ニューラルネットワーク形学習による 声道断面積を用いた連続母音認識

正員 須貝 康雄 (千葉大)

非会員 松浦 嘉宏 (明電舎)

正員 平田 廣則 (千葉大)

Recognition of Vowel Sequences by a Neural Network Using Vocal Tract Area

Yasuo Sugai, Member (Chiba University), Yoshihiro Matsuura, Non-member (Meidensha Corporation), Hironori Hirata, Member (Chiba University)

This paper proposes a simple new model for the recognition of vowel sequences by a neural network using vocal tract area. Utilizing vocal tract area as input data reduces the number of input units and requisites computational time of learning compared with utilization of frequency spectrum. Practically, back-propagation rule is applied on three-layer structure, and suitable prediction of vocal tract area makes it possible to recognize speaker-independent vowel sequences by 39 units which is comparatively small as the number of total units.

キーワード：連続母音認識，声道断面積，ニューラルネットワーク，バックプロパゲーション

1. ま え が き

最近、ニューラルネットワーク^{(1)~(3)}によって音声認識を行わせようとする試みが幾つか報告されている⁽⁴⁾⁽⁵⁾。ニューラルネットワークを音声認識に応用する場合、時間的変化の取扱いが必要なため、入力ユニット数を増加させたり、ネットワーク構造を複雑化することになり、多大な学習時間を要する。時間的変化の取扱いには、ネットワークの構造自体に時間遅れの要素を直接取入れる方法や、時間的な広がりや空間的な広がりに置換え、時間的に連続した幾つかのデータを空間的に並べて一つの入力データとして扱う方法がある。ニューラルネットのソフトウェア上での実現が主流になっている現状では、いずれの場合においても、ある程度長時間の学習が必要である。学習時間を少しでも軽減するためには、入力ユニット数になるべく少なくてすむような工夫が必要である。

一方、音声は音帯の振動や空気流による乱流が喉頭、咽頭、口腔などからなる管状の部分（声道）によ

る共振や反共振によって、変化を受けた後に口唇から放射されて生成される。連続音声の場合には、調音器官を断続的に動かすことができないために、前後の音によって影響を受け、その音響的特徴が変化してしまう調音結合の問題が生じる。そのため、音声の区分化があいまいになり、特に有声音が連続した場合の区分化が困難になる。日本語の場合、ほとんどの音声の子音+母音の形式で構成されていて、子音部では声帯振動が休止するために比較的簡単に区分化できるが、子音を含まずに母音が連続する場合には区分化が困難になる。また、そのようなあいまいな部分は、話者を不特定にした場合に更に増加する。

このような連続母音の認識を、ニューラルネットを用いて行う場合、入力として声道断面積を用いることが考えられる。音声の特徴は声道の形状に左右され、その形状を決める声道断面積を推定して利用すると、音声の特徴をとらえることができる。そして得られる声道断面積は、他の周波数スペクトルやケプストラム係数などに比較して、原波形データから直接得られる

データ数が少なくすむので、ニューラルネットワークに利用する場合、比較的少ないユニット数と学習時間で実現できると考えられる。

本論文では、声道断面積を入力データとして用いたニューラルネットによる連続母音認識モデルを提案する。本モデルはバックプロパゲーション法⁽²⁾を学習則とする3層構造のネットワークを用いる。学習時間短縮のために、ユニット数はできるだけ少なくし、39個の小規模な構成である。このなかには連続母音認識で生じる母音間の涉りの影響を考慮するために、2フレーム間の時間的変化量を表す入力ユニット1個と、5母音のいずれでもないことを示す出力ユニット1個を設けている。不特定話者（男性）の音声データを用いた計算機実験により、本モデルは連続母音の認識ができ、入力として声道断面積が有効であることがわかった。

2. 声道断面積

最初に、ネットワークの入力に使用する声道断面積データについて述べる。声道をその断面積が連続的に変化している一本の音響管とみなし、更に、声道を長さ Δx の区間ごとに断面積一定の微小音響管に分けて近似して考える(図1)。その場合、反射係数と線形予測残差の相関係数であるPARCOR係数との類似性から、PARCOR係数によって、声道断面積が推定できることが知られている⁽⁶⁾⁽⁷⁾。いま、口唇から*i*番目の微小音響管における声道断面積を A_i ($i=1, \dots, n$)、対応するPARCOR係数を k_i とし、 $A_0=1$ とすると、

$$A_i = \frac{1-K_i}{1+K_i} A_{i-1} \dots\dots\dots (1)$$

となる。

以下に本モデルでの具体的な計算手順を示す。なお、声道の分割数は16とする($n=16$)。

(1) A/D変換 話者の発した音声アナログ信号をマイクを通じてカットオフ周波数6 kHzの低域フィルタに直接通した後に、A/D変換によってデジタル化する。サンプリング周波数は20 kHz、変換

図1 声道の音響管近似

Fig.1. An approximation of vocal tract by acoustic tubes.

ビット数は12 bitである。その後、4.4 kHzで3 dB、5 kHzで20 dB以上の減衰率をもつデジタル低域フィルタを通した後に、1/2に分周して10 kHzのデジタル化音声信号を得る。

不特定話者の母音認識を考えた場合、各母音を音韻の特徴によって分離するためには、およそ4 kHz以下の成分が必要とされていて、それに対して、5 kHz以上の成分は逆に話者固有の特性を表すので⁽⁸⁾、不特定話者での認識を行うためにはかえって障害になる可能性がある。そのため、このような処理により4 kHz以下の成分は残し、5 kHz以上の成分は切捨てる。なお、上記フィルタの伝達関数は、次式のとおりである⁽⁹⁾。

$$H(z) = \frac{B_0 + B_1 z^{-1} + B_2 z^{-2} + B_3 z^{-3}}{1 + A_1 z^{-1} + A_2 z^{-2} + A_3 z^{-3}} \dots\dots\dots (2)$$

ここで、

$$A_1 = -0.9173114, \quad B_0 = 0.1619881,$$

$$A_2 = 1.034724, \quad B_1 = 0.1892079,$$

$$A_3 = -4.150209, \quad B_2 = B_1, \quad B_3 = B_0$$

(2) フレームの切出し 1回の分析で使用する区間の長さ(フレーム長)を30 msにして計算を行う。また、フレームを切出した際に、両端での急激な変化をなくすために、窓関数として(3)式で表されるハミング窓を使用する。

$$W(n) = 0.54 - 0.46 \cos\left(\frac{2n\pi}{N-1}\right) \dots\dots\dots (3)$$

フレームを進める長さ(フレーム周期)は、窓関数を掛けることで連続性が失われることがないように、10 msにして各フレームごとに20 msの重なりをもたせて計算を行う。

(3) PARCOR係数 ハミング窓を通し、適応一次逆フィルタ⁽⁶⁾⁽⁷⁾を5段階通した後に、PARCOR係数を次数16とし、Le Roux法⁽⁷⁾を使用して計算する。

(4) 声道断面積の計算 PARCOR係数から(1)式に従って声道断面積を計算する。ただし、 $A_0=1$ とする。得られた声道断面積の値は声道断面積比の形で得られることと、バックプロパゲーションルールへ適用することを考慮して、フレームごとに最大値が1になるように0~1の範囲で正規化する。このようにすることで、個人差を吸収する効果も期待できる。なお、無音部であるとみなされるフレームは強制的に0になるようにする。最終的に声道断面積は、声門から口唇まで16に分割した形で0~1の実数値で10 msごとに計算する。

3. バックプロパゲーションルール

本章では、ニューラルネットワークの学習則として使用したバックプロパゲーションルールについての簡単な説明をする。バックプロパゲーションルールは、Rumelhart らによって提案された、層構造のニューラルネットワークにおけるユニット間結合の重みを学習させる規則である⁽²⁾。

ユニット i, j の出力値をそれぞれ o_i, o_j 、ユニット j の内部状態を s_j 、ユニット i, j 間の結合の重みを w_{ij} 、ユニット j のしきい値を θ_j とし、ユニット i から j にデータが流れるとすると、各々の関係は次のようになる。

$$s_j = \sum_i o_i w_{ij} + \theta_j \quad \dots\dots\dots (4)$$

$$o_i = f(s_j) \quad \dots\dots\dots (5)$$

$$f(x) = \frac{1}{1 + \exp(-x)} \quad \dots\dots\dots (6)$$

そして、学習は w_{ij}, θ_j を次のように変更して行う。なお、各々は時間ステップ T の関数として記述してある。

$$w_{ij}(T) = w_{ij}(T-1) + \epsilon \Delta w_{ij}(T) \quad \dots\dots\dots (7)$$

$$\Delta w_{ij}(T) = \eta \delta_j(T) o_i(T) + \alpha \Delta w_{ij}(T-1) \quad \dots\dots\dots (8)$$

$$\theta_j(T) = \theta_j(T-1) + \epsilon \Delta \theta_j(T) \quad \dots\dots\dots (9)$$

$$\Delta \theta_j(T) = \eta \delta_j(T) + \alpha \Delta \theta_j(T-1) \quad \dots\dots\dots (10)$$

ただし、 $\Delta w_{ij}(0) = 0, \Delta \theta_j(0) = 0$ とし、重みとしきい値の初期値 $w_{ij}(0), \theta_j(0)$ は乱数によって与える。また、 δ_j はユニット j について計算される誤差信号で、ユニット j が出力層のユニットである場合は、ユニット j が出力すべき値（教師値）を t_j として、

$$\delta_j = (t_j - o_j) o_j (1 - o_j) \quad \dots\dots\dots (11)$$

となり、出力層よりも下の層のユニットである場合には、ユニット j より、一段上の層にあるユニット k を考え、

$$\delta_i = o_j (1 - o_j) \sum_k \delta_k w_{jk} \quad \dots\dots\dots (12)$$

として再帰的に計算する。このように、重みの修正がデータの流と逆方向に伝搬していくので、バックプロパゲーションルールと呼ばれている。なお、これらの学習規則は、ユニット j の教師値を t_j としたときの誤差 E を、

$$E = \frac{1}{2} \sum_j (t_j - o_j)^2 \quad \dots\dots\dots (13)$$

のように定義し、これに対して、

$$\Delta w_{ij} \propto -\frac{\partial E}{\partial w_{ij}} \quad \dots\dots\dots (14)$$

$$\Delta \theta_j \propto -\frac{\partial E}{\partial \theta_j} \quad \dots\dots\dots (15)$$

とする最急降下法によって得られる。ここで、 α, η, ϵ は定数で、本モデルでは実験的に定めた $\alpha = 0.3, \eta = 0.1, \epsilon = 3.0$ という値を用いる。これらは通常よく用いられている範囲の値である。

4. 定常母音の認識

定常母音認識のための学習は、以下のようにして行う。入出力の概念図を図2に示す。

(1) 学習データ 学習の入力に使用したデータは、3人の男性話者による5母音の声道断面データのうち、定常であると考えられる部分から、各母音について20フレームずつを切出して使用する。教師となる出力パターンは、5個の出力ユニットがそれぞれ5母音に対応するような形で与える。また、声道断面積を計算する際に、無音部分と考えられる音の小さな部分は、強制的に0になるようにしているので、その部分ではユニットの出力がすべて0になるようなデータも学習させる。

(2) ユニットの構成 声道断面データは1フレームごとに16個あるので、入力層のユニット数は16個とする。また、日本語の5母音、/a/, /i/, /u/, /e/, /o/を認識するために、出力層のユニット数は5個とする。そして中間層のユニット数は、10~20個の間で予備実験を行い、学習時間と認識性能とのつり合いを考慮して、16個という数を採用する。総ユニット数は37個である。

(3) 学習回数 すべてのデータを1回ずつ学習するのを1回の学習として、1500回学習させる。

実験の結果、学習に要した時間は、数値演算コプロセッサ搭載のNEC製PC-9801 VX 21で実行して約14時間45分であった。

学習させた重みに対して、学習に使用した話者（ただし、未学習データ）も含めた9人の男性話者（A

図2 3層モデル
Fig. 2. Three-layered model.

図 3 /u/の認識結果
Fig. 3. Recognition of vowel /u/.

～Iで表し、学習に使用した話者は、A, F, Iの3名)について定常母音の認識を行わせた。判定の基準は、出力層の出力値が最大であるユニットの値が0.5以上であり、かつその値がその次に大きな値を出力するユニットの値の1.5倍である場合に、その最大であるユニットが発火したとみなすようにし、それ以外の場合には判定不能であるとした。無音部分のデータはすべての出力ユニットの出力が0になるように学習しているので、そこでは判定不能となる。

実験の結果、/a/については話者B、/e/については話者H、/i/については話者F, H、/o/については話者C, D、/u/については話者B, C, D, Eにおいて、音声の部分で、判定不能となった部分がある。例として、/u/の認識結果を図3に示す。図3において、時間は左から右、上から下に、1文字あたり10ms進み、 $\cdot$ で表されたところは判定不能であることを示す。判定不能であったフレームのデータを学習データに追加し、更に20回の学習を行った。その結果、話者9人の5母音に対してほぼ完全に認識ができた。図3に対応する追加学習後の認識結果を図4に示す。

図 4 追加学習後の/u/の認識結果
Fig. 4. Recognition of vowel /u/ after a supplementary learning.

5. 連続母音の認識

〈5・1〉 連続母音認識の問題点 男性不特定話者の定常母音認識に対しては、第2章で述べたネットワークに1フレーム分の声道断面積データを入力すれば、認識が行えることを示した。しかしながら、連続母音認識を行うためには、以下の問題点を解決する必要がある。その問題点としてまず、三つ以上の母音が連続した場合、調音結合により二つの母音に挟まれた母音が前後の影響で全くその特徴を失ってしまうことが考えられる。

そこで、定常母音認識のための学習により得られた重みに対して、同じ9人の/aoi/という3連母音の認識実験を行った(図5)。その結果、/a/と/i/に挟まれた/o/の音の認識はできたものの、/o/と/i/の間に他の音に誤認識されているフレームがあることがわかった。各母音が発せられている時間はほぼ200ms程度であり、特に長いわけではない。また、/a/と/o/の間には他の母音に誤認識されている部分はなかった。残りの母音の組合せ(2連母音)についても、同様に

認識実験を行った結果、全体の傾向として/a/-/i/, /o/-/i/, /e/-/o/間に誤認される部分があることがわかった。

これらの現象は、次のことが原因であると考えられる。図6は合成音声の第一および第二ホルマントを変動させたときの人間の音韻知覚分布を表したものであ

図5 定常母音認識モデルによる連続母音/aoi/の認識結果

Fig. 5. Recognition of vowel sequence /aoi/ by the recognition model of single vowels.

図6 合成母音の音韻知覚分布
Fig. 6. Perceptual phoneme distribution of synthesized vowel-like sounds.

るが⁽¹⁾、母音を連続して発した場合の母音間での変動は、図6の上では最短距離で動くと考え、/o/と/i/の間で/u/の領域を、/o/と/e/の間で/u/または/a/の領域を、/a/と/i/の間で/e/または/u/の領域を横切ることになる。しかし、それ以外の場合では、他の領域を横切ることにはない。これらのことから、誤認されている部分はフレーム単位で考えている場合は誤認ではないとも考えられる。従って、このままでは、いくら学習を行ってもこの部分での誤認率を下げることはできないことになる。

そこで、次節ではその対策法を提案し、それに基づいて行った実験結果を示す。

〈5・2〉連続母音認識の学習モデル 前節において、母音が連続して発せられた場合、声道の形状変化によって、途中で誤認されてしまう部分が出てくることがわかった。図2の構成では時間的変化を取扱っていないため、このような結果は当然とも言える。しかしながら、時間的変化を考慮するために一般的にとられる方法、すなわち、時間的に連続したデータを空間的に並べて配置する手法では、ユニット数が増大してしまい、多大な学習時間を要することになる。そこで、ユニット数をあまり増大させずに時間的な変化をとらえるために次のようにする。

母音が連続して発声された場合、その間の部分では、実際の声道の形状は急激な変化をしていると考えられる。特に口唇部に近い部分においてその変化は顕著であると考えられる。そこで、得られた声道断面積データについて、1フレーム前のデータとの間で、口唇部から6個分のデータの差分をそれぞれとり、その絶対値の平均を入力データに加える。つまり、そのための入力ユニットを1個追加することにする。これによって、母音間の涉りの部分とそれ以外の部分との相違が検出可能になると期待できる。更に、そのような

図7 連続母音認識モデル
Fig. 7. Recognition model of vowel sequences.

〈5・3〉 連続母音の認識実験 以上のように拡張したネットワークに対して、次のような学習を行った。

(2) ユニットの構成 さきに述べた理由から、入力ユニット17個、中間層に16個、そして出力ユニット6個の計39個のユニット構成である。

話者 A
aaaaaaa00000_??eee_iiii

話者 B
aaaaaaaaaa_00000
000000000?eiiiiiiiii

話者 C
ooooooooooooo_??eiiiiiiiii

話者 D
oooooooooooooooooo_uuuuiiiiiiii

話者 E
ooooooooooooou_aauuiiiiiiii

話者 F
aaaaaaaaa0000000000000000000
?iiiiii_iiii

話者 G
aaaaaaaaaaaa_000000000000_u?ueiiiiii_u

話者 H
ooooodaauweeeiiiiiiiii

話者 J
aaaaaa_00000000000000000000_?iiiiiii

機を用いて、3,000 回学習させた。学習時間は約 30 時間 35 分であった。

File aoym1.v49		Max value = 1	
10 - 14			
15 - 19			
20 - 24			
25 - 29			
30 - 34			
35 - 39			
40 - 44			
45 - 49			
50 - 54			
55 - 59			

File a01gm.c48	Max value = 3			
10 - 14				
15 - 19				
20 - 24				
25 - 29				
30 - 34				
35 - 39				
40 - 44				
45 - 49				
50 - 54				
55 - 59				

T. IEE Japan, Vol. 110-C, No. 6, '90

もないことを示すユニットの値に対応している。

量的な議論はできないが、図5と比較して、図8では多少判定不能の部分が増えたり、?と判定される部分が現れたりしているものの、それらが生じている部分は数10ms程度の区間であり、ほぼ認識ができていえる(図10参照)。また、残りの母音の組合せについても、今まで誤認してしまっていた部分に?が現れるようになり、誤認率は減ってきていると言える。

このように追加学習をさせることにより、結果が良好になってきていることから、更に学習させるデータの種類を増やし、学習を続けることによって良い結果が得られると期待できる。

6. む す び

本論文では、声道断面積を入力に用いたニューラルネットによる連続母音認識モデルを提案した。本モデルを用いた男性不特定話者の定常母音の認識実験により、ニューラルネットワークで音声認識を行わせる場合に、入力として声道断面積データを用いることが有効であることを示した。更に、ユニット数をあまり増大させることなく連続母音の認識をさせる手法を提案し、その有効性を示した。

定常母音認識に関しては、ほぼ満足のいく結果が得られている。また、連続母音認識の結果においては、まだ完全に認識させることはできていないが、誤認していた部分をずれの母音でもないかと判断させることで、誤認率を減少させることができた。今後の課題としては、女性を含めた不特定話者のデータを用いての学習および認識実験を行うこと、あるいは母音以外の他の音声を確認する同様のシステムを組合せることによって、同時に認識可能な音声の種類を増やしていくことがあげられる。

日ごろ有益なる御討論御助言をいただいている本学電気電子工学科 倉田 是教授に感謝いたします。

(平成元年9月25日受付)

文 献

- (1) 麻生: ニューラルネットワーク情報処理 (昭63) 産業図書
- (2) D.E. Rumelhart, et al.: Parallel Distributed Processing, Vol. 1 (1986) MIT Press
- (3) T.J. Sejnowski & C.R. Rosenberg: "NETtalk: A Parallel Network that Learns to read Aloud", The Johns Hopkins Univ. Electrical Engineering and Computer Science Technical Report, JHU/EECS-86-01 (1986)
- (4) A. Waibel: 「時間遅れ神経回路網(TDNN)による音韻認

識」, 信学技報, SP 87-100 (昭62)

- (5) 迫江・磯: 「ダイナミックニューラルネットワークの検討」, 同上, SP87-101 (昭62)
- (6) 古井: デジタル音声処理 (昭60) 東海大学出版会
- (7) 斎藤・中田: 音声情報処理の基礎 (昭56) オーム社
- (8) 三浦: 聴覚と音声 (昭57) 電子通信学会
- (9) 三谷: デジタルフィルタデザイン (昭62) 昭晃堂
- (10) 安居院・中嶋: コンピュータ音声処理 (昭61) 秋葉出版
- (11) 大友・高橋・原: 「神経回路モデルによる母音知覚分布の学習と識別」, 信学論 D, J71-D, 2291 (昭63-11)
- (12) 松浦・須田・平田: 「声道断面積を用いたニューラルネットワークによる連続母音認識」, 信学技報, SP89-10 (平成元)

須 貝 康 雄 (正員)

昭和57年千葉大学大学院工学研究科電子工学専攻修士課程修了, 60年東京工業大学大学院理工学研究科情報工学専攻博士課程修了, 61年千葉大学工学部助手, 63年同大学院自然科学研究科助手となり, 現在に至る。VLSI システム, 生態システムなどの大規模システムおよびニューロコンピュータなどの分散システムの研究に従事。工学博士。電子情報通信学会, 計測自動制御学会, 情報処理学会会員。

松 浦 嘉 宏 (非会員)

平成元年千葉大学大学院工学研究科電子工学専攻修士課程修了。同年(株)明電舎に入社, 現在に至る。在学中はニューラルネットワークによる音声認識の研究に従事。電子情報通信学会会員。

平 田 廣 則 (正員)

昭和51年東京工業大学大学院博士課程電気工学専攻修了, 同年千葉大学工学部助手, 56年同助教授となり, 現在に至る。大規模システム of モデリング, 解析と設計, 特に生態システム, VLSI レイアウトならびに分散形コンピュータの基礎的理論に興味をもつ。工学博士。IEEE, ISEM, INNS, 電子情報通信学会, 情報処理学会, 計測自動制御学会会員。