ニューラルネットワークを用いた音声認識の試み

小机　わかえ[1]・宮地　秀征[1]・野村　浩史[2]・花岡　英樹[2]

[1] 機械工学科
[2] 平成９年度機械工学科４年生

Speech Recognition Using Neural Network

Wakae KOZUKUE[1], Hideyuki MIYAJI[1], Hiroshi NOMURA[2],and Hideki HANAOKA[2]

Abstract

In this paper we try to apply a neural network algorithm to human speech recognition. The numrical experiment is carried out to recognize vowels using time domain waveform and frequency domain waveform as training data. As a result, it is concluded that using the frequency domain waveform is more accurate than using the time domain waveform. Furthermore it is shown that the shaffling of training data has influence on the correct recognition of vowels.

Key Words: Neural Network, Speech Recognition, FFT, Vowel.

1.はじめに

近年、人間の脳の情報処理能力に着目した計算アルゴリズムとして、ニューラルネットワークが注目されている。ニューラルネットワークは、従来のノイマン型計算機が苦手とするパターン認識などを、効率よく処理できることから、いろいろな分野で研究されている。ニューラルネットワークに関しては、様々な構造や学習アルゴリズムが提案されており、もっともよく研究されているものは、誤差逆伝搬法(バックプロパゲーション法)を用いた階層型ネットワークである。しかし、階層型ネットワークは、入力層や出力層のニューロンの数が増えると、計算すべき重み係数の数が増加し、計算効率が悪くなることが欠点であった。これに対して、Sutherlandの提案した、ホログラフィックニューラルネットワーク[1]は、学習に要する時間が少ない、予測精度がよい、などにより、階層型ニューラルネットワークより多量の入出力データを効率的に処理できるという点で、優れているといわれている。

本研究では、このホログラフィックニューラルネットワーク(以下、H.NNと略す)を人間の音声の認識に応用することを試みる。

音声認識は、音声タイプライタや音声駆動コンピュータなどの実用技術に深く関わることから、古くから研究されてきており、実用化されたシステムもいくつか存在する。そして、従来の音声認識の手法では、スペクトル分析法や線形予測分析法などの統計的方法が、実用上きわめて優れた方法とされていた[2]。

一方、ニューラルネットワークを用いて音声認識を試みる研究は、活発に行われており、特にTime Delay Neural Network(TDNN)が音韻識別に高い能力を持つことが示され、関心が高まった[3]。また、バックプロパゲーション法による母音認識の研究も行われており、従来の統計的方法に比較して、高い認識率が得られることなどが報告されている[3]。本研究では、上で述べたように、バックプロパゲーション法よりも効率が良いと言われているH.NNをニューラルネットワークとして用い、特定話者の母音(あ、い、う、え、お)の認識という最も基本的な問題を取り上げて、検討を行う。そのために、実際に人間の音声(母音)を録音し、計算機上でH.NNによる音声認識のための数値実験を行う。

2．ホログラフィックニューラルネットワーク(HNN)の概要[1]

HNNはSutherland[1]によって開発されたNNで、その最大の特徴は、入力データと出力データを複素平面上に均一に写像することによって、両者の間に線形関係を持たせていることである。HNNではニューロンの数は１個であり、NNの構築は、入力と出力の間の伝達関数を求めることに等しい。そのため、計算が収束するまでの時間を

大幅に短縮することが可能となる。

学習にl組のm次元入力ベクトル$\mathbf{s}$とn次元出力ベクトル$\mathbf{r}$を用いるとする。入出力ベクトルの各要素は、非線形変換関数により、複素平面上に変換される。

$$f(s_{hk}) = \lambda_{hk}\, e^{i\theta_{hk}} \tag{1}$$

$$g(r_{jk}) = \gamma_{jk}\, e^{i\phi_{jk}} \tag{2}$$

ここで、iは虚数単位、θ_{hk}、ϕ_{jk}はシグモイド関数のような写像関数により、変換される位相角度で区間［0、2π)の値を持つ。λ_{hk}、γ_{jk}は入出力データが、対応する位相角領域に出現する確率を示し、区間[0、1]の値を持つ。以上の式(1)、(2)の操作により次のような入力行列[S]、教師行列[T]が得られる。

$$[S] = \begin{bmatrix} \lambda_{11}e^{i\theta_{11}} & \lambda_{12}e^{i\theta_{12}} & \cdots & \lambda_{1m}e^{i\theta_{1m}} \\ \lambda_{21}e^{i\theta_{21}} & \lambda_{22}e^{i\theta_{22}} & \cdots & \lambda_{2m}e^{i\theta_{2m}} \\ \vdots & \vdots & & \vdots \\ \lambda_{l1}e^{i\theta_{l1}} & \lambda_{l2}e^{i\theta_{l2}} & \cdots & \lambda_{lm}e^{i\theta_{lm}} \end{bmatrix} \tag{3}$$

$$[T] = \begin{bmatrix} \gamma_{11}e^{i\theta_{11}} & \gamma_{12}e^{i\theta_{12}} & \cdots & \gamma_{1m}e^{i\theta_{1n}} \\ \gamma_{21}e^{i\theta_{21}} & \gamma_{22}e^{i\theta_{22}} & \cdots & \gamma_{2m}e^{i\theta_{2n}} \\ \vdots & \vdots & & \vdots \\ \gamma_{l1}e^{i\theta_{l1}} & \gamma_{l2}e^{i\theta_{l2}} & \cdots & \gamma_{lm}e^{i\theta_{ln}} \end{bmatrix} \tag{4}$$

出力行列[A]は、伝達関数[X]を用いて、次式のように表わされる。

$$[A] = [S]\cdot[X] \tag{5}$$

[X]は、次式に表わされる[A]と[T]との差が最小となるように決定される。

$$E_{rr} = ([A] - [T])^{H.}\,([A] - [T]) \tag{6}$$

ここで、Hは共役転置を表わす。この条件より、次式が得られる。

$$[X] = ([S]^{H.}\,[S])^{-1}\cdot[S]^{H.}\,[T] \tag{7}$$

[S]、[T]は既知であるので、[X]は式(7)により直接求まる。しかし、Gauss-Jordan消去法でn次の行列の逆行列を求める場合、その演算回数はn^3回となるので、[S]が高次元になると演算に膨大な時間が必要となる。そこで、これを避けるため、次式の反復学習により[X]を求める。

$$[X]_i = \frac{1}{E}[S]^{H.}\,[T] \tag{8.a}$$

$$[X]_{i+1} = [X]_i + [S]^{H.}\left([T] - \frac{1}{E}[S]\cdot[X]_i\right) \tag{8.b}$$

ここでEは[X]のノルムを正規化するパラメータである。伝達関数[X]の収束を速くするためには、入力ベクトルを複素平面上に均一に変換することが望ましい。このため、変換関数には、次ような形を持つシグモイド関数などが用いられる。

$$\theta_k = \frac{2\pi}{1+e^{(\mu - s_k)/\sigma}} \tag{9}$$

ここで、s_kは入力、μ, σは、パラメータである。

3.実験方法

本研究では、N.Nの学習データとして、音声の時間領域における波形、FFTによって変換された周波数領域における波形を用いた。特定話者の音声をマイクロフォンで録音し、それをFFTアナライザによって処理して、波形を数値化する。時間領域波形に対しては、標本化周波数320Hz、測定時間250 msecとし、周波数領域波形では、0 から 3.125 kHzを対象範囲とした。時間領域、周波数領域共に、１つの波形を250個の離散データで表現し、それらをニューラルネットワークに入力するフォーマットに変換し、学習データ、及び結果を評価するためのテストデータを作成する。学習データの出力部分には1か0を与えた。すなわち、例えば「あ」を認識するための実験では、「あ」の波形は出力値として１を、その他の母音の波形には、出力値として０を割り当てる。

ニューラルネットワークに、上記の音声波形と出力値を学習データとして入力して学習を行う。そして、学習済みのニューラルネットワークに未知の音声の波形を入力し、出力値を求める。認識したかどうかの条件は以下のように設定した。すなわち、出力値が「出力の理論値±0.3」のとき、正解であるとした。例えば、出力の理論値が１のとき、出力値が0.7から1.3であれば正解、-0.3 から 0.3であれば不正解(誤答)とし、出力の理論値が０のときは、出力値が-0.3 -から0.3であれば正解、0.7 -から1.3であれば不正解(誤答)とした。出力値がそれ以外の数値の場合は、認識不可能とみなした。実際の実験では、150個のテストデータ(波形)をH.NNに入力して、それらの正解率、誤答率、認識不可能率を調べ、ニューラルネットワークの学習のさせ方によって、音声の認識率がどの程度変化するかを比較・検討した。

4.実験結果

本研究では、音の時間領域における波形、FFTによって変換された周波数領域における波形の２通りを数値実験の対象にした。そして、認識率を向上させるために、元の波形を以下に述べる方法で編集して、実験を行った。表１に学習データとして用いた時間領域の波形の一覧とそのID番号、表２に同様に周波数領域の波形の一覧とそのID番号を示す。各ID番号に対応した、学習データの特

徴を以下に示す。

(1)時間領域における波形を用いたの学習データ

[1]TN250

時間領域における波形をそのまま用いた。学習データは、’あ’、’い’、’う’、’え’、’お’を各50個、計250個により構成されている。データの並び方は、最初に’あ’が50個、次に’い’が50個、というようになっている。

[2]TS500

TN250のデータを、最大音圧レベルに関して、音圧レベルを正規化したものである。ここで、学習データとしては、各母音100個、計500個を作成した。

[3]TSJ500

TN250の学習データの並び方は、’あああ…　あああいいい…いいうううう…’であったが、これを’あいうえおあいうえお…’のように並び替えた。

[4]TK400

TN250では、’あ’が50個に対して、’い’、’う’、’え’、’お’が計200個あった。そこで、’あ’を200個、’い’、’う’、’え’、’お’を計200個とし、総数400個のデータを作った。すなわち、出力が1であるデータと、出力が0であるデータの数を等しくする均等化を行った。ただし、正規化と並び替えは行っていない。

[5]TSJK400

TSJ500のデータの均等化をTK400と同じように行った。

(2)周波数領域における波形を用いた学習データ

[1]SpN250

周波数領域における波形をそのまま用いた。学習データは、’あ’を50個、’いうえお’を各50個とし、計250個により構成されている。データの並び方は、TN250と同様である。’あ’の出力を1、’いうえお’の出力を0とした。

[2]SpN250i

学習データは、’あいうえお’各50個、計250個よりなる。’い’の出力を1、’あうえお’の出力を0とした。

[3]SpN250u

SpN250iと同様であるが、’う’の出力を1とした。

[4]SpN250e

SpN250iと同様であるが、’え’の出力を1とした。

[5]SpN250o

SpN250iと同様であるが、’お’の出力を1とした。

[6]SpNJ250

SpN250の学習データを、TSJ500のようにを並び替えたものである。

表　1　時間領域波形を用いた学習データ一覧

学習データのID	説明
TN250	元の波形は全く加工しないもの
TS500	音圧レベルを正規化したもの
TSJ500	学習データの並び変えを行ったもの
TK400	学習データの均等化を行ったもの
TSJK400	TSJ500の均等化を行ったもの

表　2　周波数領域波形を用いた学習データ一覧

学習データのID	説明
SpN250	元の波形は全く加工しないもの
SpN250i	元の波形で「い」の認識を行ったもの
SpN250u	元の波形で「う」の認識を行ったもの
SpN250e	元の波形で「え」の認識を行ったもの
SpN250o	元の波形で「お」の認識を行ったもの
SpNJ250	学習データを並び変えたもの

以上のようにして作成した、各学習データに対して、各母音30個からなる総計150個のテストデータを対象に、H.NNによる解析を行った。なお、学習に際しては、誤差が一定の値に収束するまで、100～300回程度の繰り返ししか要しなかった。テスト結果より得られた各母音に関する認識率を表3、表4に示す。

表3より、時間領域の波形を学習データとして用いた場合、データの正規化、並び替えは、ある程度認識率を向上させるのに有効であることがわかる。またデータの均等化によっては認識率はかなり向上する。

表4より、周波数領域の波形を学習データとして用いた場合、データの正規化や均等化などのような編集操作を行わなくても、比較的高い認識率が得られた。一方、’う’と’え’の正解率は50%以下となった。しかし、それ以外では、時間領域の波形を学習データとするときより、はるかに高い正解率を示しており、周波数領域の波形を用いる方が有利であることが示されている。

表 3 時間領域波形の認識結果

学習データのID	母音の種類	正解率	誤答率
TN250	'あ'	7.1 %	91.4 %
	'いうえお'	98.8 %	0 %
TS500	'あ'	30.0 %	50.0 %
	'いうえお'	71.3 %	13.8 %
TSJ500	'あ'	31.4 %	47.8 %
	'いうえお'	78.8 %	6.3 %
TK400	'あ'	80.0 %	17.1 %
	'いうえお'	51.3 %	47.5 %
TSJK400	'あ'	88.6 %	10.0 %
	'いうえお'	61.3 %	37.5 %

表 4 周波数領域波形の認識結果

学習データのID	母音の種類	正解率	誤答率
SpN250	'あ'	85.0 %	3.75 %
	'いうえお'	96.25 %	0 %
SpN250i	'い'	75 %	5 %
	'あうえお'	95 %	0 %
SpN250u	'う'	45 %	15 %
	'あいえお'	92.5 %	0 %
SpN250e	'え'	35 %	0 %
	'あいうお'	92.5 %	2.5 %
SpN250o	'お'	70 %	10 %
	'あいうえ'	96.25 %	2.5 %
SpNJ250	'あ'	90 %	0 %
	'いうえお'	97.5 %	1.25 %

5.結果の検討

学習データとして、時間領域の波形より周波数領域の波形を用いた方が正解率が高いのは、時間領域の波形では、データによって音声の持続時間がばらつき、短い音声については、後ろの方にゼロが続いてしまい、波形の特徴が明確に現われないからだと考えられる。周波数領域のデータでは、対象とする周波数範囲では、データが音声の特徴をある程度保っているため、認識しやすいと推察できる。

また、学習データの並び替えが認識率に影響を与えるのは、ニューラルネットワークの学習におけるシャッフリング効果、すなわち学習させる順番により精度が左右される現象が現われたためと考えられる。

なお、時間領域の学習データについて、均等化を行った場合認識率が向上する理由は、今のところ不明である。

6.まとめと今後の課題

H.NNを用いて人間の音声の認識ができるかどうかを調べた。そのために、特定話者の母音を対象に計算機を用いて数値実験を行った。N.Nの学習データとして、時間領域の波形と周波数領域の波形について解析した結果、周波数領域の波形を用いる方が、より高い認識率を得られることがわかった。また、時間領域、周波数領域の学習データの両者に関して、同じ出力の値を持つデータを引き続いて学習させるより、ある程度並び替えて学習させる方が有効であることも示された。

今後の課題としては、より複雑な単語認識、会話認識を試みることが挙げられる。また、本研究は音声を一度計算機で処理できる形のファイルに変換して解析を行っている。今後は、音声をリアルタイムに認識できるような、より実用に近いシステムを構築することも必要となろう。さらに今回は明確な基準を設けて学習データの大きさ、数を決めたわけでない。認識のために、最小限必要な学習データの規模を求めるための指針を得ることも望まれる。

参考文献

(1)J.G.Sutherland, "The Holographic Model of Memory, Learning and Expression", International Journal of Neural Systems, 1990, Vol.1 No.3, pp. 259 -267.
(2)古井著『ディジタル音声処理』(東海大学出版会)
(3)久間・中山編著『ニューロコンピュータ工学』(工業調査会)