TR-I-0177

ニューラルネットによる音素フィルタを用いた母音認識

Vowel Recognition by Phoneme Filter Neural Networks

中村雅巳、田村震一*

*Masami NAKAMURA, Shin'ichi TAMURA**

1990.9

概要

本稿ではニューラルネットによる音素フィルタ(PFN; Phoneme Filter Neural Network)を提案し、母音認識への適用実験により評価したので報告する。PFN は音素別に用意した、中間層を圧縮した多層ニューラルネットで構成されており、1つのPFN には1種類の音素パターンのみ恒等写像学習することにより、特定の音素のみ変形せずに通すような音素フィルタとなる。認識時は音声パターンをPFNに入力し、PFNの出力パターンと音声パターンの類似度をそれぞれのPFNについて計算し、その比較により認識を行なう。従来の分類型ニューラルネットによる音声認識では、2位以下の候補の出力値が0近くに抑えられ、認識スコアとして用いることができないという問題点があったが、本方法で日本語5母音の認識実験を行なった結果、2位以下の累積認識率が従来の分類型ニューラルネットより良好で、認識スコアは候補カテゴリへの近さを反映していることがわかった。また、主成分分析による方法との比較実験により、PFNの非線形写像の効果が確認できた。

Abstruct

This paper describes a vowel filter neural network (PFN) approach to vowel recognition. Most conventional speech recognition neural networks have a serious drawback: the network output values do not correspond to candidate likelihoods. The PFN is a multi-layer neural network with fewer hidden units than input units prepared for each of the phoneme categories. Each network is trained as identity mapping by speech data belonging to one phoneme category. In the recognition process, the distance between the input data and output data is computed for each network. The results of the experiment to apply the Japanese vowel recognition task showed that the PFN recognition rates for the top 2 or more choices are higher than those of a conventional 3-layer neural network. It was also confirmed that the PFN outputs represented candidate likelihoods and that, because of it's non-linearity, the performance of the 5-layer PFN wassuperior to that of the 3-layer PFN.

ATR自動翻訳電話研究所

ATR Interpreting Telephony Research Laboratories

*現在、ソニー（株）総研、AV研

1. まえがき

近年、ニューラルネットを用いたパターン認識の研究が盛んに行われているが、その多くのモデルは、入力にパターンデータ、出力に認識対象のカテゴリを局所表現で学習させる、分類型(Classificasion Type)に属している。これは入力パターンの空間を線形識別における超平面のようなもので切り、カテゴリ別に分類しようとするものである。実際の学習では一般に、入力パターンの属するカテゴリの出力ユニットに、他のカテゴリの出力ユニットと違う値（例えば、学習パターンのカテゴリの出力ユニットに対して1、他のユニットに対して0）を教師データとして与え、実際の出力との差が小さくなるように学習を行なう。この場合、カテゴリ境界付近でも0から1の間の中間値が出力されにくく、カテゴリに対する近さが出力値に現れないという問題点があった。これは、教師信号として0あるいは1などの2値信号を与えるため、出力がユニット入出力関数である非線形シグモイド関数の両端の飽和領域に追い込まれるためであると考えることができる。音声認識において、このような分類型ニューラルネットの出力値を音素単位の認識スコアとして用いた場合、単語認識や連続音声認識に発展させるのが難しくなる。例えば、カテゴリ境界付近で2位以下のカテゴリ候補の出力値が0近くに抑えられると、致命的な誤認識が発生する場合がある。この問題を解決するため、分類型ニューラルネットに対し、様々な手法が試みられてきた。例えば、入力空間中の2点および出力空間中の2点を結ぶ線分の間の対応関係を学習する、k-近傍内挿学習による方法[1]、入力パターンや中間層の出力パターンの近傍でニューラルネットの出力を積分して平滑化する方法[2]、さらに、対判定型のニューラルネットにその他のカテゴリとして0.5の値を学習させて急峻な境界面の形成を抑制する方法[3]、境界付近の近傍情報によりカテゴリへの帰属度を明示的に教師パターンとして与える方法[4]等が提案されている。ここでは、この問題を解決するために、分類型ニューラルネット以外のニューラルネットを用いて音素認識することを試みる。

本稿では、特定カテゴリの音素信号を歪ませることなく圧縮復元するような音素フィルタを、中間層を圧縮した多層ニューラルネットを用いた恒等写像により実現する方法と、このモデルを用いて音素認識する方法を提案する。単一のニューラルネットを用いた、情報圧縮を目的とした恒等写像は、画像や音声の分野では試みられているが[5][6]、それに対して本方法ではパターン認識を目的としているため、カテゴリ毎にニューラルネットを用意し、情報圧縮をフィルタリングの作用として利用している。本方法を評価するために、日本語の5母音の認識実験を行なったので報告する。さらに、本方法の非線形性の効果を調べるために、主成分分析による方法との比較を行なったので、それについても報告する。

2. ニューラルネットによる音素フィルタ（PFN）

ニューラルネットによる音素フィルタ（Phoneme Filter Network、以下PFNと呼ぶ）は、１つのカテゴリの音素フィルタを１つのニューラルネットで実現し、学習時はその音素カテゴリに属する音声パターンのみ用いるため、次のような特長を持つ。

a. それぞれのPFNの入力と出力の比較演算により、音素認識が可能である。

b. 候補カテゴリへの近さを表現した認識スコアが得られるため、単語認識、連続音声認識への拡張が容易である。

c. HMMと同様に、１つのPFNは一つのカテゴリの特徴によりモデル化されるため、他のPFNとは独立である。

d. 新たに認識カテゴリを増やす場合、すでに学習しているPFNを再学習する必要はない。

以下にPFNの構成および認識の方法について述べる。

2.1 PFNの構成

Fig. 1 にPFNの構成を示す。今回は２つのタイプのPFN を試みた。タイプA は３層のフィードフォワード型ネットワークであり、入力層と出力層は同じユニット数を持ち、中間層は入力層よりも少ないユニット数を持つ。タイプB は５層のフィードフォワード型ネットワークであり、入出力層と３番目の中間層はタイプA と同じ構成であるが、１番目と３番目の中間層は入力層よりも多いユニット数を持つ。いずれも入力層と出力層にはシグモイド関数を用いていない。このPFN を認識カテゴリの数だけ用意する。１つのカテゴリに属する音声パターンを対応するPFNに入力し、出力に入力と同じ音声パターンを教師信号として与えてバックプロパゲーション法[7]により学習する。これをすべてのカテゴリ、すなわちすべてのPFN について行なう。これにより、PFNは特定のカテゴリの音素パターンを圧縮復元するような音素フィルタとなる。

情報圧縮を伴う恒等写像では、３層ネットワークの場合は、従来の主成分分析による方法より精度を上げることができないことがすでに示されている[8]。一方、中間層のユニット数が十分多い３層ネットワークでは、任意の連続な写像が実現し得るという証明[9][10]から、圧縮、復元にそれぞれ３層を用いた５層ネットワークにより、非線形写像の効果が得られることが、円や半球等の次元圧縮により実験的に示されている[11][12][13]。ここでも、音声信号の情報圧縮、復元として３層及び５層のPFNの比較実験を行う。

2.2 PFNを用いた音素認識の方法

次に、PFNを用いた音素認識の方法について説明する。1つのPFN には1つのカテゴリに属する音声パターンのみ学習させるので、学習した音声パターンはPFN の入力空間において、ある分散を持った領域を形成している。学習音声パターンの分布の密度が高いところが、それに比例して学習がよく進むことを考えれば、最も歪みが少なく復元できる音声パターンは学習した音声パターンの領域の中心付近であり、それから遠ざかるに従い、歪むことが期待される。従って、PFN の入力パターンと出力パターンを比較し、それらの類似度を定義すれば、そのPFN に対応するカテゴリらしさの尺度を表現することができる。今回の認識対象である母音の特徴は、16チャネルのFFTメルスペクトラムを入力音声ベクトルとした場合、ベクトルの形、すなわちベクトルの向きに依存すると考えられるため、類似度S は、入力した音声ベクトルX とPFNの出力ベクトルX'とのなす角度として次式のように定義し、それを認識スコアした。

$$S = (X, X') / (\|X\| \cdot \|X'\|) \qquad (3.1)$$

ここで、(X, X')はXとX' の内積、$\|X\|$ はX のノルムである。Fig.2にPFNを用いた日本語5母音の音素認識の構成図を示す。認識対象の音素パターンXを日本語の5母音のカテゴリで学習した5つのPFNに入力して、それぞれのPFNの出力を得る。この際、例えば認識対象の音素パターンXが/u/のカテゴリの音素パターンに近ければ、/u/の音素パターンで学習したPFNの出力パターンが音素パターンXにもっとも近いものとなる。そこで、認識においては、入力パターンXとそれぞれのPFNの出力パターンとの類似度を求めて、認識スコアとする。

3. 音素認識への適用実験

3.1 実験条件

認識対象は日本語の5母音とし、一人の話者が単語発声した12kHzサンプルの波形データから切りだした各母音のデータを学習に用いた。PFNの入出力には、この波形データ256ポイントに対する16チャネルのFFTメルスペクトラムを1フレームずつ10msのシフトで各母音2,000サンプルずつ与えた。FFTメルスペクトラムは1つの連続する母音区間のなかで平均0、71.0に正規化した。認識評価用データは学習と同じ話者の文節発声データから各母音1,000サンプル用いた。また、従来の分類型ニューラルネットとの比較を行うため、3層ネットワーク（ユニット構成は16-10-5）（Phoneme Classification Network；以下PCNと呼ぶ）を用いた認識実験も行った。入力データはPFNと同じFFTデータで、出力には5母音に対応するユニットに教師信号として0、1

のデータを学習させた。PFN 同様、入力層と出力層にはシグモイド関数を用いていない。

3.2 実験結果

FFTデータによる認識結果をFig.3に示す。PFNの３層モデルは同じ３層のPCNに比べ、１位の認識率は劣るが（PFN 79.5%、PCN 84.6%）、２位以下の累積認識率では優れた結果がでている（２位、PFN 94.6%、PCN 93.8%、３位、PFN 98.9%、PCN 96.3%）。また、３層PFN と５層PFN を比較すると、Fig.3でわかるように５層PFN の方が認識率（１位、82.4%、２位、95.6%、３位、99.2%）が良い。

候補カテゴリへの”近さ”が、PFNではどのように学習されたか調べるため、１位認識でPFN、PCNとも/i/に誤認識された１つの/u/のパターンに対する認識スコアをFig.4 に示す。認識スコアはPFNでは入力パターンとそれぞれの５母音PFN出力との類似度であり、PCNでは５母音に相当する出力ユニットの値である。また、×印は全ての評価用データから計算した５母音の平均値（標準パターン）と、この/u/パターンとの類似度（Similarity to Reference Vector；以下SRV と呼ぶ）を示したものである。この/u/パターンは/i/の標準パターンに近いものであるため、PFN、PCNとも１位候補は誤認識しているが、２位以下の候補についてはPCNは教師データとして０、１を教えているため、出力値が０近くに抑えられ、いわゆるハードディシジョンの現象が起きている。これに対してPFNは、１位候補から４位候補まで相対値として、SRV に近い認識スコアを示している。これがPFN の２位以下の認識率の向上につながった要因と考えらる。全体として、PFN、PCNの認識スコアが、SRV にどの程度近いかを見るために、すべての評価用データに対するPFN対SRV、PCN対SRVの類似度の平均を計算するとそれぞれ0.897、0.696であった。すなわち、マクロ的な評価においてもPFNの認識スコアはSRVに近いパターンを示しており、PFN はPCN よりも候補カテゴリへの近さを表現していることがわかる。

4. PFN の非線形効果確認実験

ニューラルネットの中間層のユニット数を少なくすることにより、特定のカテゴリのみ恒等写像を行なうフィルタを作るという考え方は、線形な手法によっても実現できる。そこで、線形直交変換であるK-L変換による圧縮復元の方法（以下KLT と呼ぶ）と、非線形写像を行なっていると考えられる５層PFN との比較を行なった。タスクとして、識別時、混同が多いとされている/N/と/u/の認識を行なった。データは３章の５母音のデータと同じ条件である。Fig.5、6に示すように入力パターン（FFTの１６次元の係数値）の平均は/N/と/u/は非常に良く似ており、/u/の方が全般的に分散が大きい。このため、ホルマントの位置では認識が困難であると考えられる

ため、ここではPFN、KLTとも、認識の距離尺度として類似度ではなく、平均2乗誤差（MSE）を用いることにした。また、フィルタとして効果的に働くための圧縮次元数の検討と、PFNの学習が効果的になされるためのHiddenユニットの数や学習方法の検討も行なった。

4.1 K-L変換による認識

K-L変換は主成分分析によって得られる第1の主成分から第kの主成分までのk次元空間に線形直交変換する方法である。よって、KLTは次式に示すようにn次元の入力パターン（ベクトル）X（ここではn=16）をk次元（$k<n$）のベクトルYに次元圧縮し、さらに元のn次元の空間のパターンX'に復元する。

$$Y = A_k \cdot X \qquad (4.1)$$

$$X' = A^{-1} \cdot Y \qquad (4.2)$$

ここで、$A = (e_1, e_2, \ldots, e_n)^t$、$A_k = (e_1, e_2, \ldots, e_k)^t$、$e_i$は主成分分析によってえられた入力データの共分散行列の固有値 λ_i（$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_i \geq \ldots \geq \lambda_n$）に対応する$n$次元固有ベクトルである。(4.2)式では$Y$の次元数が$k$であるため、$k+1$次以上の（$n-k$）個の要素に0を入れる。認識の方法はPFNによる方法と同じである。

Fig.7に圧縮次元数kと認識率の関係を示す。/N/と/u/の認識率のピークは/N/では低次、/u/では高次にあるが、両者の平均認識率は圧縮次元数kが7から9の時に最大になる。

4.2 PFNのHiddenユニット数

K-L変換による方法と比較するため、4.1節の結果から圧縮次元数を8とする。PFNの写像能力を生かすためのHiddenユニットの数を調べるために、5層PFNをFig.8に示すように上位サブネットと下位サブネットの2つに分割して考える。次元圧縮よりも復元（次元が増える）の方が高い写像能力が要求されるため、ここでは上位サブネットについての写像能力を調べる。上位サブネットにK-L変換で求めた/N/の圧縮パターンYを入力として与え、出力に元の/N/の入力パターンXを教師パターンとして与えたときの学習時の出力誤差（MSE）をFig.9に示す。Hiddenユニット数が28以上でほぼ写像学習の能力が飽和しているのがわかる。すなわち、少なくとも圧縮次元数が8以上の場合はHiddenユニット数は28で十分であることがわかる。

4.3 PFNの非線形効果

Fig.10にKLTとPFNとの出力誤差（MSE）の平均の比較結果を示す。圧倒的にPFNの写像

能力の方が高く、非線形写像の効果が確認できる。

次に4.2節の結果により、ユニット構成が16 - 28 - 8 - 28 - 16 のPFN で/N/ と/u/ の認識を行ない、8次元の圧縮のKLT と比較した。Fig.11 に学習データと同じ話者で単語発声の未学習データに対する認識率の比較結果を示す。PFN の認識率（97.6%）がKLT の認識率（86.7%）を上回っており（誤認識率で1/5 以下）、PFN の非線形性が効果的に働いていることがわかる。

4.4 PFNの分割学習

5層PFN は強力な非線形写像を実現するが、バックプロパゲーション学習法では出力誤差を入力層に向かって逆伝播するため、層数が増加すると学習が遅くなり、期待していない局所的最小値に陥りやすくなる。これを回避する方法として、Fig.8 で示したようにニューラルネットを分割して学習する方法が考えられる。5層PFN の場合は、圧縮のための下位の3層サブネット と復元のための上位の3層サブネット に分割することができる。この場合、圧縮層（下位サブネットの出力層、上位サブネット の入力層）に教師パターンを与える必要があるが、今回はK-L 変換によりえられたパターンを用いた。分割学習のみであれば、下位サブネット の圧縮性能はK-L 変換を越えることができないが、上位サブネット の復元性能は非線形の写像効果が期待できる。また、分割学習の後は2つのサブネット を結合して元の5層PFN にし、圧縮層をフリーにして再学習を行なうため、圧縮層の教師パターンはその時の初期値とみなせる。分割学習の効果を調べるために、Fig.12 に分割学習したものとしないものとの学習状況を比較した結果を示す。エラーの収束状況は分割学習するほうが早い。ただし、分割学習後、5層PFN の学習で大きな誤差の減少が見られないことから、圧縮層のパターンはK-L 変換のパターンから大きく変化していない可能性がある。下位サブネット に積極的に非線形写像の効果を獲得させるためには、圧縮層の教師データに工夫が必要である。

5. むすび

ニューラルネットによる音声認識の新しい方法として、中間層を圧縮した多層ニューラルネットで構成された音素フィルタを用いた音素認識の方法を提案した。日本語5母音の認識実験を行なった結果、2位以下の累積認識率が従来の分類型ニューラルネットより良好で、認識スコアは候補カテゴリへの距離を反映していることがわかった。また、主成分分析による方法との比較実験により、PFN の非線形写像の効果が確認できた。今回は認識スコアとして、入力パターンと出力パターンの類似度あるいは平均2乗誤差を用いたが、より自然な認識スコアを得るために、入力空間あるいは出力誤差空間の分散を考慮した距離を今後検討していく必要がある。さらに、

入力データのパワーの正規化方法の再検討や、PFN の分割学習時の圧縮層の教師パターンとして、非線形写像をもっと積極的に生かすような中間パターンの検討も行ない、単語認識へ発展させる予定である。

謝辞

研究の機会を与えていただいたATR 自動翻訳電話研究所の樽松社長に感謝いたします。また、日頃活発に討論していただく音声情報処理研究室の嵯峨山室長、杉山主幹研究員、ならびに研究員の皆様に感謝いたします。さらに主成分分析関係のプログラムを提供していただいたデータ処理研究室の坂野研究員、慶応大学の南氏に御礼を申し上げます

参考文献

(1) 川端豪, "k-近傍内挿学習による音韻認識 ", 日本音響学会講演論文集, Vol.I, 2-P-21, pp.161-162, (1990.3).

(2) 南泰浩, 田村震一, 沢井秀文, 鹿野清宏, "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", 日本音響学会講演論文集, Vol.I, 1-3-18, pp.35-36, (1990.3).

(3) 鷹見淳一, 嵯峨山茂樹, "対判定型TDNN による音素認識", 音声研究会(SP), (1990.6 発表予定).

(4) 小森康弘, 杉山雅英, " 近傍情報を用いたファジー学習(FuNI)法と音韻識別ニューラルネットワークによるその評価", 電子情報通信学会論文誌D(現在投稿中).

(5) G.W.Cottrell, P.Munro, D.Zipser, "Image compression by backpropagation", ICS Rep.8702 (1987).

(6) 森島繁夫, 中山博文, 清水誠司, 片山泰男, 原島博,

"ニューラルネットに基づく音声情報圧縮 ", 信学技報, SP88-142 (1988).

(7) D.E.Rumelhalt, G.E.Hinton, R.J.Williams, "Learning internal representations by error propagation" in Parallel

Distributed Processing, 1, M.I.T.Press (1986).

(8) 船橋賢一, " 3層ニューラルネットワークによる恒等写像の近似的実現についての理論的考察 ", 電子情報通信学会論文誌A, Vol. J 73-A, No.1, pp. 139-145 (1990.1).

(9) 船橋賢一, "ニューラルネットワークのcapabilityについて ", 信学技報, MBE88-52 (1988.7).

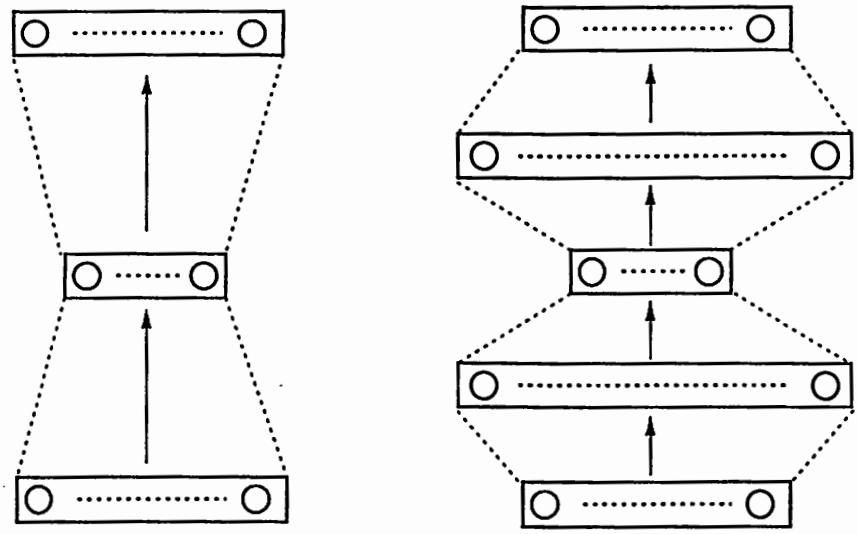
(10) B.Irie, S.Miyake, "Capabilities of Three-layered Perceptrons", Proc. ICNN88, Vol. 1, pp. 641-648 (1988).

(11) 片山泰男, 大山公一, "自己組織逆伝播ニューラルネットの諸特性 ", 信学会全国大会, SD

-1-14 (1989).

(12) 入江分平, 川人光男, "多層パーセプトロンによる内部表現の獲得 ", 信学技報, NC89-15 (1989).

(13) 森島繁夫, 中山博文, 清水誠司, 片山泰男, 原島博, "音声情報圧縮を実現する多層ニューラルネットの特性解析 ", 信学技報, SP89-121 (1989).



A. 3-Layer PFN B. 5-Layer PFN

Fig. 1 PFN ; Phoneme Filter Neural Network

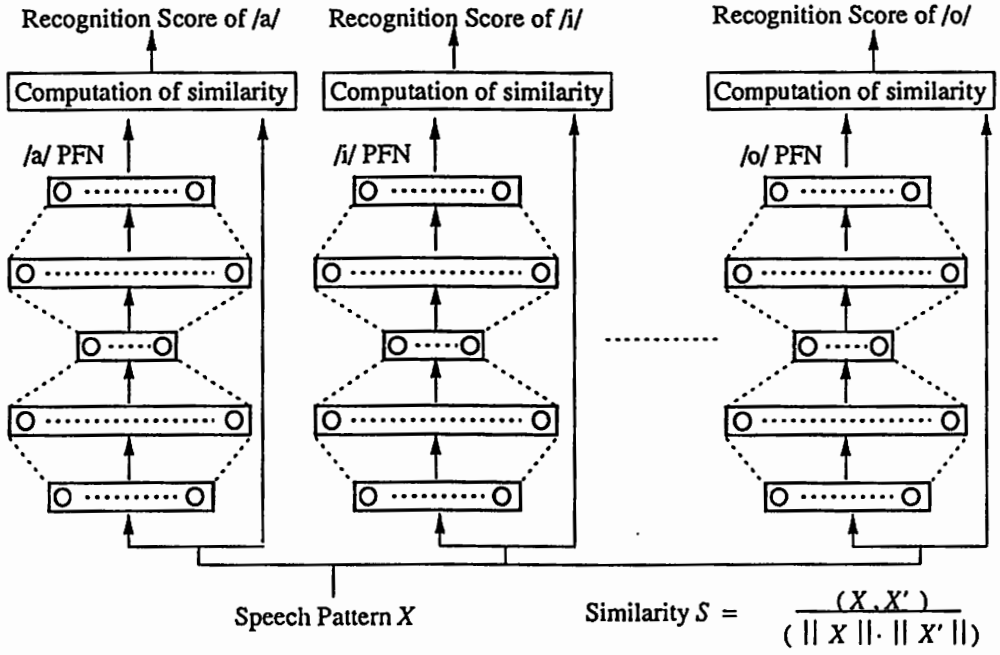


Fig.2 Japanese vowel recognition by PFN

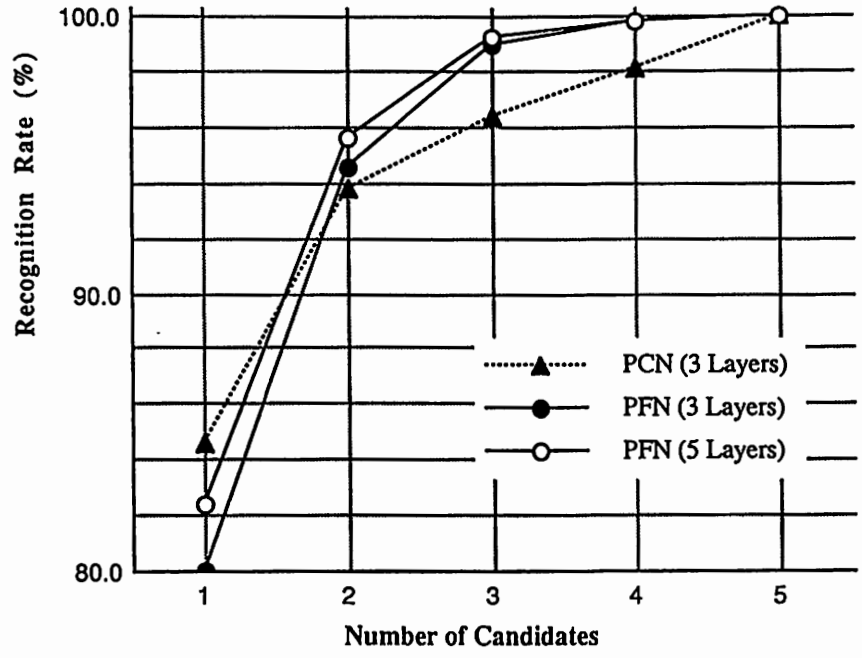


Fig.3 Vowel phoneme recognition results (PFN vs. PCN)

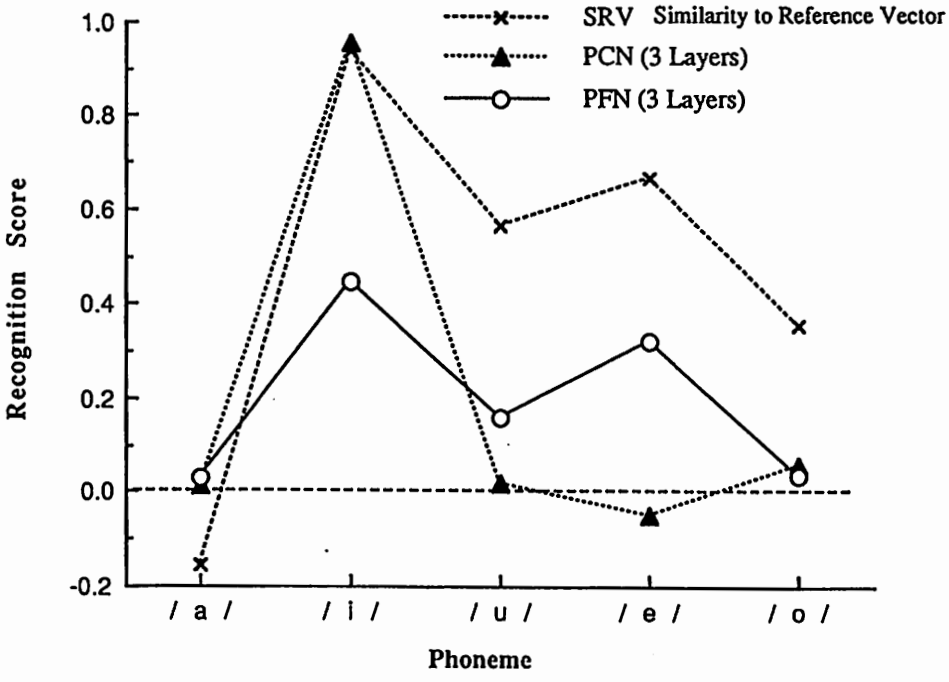


Fig.4 Recognition scores for an /u/ pattern

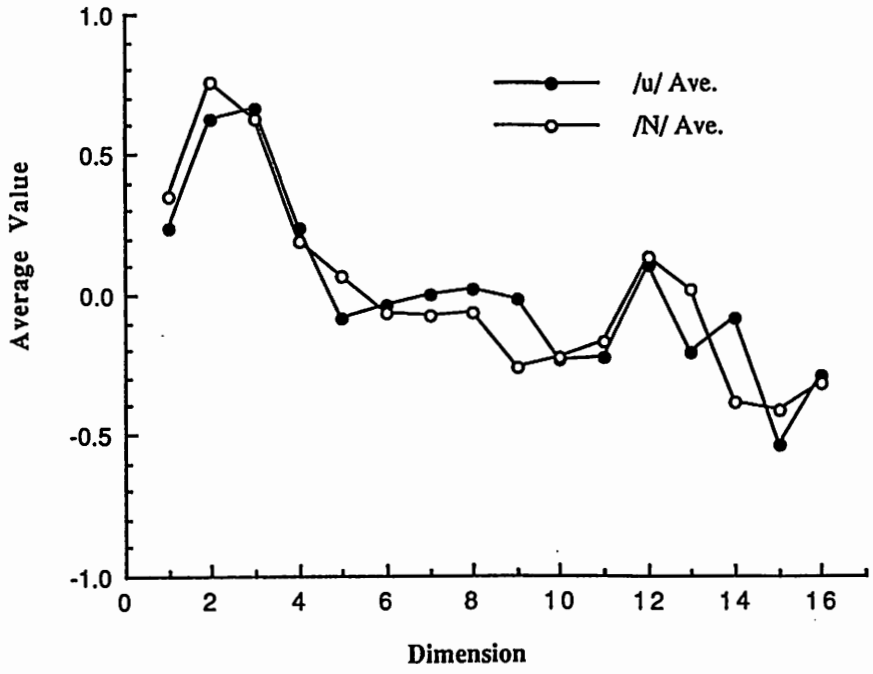


Fig.5 Averages of FFT parameters for /N/ and /u/ patterns

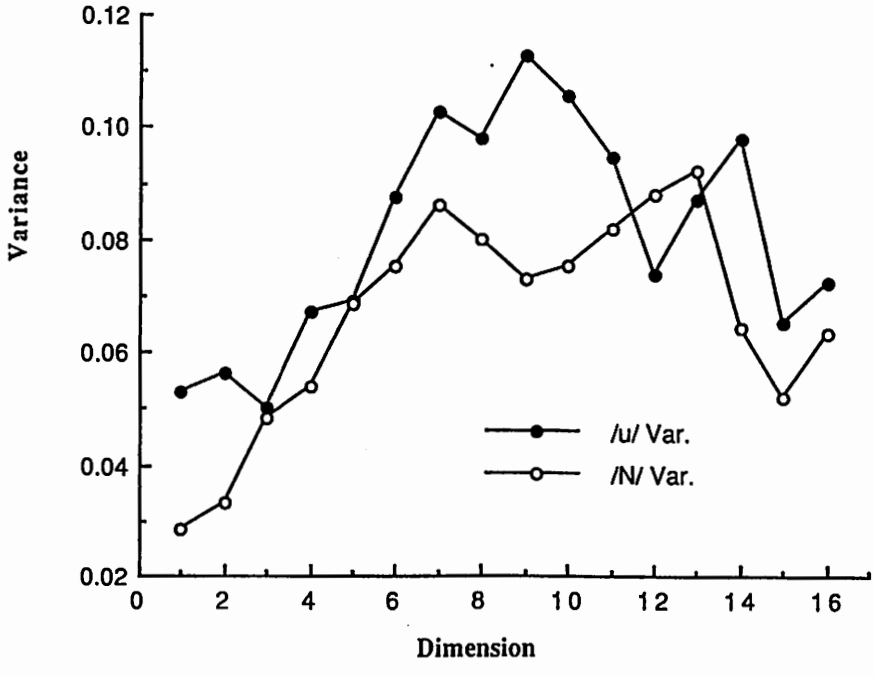


Fig.6 Variances of FFT parameters for /N/ and /u/ patterns

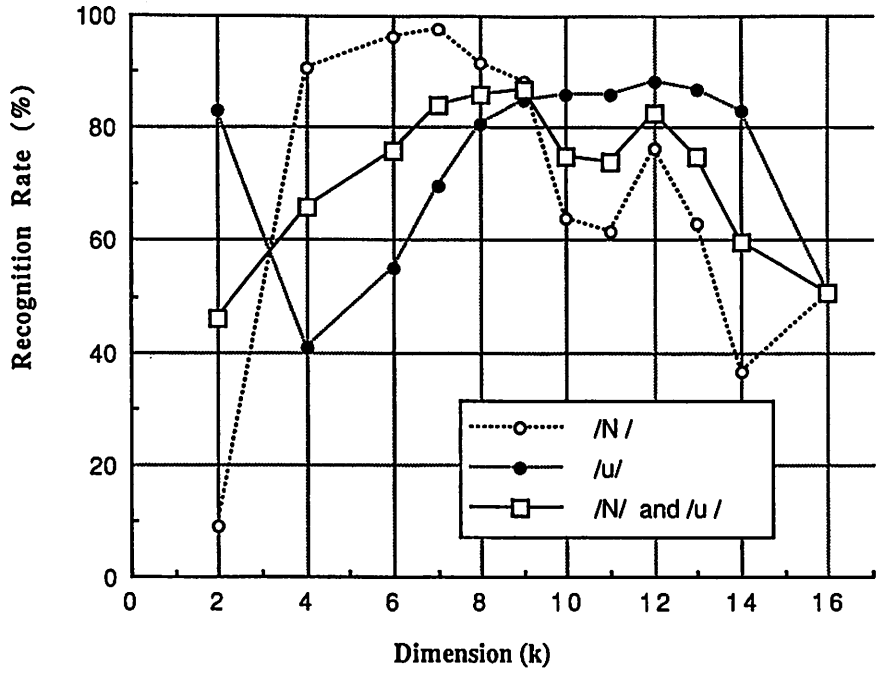
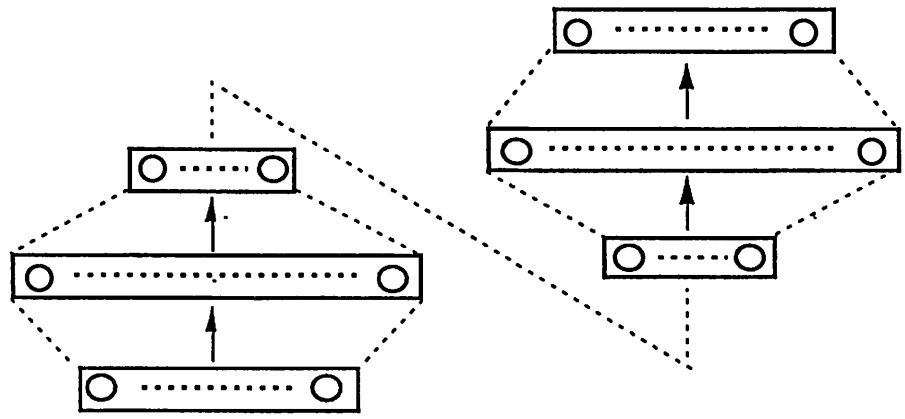


Fig.7 Recognition rates for /N/ and /u/ by KLT



a. Lower sub-network b. Upper sub-network

Fig. 8 Division of the 5-layrer PFN

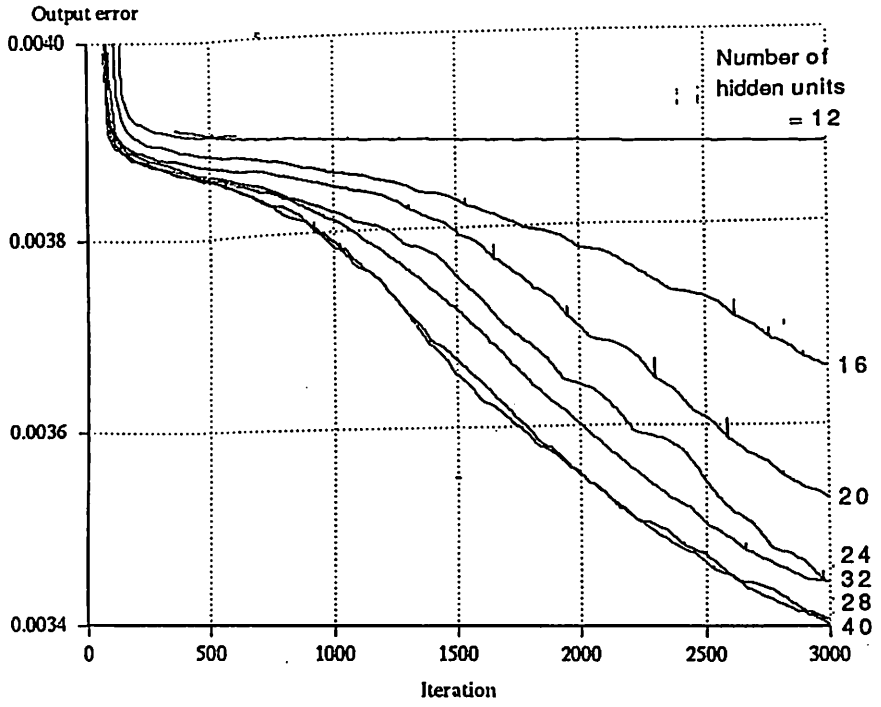


Fig. 9 Relation between the mapping ability and the number of hidden units of the upper PFN

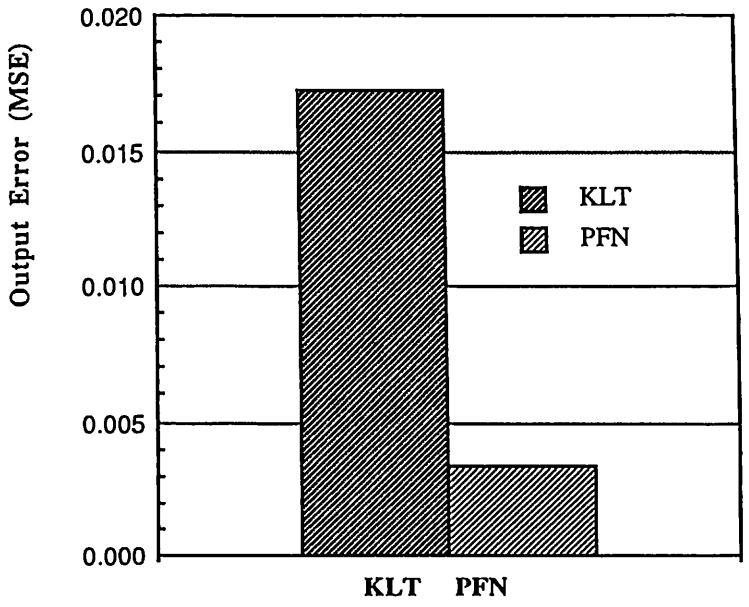


Fig.10 Output errors of KLT and PFN for /N/ patterns

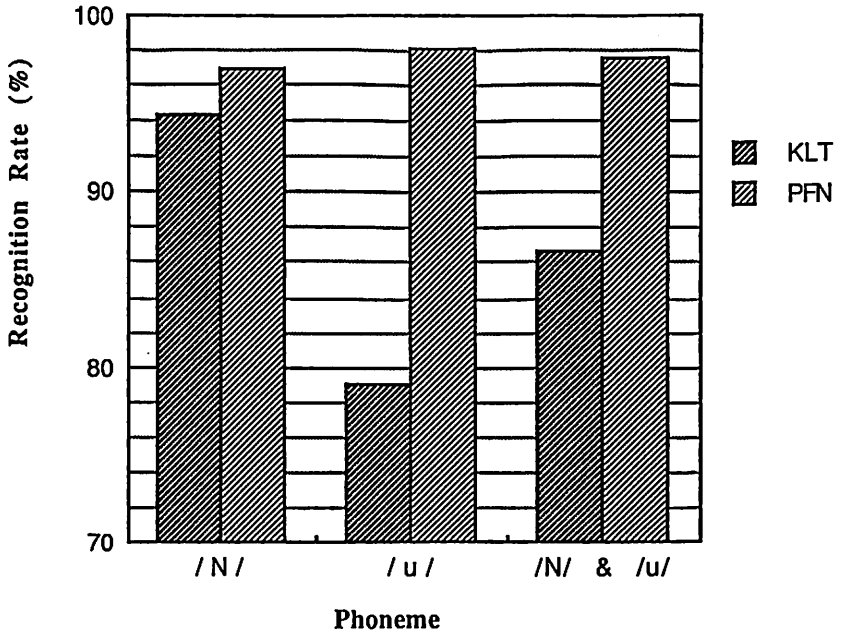


Fig.11 Recognition rates of PFN and KLT for /N/ and /u/ patterns

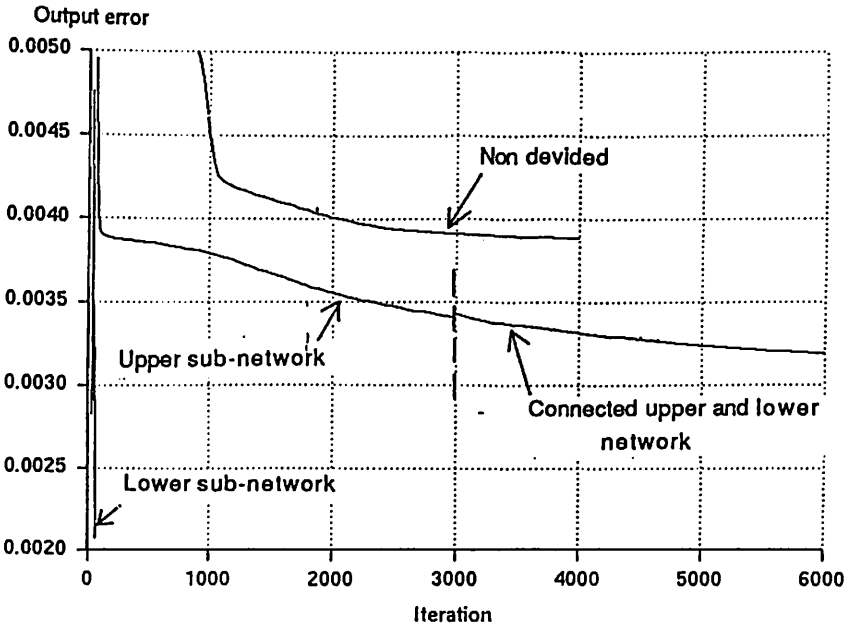


Fig. 12 Effect of the devided PFN learning

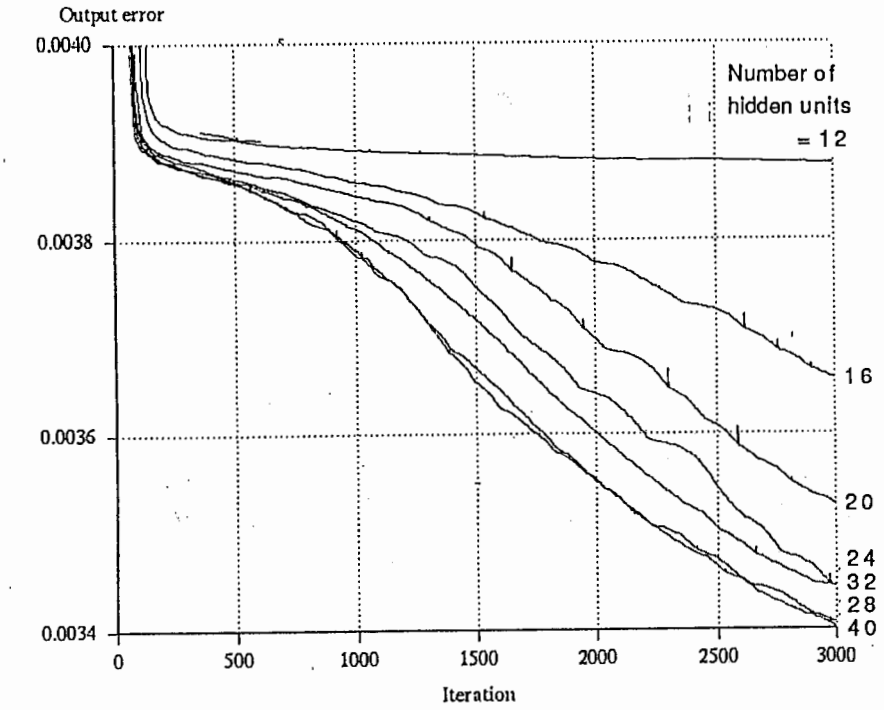


Fig. 9 Relation between the mapping ability and the number of hidden units of the upper PFN

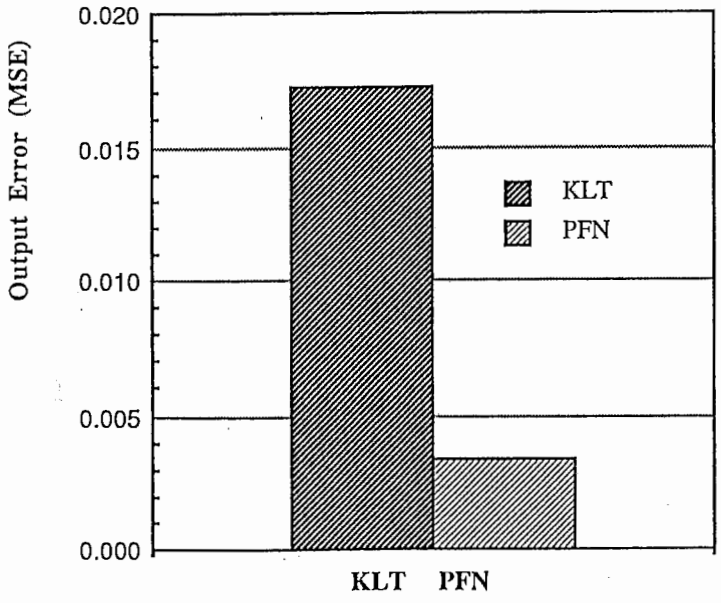


Fig.10 Output errors of KLT and PFN for /N/ patterns

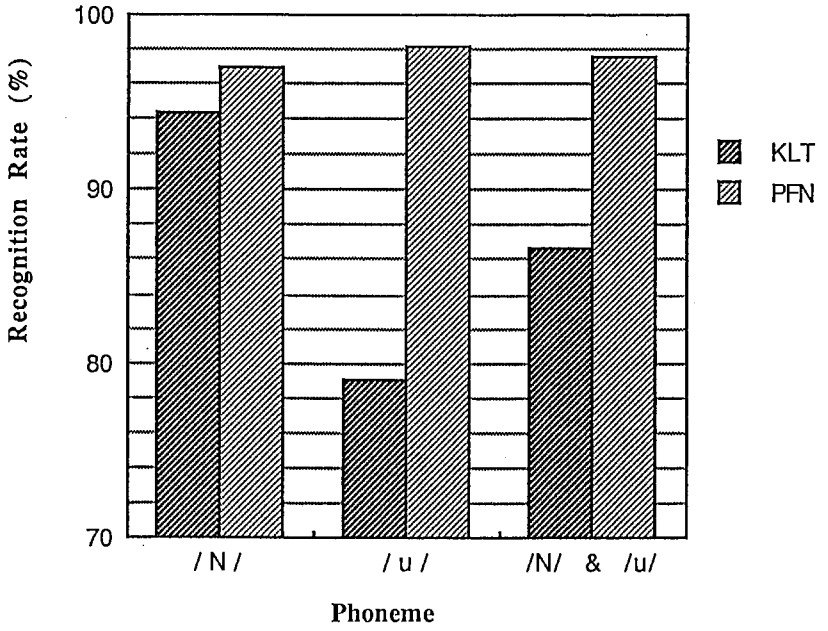


Fig.11 Recognition rates of PFN and KLT for /N/ and /u/ patterns

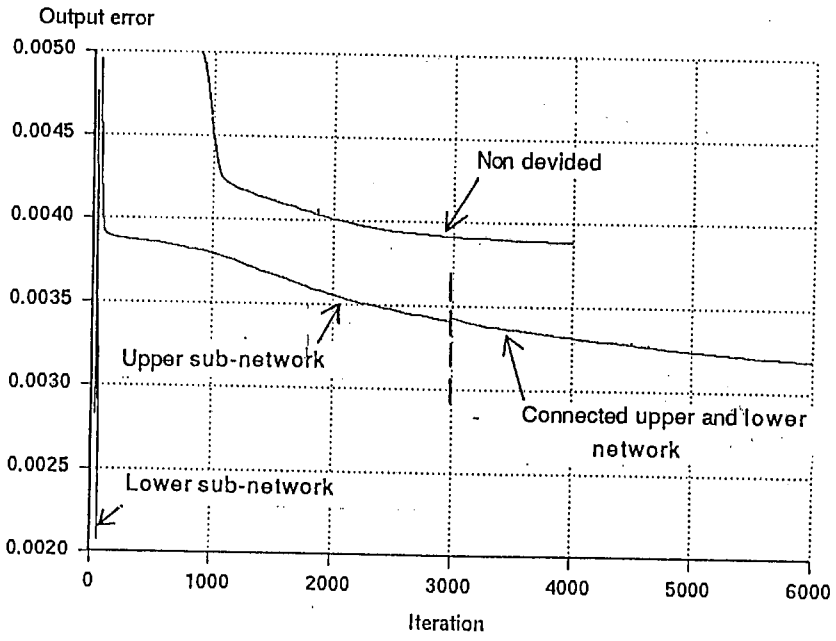


Fig. 12 Effect of the devided PFN learning