E-023

口裂周辺の筋電信号を用いた少数語彙世界における黙声単語認識

Inaudible Word Recognition on Small Lexical Set using Myoelectric Signals around a Mouth

永井 秀利† 谷口 竜太郎† 副島 小波† 中村 貞吾† 野村 浩郷†
Hidetoshi Nagai Ryotaro Taniguchi Konami Soejima Teigo Nakamura Hirosato Nomura

1. はじめに

我々は，声を出さずに発声 (いわゆる口パク) された内容を口裂周辺から頸部にかけての位置で体表から測定可能な筋電情報に基づいて認識することを目指している．我々はこれを筋電による黙声認識と呼ぶ．本技術は，他者に騒音で迷惑をかけず盗み聞きもされない音声入力や大音量下等のマイク入力困難時の音声認識の支援，発声と黙声との切替えにより対話と操作とをシームレスにした音声インターフェース，喉頭切除で声を失った人の発声代行など，多数の応用が考えられる．

従来の研究 [1] では，口裂周辺の 4 筋を計測対象とした日本語母音認識実験において，無理に大きく動かすことをしないような自然な口の動きに対して，発声開始から 100ms ずつの 3 区間の筋電情報を用いて約 87%の認識精度を得た．

日本語母音は口唇形状に特徴があるため口裂周辺の筋によってかなりのレベルでの認識が可能と言える．しかし子音の場合には，口唇形状の変化に加えて舌位置も重要な要素となるため，母音認識を目的とした筋のみでは十分ではない．とはいえ，口腔内にある舌の位置や形状を左右するすべての筋の電位を表面筋電によって計測することはほぼ不可能である．従来の研究 [2] にて，舌根の挙上・下制に関わる筋の一部を計測して子音認識に役立てる試みも行っているが，まだ子音特徴を明確に得るには至っていない．

しかしながら少数単語世界であれば，現在の日本語母音認識手法を援用することにより，ある程度は満足できる精度で認識できる可能性がある．そこで本稿では，従来の結果を踏まえ，特性の異なる 2 種の少数語彙世界において，母音認識の場合と同様の筋電信号を用いて単語認識を行った結果について述べる．

2. 筋電計測位置

口裂周辺の筋電情報に基づいた日本語母音認識を行う他の研究 [3], [4] においては，口輪筋，大頬骨筋，顎二腹筋の 3 筋を計測対象としている．大頬骨筋の計測は「い」や「え」の発声において口角を後方に引く動作を捉えようとしたものと考えることができるが，我々の従来の研究では，かなり大きく口を動かさない限りは大頬骨筋に有効な筋電波形を観測することができなかった．文献 [3] や [4] において，被験者に口唇形状をはっきりさせる発声法の訓練を要求している点からもこの問題[1] の存在は明らかと言えよう．よって日常的な自然な口の動きに対しての認識では大頬骨筋は適切とは言えず，我々は口輪筋，口角下制筋，下唇下制筋[2]，顎二腹筋の 4 筋 (表 1) を計測対象とすることを提案した [1]．

表 1: 計測対象とする筋およびその機能

筋名	機能
口輪筋	口唇の縮小，収縮，突出
口角下制筋	口角を外下方に引く
下唇下制筋	下唇の引き下げ
顎二腹筋	舌骨挙上または下顎骨引き下げ

この 4 筋を用いる場合,「い」と「え」の間や「う」と「お」の間の識別能力がやや弱い．対策としてオトガイ筋を用いることの有効性が指摘されている [6] が，実験機材の都合により 4 チャネル以下の同時計測しかできず，4 筋のいずれかを除外すると全体としての認識率が大きく低下するため，この 4 筋を用いて実験を行うこととした．

なお，英語を対象とした関連研究では，上唇の形状形成に機能する筋を含めた 7 箇所ないし 6 箇所を計測するもの [7],[8] や，口裂周辺は用いずに下顎から頸部の筋のみを対象とするもの [9] などが存在する．日本語と英語による違いというのもあるとは思うが，我々は日常的に軽く発声する場合においては上唇挙上のような動作はさほど顕著ではないと考えており，前者で挙げられた計測位置の一部は重視していない．後者の場合，頸部のみでは体表から計測できる情報は限られており，少数語彙世界を越えて適用することは極めて困難であると考える．ただし，従来の研究 [2] でも試みたように，頸部から得られる可能性がある情報には認識の際に重要となりうるものが含まれるため，より多チャネルの計測機器を利用できるようになれば同時計測対象に含みたい位置である．

3. 黙声認識手法

本稿では，我々の従来の研究で孤立母音の認識を行った際に用いた手法に準じて筋電波形計測および認識実験を行った．

従来の研究では，4 チャネルの生体計測器を使用し，図 1 の位置に Ag-AgCl 皿電極を貼り付けて計測を行った．電極は，装着位置の皮膚をアルコールで清拭した後，導電性ペーストを用いて装着した．計測は，軽く口唇を閉じた状態 (安静状態) で筋電信号が安定した状態から始め，1 母音を発生後に軽く口唇を閉じた状態に戻すという過程で行った．筋電波形データは，解像度 12bit，周期 50μs(20000Hz) でサンプリングして 1 母音発声 (測定時間 2 秒間) ごとに獲得した．

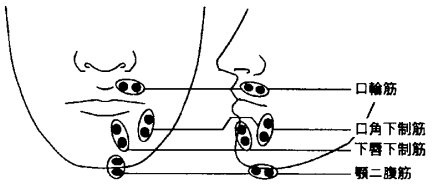

図 1: 表面筋電計測用電極の貼付位置の概略

本稿においても，1 母音の代わりに 1 語を発声するようにして，同じ計測位置，同じ手順，同じ発声過程で筋

†九州工業大学, Kyushu Institute of Technology

[1]もちろん，訓練によって認識に都合が良い発声を行い，精度を向上させることは十分に意義のあることである．だが試してみるとわかるが，それを長時間安定して続けるのはかなりの苦痛と困難とを伴うため，あまり実用的とは言えない．

[2]文献 [5] では母音ではなく子音判別を目的として計測している．

電波形データを獲得した．

獲得した生のデータは多くのノイズを含むため，ウェーブレット縮退を利用したノイズ低減手法 [10] を適用し，さらにチャネル間感度差および発声強度差[3] に関する正規化 [1] を行ったものを実験用データとした．

認識パラメータの抽出に際しては，いずれかのチャネルの電位が設定した閾値を越えた時点を発声開始時と推定し，その位置を基準として区間を切り出すこととする．

我々の従来での孤立母音認識の際の切り出し区間は，発声開始時の口唇形状形成動作と発声時の筋の定常活動との特徴を捉えることを目指して発声開始時から 100ms ずつの 3 区間とした．その区間の電位の絶対値の総和を認識パラメータ (計 12 個) とし，3 階層のフィードフォワード型ニューラルネットワークに与えて学習させた．

本稿で行った孤立単語認識の場合には，発声区間全体を認識パラメータ化する必要があるために区間の切り出し方が異なることになるが，電位の絶対値の総和を認識パラメータとしてフィードフォワード型ニューラルネットワークで学習させる点は同様とした．

4. 少数語彙世界での単語認識実験

4.1 実験対象とした少数語彙世界

本稿では 2 種類の少数語彙世界を実験対象とする．一つはビデオカメラの操作[4]を想定した語彙集合であり，もう一つは「ぜろ」から「きゅう」までの孤立数字である．

前者は，語を構成する母音系列の特徴差が比較的大きい集合であり，子音をうまく捉えることができなくても十分な認識精度が得られることを期待しうる世界である．それに対し，後者は，語長が短く母音系列の特徴差も乏しい集合であり，子音の特徴差をある程度以上捉えることができなければ認識精度を期待することができないと考えられる世界である．

我々は，それぞれの世界の特性に応じてパラメータ抽出方法を定め，孤立単語の認識を試みた．

従来の研究 [10] において連続発声から個々の母音の切り出しの試みを行ってはいるが，まだ手法として確立されたものではない．そのため今回は，一つの語の発声全体の波形データを単純に与えることとしたした．

単純に発声全体の波形データを用いる場合，一般的には発声速度のばらつきの影響を避けるための時間軸方向の正規化が必要であろう．しかし今回は，できるだけ同じペースで発声する[5] ようにして，時間軸方向の正規化は行わないこととした．

なお，日本語の孤立数字に関する他の研究 [12] では，1 語につき 2〜5 秒と，日常的な発声に比べて非常にゆっくりと発声したデータに基づいて認識を試みている．それに対し，本研究では，ゆっくりめではあるものの日常的な速度で発声したものを認識対象としていることを注記しておく．

4.2 ビデオカメラ操作用語彙の認識

認識対象は「ズーム」,「ワイド」,「明るく」,「暗く」,「録画」,「停止」,「逆光オン/オフ」,「ライトオン/オフ」,「手ぶれオン/オフ」の 12 語とし，各語 26 回で計 312 個の波形データを獲得した．

今回獲得したデータでは，4 筋のいずれかの電位が閾値を越えた時点を発声開始とみなした場合，最も発声時間が長い「逆光オン/オフ」でも発声開始からおよそ 1300ms 以内には発声を終了していた．母音系列を捉えるという点では，従来と同様の 100ms 単位の区間切り出しで十分に特徴を得られると考えたため，発声開始から 1500ms までを 100ms ごとに切出して電位の絶対値の総和を求めたものを認識パラメータとした．

データは，各語 2 個ずつからなる 24 個のデータを 1 セットとするように分割し，13 個のデータセットを作成した．これらのデータセットに対し，選択した 1 セットをテストデータとし，残りを学習データとする認識実験をすべてのデータセットについて行った (計 13 回)．認識には，入力層 60 ユニット，中間層 100 ユニット，出力層 12 ユニットの 3 階層構造のニューラルネットワークを用いた．結果を表 2 および表 3 に示す．

表 2: ビデオカメラ操作用語彙の認識精度

単語	学習データ	テストデータ
全体	98.1%	87.2%
ズーム	99.1%	84.6%
ワイド	92.3%	80.8%
明るく	98.5%	92.3%
暗く	99.4%	84.6%
録画	98.8%	96.2%
停止	100.0%	100.0%
逆光オン	98.8%	88.5%
逆光オフ	99.1%	88.5%
ライトオン	98.5%	80.8%
ライトオフ	97.8%	88.5%
手ぶれオン	97.8%	80.8%
手ぶれオフ	96.6%	80.8%

実験結果における誤認傾向を調べると,「オン」,「オフ」を互いに誤認しているケースが多く見られた．これは，自然な発声の場合には末尾の母音「う」をあまりはっきりとは発声しないケースが多いことが原因であろう．直前の母音「お」の発声が強いこともあり，正規化された際に弱い発声の特徴差が不明確になったものと考える．

4.3 孤立数字の認識

認識対象は,「ぜろ」,「いち」,「に」,「さん」,「よん」,「ご」,「ろく」,「なな」,「はち」,「きゅう」の 10 語とし，各語 18 回で計 180 個の波形データを獲得した．

この語集合の場合，従来の 100ms 区間での切出しでは，例えば「いち」と「に」の母音系列「いい」と「い」との区別ができなくなるなどの可能性が高いと予測されるため，より短い区間での切出しを行った．細分によって特徴が薄れることの回避も考慮し，発声開始から 1000ms までを 40ms 幅で 20ms ずつずらしながら切出して認識パラメータを得た．

データセットは，各語 2 個ずつからなる 20 個のデータを 1 セットとし，9 個のデータセットを作成した．これらの内，1 セットをテストデータ，残りを学習データとして認識実験を行った (計 9 回)．

まずは従来と同じく 3 階層のニューラルネットワーク (入力層 196 ユニット，中間層 500 ユニット，出力層 10 ユニット) を用いて実験した結果，認識精度が非常に低く (テストデータ全体で 39.5%)，一部の語に誤認される傾向が強いなど語ごとの認識精度のばらつきが極端 (テストデータで 10%から 75%) であり，識別器としての能力が不足であると判断された．

そこで入力層 196 ユニット，中間層 250，250，50 ユニットの 3 層，出力層 10 ユニットの 5 階層のニューラ

[3] 発声の強度と筋電信号の強さとには相関が見受けられる (文献 [11]) ため，この正規化は発声の特徴を一部捨てていることになる．

[4] 頭部固定カメラを黙声で操作することにより，操作音声の混入なく両手を自由に使えるという状況を想定している．

[5] 現在の計測環境では筋電波形の走査を目視しながら発声するため，同じペースを維持することは比較的容易である．

表 3: ビデオカメラ操作用語彙の認識結果

正解	分類結果											
	A	B	C	D	E	F	G	H	I	J	K	L
ズーム (A)	22		1	1		1		1				
ワイド (B)	1	21		2			1	1				
明るく (C)			24	1		1						
暗く (D)	1	1	2	22								
録画 (E)					25						1	
停止 (F)						26						
逆光オン (G)			1				23	2				
逆光オフ (H)							2	23		1		
ライトオン (I)		1	1				1		21	2		
ライトオフ (J)		2						1		23		
手ぶれオン (K)		1									21	4
手ぶれオフ (L)							1				4	21
合計	24	26	29	26	25	28	28	28	21	26	26	25

表 4: 孤立数字の認識精度

数字	学習データ	テストデータ
全体	59.1%	47.7%
ぜろ	60.0%	55.5%
いち	75.7%	55.5%
に	62.9%	66.6%
さん	52.9%	16.6%
よん	50.0%	38.8%
ご	60.0%	55.5%
ろく	48.6%	55.5%
なな	61.4%	50.0%
はち	58.6%	44.4%
きゅう	61.4%	38.8%

表 5: 孤立数字の認識結果

正解	分類結果									
	A	B	C	D	E	F	G	H	I	J
ぜろ (A)	10	1	4					1	1	1
いち (B)	3	10	2	2				1		
に (C)	2	2	12	1				1		
さん (D)	2	4	2	3			2	4	1	
よん (E)	4		1		7	1		1	1	3
ご (F)				1	2	10	2	1	1	1
ろく (G)	1			1		5	10		1	
なな (H)		1				5		9	2	1
はち (I)	1		4				1	3	8	1
きゅう (J)					1	4	1		5	7
合計	23	18	25	8	10	25	16	21	20	14

ルネットワークを用いて実験を行った．実験結果を表 4 および表 5 に示す．

認識精度は低いものの，中間層を増したことで精度が向上し，語ごとのばらつきも改善された．しかし，誤認傾向を分析してみると，母音系列としての共通点がない誤認も多く見られた．これは，切出し区間が狭くなったことによって，母音発声の特徴が薄れたり，発声前の準備動作や発声後の復帰動作の影響が大きくなったりしたこと，時間軸方向での正規化を行っていないことの悪影響が増大したことが原因と思われる．

5. おわりに

黙声母音認識の成果に基づき，少数語彙世界での単語認識を行った．音節数が多く，母音系列としての特徴差が大きい場合には高い精度を得ることもできるが，子音の特徴差の識別が必要な場合には，従来の手法の単純な延長では高い精度を得ることは難しいと言える．

認識精度を大きく向上させるには子音の認識技術の発展が不可欠であり，筋電波形の短時間の推移として現れるはずの子音の特徴の抽出方法の開発や舌の位置を左右する舌骨の動きに影響する筋を加えての認識実験などが今後の課題である．

参考文献

[1] 永井秀利, 中山浩之, 中村貞吾, 野村浩郷:, “筋電に基づく黙声認識におけるニューラルネットワークを用いた母音認識”, 電気関係学会九州支部大会 12-1P-05 (2004)

[2] 永井秀利, 南誠子, 中村貞吾, 野村浩郷:, “筋電に基づく黙声認識における子音認識のための基礎的調査”, 電気関係学会九州支部大会 12-1P-06 (2004)

[3] 角田耕一, 杉江昇: “音声合成方式発声代行システム —筋電位信号からの母音の判別と発声—”, 電気学会論文誌 105-C, pp.25–32(1985)

[4] 真鍋宏幸, 平岩明, 杉村利明: “無発声音声認識 : 筋電信号を用いた声を伴わない日本語 5 母音の認識”, 電子情報通信学会論文誌 D-II, Vol.J88-D-II, No.9, pp.1909–1917(2005)

[5] 村木茂, 角田耕一, 杉江昇: “代用発声のための子音判別法に関する基礎的研究”, 電子通信学会技術報告, MBE83-108(1983)

[6] 江崎友治: “筋電を用いた黙声認識 —「い」と「え」,「う」と「お」の判別 —”, 2004 年度九州工業大学情報工学部知能情報工学科卒業論文, p.38(2005)

[7] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel: “Session Independent Non-Audible Speech Recognition Using Surface Electromyography”, Proc. ASRU, pp.331–336(2005)

[8] S.-C. S. Jou, T. Schultz, and A. Waibel: “Continuous Electromyographic Speech Recognition with a Multi-Stream Decoding Architecture”, Proc. ICASSP, Vol.4, pp.401–404(2007)

[9] B.J. Betts and C. Jorgensen: “Small Vocabulary Recognition Using Surface Electromyography in an Acoustically Harsh Environment”, tech. memo TM-2005-213471, NASA (2005)

[10] 永井秀利, 中村貞吾, 野村浩郷: “無発声ないし微発声音声認識のための表面筋電波形からのノイズ低減手法”, 情報処理学会九州支部「火の国シンポジウム 2003」, pp.1–8(2003)

[11] 永井秀利, 中村貞吾, 野村浩郷: “自然言語インターフェースのための無発声音声認識への活用を目的とした表面筋電波形の分析”, 電子情報通信学会技術報告 Vol.102, No.688, pp.25–32(2003)

[12] 張志鵬, 真鍋宏幸, 平岩明, 杉村利明: “HMM 及びケプストラム係数特徴による筋電信号を用いた無発声音声認識”, 電子情報通信学会技術報告 Vol.103, No.401, pp.7–12(2003)