ニューラルネットワークによる音声パタン認識

西 正明* ・ 林川基治**

Voice Recognition with Neural Networks

Masaaki NISHI* , and Motoharu HAYASHIKAWA**

(Received October 27, 2000)

1. まえがき

人の脳を形成している神経系の仕組みを工学的に模し，情報処理を行なおうとするのがニューラルネットワークである。神経生理学により，生物の脳神経系の構成要素はニューロンと呼ばれる神経細胞であり，人の大脳の場合は１００億を越える神経細胞のひとつひとつが互いに１万もの結合をしていることがわかってきた。ニューラルネットワークではこれをもっと簡単にモデル化して用いている[1)]。

本論文ではニューラルネットワークの応用分野のひとつである音声パタン認識を行なう。扱うニューラルネットワークはディレー素子を内包するニューラルネットワークで，クロストークリンクを備えたBPD[2)]を用いる。BPDは幾つかの時系列処理に有効であることが報告されている[3)]。本論文では，特定もしくは不特定の話者の発語の認識に有効であるかどうかを検討する。

2. ネットワーク構成と学習アルゴリズム

本論文では，3層構成のBPDニューラルネットワークを扱う。ネットワーク構成を図１に示す。ディレー素子を介して前の時刻の出力を入力データとして再入力するようにフィードバックすることで，時系列処理を可能にしている。BPDニューラルネットワークの学習アルゴリズムを以下に簡単に述べる[3)]。図１に示したクロストークリンク付きBPDで出力層のk番目ニューロンの内部状態$s_k(t)$と出力$o_k(t)$は次式で示される。

$$s_k(t) = \sum_j w_{kj}(t)o_j(t) + \sum_r \sum_n v_{krn}(t)o_r(t-n) \quad (1)$$

$$o_k(t) = f[s_k(t)] = \frac{1}{2}\{1+\tanh[\frac{s_k(t)}{u_0}]\} \quad (2)$$

* 信州大学教育学部　　** 山口大学教育学部

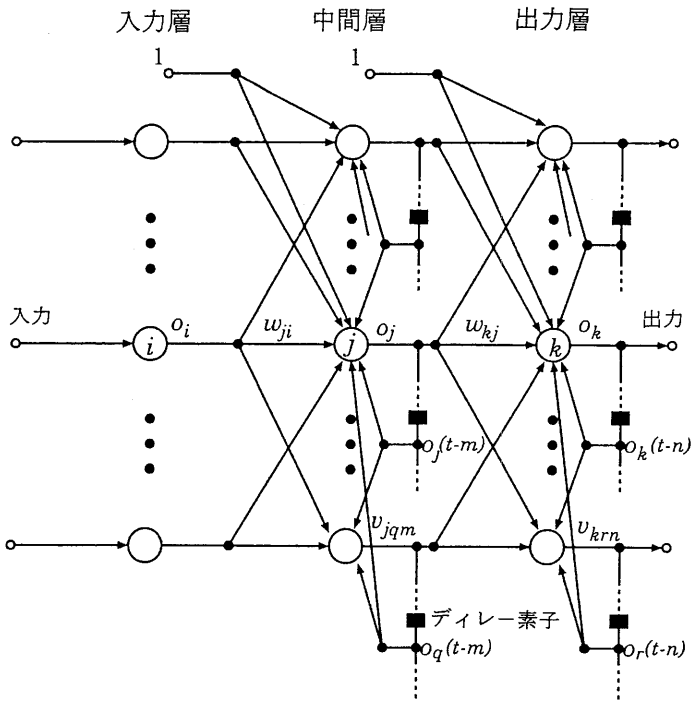

図 1 クロストークリンク付き BPD

結合荷重の修正量は教師信号を $d_k(t)$ として，平均自乗誤差を次式 (3) とし，学習係数を α とすれば，次式 (4) で計算することができる。

$$E(t) = \frac{1}{2}\sum_k \{d_k(t) - o_k(t)\}^2 \tag{3}$$

$$\Delta w_{kj}(t) = -\alpha\frac{\partial E(t)}{\partial w_{kj}(t)} \ , \ \Delta v_{krn}(t) = -\alpha\frac{\partial E(t)}{\partial v_{krn}(t)} \tag{4}$$

3. シミュレーション

図 1 に示した 3 層構造のクロストークリンク付き BPD を用いて，人間の発語「あ」「い」「う」「え」「お」を認識するシュミレーションを行なう。以下に入力データ，認識率の扱い，およびシミュレーション結果を述べる。

3.1 入力データ

音声データは複数の話者が発語した「あ」「い」「う」「え」「お」のデータを話者ごとにランダムに再配置したものを 1 サンプルとした（例：話者A「うえいあお」，話者B「いおあうえ」など)。教師信号は，例えば「あ」の場合は「あ」に対応したニューロンが 0.75，その他のニューロンが 0.25 を出力するように数値を与えた。「い」「う」「え」「お」の場合も同様とした。学習用の入力データは，1 つの入力信号と 5 つの教師信号（識別すべき発語種数に

対応して出力層ニューロンが5個のため）を1組として，複数個の組で構成した。認識用の入力データは，学習に用いた話者とは別の話者の音声データを用い，一語ずつとした（例：話者X「あ」，話者X「い」... ，話者Y「あ」...)。話者は不特定多数の10代後半から20代前半までの男性を集めることにした。最終的に学習用に20人，認識用に10人分のデータを用意することができた。

音声データはMicrosoft Windows3.1付属のサウンドレコーダを用いてWAVEファイルで採取した。サンプリングレートを当初は22.050kHz(解像度16bit)としたが，音声データ「あ」一語でもサンプリング点数が4000個で，ニューラルネットワークの学習処理に2000回学習で22時間という膨大な処理時間が必要とされたため，Creative Technology Ltd.製サウンドボードSound BLASTER 16の付属ソフトであるCreativeWaveStudioを用いてサンプリング点数を半分に間引いて，サンプリングレート11.025kHzのデータに変換した。また，通常の会話中の「あ」一語の発声時間は0.2秒にも満たないが，各時刻に対応するサンプリングデータの数は2000個を越すため，CreativeWaveStudioを使い，「あ」「い」「う」「え」「お」それぞれの波形において，発声直後から0.02～0.03秒以内の部分だけを切り出すことにした。これを3区画分つないで，学習・認識データとした。図2にその一例を示す。一語あたり0.02秒から0.03秒の短いものになったが，人間の耳で聞いて識別可能な範囲であった。これでサンプリング点数が平均200程度のデータ量に抑えることができた[4]。最後に，中山裕基氏製作のWAVEファイラー（WAV-COMP）を用いて，WAVEファイルをテキスト形式のデータに変換した。

3.2 認識率の扱い

図3に認識処理結果の一例を示すように，出力値が教師信号の値(0.25または0.75)に一致することはほとんどない。そこで，認識処理をしたとき，入力で与えた語に対応するニューロンの出力が，他のニューロン出力より大きければ正しく認識したとみなすことにする。図3に示すように，「あ」～「お」に対応するニューロンの出力は時刻と共に変化するので，出力層の各ニューロンごとに最大の出力値を出した時刻数（対応する語ごとにT_a, T_i, T_u, T_e, T_oとする）を求め，以下の方法で認識率を算出することとする。

- 入力した語に対応する出力層ニューロンが最大の出力値を出した時刻数（$T_v : v = a, i, u, e, o$）を，その認識処理をしたときの全時刻数（$T = \Sigma_v T_v$）で割った値をその語の認識率（$R_v : v = a, i, u, e, o$）とする。

$$R_v = \frac{T_v}{T} \qquad : v = a, i, u, e, o$$

- ある話者pの「あ」から「お」までの5つの認識率の平均に100をかけたものをその話者pに対する認識率R_pとする。

$$R_p = \frac{\Sigma_v R_v}{5} \times 100(\%)$$

- また認識用の10人の音声データのそれぞれを，特定のニューラルネットワークnで認識させたときの認識率の平均をそのネットワークnの認識率R_nとする。

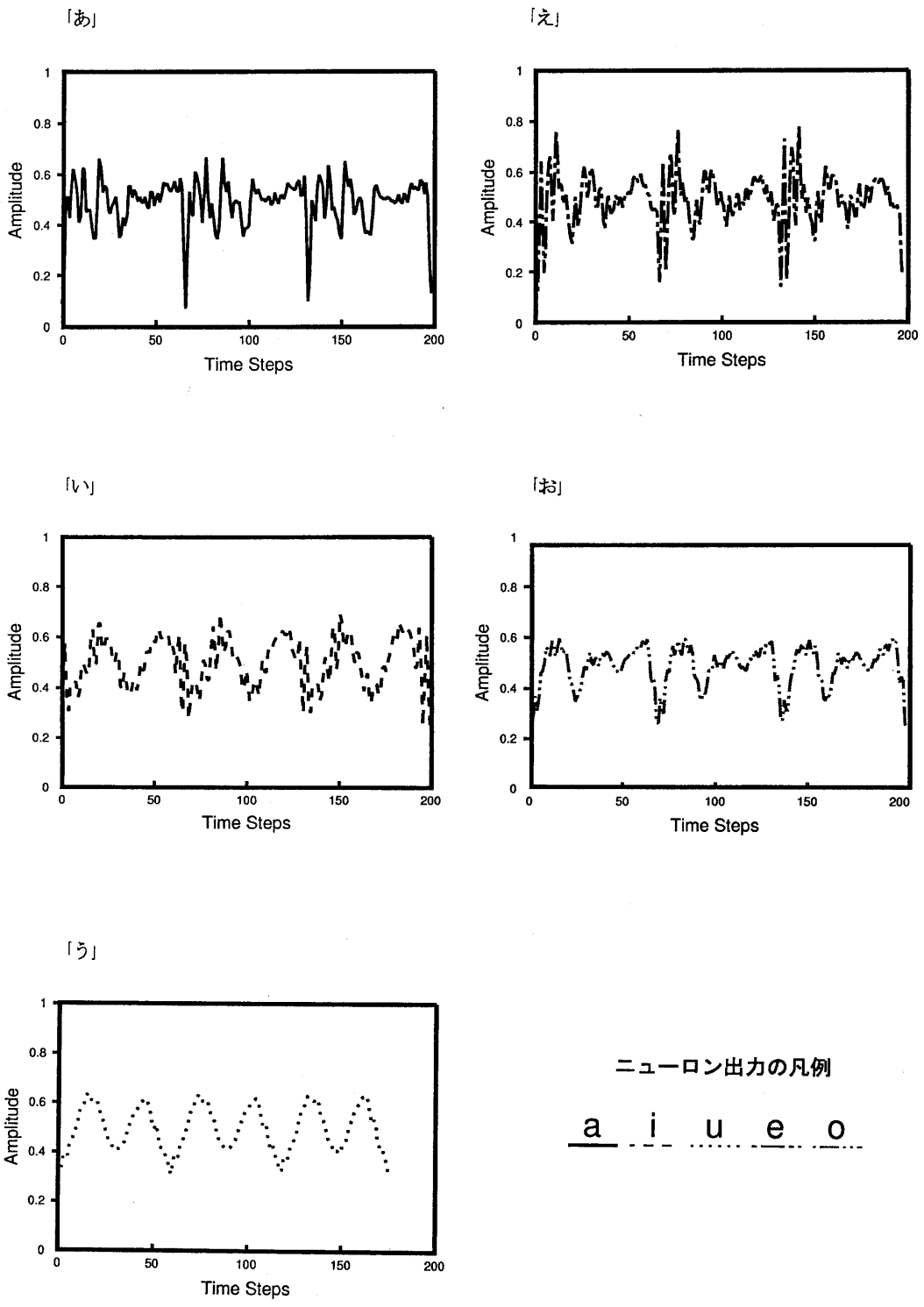
「あ」
「え」
「い」
「お」
「う」
Amplitude
Time Steps
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
ニューロン出力の凡例
a i u e o

図 2 「あ」～「お」の音声波形例

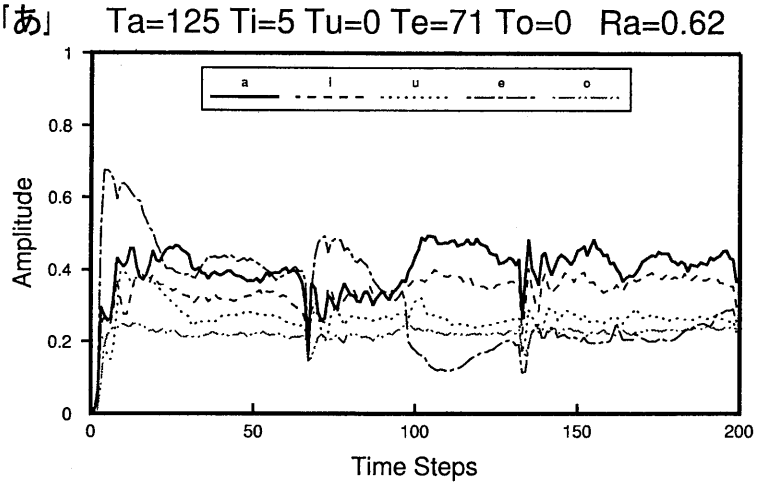

図 3 認識処理結果の例

図 3 の例では，T_a，T_i，T_u，T_e，T_o の値はそれぞれ，125，5，0，71，0 である。従って「あ」に対する認識率は $R_a = 0.62$ と求められる。この話者における「あ」以外の認識率は $R_i = 0.85$，$R_u = 0.72$，$R_e = 0.82$，$R_o = 0.51$ であった。よって，この話者 p に対する認識率は $R_p = 70.5\%$ である。

3.3 シミュレーション結果

まず，ネットワークのユニット数とディレー素子段数を決定し，次に認識率をシミュレーションにより評価検討する。

入力データに話者 1 の発語した「あ」～「お」を用い，初期値をランダムに与えて学習を 2000 回行う。これを 10 篇行って，それぞれの最小達成誤差値を平均した値をそのネットワークの誤差値とする。入力層のニューロン数を i，中間層でのニューロン数とリカレントリンク段数とクロストークリンク段数をそれぞれ j,m_r,m_c とし，出力層でのニューロン数とリカレントリンク段数とクロストークリンク段数をそれぞれ k,n_r,n_c として，BPD ニューラルネットワークの構成を $i-j(m_r)(m_c)-k(n_r)(n_c)$ と表すことにする。

まず，$1-j(4,3)-5(4,3)$ を初期の構成として中間層でのユニット数 j を 1，2，3，4，5，10，15，20 とした場合で学習させた。その場合の誤差を図 4 に示す。図 4 により，j が 3 のとき誤差値は最小になるので，$j = 3$ が適切であると判断した。他のパラメータの確定後，$1-j(0,1)-5(95,5)$ として再び j の値を変えて誤差を求めたところ，結果は同じく $j = 3$ のときに最小の誤差になった。$1-3(m_r,3)-5(4,3)$ の構成で中間層でのリカレントリンク段数 m_r を 0，1，2，3，7，10，16，20 として学習させた場合の誤差を図 5 に示す。図 5 により，$m_r = 0$ が適切であると判断した。$1-3(0,3)-5(n_r,3)$ の構成で出力層のリカレントリンク段数 n_r を 0，1，3，10，20 として学習させた場合の誤差を図 6 に示す。図 6 により，$n_r = 95$ が適切であると判断した。他のパラメータの確定後，$1-3(0,1)-5(n_r,5)$ として再び n_r の値を変化させて誤差を求めたところ，結果は同じく $n_r = 95$ のとき誤差は最小になっ

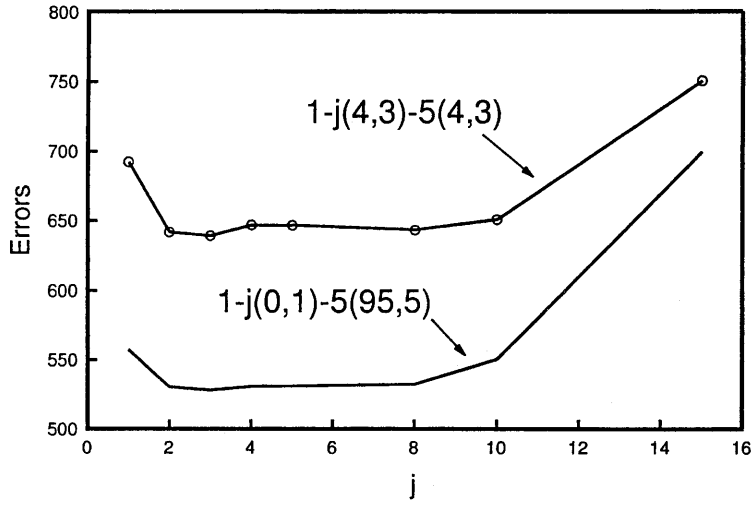

図 4　中間層のニューロン数 (j) と誤差

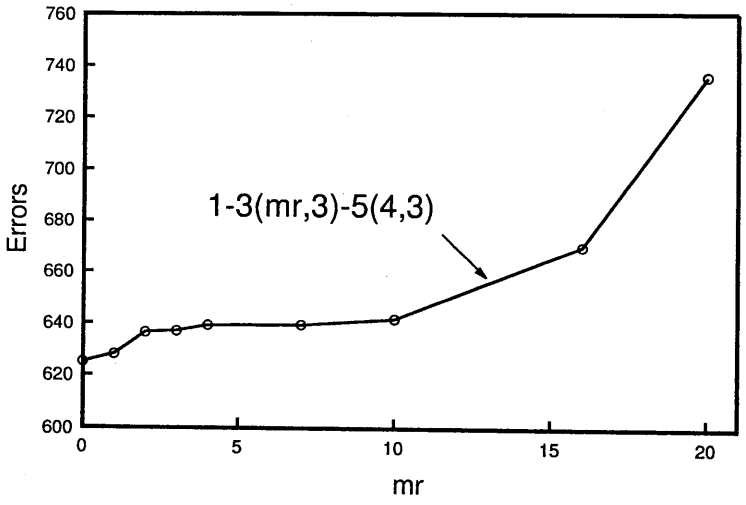

図 5　中間層のリカレントリンク段数 (m_r) と誤差

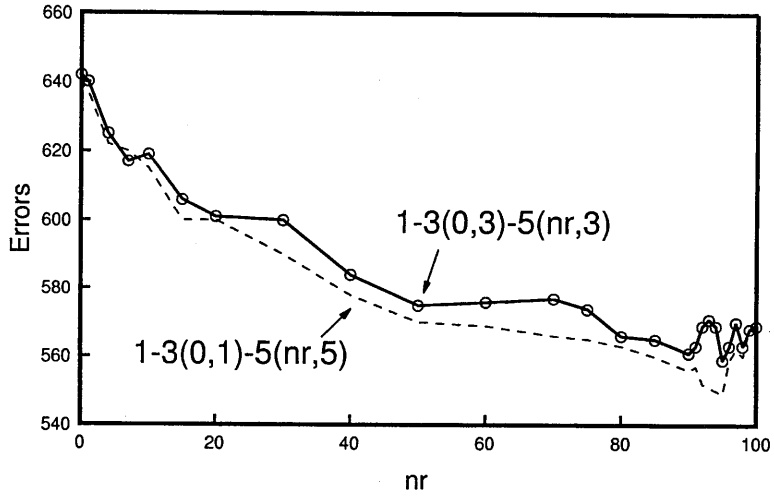

図 6 出力層のリカレントリンク段数 (n_r) と誤差

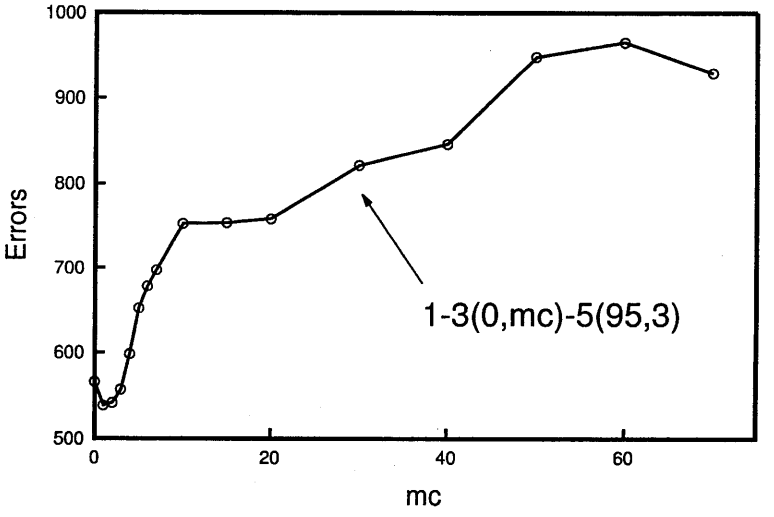

図 7 中間層のクロストークリンク段数 (m_c) と誤差

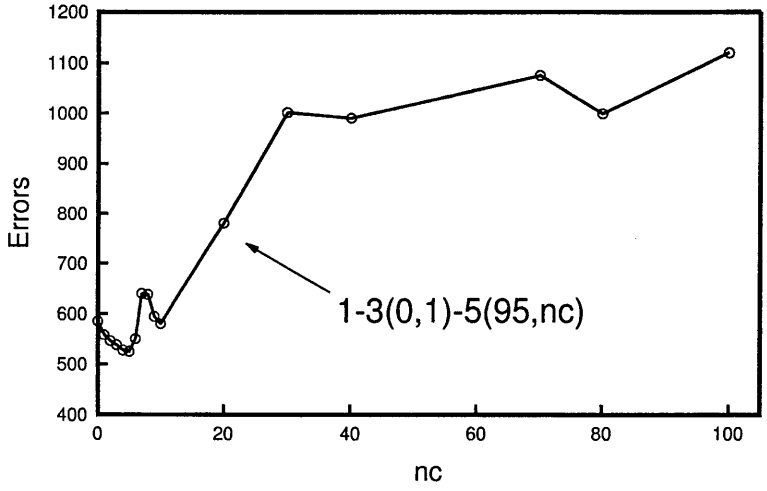

図 8 出力層のリカレントリンク段数 (n_c) と誤差

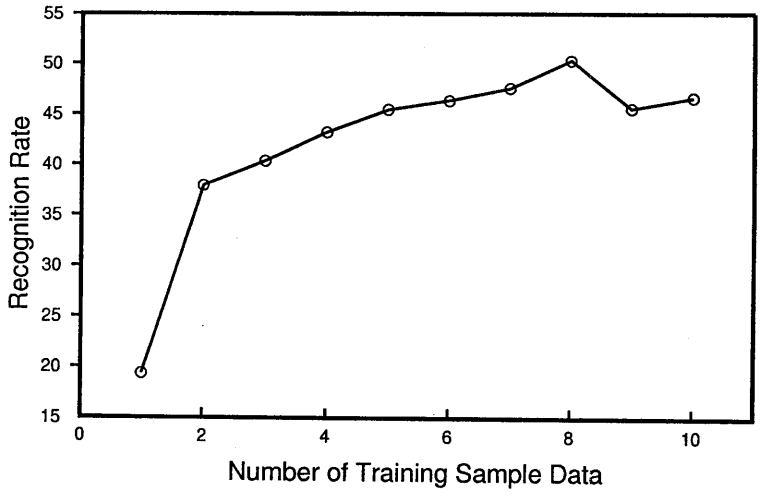

図 9 認識率の推移

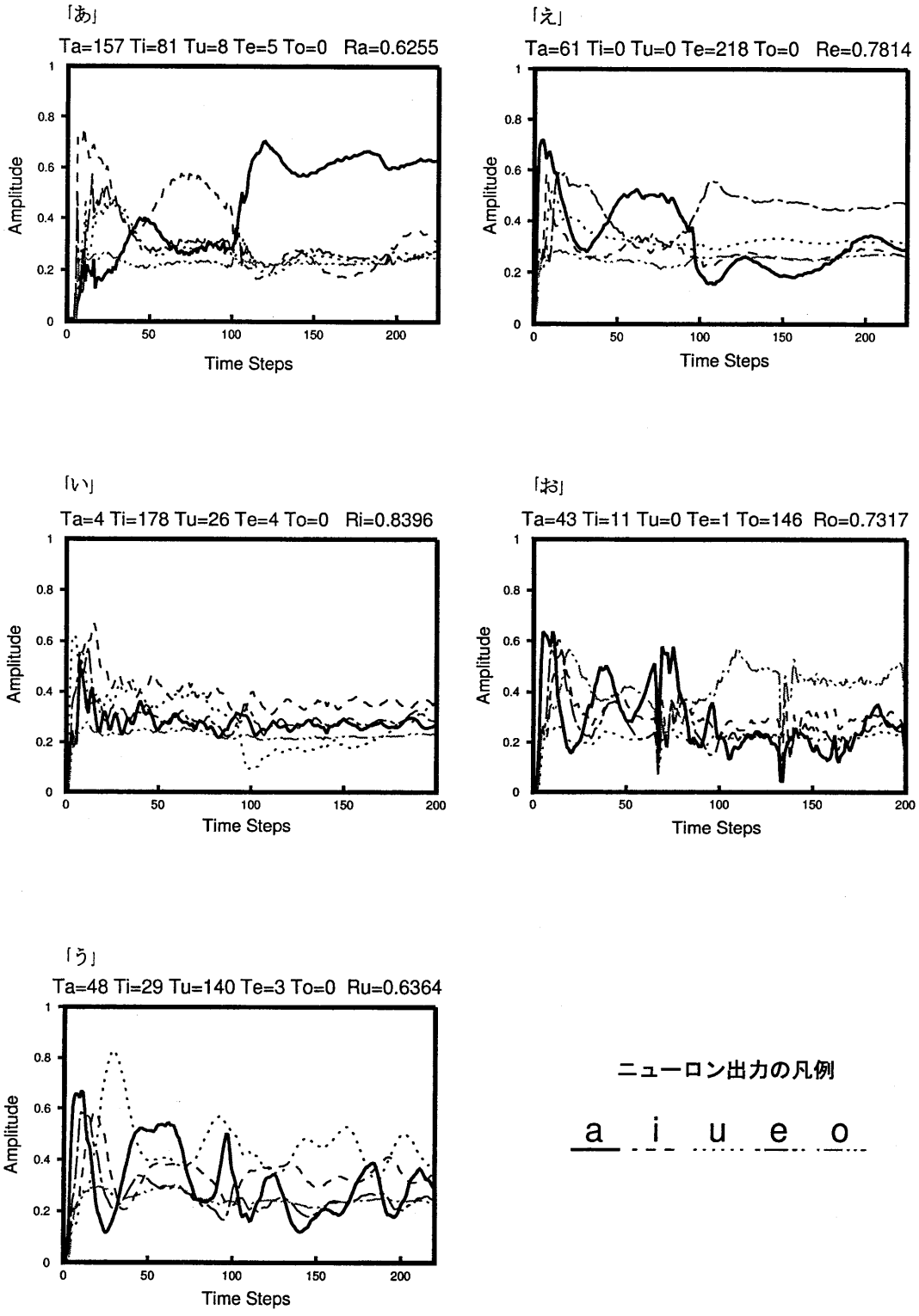

図 10 話者 10 の認識時の処理出力例（$R_{10} = 72.29\%$）

た。$1-3(0,m_c)-5(95,3)$ の構成で中間層のクロストークリンク段数 m_c を 0, 1, 2, 4, 5, 6, 7, 10, 15, 20, 30, 40, 50, 60, 70, 80 として学習させた場合の誤差を図 7 に示す。図 7 により，$m_c=1$ が適切であると判断した。$1-3(0,1)-5(95,n_c)$ の構成で出力層のクロストークリンク段数 n_c を 0, 1, 2, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 70, 80 として学習させた場合の誤差を図 8 に示す。図 8 により，$n_c=5$ が適切であると判断した。従って，本論文では音声パタン認識用の BPD ニューラルネットワーク構成として，$1-3(0,1)-5(95,5)$ を扱うことにする。

このネットワーク $1-3(0,1)-5(95,5)$ に，学習用に用意した 20 人分の音声データを順次入力して 2000 回ずつ学習して 1 人分の学習が終了する毎に認識率を評価した。その誤差の推移を 10 人分について図 9 に示す。学習データを 10 人分以上に増やしても認識率はほとんど向上しなかった。認識率は同じネットワークを用いても，話者により大きく認識率は変化した。最良の時点でのこのネットワークの話者ごとの認識率は，$R_1=70.5\%$，$R_2=63.83\%$，$R_3=53.6\%$，$R_4=65.99\%$，$R_5=40.01\%$，$R_6=61.71\%$，$R_7=50.45\%$，$R_8=51.55\%$，$R_9=50.79\%$，$R_{10}=72.29\%$ であった。これらを平均して求めたネットワーク自体の認識率は 58.07% となった。図 10 にこのネットワークを用いて話者 10 の音声データを認識処理させた場合の出力例を示す。図 10 は最も認識率の高かった話者についてのもので，72.29% が達成されている。図 10 において時間軸の $\frac{1}{3}$ の時点と $\frac{2}{3}$ の時点で，特に「お」に顕著に表れているが，ネットワークの出力が振動しているのが確認できる。これは音声データを抽出する際，各発語を切りとって 3 区画分で並べたためで，その切れ目の部分にニューラルネットが反応を示しているためだと思われる。

4. むすび

クロストークリンクを備えたディレー素子内包型ニューラルネットワーク BPD を用いて音声認識を行ない，BPD の有効性を検討した。その結果，認識率は話者の個人差でかなり大きくばらつくが，クロストークリンクを備えた BPD で 70% 程度の音声認識ができる感触を得られた。今後，情報教育に応用して行くことが可能になると思われる。

人間の発する音声とは，音響工学の見地から有声／無声，破裂性，フォルマント周波数（声道の共鳴周波数），ピッチ周波数（声の高さを表す），振幅などの諸情報が複雑に絡み合ったものとして定義される [5]。今回はこれらの諸情報について解析することはせず，単純に音声の波形のみを入力データとして用い，ニューラルネットワークの能力を検討した。今後は音響工学面からの解析も試みる必要があると考える。また，時系列処理のできる他のニューラルネットワークとの比較検討もしていきたい。

参考文献

1) 中野馨, 飯沼一元, “入門と実習ニューロコンピュータ”, 技術評論社, 1989.

2) 西正明, 降矢順治, 滝沢寛之, 中村維男, “ニューラルネットワーク（クロストークリンク付き BPD）の FSK 復調への応用,” 日本産業技術学会教育学会誌, 第 41 巻 1 号, pp.9–16, 1999.

3) 西正明, 降矢順治, 中村維男, “ディレー素子内包型バックプロパゲーションニューラルネットワーク（BPD）の一構成,” 信学論 (D-II), Vol.J78-D-II, No.10, pp.1522–1530, 1995.

4) 雨宮好文, 佐藤幸男, “信号処理入門”, オーム社, 1987.

5) 中川聖一, “確立モデルによる音声認識”, 電子情報通信学会, 1988.