] rnav ,K. ,Shreyas ,K. ,Pattie ,M.: *AlterEgo: A Personalized Wearable Silent Speech Interface*, ACM IUI (2018).
[3] Naoki ,K. , Michinari ,K., Jun ,R.: *SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks*, ACM CHI (2019).
[4] Terry ,T., Franz ,M., Daniel ,P. , Satoshi ,E. , Muriel ,L. , Sandra ,H., Urban ,M., Dana ,K.: *Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks*, ICMI (2017).
[5] Diederik,P. Kingma,J.: *Adam: A Method for Stochastic Optimization*, ICLR (2015).
[6] 大内 慶久: *Hybrid HMM/SVM* を利用した筋電位に基づく日本語黙声認識, 筑波大学修士 (情報学) 学位論文 (2013).
[7] Gopala, K., Anumanchipalli, Josh, C and Edward, F.: *Speech synthesis from neural decoding of spoken sentences*, Nature volume 568, pages493–498(2019).
[8] Ian, J., Goodfellow,D., Warde-Farley,M. ,Aaron, C. and Yoshua, B.: *Maxout Networks*, ICML(2013).
[9] 福 見 稔: 深層学習を用いる口唇 *EMG* によるサイレント音声認識と個人認証, 電気学会・電子情報システム 部門大会 (2018)
[10] Lopez-Larraz,E., Mozos,O. M. , Antelis,J. M. and Minguez,J.: i*yllable-based speech recognition using EMG*, Proc. Ann. Int. Conf. Eng. Med. Biol. Soc., pp. 4699-4702(2010).

ⓒ 2021 Information Processing Society of Japan

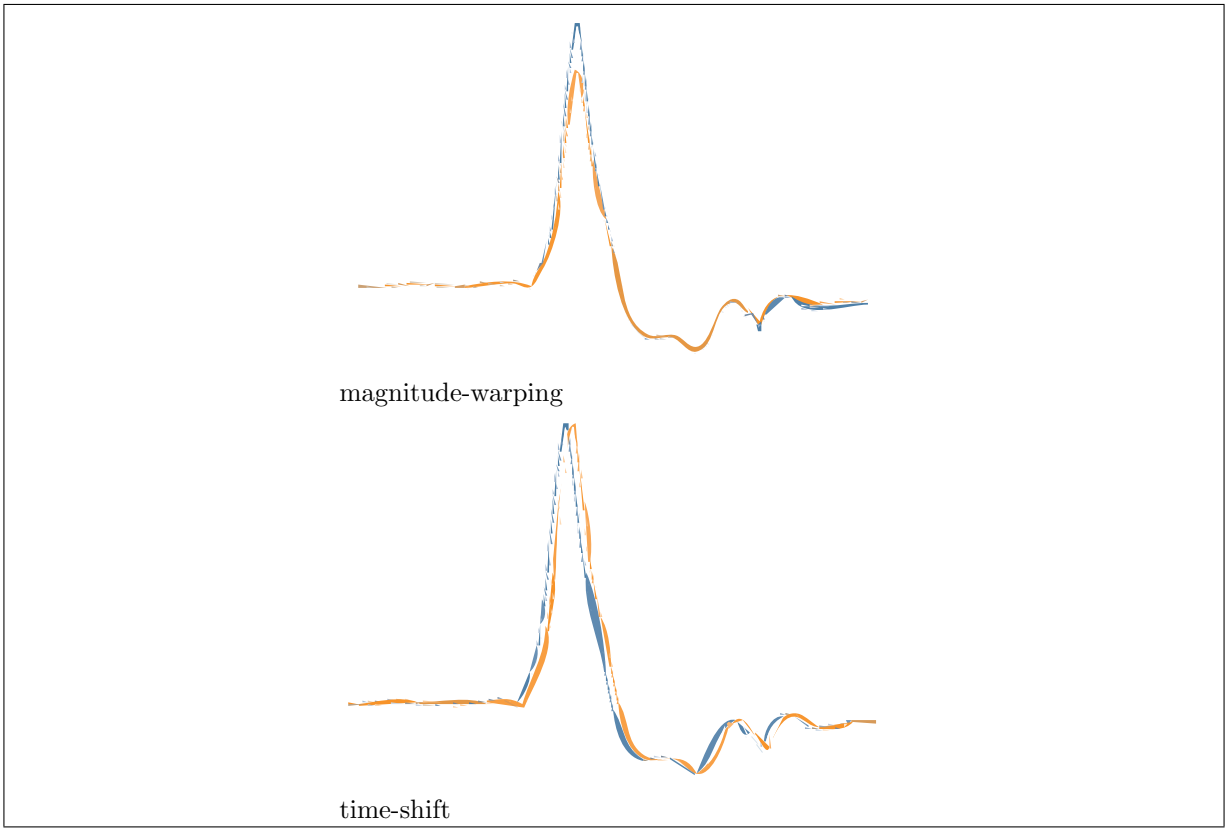


図 3 オーグメンテーションの例

基本の単語	長さの異なる単語		
わるい	わるーい	わーるい	
どんな	どーんな	どんーな	
たかい	たかーい	たかーーい	たかーーーい
ない	なーい	なーーい	なーーーい
ド	ドー	ドーー	ドーーー
レ	レー	レーー	レーーー
ミ	ミー	ミーー	ミーーー

図 4 使用した単語一覧

単語	音節セグメント予測による分類			ラベル予測による分類		
	nothing	MW	TS	nothing	MW	TS
わるい	35	75	55	65	75	30
どんな	60	50	30	75	70	55
たかい	40	55	30	40	45	30
ない	35	35	30	45	50	55
ド	40	40	25	60	65	70
レ	55	60	45	80	65	40
ミ	25	50	25	60	40	25

図 5 単語の複数分類の正答率 [%]

ⓒ 2021 Information Processing Society of Japan

単語1	単語2	音節セグメント予測による分類	ラベル予測による分類
		MW	MW
わるい	わるーい	95	95
わるい	わーるい	60	60
どんな	どーんな	75	80
どんな	どんーな	65	50
たかい	たかーい	55	65
たかい	たかーーい	80	90
ない	なーい	75	75
ない	なーーい	85	85
ド	ドー	80	85
ド	ドーー	85	85
ドー	ドーーー	55	65
レ	レー	65	80
レ	レーー	85	90
レー	レーーー	90	95
ミ	ミー	75	65
ミ	ミーー	55	90
ミー	ミーーー	65	75

図 6 2 種類分類の正答率 [%]

ⓒ 2021 Information Processing Society of Japan

正誤表

本稿の提出原稿のバージョンミスが発覚しました．筆者の不注意を深くお詫び申し上げますとともに，本稿の訂正をさせて頂きます．また，訂正後の原稿についても掲載させて頂きます．

該当箇所	誤	正
p1　2 章	EMG から音声を生成するために推定を 2 段階にわけて行う．	図 1 のように EMG から音声を生成するために推定を 2 段階にわけて行う
P1　3 章	本研究では以下の 11 の単語と一部の音節の長さが異なる単語の EMG と音声の測定を行った．	本研究では図 4 の 7 の単語と一部の音節の長さが異なる単語の EMG と音声の測定を行った．ここで，「ー」はおよそ 300ms の長音符である．
P2　3 章	図の 6 箇所の EMG の測定をサンプリング周波数 1kHz で行った	図 2 の 6 箇所の EMG の測定をサンプリング周波数 1kHz で行った
P2　4.2 最後に追加		例として図 3 に EMG オーグメンテーションをした例を示す．
P2　5.2	分類による方法のほうが複数の分類の正答率は良くなった	図 5 にラベル予測による分類の正答率を示している．ラベル予測による分類のオーグメンテーション方法の比較では TS は音節アライメント予測による方法と同様の傾向がみられたが，オーグメンテーションを行わない方法の正答率が 1 つを除いて向上した．MW による方法でにおいても 1 つの単語のみ正答率が下がるという結果となった
P2　6 章	どのような音節を伸ばした場合，分類しやすいのかを調べた．	その結果を図 6 に示す．図 6 は単語 1 と単語 2 での分類を行った結果である．オーグメンテーション方法は代表として MW を行った結果を示している． 「わるい」と「わるーい」の分類では 95%であるのに対し「わるい」と「わーるい」では，60%の結果と，伸ばす音節によって，分類しやすさが大きく変わることが分かった．75%以上分類できた単語は以下である．•「わるい」「わるーい」•「どんな」「どーんな」•「ない」「なーい」•長音符 2 個以上の差のある単語
P2　7 章	長音同士の分類は難しい．短音と長音符 2 個分の長音は精度高く分類できた．るい，かい，ない，のような自然と口が大きく動く単語は精度高く分類できた． EMG のオーグメンテーションでは，オーグメンテーションをしない場合よりもほとんどの場合，magnitude-warping を行った場合の方が良い結果となった．一方で，時間方向に変化を加えた場合はオーグメンテーションをしない場合よりも悪い結果となった．音節の長さのように	類がよくできた音節の母音に着目すると，「るーい」，「なーい」などの，単語は母音の組み合わせが日本語では発声時舌が前後に動く．一方では「わーる」ではどちらの母音の下を後ろで発音する．このように，母音の組み合わせごとの舌や口の動きにより，分類しやすさが変わるという仮説が考えられる． 長さの比較では，ラベル予測による長音符の数が 0 と 1 の分類の場合と 0 と 2 の分類の場合の比較では平均すると 74%と 88%と，伸ばす長さが増えることによって高

		い正答率となった．一方で長音符の数が 1 と 3 のでは平均 78%と，音を伸ばす同士の場合はそうでない場合と比べて低い結果となった．予測方法による正答率を比較すると，大きな差は認められなかったが，1 つの単語を除いてラベル予測による分類が音節セグメント予測による分類以上の結果となった
P3　8 章	実験の結果三角関数を用いた magnitude-warping はすくない周波数や変化量で同じ位相の三角関数しか用いなかったが，どのような与える変化の量を増やすことによって，より，学習効率が良くなる可能性がある．一部の音節の組み合わせは音節アライメントと EMG との関係性が高いことがわかった．より，	本研究では EMG による音節の長さの分類は長音符がない場合と長音符が 2 個の場合（およそ 600ms の長さ）で，85%以上の精度で分類が可能であるという結果が得られた．また，母音の組み合わせについても，大きく口や舌を動かすものが精度が高くなった．精度が低い組み合わせについても，意識して大きく動かすことによって，正答率が改善される可能性がある．予測方法による比較ではラベル予測による分類のほうが良い結果となった．しかし，大きな差ではなく，ラベル予測による方法では，常に 2 種類や 3 種類などの決まった長さの音声しか出力できないという問題があるため，学習データを増やすなど精度を上がると，音節アライメントによる予測方法が良くなる．オーグメンテーション方法の比較では三角関数を用いた magnitude-warping による方法が良い結果となった．本研究ではすくない周波数や変化量で同じ位相の三角関数しか用いなかったが，どのような与える変化の量を増やすことによって，より，学習効率が良くなる可能性がある．EMG による長さ情報の付与は可能であることがわかったが課題として，分類精度を向上させるために，データ測定方法の工夫や，よりよいオーグメンテーション方法の研究，どのような音節が分類しやすいのかという研究が必要である．

筋電位によるCNNを用いた黙声の音素セグメント情報の推定

宮坂 清貴[1,a)]　坂本 雄児[1,b)]

概要：騒音環境など声を出せない環境で，あるいは、声をうまく出せない人が音声コミュニケーションをとるための手法として，音以外の情報から発話動作を認識する方法（黙声認識）がある．黙声時の顔の筋電位図 (Electromygram:EMG) から黙声認識する手法は微小な動きから認識が可能である他，発話する単語の認識が高い精度で可能である．しかし，発話する単語の長さなどの情報が失われてしまうため表現が乏しくなる．そこで，本研究では，推定した単語情報と EMG から黙声する単語の音素セグメントを CNN を用いて推定し，既存のシステムによって生成した音声に長さの情報を加えることにより，EMG から長さの情報の加わった音声を生成する手法を提案する．EMG から音素セグメントを推定する実験の結果，提案手法を用いることにより，音声に長さの情報を加えることが可能であることがわかった．

キーワード：EMG，黙声認識，CNN，筋電位

1. はじめに

会話などの音声を使ったコミュニケーションは誰でも，直感的に情報を伝えることができるため，非常に便利なコミュニケーション方法である．一方で，音声を自由に使用できない状況も存在する．秘匿性が求められる内容や，電車の中などの騒音環境，公共の場などの静かさが求められる環境など，声を出せない環境では音声を使用することはできない．また，喉頭がんなどが原因で声をうまく出せない人も音声を使用できない．音以外の情報から発話動作を認識する方法（黙声認識）がある．黙声認識の方法には口の映像から発話動作を認識する方法 [1] や，口内の超音波画像から認識する方法 [2]，脳波や顔の筋電位図（Electromygram:EMG）などの生体情報から認識する方法 [3], [4] などがある．なかでも EMG から認識する手法は，EMG が実際に筋肉が動くよりも少し前に発生するため，微小な動きから認識が可能である他，リアルタイム認識の可能性があるため，黙声認識によるコミュニケーションに向いている [5], [6]．現在まで，EMG による黙声認識の研究では，90%以上の精度で単語の分類を行うことができている [7]．しかし，現在の方法でコミュニケーションを行うと，長さなどの発話の表現の情報が失われてしまう問題があり，EMG と発話の長さとの関係については研究が行われていない．本研究では．EMG から長さの情報を取り出すことを目的として．CNN を用いた長さの推定方法と，学習データの測定，処理方法を提案する．この研究が進めば，実際には声をださずに好きな声で表現豊かなコミュニケーションができることが期待される．

[1] 北海道大学
Hokkaido University
[a)] miyasaka_meijin@ist.hokudai.ac.jp
[b)] yuji@ist.hokudai.ac.jp

2. 提案システムの概要

図 1 のように EMG から音声を生成するために推定を 2 段階にわけて行う．1 段階目は単語の推定を行う部分であり，十分な量の学習データを用意すれば高い精度で分類が可能であることがわかっている．2 段階目は音節の長さを推定する部分である．本研究では 2 段階目について研究を行う．これによって推定された単語情報を音節アライメント情報から既存の音声生成の手法を用いて様々な声の音声を出力することができる．

3. 学習に使用するデータの測定

音節の長さを学習するために EMG と対となる音声の長さの情報が必要となる．そのために，学習に使用する単語を 40 回ずつ声をだし発話し，その際の顔の EMG の測定と音声の録音を同時に行うことにより，EMG と音声の対のデータを得た．本研究では**表 1** の 7 の単語と一部の音節の長さが異なる単語の EMG と音声の測定を行った．ここで，「ー」はおよそ 300ms の長音符である．

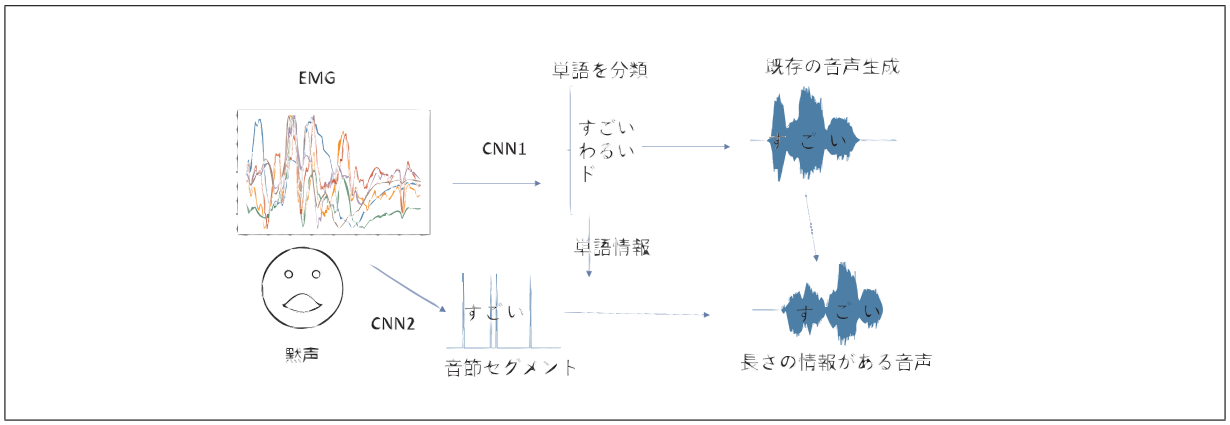


図 1　システムの概要

表 1　使用した単語一覧

基本の単語	長さの異なる単語		
わるい	わるーい	わーるい	
どんな	どーんな	どんーな	
たかい	たかーい	たかーーい	たかーーーい
ない	なーい	なーーい	なーーーい
ド	ドー	ドーー	ドーーー
レ	レー	レーー	レーーー
ミ	ミー	ミーー	ミーーー

測定は**図 3** のように読み上げるテキストと音声が再生される動画を実験者がみながら行った．再生されるテキストと音声連続で 2 回ずつ再生され，1 回目は発声せず，2 回目の音声にあわせて実験者が発声を行う．EMG の測定にはミユキ技研の PolymateⅡ 使用し，**図 2** の 6 箇所の EMG の測定をサンプリング周波数 1kHz で行った．また，再生する音声には Google の Text-to-Speech[8] を使用した．

4.　音節アライメントの推定

4.1　前処理

発話する時間は最大 1500ms であるが，EMG は発話の前にも口の動きに伴い発声するため，発声前 500ms を含めた 2000ms ぶんの EMG データを切り出す．さらに， EMG はノイズ除去のため 30Hz のローパスフィルタをかける．音声データは教師データとなる形式に変換する．教師データには音節アライメントの位置を使用する．音節アライメントの推定には julius[9] を使用した．例えば/warui/の場合/wa/と/ru/と/i/のそれぞれの始まりと終わりの時間が教師データとなる．教師データの大きさは/wa/の終わりと/ru/の始まりが同じになるため (音節の数+1) 個の数のデータの大きさとなる．

4.2　データオーグメンテーション

学習データに起こりうる変化を加え，新たな学習データとし，データ量を増やして学習に使用する方法がデータオーグメンテーションである [10]．どのような変化が音節の長さの情報の推定に有効かを調べるため，2 種類のオーグメンテーションを行った．1 つ目は **magnitude-warping(WM)** で，EMG に以下のような変化を加え，新たなデータ EMG’ を得る．

$$EMG'[t] = EMG[t] * (0.1 \sin(2\pi f t) + 0.9)$$

$$f = \frac{1}{f_s + rand(0, f_s/20)}$$

ここで，t:時間，f_s:サンプリング周波数，rand(a,b):a から b-1 の範囲の乱数，である．

2 つ目は **time-shift(TS)** で，EMG に時間方向に以下のような変化を加え，新たなデータ EMG” を得る．

$$EMG'[t] = EMG[t + rand(-f_s/40, f_s/40)]$$

例として**図 4** に EMG オーグメンテーションをした例を示す．

4.3　学習

音節アライメントの EMG による推定に使用する CNN は 3 層の畳み込み層と，2 層の全結合層からなる．畳み込み層は活性化関数にランプ関数と 20%の dropout[11] を適用している．全結合層では活性化関数に線形関数を適用

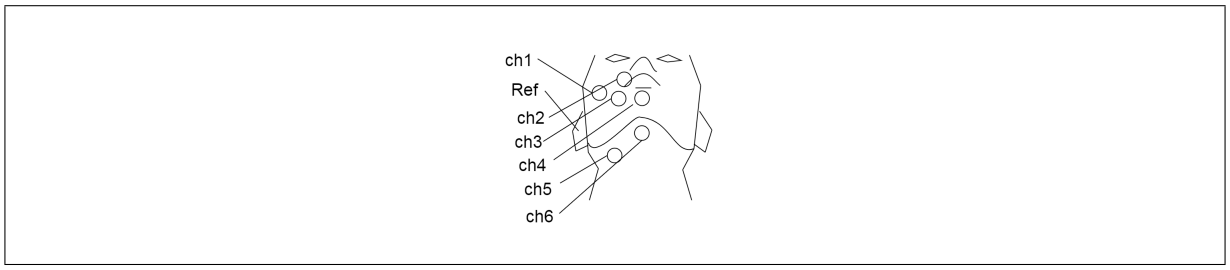


図 2　電極位置

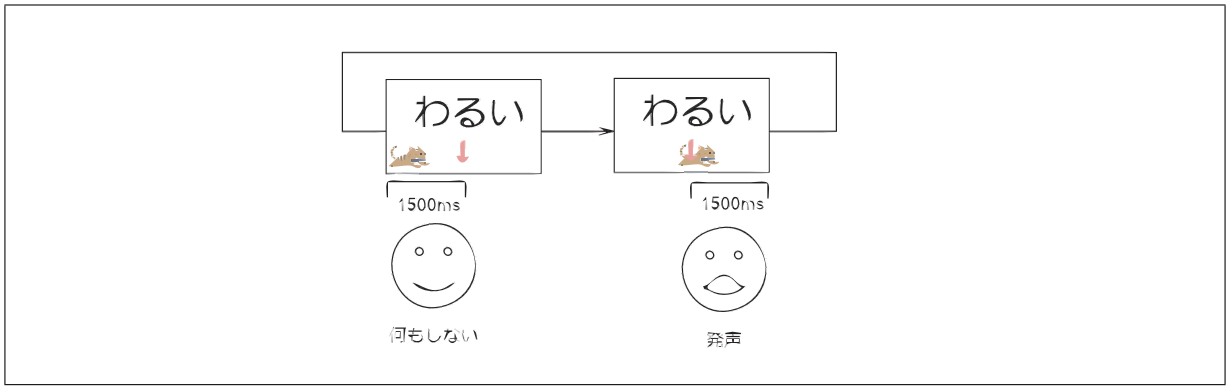


図 3　EMG の測定

している．損失関数は平均二乗誤差を用いて adam[12] で最適化を行った．1 種類の単語のデータのうち 20 個は評価用に学習には使用せず，残りのデータのみを学習に使用し，評価を行った．オーグメンテーションしない学習では 100epoch，オーグメンテーションでは 20epoch 学習を行った．

4.4　複数の種類の音節アライメントの推定

表 2 は 3 種類以上の音節の長さの単語の分類が 3 種類のオーグメンテーション方法でできた確率を示している．「わるい」と「たかい」，は表 1 に示すように 3 種類の分類で，その他の単語は 4 種類の分類である．音節アライメントの推定による方法では，音節アライメントの位置を予測したあと，学習に使用したデータの中から予測したアライメントの位置が最も近い単語を予測する音節の長さの単語とみなして正答率を求めている．

オーグメンテーション方法を比較すると，「どんな」以外の場合で，オーグメンテーションを行わない場合よりも MW によるオーグメンテーションを用いた方法が良い結果となった．一方で，時間方向に干渉する TS によるオーグメンテーションは「わるい」以外の単語でオーグメンテーションを行わない場合よりも悪い結果となった．

分類を全く行えていない場合，3 種類の単語では 33%，4 種類の場合は 25%となるが，WM による方法ではいずれもその数値を大きく上回っており，正答率が課題ではあるが，EMG による音節の長さの分類が可能であることがわかる．

5.　ラベル予測による分類

音節アライメントの推測は，音節アライメントの位置を最小にする学習であるが，分類問題としてみた場合正答率には課題があった．そこで，次に分類問題としての正答率を最小にする方法で，実験を行った．分類を行うために音節の長さの異なる単語を，別の単語として考え，EMG から，どの音節の伸ばし方なのかという分類を学習させた．このとき，前処理やオーグメンテーションなどは同じであるが，CNN の構造が異なり，出力されるのは分類する種類の大きさのラベルデータとなる．

5.1　学習

長さの異なる単語の分類の EMG による推定に使用する CNN は 3 層の畳み込み層と，2 層の全結合層からなる．最終層以外は活性化関数にランプ関数と 20%の dropout を適用している．全結合層では活性化関数に softmax 関数を適用している．損失関数はクロスエントロピー関数を用いて adam で最適化を行った．

5.2　複数種類のラベル予測による分類

表 2 にラベル予測による分類の正答率を示している．ラベル予測による分類のオーグメンテーション方法の比較では TS は音節アライメント予測による方法と同様の傾向がみられたが，オーグメンテーションを行わない方法の正答率が 1 つを除いて向上した．MW による方法でにおいても 1 つの単語のみ正答率が下がるという結果となった．

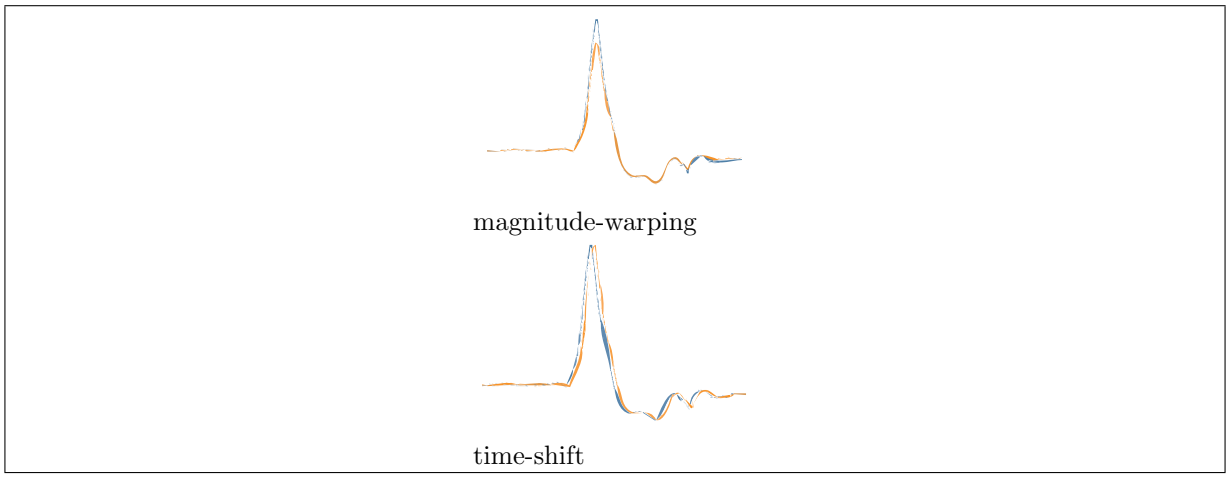


図 4　オーグメンテーションの例)

表 2　単語の複数分類の正答率 [%]

単語	音節セグメント予測による分類			ラベル予測による分類		
	nothing	MW	TS	nothing	MW	TS
わるい	35	75	55	65	75	30
どんな	60	50	30	75	70	55
たかい	40	55	30	40	45	30
ない	35	35	30	45	50	55
ド	40	40	25	60	65	70
レ	55	60	45	80	65	40
ミ	25	50	25	60	40	25

6.　2 種類の長さの異なる単語の分類

一度に複数の種類の音節の伸ばし方をする，場合を扱ったが，伸ばす音節や長さによって分類しやすさが，異なる可能性も考えられる．また，日本語として不自然な伸ばし方のものも多く扱っていた．実際には，自然な音節の伸ばし方が 2 種類の単語は多い．そこで，2 種類の単語の分類をそれぞれ行い，どのような音節を伸ばした場合，分類しやすいのかを調べた．その結果を**表 3** に示す．表 3 は単語 1 と単語 2 での分類を行った結果である．オーグメンテーション方法は代表として MW を行った結果を示している．

「わるい」と「わるーい」の分類では 95%であるのに対し「わるい」と「わーるい」では，60%の結果と，伸ばす音節によって，分類しやすさが大きく変わることが分かった．75%以上分類できた単語は以下である．

- 「わるい」「わるーい」
- 「どんな」「どーんな」
- 「ない」「なーい」
- 長音符 2 個以上の差のある単語

分類がよくできた音節の母音に着目すると，「るーい」，「なーい」などの，単語は母音の組み合わせが日本語では発声時舌が前後に動く．一方では「わーる」ではどちらの母音の下を後ろで発音する．このように，母音の組み合わせごとの舌や口の動きにより，分類しやすさが変わるという仮説が考えられる．　長さの比較では，ラベル予測による長音符の数が 0 と 1 の分類の場合と 0 と 2 の分類の場合の比較では平均すると 74%と 88%と，伸ばす長さが増えることによって高い正答率となった．一方で長音符の数が 1 と 3 のでは平均 78%と，音を伸ばす同士の場合はそうでない場合と比べて低い結果となった．

予測方法による正答率を比較すると，大きな差は認められなかったが，1 つの単語を除いてラベル予測による分類が音節セグメント予測による分類以上の結果となった．

7.　まとめ

本研究では EMG による音節の長さの分類は長音符がない場合と長音符が 2 個の場合（およそ 600ms の長さ）で，85%以上の精度で分類が可能であるという結果が得られた．また，母音の組み合わせについても，大きく口や舌を動かすものが精度が高くなった．精度が低い組み合わせについ

表 3 2 種類分類の正答率 [%]

単語1	単語2	音節セグメント予測による分類	ラベル予測による分類
		MW	MW
わるい	わるーい	95	95
わるい	わーるい	60	60
どんな	どーんな	75	80
どんな	どんーな	65	50
たかい	たかーい	55	65
たかい	たかーーい	80	90
ない	なーい	75	75
ない	なーーい	85	85
ド	ドー	80	85
ド	ドーー	85	85
ドー	ドーーー	55	65
レ	レー	65	80
レ	レーー	85	90
レー	レーーー	90	95
ミ	ミー	75	65
ミ	ミーー	55	90
ミー	ミーーー	65	75

ても，意識して大きく動かすことによって，正答率が改善される可能性がある．

予測方法による比較ではラベル予測による分類のほうが良い結果となった．しかし，大きな差ではなく，ラベル予測による方法では，常に 2 種類や 3 種類などの決まった長さの音声しか出力できないという問題があるため，学習データを増やすなど精度を上がると，音節アライメントによる予測方法が良くなる．

オーグメンテーション方法の比較では三角関数を用いた magnitude-warping による方法が良い結果となった．本研究でははすくない周波数や変化量で同じ位相の三角関数しか用いなかったが，どのような与える変化の量を増やすことによって，より，学習効率が良くなる可能性がある．

EMG による長さ情報の付与は可能であることがわかったが課題として，分類精度を向上させるために，データ測定方法の工夫や，よりよいオーグメンテーション方法の研究，どのような音節が分類しやすいのかという研究が必要である．

参考文献

[1] Yannis,M.,Brendan ,S. , Shimon ,W. and Nando ,F.: *LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING*, ICLR (2017).

[2] Naoki ,K. , Michinari ,K., Jun ,R.: *SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks*, ACM CHI (2019).

[3] Gopala, K., Anumanchipalli, Josh, C and Edward, F.: *Speech synthesis from neural decoding of spoken sentences*, Nature volume 568, pages493–498(2019).

[4] Lopez-Larraz,E., Mozos,O. M. , Antelis,J. M. and Minguez,J.: i*yllable-based speech recognition using EMG*, Proc. Ann. Int. Conf. Eng. Med. Biol. Soc., pp. 4699-4702(2010).

[5] 大内 慶久: *Hybrid HMM/SVM* を利用した筋電位に基づく日本語黙声認識, 筑波大学修士 (情報学) 学位論文 (2013).

[6] 福 見 稔: 深層学習を用いる口唇 *EMG* によるサイレント音声認識と個人認証, 電気学会・電子情報システム 部門大会 (2018)

[7] rnav ,K. ,Shreyas ,K. ,Pattie ,M.: *AlterEgo: A Personalized Wearable Silent Speech Interface*, ACM IUI (2018).

[8] Google: Text-to-Speech: 自然な音声合成 Google Cloud, 入 手 先 ⟨https://cloud.google.com/text-to-speech⟩, (2020.11.25).

[9] Lee, A. and Kawahara, T.: *Recent Development of Open-Source Speech Recognition Engine Julius*, APSIPA ASC (2009).

[10] Terry ,T., Franz ,M., Daniel ,P. , Satoshi ,E. , Muriel ,L. , Sandra ,H., Urban ,M., Dana ,K.: *Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks*, ICMI (2017).

[11] Ian, J., Goodfellow,D., Warde-Farley,M. ,Aaron, C. and Yoshua, B.: *Maxout Networks*, ICML(2013).

[12] Diederik,P. Kingma,J.: *Adam: A Method for Stochastic Optimization*, ICLR (2015).