TR-I-0214

対判定型ニューラルネットワークの原理と時間遅れ神経回路網との統合による ロバストな音素認識

鷹見 淳一　嵯峨山 茂樹
JuNichi TAKAMI　Shigeki SAGAYAMA

1991.5.1

概 要

本稿では、分類型ニューラルネットワークの頑健性の向上を図った対判定型ニューラルネットワークの原理、およびこの原理と時間遅れ神経回路網(TDNN)とを統合した対判定型TDNNによる音素認識手法について述べる。従来より、分類型ニューラルネットワークを用いた音声認識では、ネットワークが過剰に学習されやすいことや、未学習データに対する汎化能力が低い等の理由から、学習データ以外の発話様式の音声データに対して認識率が大きく低下するという問題があった。我々はこのような問題を解決するために、各カテゴリ間における緩やかな識別境界面の形成と、1つのカテゴリに対して異なる判定基準を持つ複数の識別境界面の形成を同時に実現し、複数の識別境界面での多数決によりカテゴリの識別を行うという対判定型ニューラルネットワークの原理を提案する。さらにこの原理とTDNNとを統合した対判定型TDNNによる音素認識実験を行い、この手法の有効性を示す。単語発声データで学習したネットワークを用いて、各種発話様式のデータ中の音素認識実験を行ったところ、連続音声に対して、/b,d,g,m,n,N/の6音素については81.6%(7フレーム入力の一括判定型TDNNと比べて3.8%の向上)、18子音については76.8%(15フレーム入力のモジュール構成型TDNNと比べて20.2%の向上)という高い認識率が得られた。

ATR Interpreting Telephony Research Laboratories
ATR自動翻訳電話研究所

目　　次

1. はじめに 3

2. 対判定型ニューラルネットワークによるパターン認識の原理 5

2.1 対判定型ニューラルネットワークの概要 5

2.2 対判定型ニューラルネットワークの学習メカニズム 5

2.3 対判定型ニューラルネットワークの認識メカニズム 7

3. 対判定型ニューラルネットワークの原理とTDNNの統合 8

4. 対判定型TDNNによる音素認識実験 9

4.1 実験条件 9

4.2 音声資料 9

4.3 入力パラメータ 9

4.4 認識実験1 (/b,d,g,m,n,N/の認識) 10

4.4.1 ネットワークの学習 10

4.4.2 比較実験1 (中間値学習を行わない対判定型TDNNによる方法) 10

4.4.3 比較実験2 (従来のTDNNによる認識) 10

4.4.4 実験結果 10

4.4.5 シフト･トレラント性の確認 11

4.5 認識実験2 (18子音の認識) 12

4.5.1 ネットワークの学習 12

4.5.2 実験結果 12

5. 対判定型TDNNの認識性能の考察 13

5.1 中間値を用いた学習の効果 13

5.2 対判定型TDNNの尤度表現能力 14

6. まとめ 15

謝 辞 15

参考文献 16

図･表 18

1. はじめに

ニューラルネットワークの効率的な学習法である逆伝搬学習法[1](バックプロパゲーション)の登場以来、ニューラルネットワークはさまざまな分野に応用され、多くの報告によってその有効性が示されている。さらに近年のコンピュータのめざましい進歩や、バックプロパゲーションの高速化手法[2]の提案などにより、大規模なネットワークや複数のネットワークを用いるアプローチが可能となった。

音声認識の分野においてもニューラルネットワークの応用が盛んに研究され、多くの研究者たちの関心を集めている。その中でも1987年にA.Waibelによって提案された時間遅れ神経回路網[3][4](Time-Delay Neural Network:TDNN)は、ユニット間の結合重み係数(link-weight)の結び(tied-connection)によって、音素抽出位置の違いから生じる入力パターンの時間軸方向のずれを吸収する能力を持ち、単語発声データ中の音素に対して、非常に高い認識性能を得られることが示されている。

しかしその一方で、TDNNに代表されるような分類型ニューラルネットワーク(classification type neural network:入力データの属するカテゴリの情報を教師信号として与えて学習し、認識時には出力値から直接そのカテゴリ候補を決定することを目的に形成されたネットワーク)では、その学習に用いるデータとは性質の異なるデータ(発話様式の異なる音声データ、あるいは別の話者よって発声された音声データ等)に対してその認識性能が大きく低下し、第1位の認識率のみでなく、第2位以下の認識率の低下にもつながる致命的な誤認識が起き易いという、いわゆるネットワークの頑健性(robustness)の問題が多く取り上げられるようになった[5][6][7][8][9][10][11]。これは、ネットワークが学習用データに対して過度に適応し、データの変動に対して敏感に反応するような急峻な識別境界面が形成されることや、音声空間全体をカバーしうるだけの十分な学習用データが得られないために、未学習のデータに対するネットワークの出力値が望んでいる出力値とはかけ離れたものになることに起因すると考えられる。このうち、前者は過剰学習(over-training)の問題として、また後者は汎化(generalization)の問題として、互いに密接に関連し合いながら、ニューラルネットワークにおける1つの大きな研究対象分野を形成している。

これらの問題に対しては、次のような対策がすでに提案されている。

○ 学習時における対策例

- 教師信号の変更
  教師信号を[0,1]より狭い範囲に設定し、過学習を回避する。
- 学習用データへのノイズの混合[5]
  学習用データにノイズを重畳し、見かけ上のデータ数を増やす。
- KNIT法[6]
  点から点への写像でなく、線分から線分への写像をネットワークに教える。
- ファジー学習法[7]
  学習用データの尤らしさを予め算出し、それを教師信号として使用する。
- 新しい評価関数の導入[8]
  識別境界面の平均曲率最小化という評価基準を導入し、過剰学習を回避する。

○ 認識時における対策例

● 近傍情報による出力値の平滑化[9]

入力層や中間層での値をベクトルとみなし、この近傍でネットワークの出力を積分することにより出力値を平滑化する。

以上の方法は、いずれも学習時あるいは認識時のどちらか一方での対策により頑健性の向上を図ったものである。そこで今回我々は、

[学習時]	出力値に認識候補の尤らしさを反映させることができるような、比較的緩やかな識別境界面を持つネットワークを形成するためのメカニズムの導入。
[認識時]	それぞれ異なる判定基準を持つ複数のネットワークでの多数決による、未知入力データに対する出力値のばらつきの平滑化。

という、学習時および認識時それぞれにおける対策を同時に実現することが有効であると考えた。

そしてこの2つの対策を実現するために、複数カテゴリの音素認識問題を2カテゴリ間の対立問題に分割して解く[12]という、対判定型ニューラルネットワークの原理を提案する[13][14][15][16]。

本稿では、対判定型ニューラルネットワークの原理と、この原理をTDNNと統合した対判定型TDNNの概要、およびこれを用いた音素認識実験とその結果について示し、対判定型TDNNの認識性能に対する検討を行う。

2. 対判定型ニューラルネットワークによるパターン認識の原理

2.1 対判定型ニューラルネットワークの概要

全認識対象カテゴリの中から任意に選択した2つの異なるカテゴリc_i、c_jをカテゴリ対と呼び、$(c_i:c_j)$と表すことにする。対判定型ニューラルネットワークとは、入力データがc_iに属する場合には1、c_jに属する場合には0、そのどちらでもない場合は0.5をそれぞれ出力するように学習された、出力層上にただ1つのユニットを持つニューラルネットワークである。このようなネットワークを$\mathrm{Net}(c_i/c_j)$と表すことにする。$\mathrm{Net}(c_i/c_j)$の構造を図1に示す。

ある入力データに対する$\mathrm{Net}(c_i/c_j)$の出力値は、そのデータを2カテゴリ間の対立という観点から見たときの、c_iに対する対判定スコアであると考えることができる。さらに、$\mathrm{Net}(c_i/c_j)$におけるc_iとc_jの出力値の関係は、0.5をはさんで互いに対称であると考えられる。そのため、$\mathrm{Net}(c_i/c_j)$の出力値を1から引いた値はc_jに対する対判定スコアとみなせる(図2)。

ここで、出力ユニット内の非線形関数として通常のシグモイド関数を用いた場合、学習時における教師信号と実際のネットワーク出力の間の誤差を収束させるまでに、膨大な計算時間が必要となる。これは、シグモイド関数では出力が0.5となる位置で微分係数が最も大きく、0.5という教師信号を持つ学習用データにおける誤差に対して感度が高くなるため、全体としての平均誤差が振動してしまうことに原因があると考えられる。そこで0.5の出力が要求されるデータに対して学習時の収束性を向上させ、かつ認識時の頑健性を増すために、出力ユニットについてのみ式(1)で示される非線形関数を導入する。

$$f(x)=\begin{cases}\dfrac{g(x+\alpha)}{2g(\alpha)} & (x<0)\\[2ex] 1-\dfrac{g(-x+\alpha)}{2g(\alpha)} & (x\geqq 0)\end{cases} \qquad (1)$$

但し、$g(x)=\dfrac{1}{1+e^{-x}}$

この概形を図3に示す。ここでαは、出力が0.5となる位置の微分係数の大きさをコントロールするためのパラメータである。αを大きく設定する程、0.5の出力が要求されるデータに対する出力ユニット上での分布の許容幅は広がり、またαを0とすると通常のシグモイド関数に一致する。

2.2 対判定型ニューラルネットワークの学習メカニズム

通常の分類型ニューラルネットワークの学習では、学習用データがある特定のカテゴリに属するか否かによって、1または0という2通りの教師信号が用いられる。ここで、1の教師信号を持つ学習用データの集合をS_1、0の教師信号を持つ学習用データの集合をS_0とする。

ネットワークの学習がある程度進行し、S_1のすべての要素に対する出力値が0.5より大きく、S_0のすべての要素に対する出力値が0.5より小さくなった場合を考える。このときの出力ユニット上でのS_1およびS_0の分布は図4のようになっていると考えられる。(この時点で、学習用データに対する識別率は100%である。)

この時点において、出力ユニット内のシグモイド関数の傾斜を急峻にすることによって、つまり、

$$g(x)=\frac{1}{1+e^{-kx}} \tag{2}$$

において、kの値を1より大きくすることによって、各データに対する誤差を確実に減らすことができる。式(2)は、見かたを変えれば、シグモイド関数の傾斜は一定のまま、出力ユニットへの入力値xをk倍したものと考えることができる。つまり、

$$g(x)=f(kx)\ ,\quad f(x)=\frac{1}{1+e^{-x}} \tag{3}$$

となる。さらに、出力ユニットへの入力値xは、その前の中間層上のj番目のユニットからの出力値y_jとその間の結合重み係数w_jにより、

$$x=\sum_j w_j y_j \tag{4}$$

と表すことができる。従って、出力ユニットへの入力値xをk倍することは、その間の結合重み係数w_jをk倍することと等価であり、実際のバックプロパゲーションにおいては、このw_jをそれらの間の比率の関係を保ったまま増加(w_jが負の場合には減少)させることによって、全学習用データに対する平均誤差を幾らでも減らすことが原理的に可能である。(実際には極小解に陥り、学習がそれ以上進行しない場合もある。)

このように、ネットワークの出力値として1または0のみを要求したような学習では、各カテゴリ間に形成される識別境界面は急峻になりやすく、ほとんどの入力データに対する出力値が出力ユニット内のシグモイド関数の飽和領域(微分係数が十分に小さい領域)における写像により実現されるという傾向が強い。

一方、対判定型ニューラルネットワークで行われている、1や0の他に0.5という教師信号を用いた場合の学習を考える。この場合、S_1、S_0に加えて、0.5の教師信号を持つ学習用データの集合$S_{0.5}$が存在することになる。

ここで、ネットワークの学習がある程度進行し、ある時点での出力ユニット上でのS_1、S_0および$S_{0.5}$それぞれの分布が図5のようになったとする。このとき、S_1およびS_0に対しては、上記の理由から、識別境界面は急峻なほど各データに対する誤差は小さくなるが、逆に$S_{0.5}$に対しては誤差が大きくなってしまう。このために、識別境界面の傾斜は急峻になり過ぎることなく、適当な値に落ち着くことが予想される。

このように、対判定型ニューラルネットワークの学習法は、急峻な識別境界面の形成を抑制する働きを持つ0.5という教師信号の導入により、過剰学習を回避するメカニズムを持っているといえる。

2.3 対判定型ニューラルネットワークの認識メカニズム

N個のカテゴリを持つ認識問題を解くために、${}_NC_2$個のカテゴリ対それぞれに対する対判定型ニューラルネットワークが形成される。このうち、ある特定のカテゴリc_iをカテゴリ対の一方に含んでいるようなカテゴリ対の集合をC_iと表す。つまり、$C_i=\{(c_i:c_j)|1\leqq j\leqq N, j\neq i\}$であり、要素数は$N$-1個である。

ここで、C_iに属するN-1個のカテゴリ対それぞれに対応する対判定型ニューラルネットワークNet(c_i/c_j)を考える。これらN-1個のネットワークは、すべてカテゴリc_iについての識別を行うためのものであるが、c_iと対をなすもう一方のカテゴリが異なっているために、各ネットワークにおけるカテゴリc_iの識別境界面は、それぞれ異なる判断基準に基づいて形成されるものと考えられる。そのために、未学習のデータに対する出力結果も各ネットワークでは違ったものになる。従って、カテゴリc_iの認識の際に、これらN-1個の識別境界面による識別結果をすべて用いて、多面的な観点から総合的に判断することによって、未学習のデータに対する個々のネットワーク出力のばらつきは平滑化され、精度の高い認識が可能になると考えられる。

カテゴリc_iに対するこのような総合判定スコア$T(c_i)$は、N-1個のネットワーク出力の平均として、式(5)により求めることができる。

$$T(c_i)=\frac{1}{N\text{-}1}\sum_{j\neq i}S(c_i|c_i:c_j) \tag{5}$$

ここで、$S(c_i|c_i:c_j)$はカテゴリc_iとカテゴリc_jの対立の観点から見たときの、カテゴリc_iに対する対判定スコアであり、

$$S(c_i|c_i:c_j)=\begin{cases}\mathrm{Out}(c_i/c_j)\\ \quad\text{or}\\ 1\text{-}\mathrm{Out}(c_j/c_i)\end{cases} \tag{6}$$

但し、Out(c_i/c_j)はNet(c_i/c_j)からの出力値を示す。

により得られる値である。

以上のように、対判定型ニューラルネットワークによる認識処理は、個々のネットワークにおける出力結果のばらつきを平滑化するメカニズムを持っているといえる。

3. 対判定型ニューラルネットワークの原理とTDNNの統合[13][14][15][16]

1987年にA.Wibelによって考案された時間遅れ神経回路網[3][4](Time-Delay Neural Network: TDNN)は、学習用データと同種のデータに対して、非常に高い識別性能を持っていることが知られている。しかしこのTDNNも分類型ニューラルネットワークの一種であり、他の分類型ニューラルネットワークと同様に、過剰学習の問題や汎化能力の問題を避けて通ることができない。そのため、実際の音素認識の場面でも、単語発声データ中の音素で学習したネットワークを用いて文節発声データ中の音素や連続発声データ中の音素を認識しようとした場合、その認識性能が大きく低下してしまうということが最近の研究によって明らかになってきた[10]。

そこで今回、対判定型ニューラルネットワークの原理とTDNNを統合することにより、対判定型ニューラルネットワークのメカニズムを活かしながらTDNNの優れた識別能力を用いて、高性能の音素認識を行うことを試みた。こうして統合されたネットワークを"対判定型TDNN (Pairwise Discriminant TDNN:PD-TDNN)"と呼ぶ。

対判定型TDNNは従来のTDNN同様、ユニット間の結合重み係数にtied connectionの構造を持っており、音素切り出し位置の違いにより生じる入力パターンの時間軸方向のずれを吸収する能力を保持している一方で、対判定型ニューラルネットワークの特徴である、ただ1つの出力ユニットと、その中に設けられた式(1)で表される非線形関数を持ったネットワークである。この構造を図6に示す。

なお従来のTDNNでは、入力フレーム数を15としていたが、これを7にした方が音素中心部での認識率は向上することが予備実験により確認でき、またネットワークの小規模化により計算時間も短縮できることから、今回は入力フレーム数を7としている。

4. 対判定型TDNNによる音素認識実験

4.1 実験条件

対判定型TDNNの音素認識性能を評価するために、日本語子音の認識実験を行った。今回の実験では、誤りを起こしやすいといわれている/b,d,g,m,n,N/の6子音、および/b,d,g,p,t,k,m,n,N,s,sh,h,z,ch,ts,r,w,y/の18子音の2通りのタスクを用いた。どちらのタスクに対しても、図6に示した構造のネットワークを使用した。また出力ユニット内の式(1)で表される非線形関数におけるαの値は、これをどのくらいの大きさに設定した場合にカテゴリ対間の分離度が最も高くなるかを調べた簡単な予備実験により、$\alpha=3.0$として用いた。

4.2 音声資料

今回の認識実験には、日本人男性話者1名(MAU)により発声され、サンプリング周波数12kHz、16bit量子化により収録された音声データベース中の音声を使用した[17][18]。ネットワークの学習および認識実験には、それぞれ以下に示す音声中から、視察により付けられているラベル情報に基づいて切り出した音素データを用いる。

[ネットワークの学習]
- 孤立発声された重要語5240単語のうちの偶数番目(5.68モーラ/秒)。

[認識実験]
- 孤立発声された重要語5240単語のうちの奇数番目(単語発声、5.68モーラ/秒)。
- 国際会議参加問い合わせに関する会話文を、文節単位で区切って発声したもの(文節発声、7.72モーラ/秒)。
- 国際会議参加問い合わせに関する会話文を、複合語を許さない文節単位で区切って発声したもの(短い文節発声、7.14モーラ/秒)。
- 国際会議参加問い合わせに関する会話文を、区切り指定を行わずに発声したもの(自由発声、9.56モーラ/秒)。

4.3 入力パラメータ

ネットワークへの入力パラメータとしては、各音声中から、音素セグメントの終端が入力フレームの中央に位置するように切り出された音素データから算出された7フレーム分(フレーム周期10ms)の16チャンネルメルスペクトラムを用いた。

入力パラメータの算出は以下の手順で行う。まず、分析周期5ms毎に256点のハミング窓を使用してFFTを行い、16チャンネルのメルスペクトラムに変換する。次に2フレーム毎にその値を平均し、最後に7フレーム分のデータに対して、平均値が0、絶対値の最大が1となるよう正規化を行う。こうして得られた7フレーム分の16チャンネルのメルスペクトラムを1サンプルとし、ネットワークの学習および認識実験に用いる。

4.4 認識実験1 (/b,d,g,m,n,N/の認識)[13][14][15]

4.4.1 ネットワークの学習

カテゴリ数が6個の場合、15個(${}_6C_2$)の対判定型TDNNが必要となる。個々のネットワークをNet(p_i/p_j)と表したとき、Net(p_i/p_j)の学習は、カテゴリp_iに属するデータには1、カテゴリp_jに属するデータには0、そのどちらにも属さないデータには0.5を教師信号として与え、バックプロパゲーションにより行う。学習用データ数は各カテゴリ毎に最大500個とし、そのときの学習用データに対する平均誤差が0.01以下になるまで学習を繰り返した。この結果、学習用データに対する第1位の認識率は98.8%となった。

4.4.2 性能比較実験1 (中間値学習を行わない対判定型TDNNによる方法)

対判定型TDNNでは、カテゴリ対のどちらにも属さないデータに対しては、0.5という中間値を教師信号として与えて学習している。そこでこのような中間値学習の効果を調べるために、どちらのカテゴリにも属さないデータは用いず、カテゴリ対に属する2カテゴリのデータのみに対し、1および0の2通りの教師信号を用いて学習したネットワークによる音素認識実験を行った[14]。中間値学習を省略したために、出力ユニット内の非線形関数として従来通りのシグモイド関数を用いている以外、ネットワーク構造や認識法等は通常の対判定型TDNNと全く同じである。

4.4.3 性能比較実験2 (従来のTDNNによる認識)

従来のTDNNによる認識手法との性能の比較を行うために、図7に示すような構造のネットワーク(一括判定型TDNN)を用いた場合の実験も合わせて行った[13][14]。なお入力フレーム数は、対判定型TDNNと同じ7フレームとした。このネットワークの学習は、対判定型TDNNの学習に用いたものと同じデータを使用して、正解カテゴリに対応する出力ユニットのみに1を与え、その他のユニットには0を与えるという従来通りの方法により行った。なお今回の学習では、全学習用データに対する平均誤差が0.005となったとき、学習用データに対する認識率が、対判定型TDNNを用いた場合とほぼ同等となったため、この時点で学習を打ち切った。このときの学習用データに対する第1位の認識率は98.9%であった。

4.4.4 認識実験結果

4通りの発話様式のデータに対する認識実験結果を表1に示す。また、性能比較のために行った、中間値学習を省略した場合の認識実験結果、および従来のTDNNによる認識実験結果をそれぞれ表2、表3に示す。また、これらを発話様式毎にまとめたグラフを図8の(a)~(d)に示す。

まず中間値学習を行わない対判定型TDNNを用いた場合の結果と従来の一括判定型TDNNを用いた場合の結果を比較する。第1位認識率では従来の一括判定型TDNNの方が

文節発声データで4.2%、短い文節発声データで7.3%、自由発声データで4.6%それぞれ中間値学習を行わない対判定型TDNNを用いた場合に比べて高い認識率を示しているが、第3位までの累積認識率では中間値学習を行わない対判定型TDNNを用いた場合の方が文節発声データで1.6%、短い文節発声データで2.7%、自由発声データで4.1%それぞれ高い認識率を示している(単語発声データに対しては両者の間の差はほとんど無い)。この結果より、複数のネットワークによる多数決が累積認識率向上のために有効であると考えられる。さらに0.5という中間値も教えた対判定型TDNNを用いた場合の結果を見ると、これはすべての発話様式のデータに対して、第1位認識率、第3位までの累積認識率共に最も高い値を示しており、自由発声データに対しては一括判定型TDNNに比べて第1位認識率で3.8%の向上、第3位までの累積認識率で4.5%の向上がみられた。

以上の結果より、複数のネットワークによる多数決は、累積認識率の向上に有効であることが確認でき、さらにこのような複数のネットワークによる多数決に加えて、個々のネットワーク内部に形成される識別境界面を緩やかにし、音素対のどちらにも含まれないカテゴリのデータに対する出力値を0.5付近に集中させる効果を有する中間値学習を行うことで、第一位の認識率の向上も実現できることがわかった。

4.4.5 シフト・トレラント性の確認

従来のTDNNでは、入力として16チャンネルのメルスペクトラム15フレーム分を使用しており、この条件の下で高いシフト・トレラント性が得らることが確認されている。しかしここで使用している対判定型TDNNでは、入力フレーム数を7としている。そこで、入力フレーム数7という条件下でのシフト・トレラント性を確認するため、単語発声データに対する音素切り出し位置を、通常の切り出し位置(音素セグメントの終端を入力フレームの中央に置く)に対して-20msから+20msまで10ms毎にシフトして得られるデータを用いて、対判定型TDNNを用いた場合と一括判定型TDNNを用いた場合との認識率の比較を行った[13]。この結果を表4に示す。

この結果より、対判定型TDNNは±10ms程度のずれに対しては十分な吸収能力があり、±20ms程度のずれに対しても一括判定型TDNNを上回る認識率が得られることがわかった。

4.5 認識実験2 (18子音の認識)[16]

4.5.1 ネットワークの学習

カテゴリ数が18個の場合、153個($_{18}C_2$)の対判定型TDNNが必要となる。学習用データ数は、p_iおよびp_jに属するデータはそれぞれ最大500個、どちらにも属さないデータはそれぞれ最大100個を使用して、そのときの学習用データに対する平均誤差が0.01以下になるまで学習を繰り返した。

4.5.2 認識実験結果

4通りの発話様式のデータに対する認識実験結果を表5に示す。また性能比較のために、従来より18子音の認識に用いられてきた図9に示すような構造のネットワーク[19](15フレーム入力のモジュール構成型TDNN)を使用した場合の認識率[19]を表6に示す。さらに両者の認識率を発話様式毎にまとめたグラフを図10の(a)~(c)に示す。

この結果より、従来のモジュール構成型TDNNは、単語発声データに対しては非常に高い認識率を示しているが、短い文節発声データ、自由発声データと発話速度が速くなるにしたがって、その認識率は大きく低下している。

一方対判定型TDNNでは、単語発声データについては従来法に比べて第1位認識率は2.3%低いが、第3位までの累積認識率では逆に0.2%上回っている。さらに短い文節発声データに対しては第1位認識率で9.9%低いが、第3位までの累積認識率で5.6%、また自由発声データに対しては第1位認識率で20.2%低いが、第3位までの累積認識率で14.4%それぞれモジュール構成型TDNNを上回っており、と比較して、全体的に高い認識率を示しており、発話速度の変化による認識率の変動が大幅に減少していることがわかる。

以上、本手法の有効性を再確認した。

5. 対判定型TDNNの諸性能の考察

5.1 中間値を用いた学習の効果

対判定型TDNNの大きな特徴は、学習時に1や0の他に0.5という中間的な値を教師信号に持つデータが存在するということである。従来より行われているTDNNの学習では、ネットワークに対して1か0の値のみを教師信号として与え、中間的な出力値を要求していない。そのため、ネットワークは、学習用データに関してそのパラメータ空間を急峻な識別境界面で切り分けるように形成される。その結果、出力ユニットに設けられたシグモイド関数の飽和領域(微分係数が十分に小さい領域)を多用してエラーの収束を実現するという傾向があった。このようなネットワークでは、パラメータ空間内の大部分のデータに対し、1や0付近の値が出力される。

これに対して、今回のように、パラメータ空間内の広域にわたって分布する学習用データに、0.5という教師信号を与えて学習を行う場合を考えてみる。この場合、従来の方法により形成されるような急峻な識別境界面を持つネットワークでは、広域に分布する学習用データに対する出力値を、0.5付近に集中させることが非常に困難であると予想される。従って、1や0を教師信号に持つ学習用データが、識別境界面をより急峻にしようとする方向に働くのに対し、0.5の教師信号を持つ学習用データは、それを抑制する方向に働くと考えられる。さらに音素対のどちらのカテゴリにも属さないデータに対して0.5という値を与えて学習することにより、そのようなデータに対する出力値を0.5付近の値に集中させることができるため、認識時において個々のネットワークが対象とする音素対以外のデータを入力したときの出力値が音素対のどちらか一方の対判定スコアを水増ししてしまうといった不都合を回避することができる。

ここでその効果を確認するために、/b,d,g,m,n,N/の6音素の認識実験に用いた3通りのネットワーク(対判定型TDNN、中間値学習を省略した対判定型TDNNおよび一括判定型TDNN)それぞれの出力ユニットを対象に、学習用データ中の音素データに対する出力値が、ユニット内に設けられた非線形関数のどの領域を用いて表現されているかを調べた。出力値としてtが望まれているデータの集合をS_tと表すとき、対判定型TDNNに対しては、15個それぞれのネットワークにおけるS_1、S_0および$S_{0.5}$(中間値学習を省略した場合には、0.5の教師信号を持つ学習用データ$S_{0.5}$が存在しないが、ここでは通常の対判定型TDNNの場合にあわせて、S_1でもS_0でもないデータの集合を$S_{0.5}$と表す)の分布の総和を、また一括判定型TDNNに対しては、6個の出力ユニットそれぞれにおけるS_1およびS_0の分布の総和を求めた。それぞれの場合の結果を図11に示す。

この結果より、一括判定型TDNNでは大部分のデータに対して、非線形関数の飽和領域を使用して出力値を生成しているために、分布が広範囲に広がっており、各カテゴリ間に急峻な識別境界面が形成されている様子を示している。

中間値学習を省略した対判定型TDNNでは、S_1およびS_0に対する分布は一括判定型TDNNと比較するとかなり狭い範囲に納まっているが、0.5を教師信号に持つカテゴリのデータを学習していないため、$S_{0.5}$の分布が広範囲に広がってしまっている。

それに対して中間値学習を行った通常の対判定型TDNNでは、3通りすべての分布が狭い範囲に納まっており、S_1やS_0は主に非線形関数の飽和する直前の部分を使用して表

現されていることがわかる。

これらの結果より、対判定型TDNNの特徴である中間値を用いた学習方法が、急峻な識別境界面の形成の抑制に対して効果的であることが確認できた。

5.2 対判定型TDNNの尤度表現能力

音素認識手法を予測LRパーザ[20][21]等の手法と組み合せて、単語あるいは文節等の認識を行う場合、音素認識部における頑健性が全体の認識性能を左右する大きな要因となる。音素認識部において、正解カテゴリに対する出力値が0となり、これをそのカテゴリに対する尤度として使用した場合には、もはやトップダウンの処理を用いても救うことのできない、致命的な誤りとなってしまう。

従来のTDNNではこのような致命的な誤りを起こす傾向が強く、正解か否かに関係なく、第一位候補の尤度のみ1で残りは全て0付近の値になってしまうという場合が多かった。これは各カテゴリ間に形成された急峻な識別境界面や、ネットワークの汎化能力の乏しさに起因するものと考えられる。

しかし対判定型TDNNでは、緩やかな識別境界面と多数決の効果により、出力値には認識候補の尤もらしさが反映されているため、そのような致命的な誤りは起きにくくなっていることが予想される。

そこでここでは、/b,d,g,m,n,N/の6音素の認識実験に用いた3通りのネットワークを用いて、各ネットワークの出力値がどのような傾向を持っているかを調べた。自由発声データに対してそれぞれのネットワーク計算を行い、正解カテゴリに対する出力値を横軸に、また不正解カテゴリに対する出力値の中の最大値を縦軸にとって出力値の傾向を2次元図として表したものを図12に示す。(このような図をスキャタープロットと呼ぶ。) この図では、左下から右上に引かれた点線より右下にプロットされたもの(下三角の領域)が音素認識のレベルでの正解となる。

この結果を見ると、対判定型TDNNと一括判定型TDNNとでは、スキャタープロットの傾向が全く違っていることがわかる。つまり一括判定型TDNNでは、点が縦軸と横軸の周辺に分布しており、中央付近にはほとんど分布していない。このうち、左上隅に分布している点は、正解カテゴリの出力値がほとんど0、不正解カテゴリの出力値の中の最も大きい値がほとんど1ということであり、これは致命的な誤りに結び付く場合が多い。

これに対して、中間値学習を省略した対判定型TDNNでは、分布のダイナミックレンジは異なるものの、分布の形状は中央部に集まっており、どのカテゴリに属するかがあいまいで明確に決定することができないデータに対しては、出力値にもそのあいまいさが反映され、中間的な値となっている様子がわかる。さらに中間値学習を行った通常の対判定型TDNNでは、この傾向を保ったまま正解カテゴリに対する出力値がより大きく、不正解カテゴリに対する出力値がより小さくなっている。

以上の結果より、対判定型TDNNでは一括判定型TDNNよりも致命的な誤りを起こしにくいことが確認できた。

6. まとめ

本稿では、対判定型ニューラルネットワークの原理について述べ、これをTDNNと統合した対判定型TDNNによる音素認識性能について示し、その有効性を確認した。

対判定型ニューラルネットワークでは、

- 比較的ゆるやかな識別境界面を持つネットワークの形成、
- それぞれ形状の異なる識別境界面を持つ複数のネットワークの形成、

の2つを同時に実現することにより、分類型ニューラルネットワークにおける頑健性の向上を図っている。そしてこの原理を、シフト・トレラント性の点で優れた性能を持つTDNNと統合することにより、発話速度の異なる各種発話様式のデータに対して、従来の一括判定型TDNNと比べて、より優れた認識性能と頑健性を示すことが確認できた。

今後の課題としては、ここで述べた効果をより明らかにするために、予測LRパーザとの統合による連続音声認識実験を行い、その性能をより明確に示すことが考えられる。また実用の場面を考えて、計算時間の短縮のための検討も必要であろう。対判定型TDNNによる音素認識法は、各ネットワークが完全に独立しており、個々のネットワーク規模がそれほど大きくないため、並列処理マシン上での実現やハードウェア化等により、そのメリットが十分に活かされることが予想されるが、現在の計算機能力を考えると、計算時間の点で問題点が残る。この問題の対策としては、例えば、ネットワークの階層化やコンフュージョンの起こりにくい音素間での対判定の省略等の方法が考えられる。今後、このような可能性に対する検討も行っていきたい。

謝　辞

本研究の機会を与えて下さったATR自動翻訳研究所の樽松明社長、御討論頂いた杉山雅英主幹研究員、沢井秀文主任研究員、田村震一主任研究員、中村雅己研究員、小森康弘研究員をはじめとするATR自動翻訳電話研究所の皆様、慶応義塾大学の南泰浩君に深く感謝の意を表します。

参考文献

[1] D.E.Rumelhart, J.L.McClelland, "Parallel Distributed Processing ; Explorations in the Micro Structure of Cognition", MIT Press (1986).

[2] P.Haffner, A.Waibel, K.Shikano, "Fast Back-Propagation Learning Methods for Neural Networks in Speech", 音響学会講演論文集, 2-P-1 (1988-10).

[3] A.Waibel, "時間遅れ神経回路網 (TDNN)による音韻認識", 音声研究会資料, SP87-100 (1987-12).

[4] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano and K.Lang, "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans., Acoust., Speech, Signal Processing, vol.37, pp. 328-339 (1989-3).

[5] 沢井, "時間・周波数変動に強い時間遅れ神経回路網(TDNN)", 音響学会講演論文集, 2-P-12 (1990-9).

[6] 川端, "k-近傍内挿学習による音韻認識", 音響学会講演論文集, 2-P-21 (1990-3).

[7] 小森, ワイベル, 嵯峨山, "音素識別ニューラルネットにおけるファジー学習法", 音響学会講演論文集, 1-5-15 (1991-3).

[8] 鈴木, 河原, "平均曲率を用いた神経回路網の評価基準について", ニューロコンピューティング研究会資料, NC89-103 (1990-3).

[9] 南, 田村, 沢井, 鹿野, "入力層・中間層におけるベクトルの近傍の情報を利用したTDNN出力の平滑化", 音響学会講演論文集, 1-3-18 (1990-3).

[10] 中村, 田村, "ニューラルネットによる音韻フィルタ", 音響学会講演論文集, 2-P-24 (1990-3).

[11] Y.Minami, T.Hanazawa, H.Iwamida, E.McDermott, K.Shikano and M.Nakagawa, "On Sensitivity and Robustness of HMM and Neural Network Speech Recognition Algorithms", ICSLP'90, S31.3, pp.1345-1348 (1990-11).

[12] 天野, 畑岡, 矢島, 市川, "対判定による子音認識の検討", 音響学会講演論文集, 1-1-13 (1986-3).

[13] 鷹見, 嵯峨山, "対判定型TDNNによる音素認識", 音声研究会資料, SP90-10 (1990-6).

[14] 鷹見, 嵯峨山, "対判定型TDNNにおける中間値学習の効果", 音響学会講演論文集, 2-P-15 (1990-9).

[15] J.Takami and S.Sagayama, "Phoneme Recognition by Pairwise Discriminant TDNNs", ICSLP'90, S16.5, pp.677-680 (1990-11).

[16] J.Takami and S.Sagayama, "A Pairwise Discriminant Approach to Robust Phoneme Recognition by Time-Delay Neural Networks", ICASSP'91, S8.13 (1991-5).

[17] 武田, 匂坂, 片桐, 桑原, "研究用日本語音声データベースの構築", 音響学会誌, p747 - p754 (1988-10).

[18] K.Takeda, Y.Sagisaka and S.Katagiri, "Acoustic-Phonetic Labels in a Japanese Speech Database", European Conference on Speech Technology, pp.13-16 (1987-8).

[19] 南, 沢井, "発話変動にロバストなTDNNの検討", ATR Technical Report TR-I-0183 (1990-10).

[20] H.Sawai, "The TDNN-LR Large-Vocabulary and Continuous Speech Recognition System", ICSLP'90, S31.4, pp.1349-1352 (1990-11).

[21] 沢井, "TDNN-LR文節音声認識システムにおける追加学習の効果", 音響学会講演論文集, 2-P-11 (1990-9).

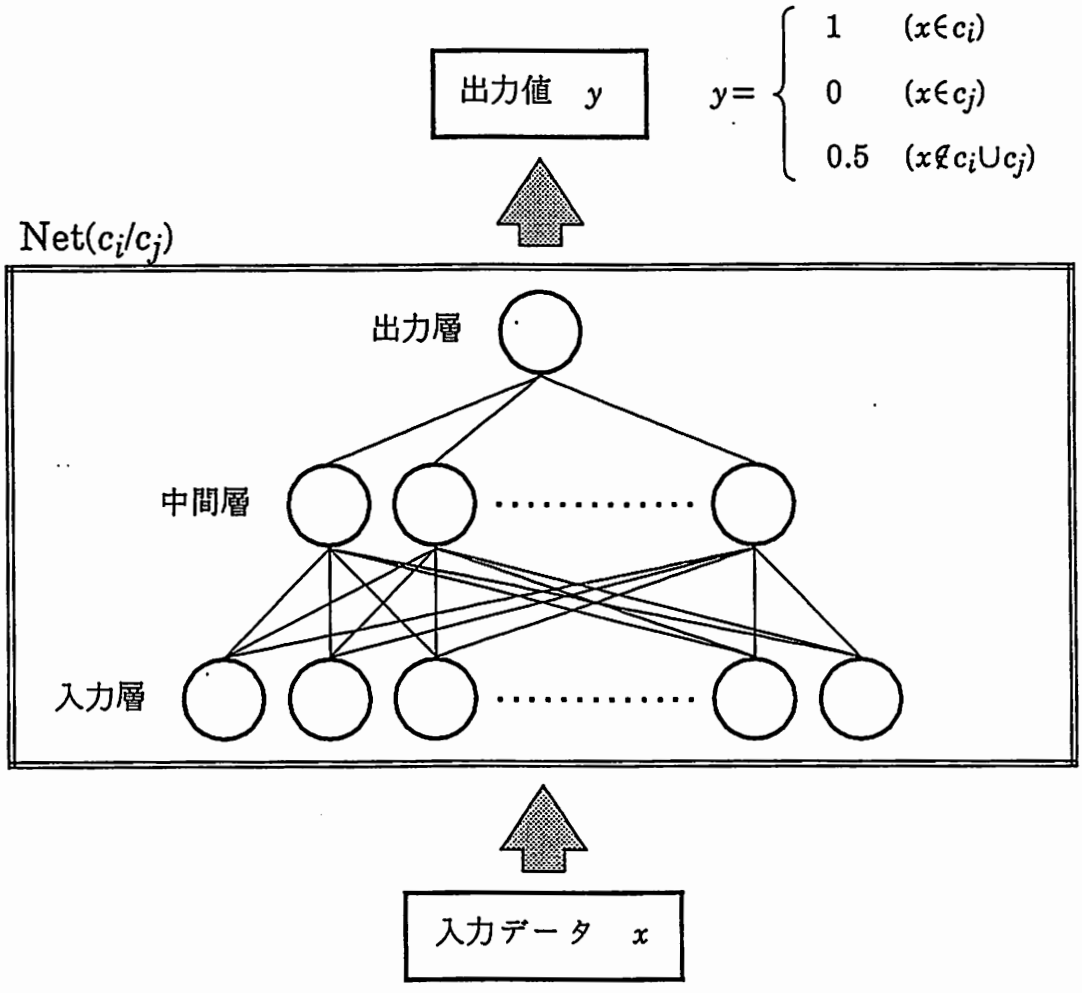


図1 Net(c_i/c_j)の構造

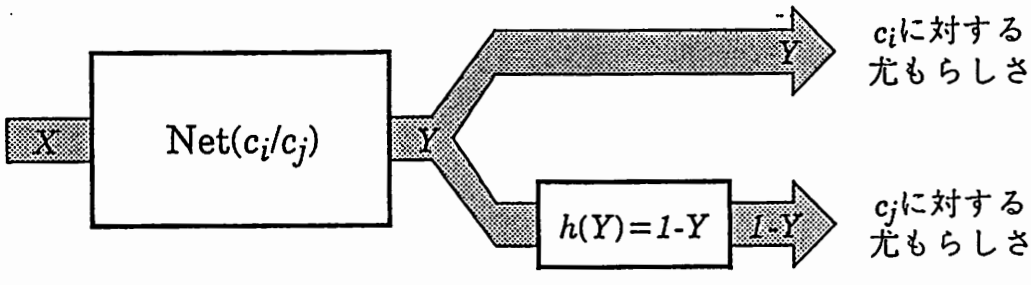


図2 Net(c_i/c_j)による対判定

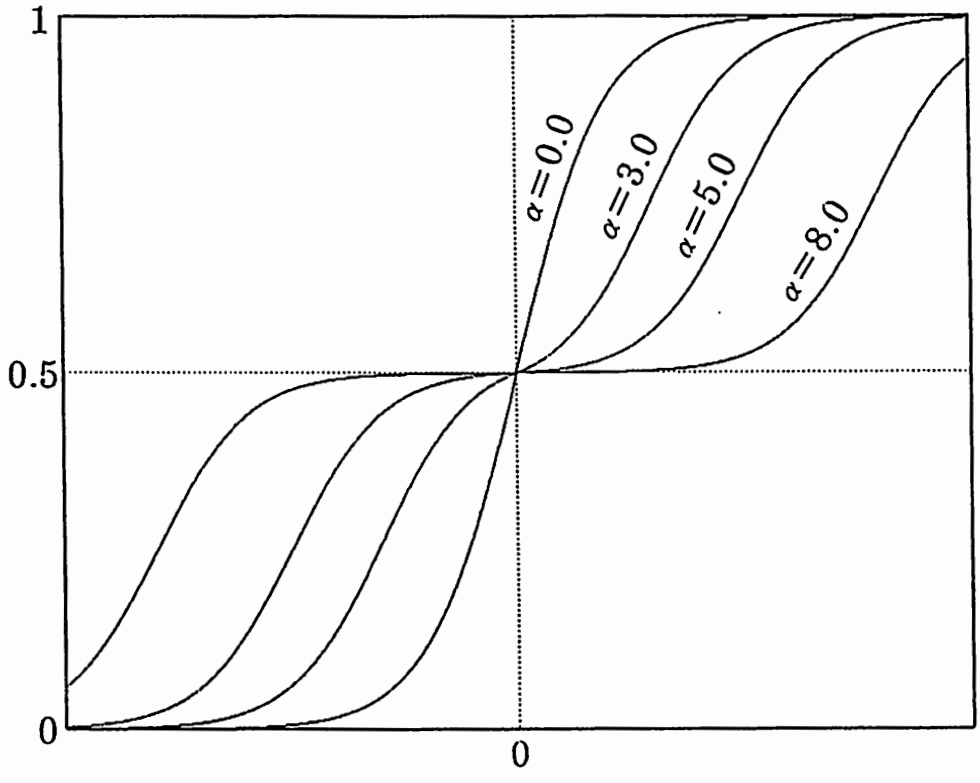


図3　出力ユニット内の非線形関数の概形

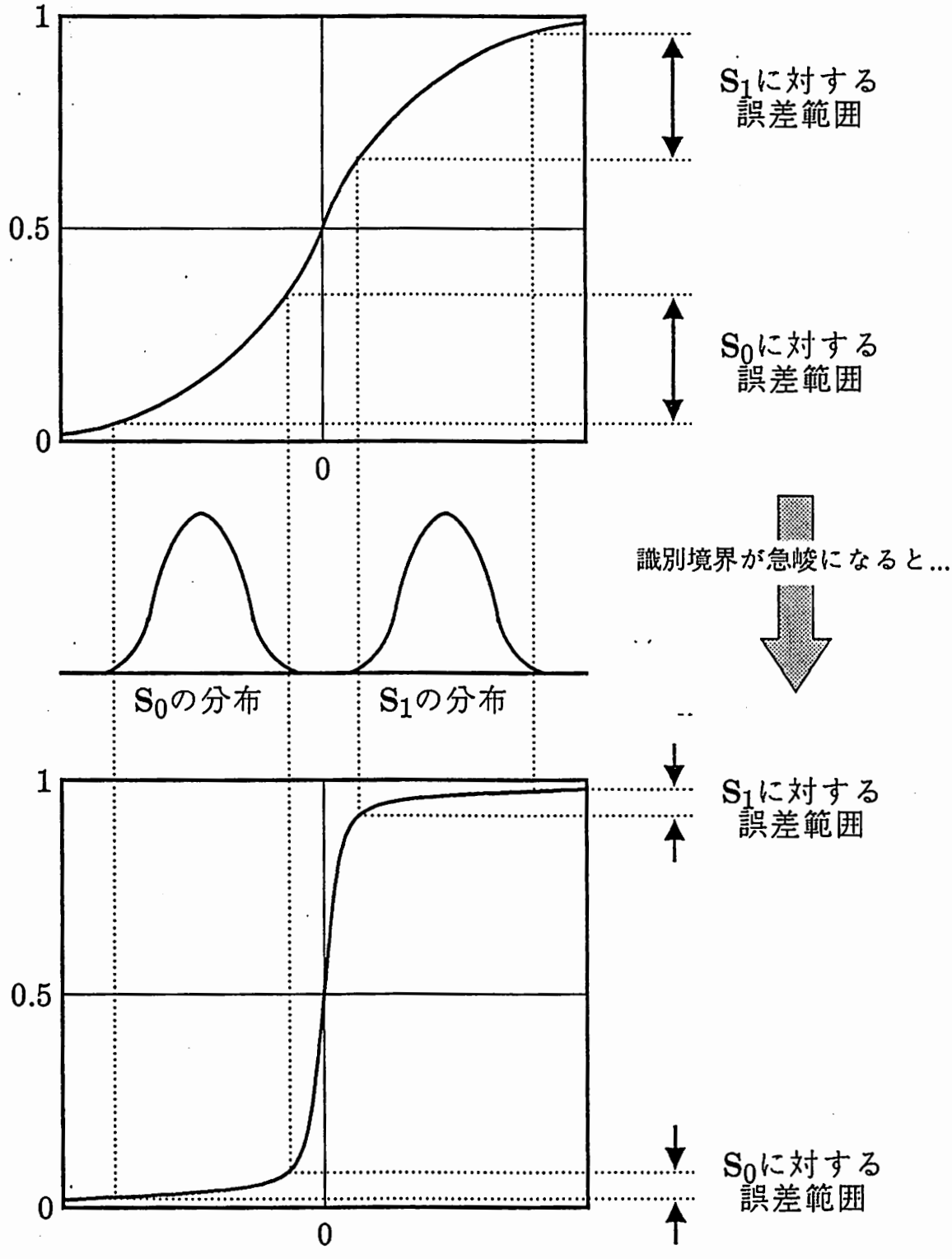


図4　1および0のみを教師信号に用いた場合の学習メカニズム

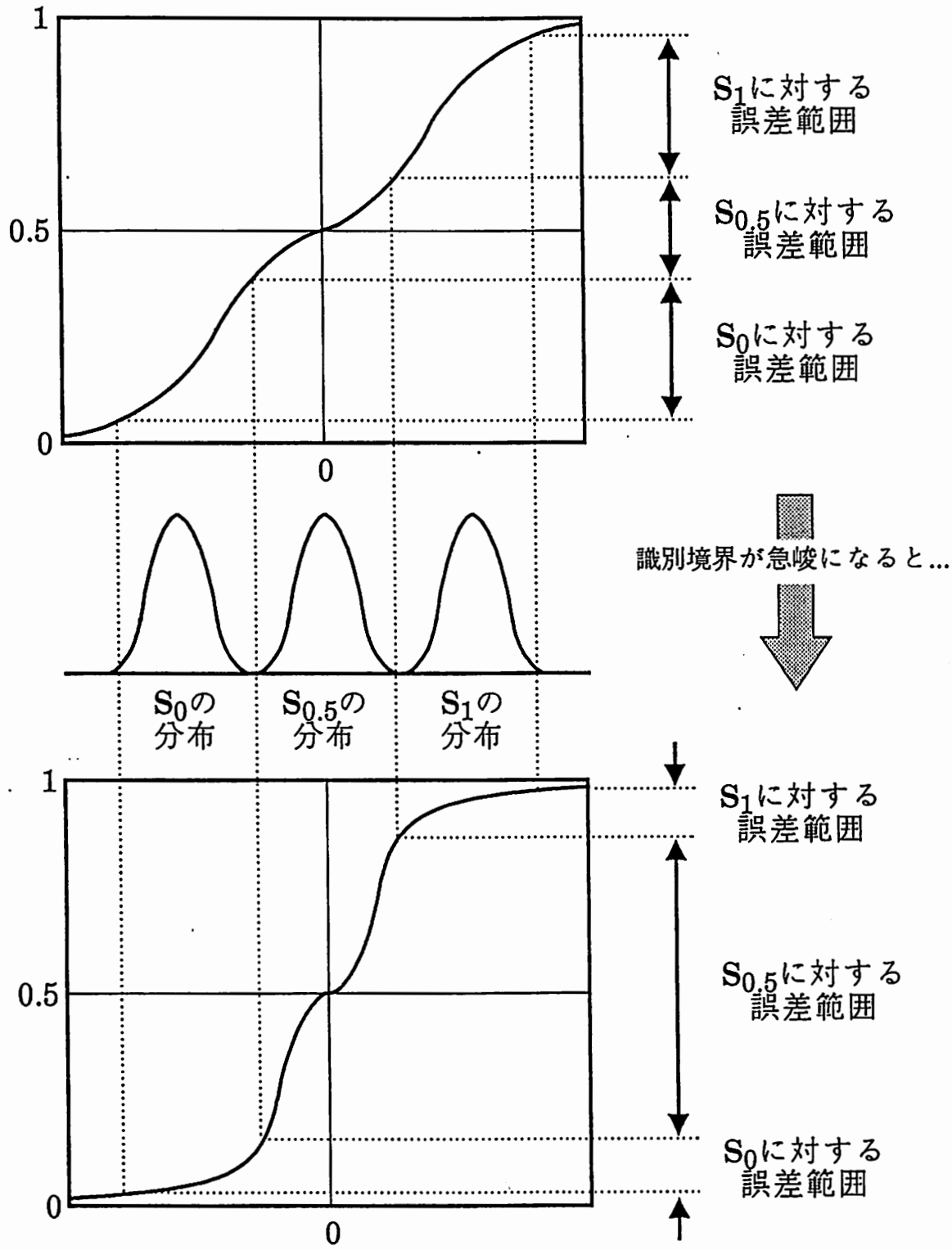


図5　1, 0, 0.5の3値を教師信号に用いた場合の学習メカニズム

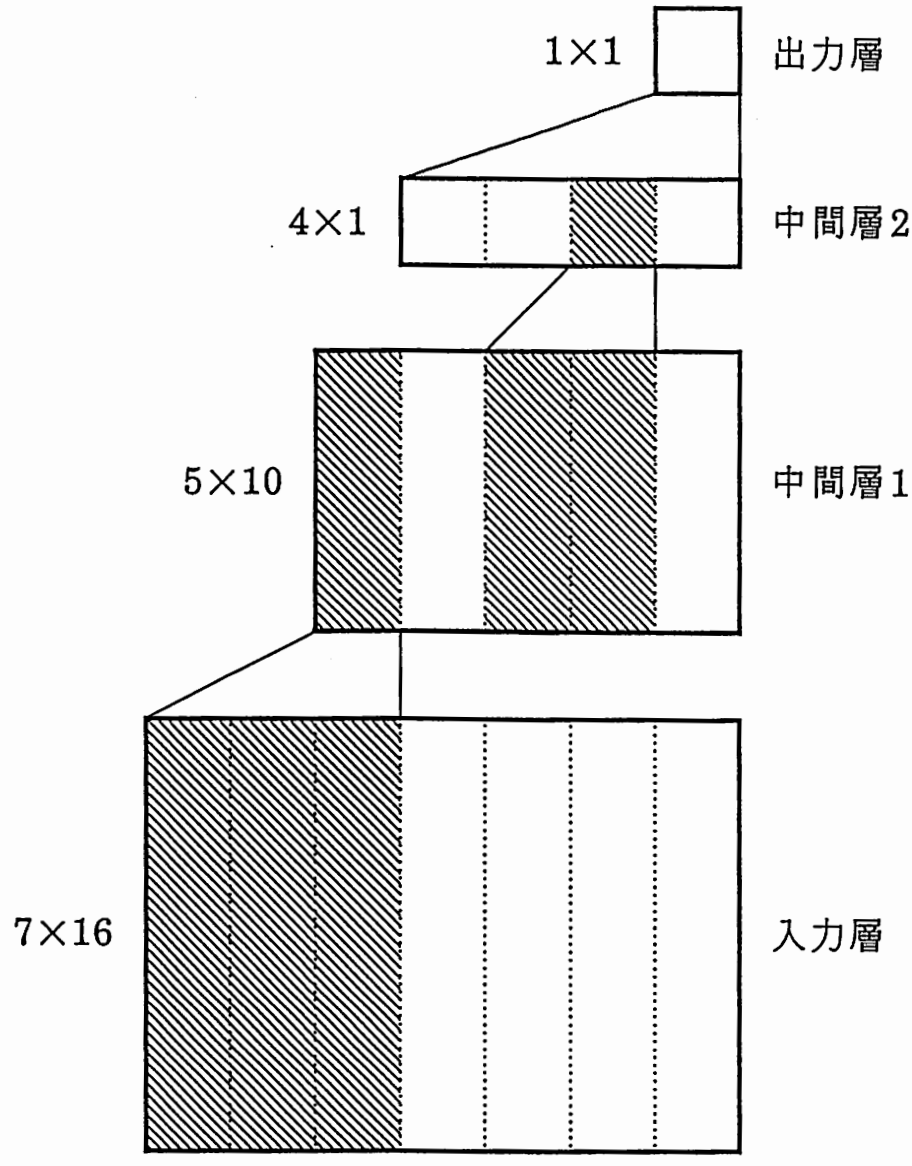


図6　PD-TDNNの構造

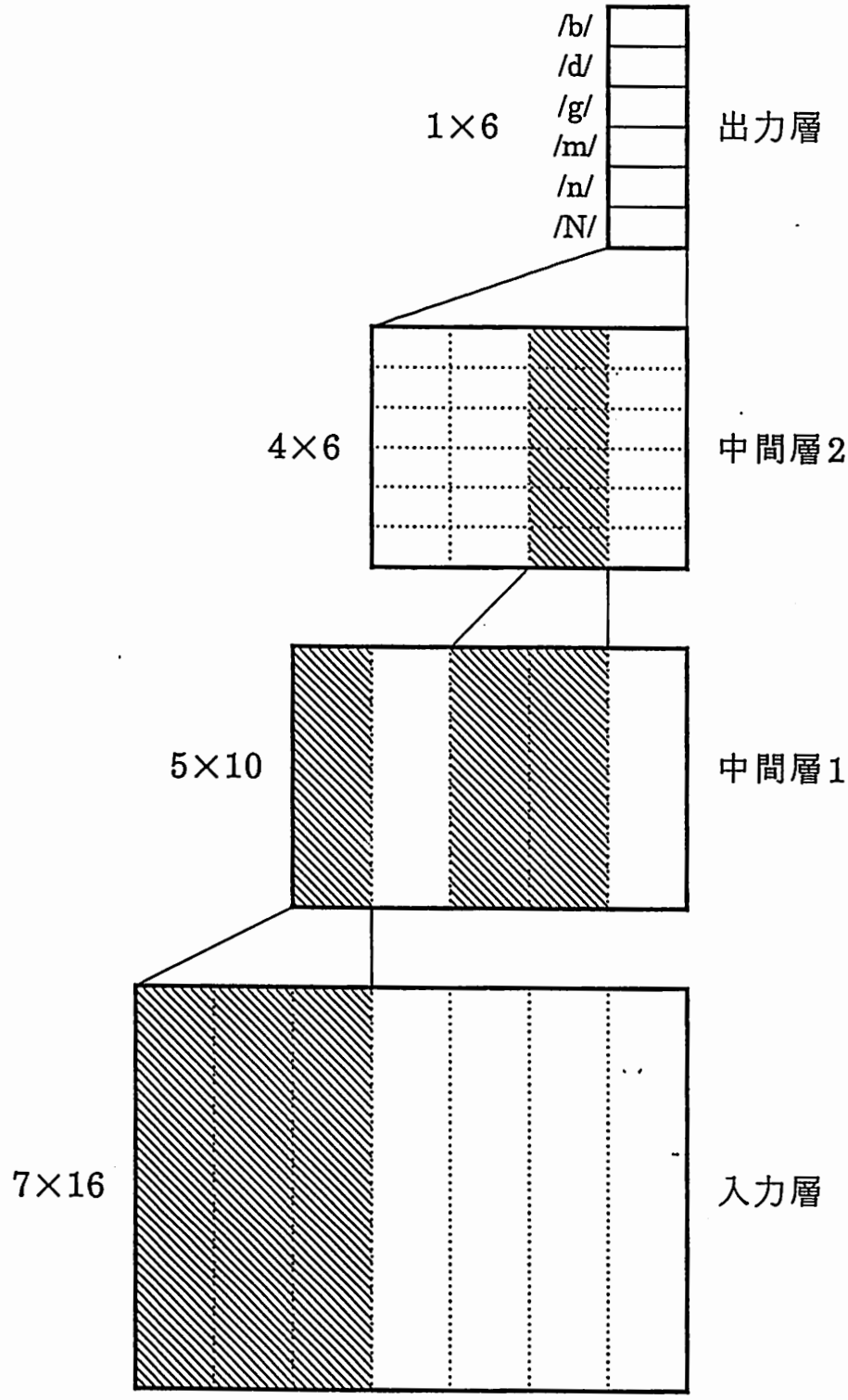


図7　一括判定型TDNNの構造(比較用)

表1　対判定型TDNNによる6音素認識率　　(単位:%)

音素	単語発声	文節発声	短い文節発声	自由発声
/b/	99.1 (100.0)	97.9 (100.0)	85.4 (100.0)	79.2 (91.7)
/d/	97.2 (100.0)	88.1 (99.2)	91.9 (97.4)	90.2 (98.7)
/g/	95.2 (100.0)	95.0 (100.0)	85.6 (100.0)	83.1 (100.0)
/m/	98.1 (100.0)	85.8 (99.0)	81.6 (97.9)	71.6 (95.3)
/n/	97.4 (100.0)	89.5 (100.0)	86.5 (99.6)	83.0 (96.4)
/N/	96.5 (99.2)	83.5 (96.5)	82.9 (93.2)	77.3 (94.1)
平均	97.3 (99.8)	88.8 (99.2)	86.2 (98.1)	81.6 (96.7)

(カッコ内は第3位までの累積認識率)

表2　中間値学習を省略した場合の6音素認識率　　(単位:%)

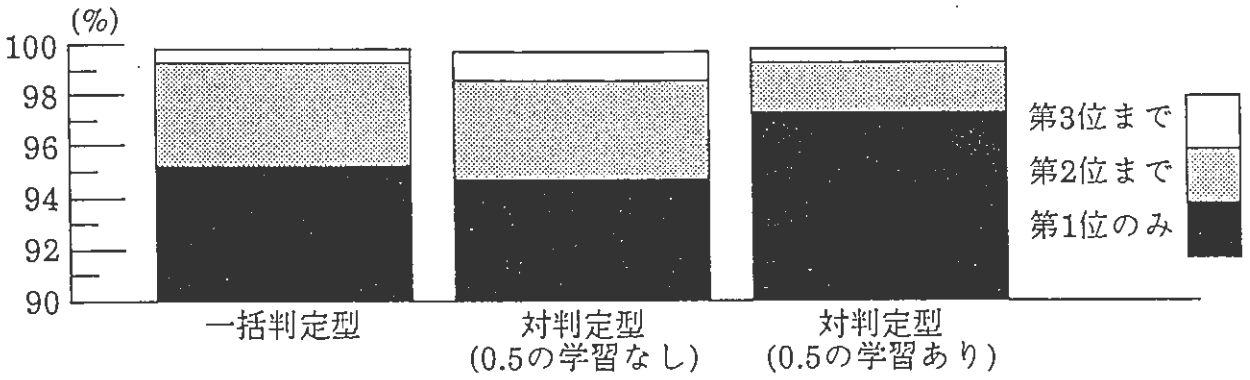
音素	単語発声	文節発声	短い文節発声	自由発声
/b/	96.0 (100.0)	83.0 (100.0)	72.9 (100.0)	58.3 (89.6)
/d/	97.8 (100.0)	87.2 (97.5)	85.9 (97.0)	80.8 (96.6)
/g/	90.5 (99.6)	88.3 (100.0)	83.1 (99.2)	83.9 (100.0)
/m/	94.2 (100.0)	70.0 (100.0)	65.3 (99.0)	55.3 (100.0)
/n/	94.3 (100.0)	77.5 (98.6)	74.1 (96.4)	76.9 (98.2)
/N/	95.7 (99.0)	82.6 (91.3)	82.1 (92.3)	74.0 (84.0)
平均	94.7 (99.7)	80.6 (98.1)	77.2 (97.0)	73.2 (96.3)

(カッコ内は第3位までの累積認識率)

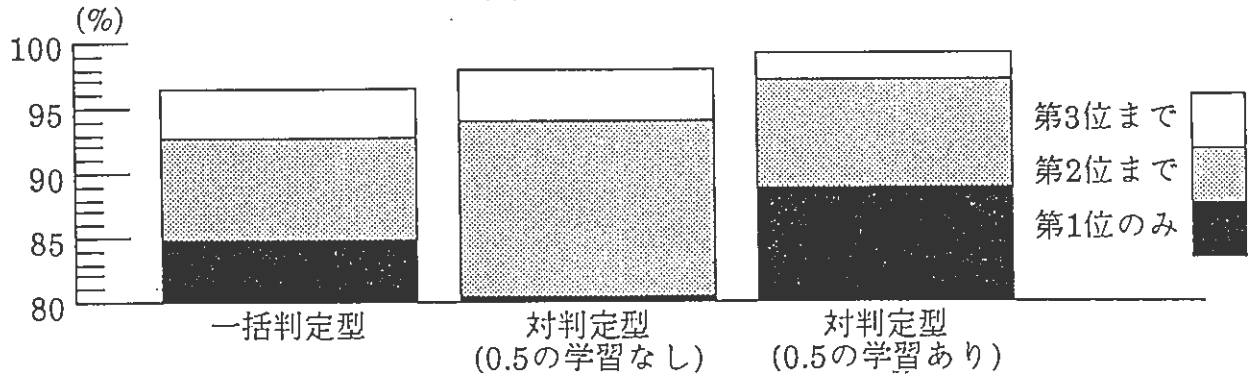
表3　従来のTDNNによる6音素認識率　　(単位:%)

音素	単語発声	文節発声	短い文節発声	自由発声
/b/	95.2 (99.6)	97.9 (100.0)	85.4 (93.8)	68.8 (81.3)
/d/	98.3 (100.0)	89.8 (97.5)	92.3 (97.9)	82.9 (94.4)
/g/	91.2 (99.6)	91.7 (99.2)	91.5 (100.0)	86.4 (100.0)
/m/	94.8 (100.0)	88.4 (99.0)	87.9 (97.9)	82.6 (97.4)
/n/	95.5 (99.3)	72.1 (90.9)	70.8 (88.0)	65.7 (82.3)
/N/	96.3 (100.0)	87.0 (100.0)	88.0 (99.2)	83.2 (99.2)
平均	95.2 (99.8)	84.8 (96.5)	84.5 (95.3)	77.8 (92.2)

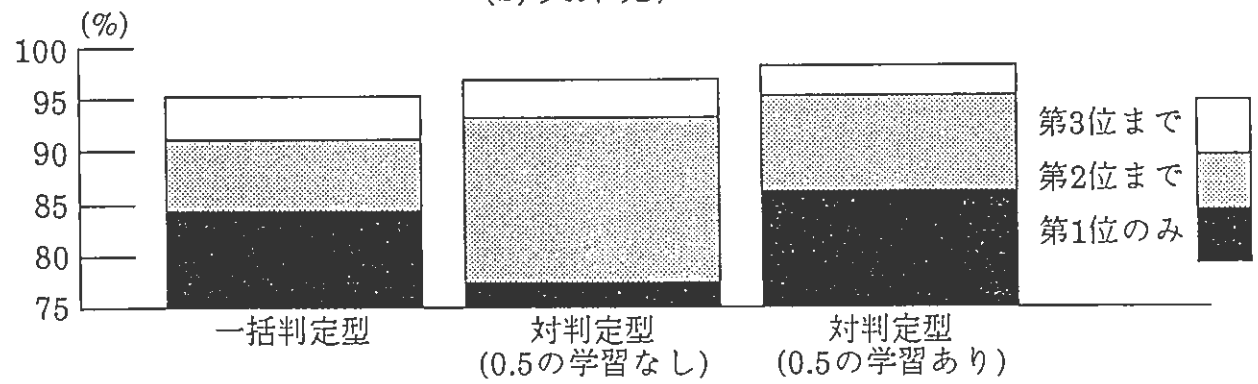
(カッコ内は第3位までの累積認識率)



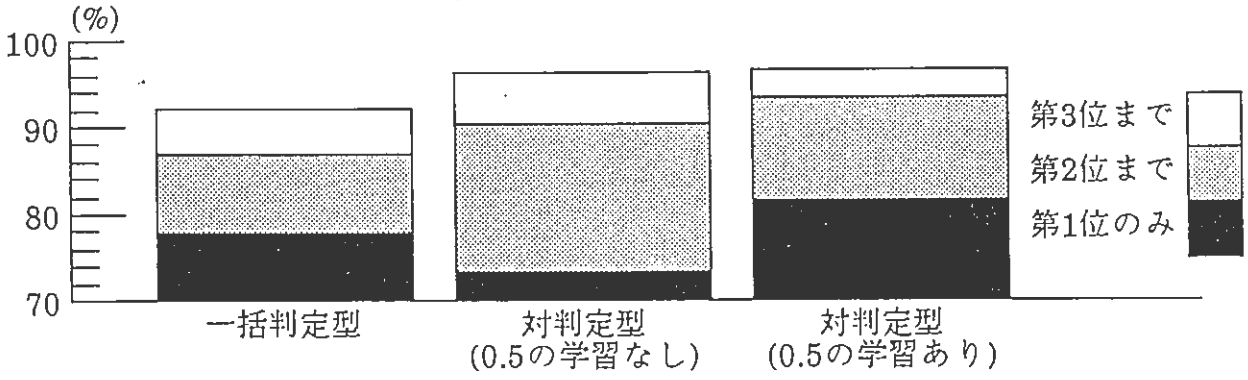
(a) 単語発声



(b) 文節発声



(c) 短い文節発声



(b) 自由発声

図8　各発話様式のデータ中の/b//d//g//m//n//N/に対する音素認識率

表4　シフト・トレラント性の確認結果(6音素)　(単位:%)

シフト(ms)	対判定型TDNN	一括判定型TDNN
-20	72.5 (96.0)	62.4 (85.7)
-10	94.8 (99.4)	87.4 (98.4)
0	97.3 (99.8)	95.2 (99.8)
10	94.2 (99.6)	88.4 (98.4)
20	70.0 (95.1)	60.7 (93.6)

(カッコ内は第3位までの累積認識率)

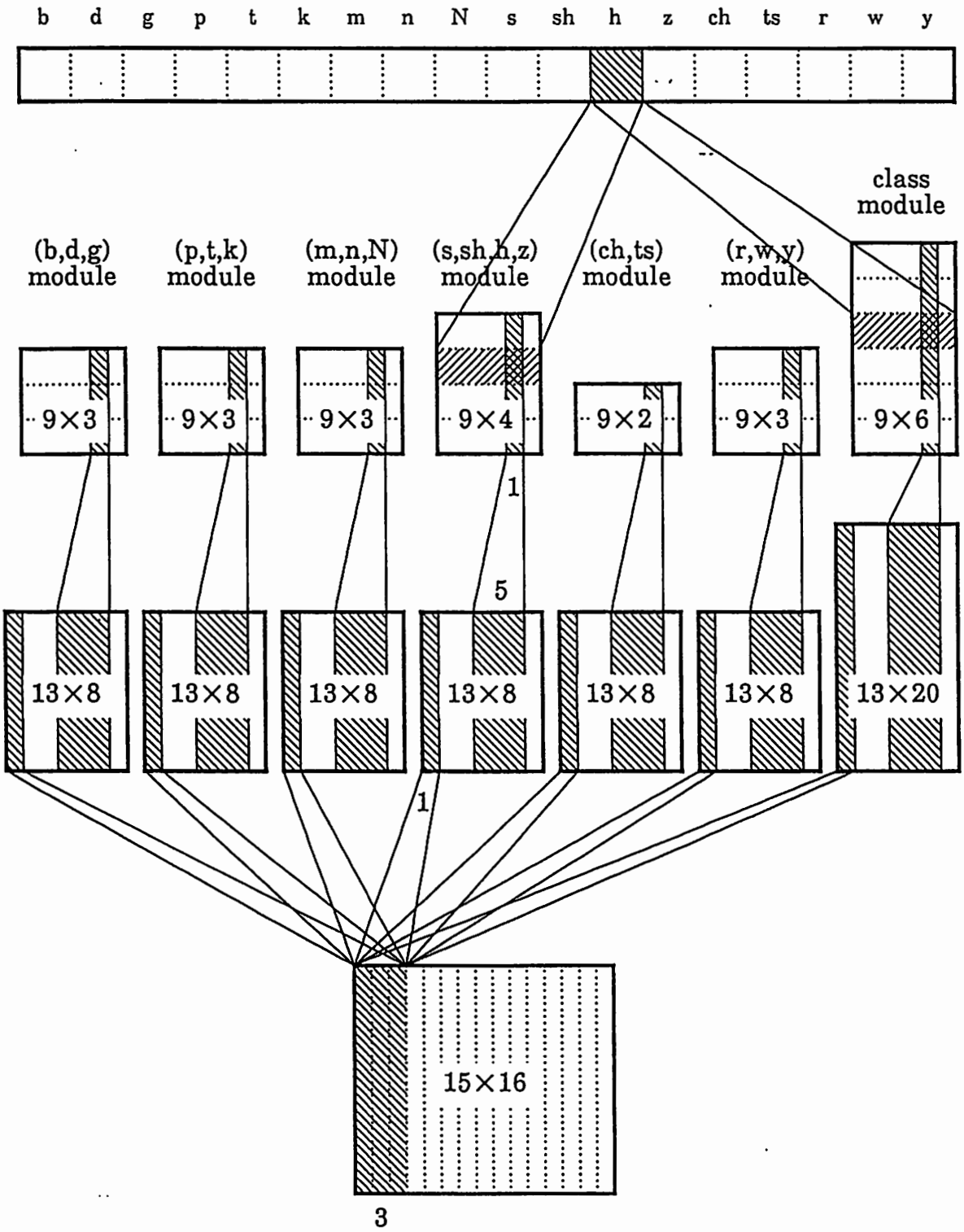


図9　モジュール構成型TDNNの構造 (比較用)

表5　対判定型TDNNによる18子音認識率　　(単位：%)

音素	単語発声	短い文節発声	自由発声
/b/	96.0 (100.0)	87.5 (100.0)	70.8 (87.5)
/d/	97.2 (100.0)	85.0 (97.4)	74.4 (91.0)
/g/	94.0 (100.0)	83.9 (99.2)	84.7 (99.2)
/p/	80.0 (100.0)	50.0 (50.0)	100.0 (100.0)
/t/	98.4 (100.0)	90.8 (97.2)	59.4 (79.9)
/k/	93.4 (99.2)	95.1 (98.9)	94.9 (99.1)
/m/	97.3 (100.0)	86.3 (97.9)	77.9 (96.8)
/n/	97.0 (100.0)	78.5 (94.2)	79.4 (96.4)
/N/	93.1 (99.8)	84.6 (94.0)	78.2 (92.4)
/s/	81.6 (100.0)	77.8 (99.3)	67.6 (99.3)
/sh/	89.8 (100.0)	85.7 (95.7)	88.4 (97.1)
/h/	94.3 (99.6)	68.7 (91.0)	22.1 (61.8)
/z/	96.5 (99.6)	100.0 (100.0)	97.9 (100.0)
/ch/	69.0 (100.0)	30.3 (78.8)	14.3 (53.6)
/ts/	94.4 (100.0)	95.7 (100.0)	82.9 (100.0)
/r/	95.3 (99.9)	87.6 (99.1)	86.7 (98.3)
/w/	100.0 (100.0)	81.3 (96.3)	66.7 (88.5)
/y/	99.4 (99.4)	92.2 (95.6)	54.3 (82.7)
平均認識率	93.9 (99.8)	86.1 (97.1)	76.8 (92.9)

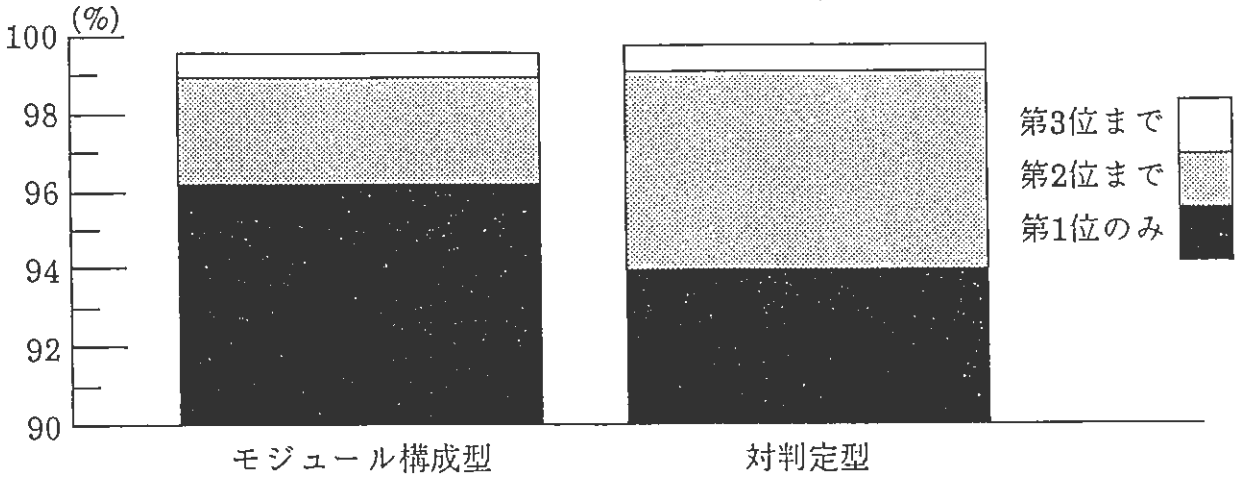
(カッコ内は第3位までの累積認識率)

表6　モジュール構成型TDNNによる18子音認識率

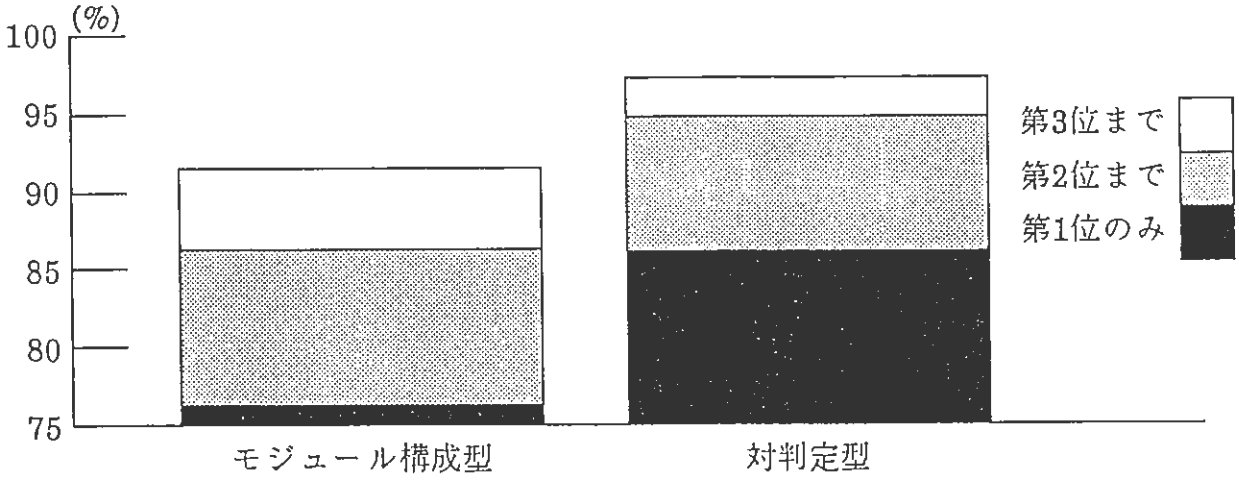
(単位：%)

	単語発声	短い文節発声	自由発声
平均認識率	96.2 (99.6)	76.2 (91.5)	56.6 (78.5)

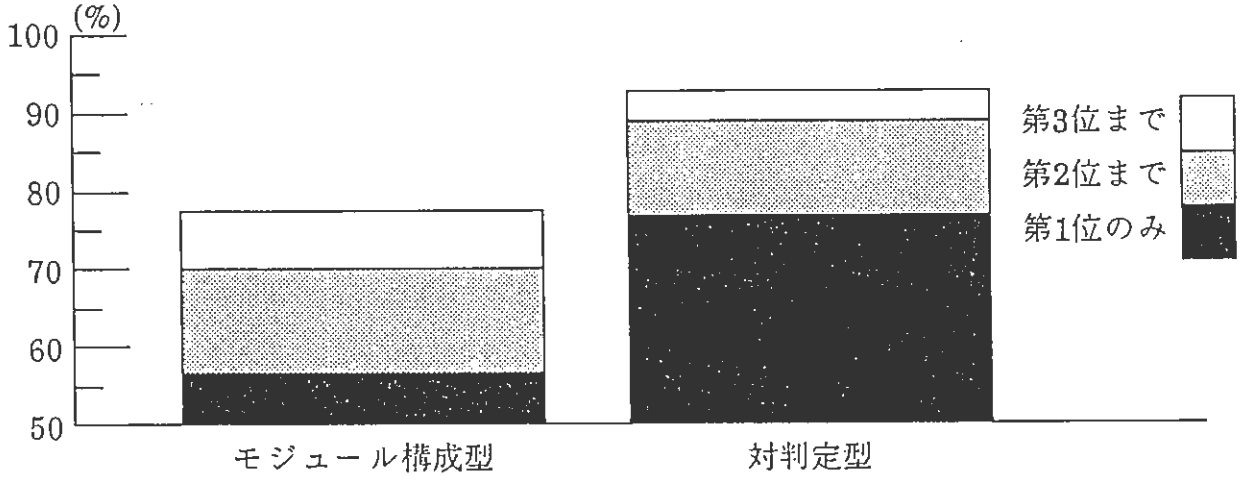
(カッコ内は第3位までの累積認識率)



(a) 単語発声

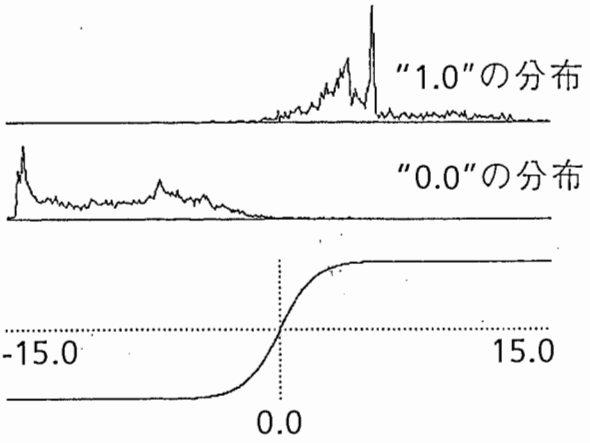


(b) 短い文節発声

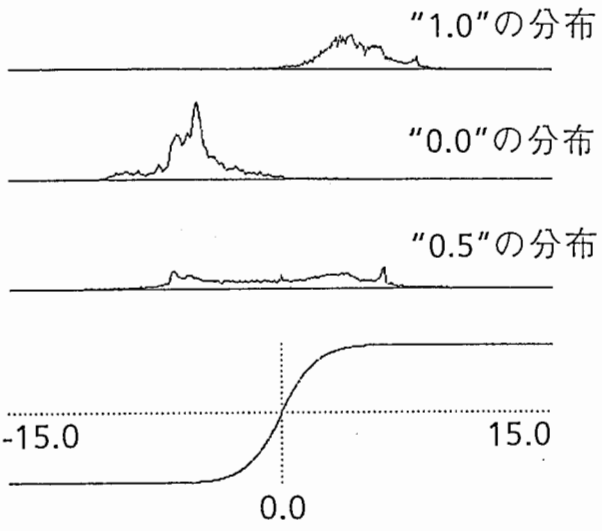


(c) 自由発声

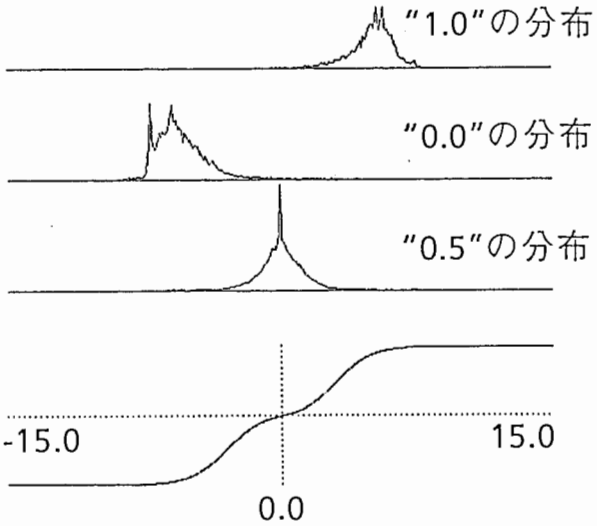
図10　各発話様式のデータ中の18子音に対する音素認識率



(a) 一括判定型TDNN

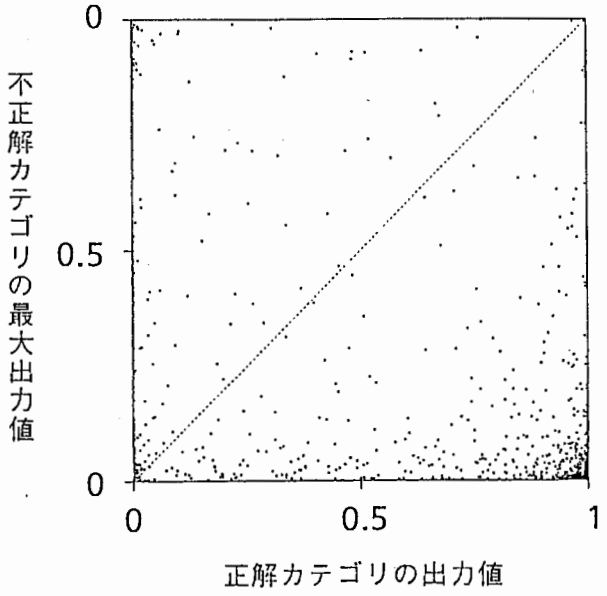


(b) 中間値学習を省略した対判定型TDNN

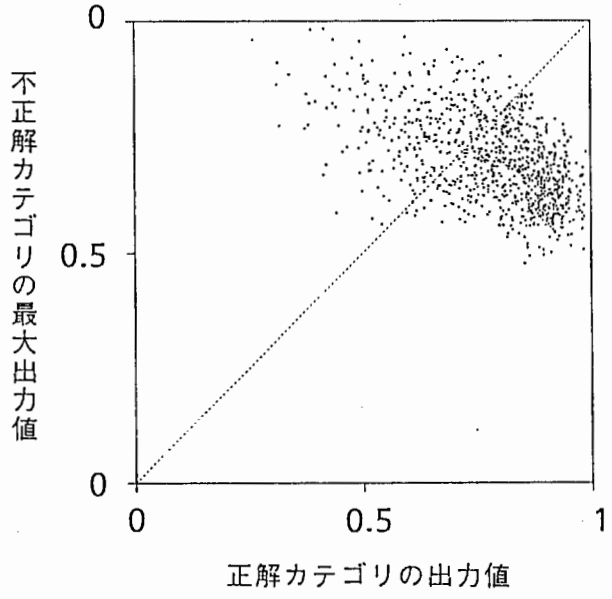


(c) 対判定型TDNN

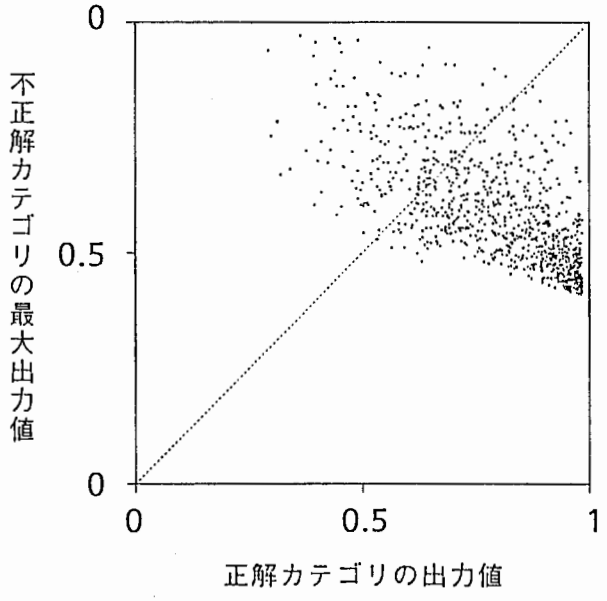
図11 学習用データに対する出力ユニット上での出力値の分布



(a) 一括判定型TDNN



(b) 中間値学習を省略した対判定型TDNN



(c) 対判定型TDNN

図12 自由発声データに対するスキャタープロット