TR-I-0173

# ATR における Neural Network を用いた音声情報処理
# Neural Networks applied to Speech Processing in ATR

杉山 雅英
Masahide SUGIYAMA

## 概要

Neural Network Model が音声情報処理に応用され、種々の分野においてその有効性が報告されている。本資料では ATR における Neural Network Model の研究を概観し、今後の研究課題などを述べる。

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

# 目次

# 表目次



# 図目次

## 1 まえがき

Neural Network Model が音声情報処理に応用され、種々の分野においてその有効性が報告されている。本資料では ATR における Neural Network Model の研究を概観し、今後の研究課題などを述べる。さらに付録に最近の他の研究機関の NN 研究に関する発表論文の一覧を載せる。本報告では主に 1989 年及び 1990 年に発表された論文について概観している。それ以前の研究に関する概観については論文 [2], [3] を参照されたい。また言語処理への応用 [7], [41], [22]、エキスパートシステムとの融合 [21],[25]、理論的な解析に関する研究成果 [11], [12], [57] については割愛した。

## 2 ニューラルネットワークによる音声認識システム

### 2.1 TDNN による音素認識

CMU からの滞在研究員である A.Weibel 等によって始められたニューラルネットワークを用いた音声情報処理の研究は Time-Delay Neural Network (TDNN) として定式化され、その有効性が示された。TDNN の構造を図1に示す。分析条件などを以下の表1に示す。従来から研究されてきた HMM に基づく音素認識系に比べて高い音素認識率を与えることを示したことは、多くの研究者を刺激し国内での音声認識に対するニューラルネットワーク研究の火をつけることになった [1], [9], [24], [70]。

表 1: TDNN 認識系における音声分析条件

| | |
|---|---|
| 発声者 | 男性 1 名 (MAU) |
| 標本化周波数 | 12kHz |
| 窓 | 256 点 Hamming 窓 |
| 更新周期 | 10ms |
| 特徴量 | 16 帯域 FFT メルスペクトル |
| 正規化 | 15 フレーム内で平均を 0 にした後、最大最小の絶対値の大きい方で正規化 (最大・最小は +1,-1 の範囲内) |

### 2.2 TDNN + LR 音声認識系

TDNN を用いた音素スポッティング法と拡張 LR パーザとを組合せ、文節認識に適応し評価を行なった。これは基本的に HMM による音素認識と LR パーザとの組合せと同種のものである。前者はスポッティングを基本にしている点で動作が若干異なる。表2にその認識性能を示す [20]。他の方法の音素認識率の比較を表3に示す [70]。

表 2: TDNN-LR による文節認識率 (279 文節)

| 候補順位 | 文節認識率 (%) | | |
|---|---|---|---|
| | TDNN | HMM Single VQ | HMM Separate VQ |
| 1 | 63.4 | 72.0 | 83.2 |
| $\leq 2$ | 74.6 | 85.3 | 93.9 |
| $\leq 3$ | 78.9 | 91.8 | 96.4 |
| $\leq 4$ | 82.1 | 94.3 | 97.5 |
| $\leq 5$ | 83.2 | 95.3 | 98.6 |

Single VQ: Power + WLR
Separate VQ: Power + WLR + DCEP
LR 探索 beam 幅 $\leq$ 18(local); $\leq$ 250(global)
TDNN-LR 探索 beam 幅 $\leq$ 100

表 4: 音声資料の仕様

| | 発話スピード | 記事 |
|---|---|---|
| 単語音声 | 5.68 モーラ | 5240 単語の奇数番目 |
| 短い文節 | 7.14 モーラ | 複合語を許さない文節発声 |
| 長い文節 | 7.72 モーラ | 文節発声 |
| 自由発声 | 9.56 モーラ | 区切り指定なしの連続発声 |

参考のために評価に使用した音声資料の仕様を表4 に示す。

## 3 TDNN を越えて ( 問題点克服の試み )

HMM に基づく方法に比べて累積認識率という点に関して低くなっている。学習の結果、NN の出力値が教師信号と同様に 0-1 と離散的となり、2位・3位の認識率に情報がなくなってしまう。そのため累積認識率が低くなると考えられる。教師信号に大きく依存してしまうこの現象は過学習と呼ばれる。一方、HMM の学習においては category の再現性を主眼としており、基本的な定式化においては他の認識 category の影響を考慮しておらず、出力される確率 (歪み) の大小には再現性の観点からの確率的な意味を持っている。そのため2位・3位の認識率にも情報を持つと考えられる。

### 3.1 平滑化による robustness の向上

NN におけるこの過学習の問題を解決するための平滑化の方法として、以下のような検討がなされてきた。

- 出力における平滑化 [29]
- 学習における平滑化

表 3: 各種の方法による音素認識率 (18 子音)

| 認識方法 | rank | 単語発声 | 短い文節 | 自由発声 |
|---|---|---|---|---|
| | | 5.7 モーラ /sec | 7.1 モーラ /sec | 9.6 モーラ /sec |
| Discrete HMM | $=1$ | 93.1 | 81.4 | 71.6 |
| | $\leq 3$ | 99.5 | 95.9 | 92.4 |
| Continuous HMM | $=1$ | 98.1 | 79.7 | 66.6 |
| | $\leq 3$ | 99.8 | 94.6 | 86.9 |
| TDNN | $=1$ | 96.2 | 76.2 | 56.6 |
| | $\leq 3$ | 99.6 | 91.5 | 78.5 |
| LVQ | $=1$ | 97.9 | 81.7 | 61.6 |
| | $\leq 3$ | 99.9 | 96.8 | 87.7 |
| LVQ+HMM | $=1$ | 97.2 | 80.6 | 69.6 |
| | $\leq 3$ | 99.6 | 94.6 | 89.3 |
| Fuzzy LVQ+HMM | $=1$ | 94.4 | 80.8 | 74.0 |
| | $\leq 3$ | 99.7 | 96.8 | 93.3 |

- 分布を平滑化した学習 [30]
- 連続値教師信号による学習 [27]

以下でこれらの研究の概説を行なう。

#### 3.1.1 出力における平滑化

[29] において、 NN の出力の平滑化を検討している。主成分分析 ( 5 次元) による直交基底ベクトルを算出しその方向の代表点に対する TDNN 出力値を用いた出力の平滑化を行なう。その概念を図 2に示す。TDNN の入出力関数を $f(x)$ と表す時、その平滑化関数 $\hat{f}(x)$ は以下のように定義される。

$$\hat{f}(x) = \frac{\int_{|x-y|<\epsilon} f(y)dy}{\int_{|x-y|<\epsilon} dy} \quad (1)$$

図 3に示すように平滑化により通常の TDNN に比べて 2% 程度識別率の向上が得られている。さらに、累積識別率においても通常の方法に比べて優れていることが示されている。積分する領域については、入力層において積分により平滑化するよりも第 1 中間層での平滑化の方が有効であることが示されている。主成分分析の次元の取り方、計算量の評価などの検討項目が残されている[1]。

#### 3.1.2 学習における平滑化 (ファジー (Fuzzy) 学習)

[30] において学習における平滑化を検討している。図 4に示すように従来の NN の学習が入力空間から出力空間への点対応の学習であったのに対して、入力空間の 2 点及びその対応する出力の 2 点を結ぶ線分の間の対応関係を学習させる。連続体 (線分等) に対する学習を実現するためには連続体をいくつかの点で覆い尽くし、その点における対応関係を学習させる方法が考えられるが、そのためには学習時間が膨大になる。それに対して、無限小の近傍の 2 点に対する近似化した方式を用いることにより従来方法の数倍程度の計算量で、線分の学習を実現している。 KNIT 法を用いた母音認識結果を表 5に示す。積分の持つ意味、積分区間を線分の中点としている点、 等の検討項目が残されている。

表 5: KNIT 法を用いた母音認識実験結果

| 学習方法 | 学習 sample | 認識率 |
|---|---|---|
| 従来法 | closed | 100.0 |
| 従来法 | open | 97.6 |
| KNIT | closed | 99.6 |
| KNIT | open | 99.2 |

[27] において図 5に示すような学習における平滑化を検討している。学習に用いる教師信号に確率のような連続値を与えることによって、より正確な学習を目指している。連続値を持つ教師信号は、学習点の近傍に含まれる点の情報によって決定される。図 6に示すように、この方法によって識別率が向上する。

### 3.2 Network 構造の検討

TDNN のネットワーク構造について文献 [54] において検討を行なっている。検討項目は以下の通りである。

[1] $\epsilon$ 近傍の取り方は、学習音声を認識するとき、認識率が最大になるように設定している。

- ネットワークモジュール構成の比較
  - 分散型 (図 7)
  - 集中型 (図 8)
- 出力、中間層間の接続法の比較
  - 自由な接続
  - 前部 (4 個) と後部 (5 個) とに分割
- 入力、中間層間の接続法の比較
- 出力、中間層間の重み係数に窓掛けする方法
- 周波数軸方向の重み係数の平滑化
- クラス分けネットワークの効果

これらの検討から以下の知見が得られている。

- クラス分けネットワークの有効性
- 入力層、中間層および出力層、中間層の間の shift invariant 接続の有効性

## 3.3 対判定 NN

TDNN を用いた音素識別において第 1 位の識別率は非常に高いが、第 2 位以下の累積識別率はあまり向上されず、認識システムの中に組み込む場合の問題点となることが指摘されている。これを解決する方法として、論文 [34],[40] では TDNN の学習により構成される critical な識別境界面をより堅牢なものとするために、図 9 に示すような音素対判別 NN を提案している。これは音素 $/p_i, p_j/$ を判定する小規模な TDNN を作成し、$/p_i/$ に対する NN の出力値を

$$\sum_{j \neq i} S(p_i|p_i : p_j) \tag{2}$$

のように対判定 NN 出力の加算で定義する。ここで、$S(p_i|p_i : p_j)$ は音素対 $/p_i, p_j/$ を識別する NN の $/p_i/$ に対する出力であり、入力音素が $/p_i/$ のとき 1, $/p_j/$ の時、0 であるように学習されている。さらに、$/p_i, p_j/$ とそれ以外の音素に対して 0.5 を教師信号とし、0.5 での学習を円滑に進めるために図 10 に示すような 2 つの sigmoid 関数を接続した非線形関数 $f(x)$ を用いている。

$$f(x) = \begin{cases} \dfrac{g(x+\alpha)}{2g(\alpha)} & (x < 0) \\ 1 - \dfrac{g(-x+\alpha)}{2g(\alpha)} & (x \geq 0) \end{cases} \tag{3}$$

この音素対判別 NN を用いた $/b, d, g, m, n, N/$ に対する識別率を表 6 に示す。比較のために従来の TDNN, 中間値学習なしおよび学習ありの各々の構造を評価した。その結果、中間値を教師信号とすることにより認識率が向上すること、また第 2 位以下の累積認識率が改善されることが示されている。

表 6: 対判定 NN を用いた音素認識結果

$(/b, d, g, m, n, N/)$

| 発話様式 | 第 1 位 | 第 2 位 | 第 3 位 |
|---|---|---|---|
| 単語発声 | 95.2 | 99.2 | 99.8 |
| | 94.7 | 98.6 | 99.7 |
| | 97.3 | 99.2 | 99.8 |
| 文節発声 | 84.8 | 92.7 | 96.5 |
| | 80.6 | 93.9 | 98.1 |
| | 88.8 | 96.9 | 99.2 |
| 短い文節発声 | 84.5 | 91.2 | 95.3 |
| | 77.2 | 93.2 | 97.0 |
| | 86.2 | 95.4 | 98.1 |
| 自由発声 | 77.8 | 87.1 | 92.2 |
| | 73.2 | 90.2 | 96.3 |
| | 81.6 | 93.3 | 96.7 |

上段: 一括判定型 TDNN
中段: 中間値学習なし
下段: 中間値学習あり

## 3.4 時間構造考慮 NN

TDNN は音素識別性能が高く、シフトトレラントであることが示されているが、学習時と異なる発話様式の音声に対する識別能力はあまり高いとは言えない。音素の時間構造の特徴はより robust であると考えられる。時間構造考慮 NN の構造を図 11,12に示す。入力層に対して、第 1 隠れ層 (もしくはさらに第 2 隠れ層) に対して時間変化に対応する構造を導入する。音素の時間構造を中間層の 4 つの状態の連鎖で表現する。入力層と 4 つの第 1 隠れ層との結合は TDNN と同様である。ただし、そのシフトの範囲は 15 フレーム全体に渡るのではなくその部分に対応する。その他の結合は全結合とする。表 7に述べるように実験の結果、連続発声の文節中の音素識別 $(/b, d, g, m, n, N/)$ に対して有効であることがわかった。ただし、シフトトレランスに対しては robust でなくなる問題点も指摘されている [26]。

## 3.5 時空間 Block 統合 NN

TDNN はその構造に shift invariant (重み係数の tied connection) を導入することにより、時間軸におけるずれに対する耐性を高めている。音声における音素の変動は時間的なものと周波数軸におけるずれとがあり、認識を難し

表 7: 時間構造を考慮した NN による音素認識率

$(/b, d, g, m, n, N/)$

| 入力音声 | TDNN | TS | STS |
|---|---|---|---|
| 短い文節 | 76.6 | 82.7 | 79.0 |
| 長い文節 | 75.9 | 77.6 | 77.7 |
| 自由発声 | 61.8 | 70.7 | 69.0 |
| -20ms | 91.2 | 52.7 | 58.7 |
| -10ms | 94.7 | 86.7 | 92.6 |
| 0ms | 95.7 | 97.6 | 97.0 |
| +10ms | 94.1 | 84.4 | 89.9 |
| +20ms | 85.9 | 56.5 | 65.8 |

TS: Temporal Structure 型
STS: Shifted Temporal Structure 型

表 8: 時空間 Block 統合 NN を用いた音素認識率

$(/b, d, g, m, n, N/)$

| 入力音声 | TDNN | TF | BW |
|---|---|---|---|
| 単語発声 | 95.9 | 96.7 | 98.2 |
| 短い文節 | 78.7 | 83.6 | 84.1 |
| 長い文節 | 79.8 | 84.3 | 83.9 |
| 自由発声 | 67.2 | 80.8 | 82.8 |

TF:Time Frequency TDNN
BW:Block Window NN

くさせている。この shift invariant 構造を周波数軸にも拡張しようとする試みである。論文 [37] では 2 つの方法を提案している。第一の方法は図 13 に示すような周波数軸におけるずれを吸収する構造をいれ、時間軸でのずれを吸収する従来の TDNN と上位の層において統合する構造 (TF) を考えている。また第 2 の方法は図 14 に示すような時空間パターンの 1 つのブロックを上位の層に統合する構造 (BW) である。入力層、第 1 、第 2 隠れ層において、時間 x 周波数を 3x4 , 5x5, 5x5 ごとに上位の層に統合させている。TF においては TDNN と同様に tied connection を周波数軸方向にもいれている。また一方、BW においては重み係数はすべて free connection としている。これら 2 つの構造と従来の TDNN との性能の比較を表 8 に示す。この結果から上の 2 つの構造は単語発声においても有効であるばかりでなく、異なる発話様式の音声に対しても顕著な認識率の改善を実現している。

表 9: 各種の雑音抑圧手法を用いた音素認識の比較

$(/b, d, g/)$

| SNR | 入力 | 評価方法 | | | |
|---|---|---|---|---|---|
| | | 無処理 | NN | CM | NN+CM |
| 5 (dB) | M1 | 43.8 | 59.9 | 58.3 | 62.3 |
| | M2 | 35.5 | 50.9 | 63.6 | 57.2 |
| | M3 | 33.7 | 64.4 | 61.5 | 59.6 |
| | M4 | 33.9 | 52.7 | 60.4 | 55.2 |
| 20 (dB) | M1 | 61.0 | 59.9 | 68.9 | 62.1 |
| | M2 | 62.3 | 55.1 | 56.2 | 61.1 |
| | M3 | 56.3 | 64.3 | 72.7 | 63.9 |
| | M4 | 53.1 | 60.2 | 61.7 | 61.3 |

NN: 雑音抑圧 NN を使用
CM: コードブックマッピングを使用
M1: WLR+DCEP+POW, M2:WLR+DCEP
M3:WLR+POW, M4:WLR

## 4 離散値写像 (discrete mapping) NN と 連続値写像 (analog mapping) NN

上で述べた NN は音素カテゴリーを認識するいわゆる識別型の Neural Network であった。それに対して、 NN の持つ非線形性を利用した非線形写像による音声処理の研究が行なわれている。雑音抑圧への応用、恒等写像を用いた音素認識、話者適応について述べる。

### 4.1 波形上での雑音抑圧

雑音抑圧への応用は NN による非線形写像の 1 つの例としてはやくから研究された [5], [10]。入力に雑音下の音声、出力に雑音除去音声を提示することにより、写像を構成する [19],[31]。雑音抑圧 NN を図 15 に示す。雑音抑圧の性能を被験者による聞き取り実験によって評価している。その結果、従来の spectral subtraction 法よりも優れていることが示された。処理量に若干問題があるが、認識においては、波形上での写像ではなく特徴量上での写像を構成することにより、高速化が図れるものと思われる。さらに、表 9に示すように VQ codebook マッピングによる方法との比較検討もなされており、それらの組合せにより、性能が向上することが得られている [39]。

### 4.2 音素ニューラルフィルター (PFN)

各カテゴリ毎に多層ニューラルネットワークを用いて、恒等写像を構成することにより、そのカテゴリ特徴をネットワークに学習させるものである。TDNN による認識においては第一位での認識率は高いが第 2 位以下の累積認識率が他の認識モデルに比べて低いという問題点が指摘され、

表 10: セグメント話者適応 NN を用いた音素認識

(/b,d,g/)

| | 未知話者 | 標準話者 |
|---|---|---|
| 適応前 | 83.6 | 98.7 |
| 適応後 1 | 86.6 | - |
| 適応後 2 | 67.5 | - |

適応後 1: 恒等写像のネットワークを用いた話者適応
適応後 2: ランダムな weight 値からの話者適応
未知話者:MHT, 標準話者:MAU

表 11: TDNN による不特定話者音素認識率

(/b,d,g/) 評価話者: 4 名

| 構造 | 学習用話者 | | | |
|---|---|---|---|---|
| | 6 話者 | | 12 話者 | |
| | closed | open | closed | open |
| Monolithic | 93.2 | 85.1 | - | - |
| Modular | 97.7 | 92.1 | 97.1 | 95.5 |
| SID | 96.4 | 80.0 | 85.8 | 85.8 |
| Meta-pi | 96.9 | 82.0 | 95.4 | 85.9 |
| Modular SID | 97.2 | 89.7 | 97.3 | 95.6 |

その解決方法として、提案されたものである [32],[33]。PFN の構造を図 16に示す。入力音声 $X_{in}$ は音素カテゴリー $(j)$ 毎に作成された NN に入力されその出力 $X_{j,out}$ との類似度 $S_j$ が計算される。

$$S_j = \frac{(X_{in}, X_{j,out})}{\|X_{in}\| \, \|X_{j,out}\|} \tag{4}$$

その類似度を用いて入力のカテゴリーを判定する。PFN を用いた母音認識実験の結果を図 17に示す。

この構造は磯による NN 予測問題 [60] と関連が深く、過去の点を用いない、即ち縮退した場合と見ることもできる。新たな枠組という点で注目される。NN の持つ識別率を最大にする学習という枠から出てしまう点で性能の劣化が懸念される。

### 4.3 話者適応

話者適応は教師ありと教師なしとに分類される。教師なしの方法は発声者に対する負担が小さいという点で優れているが、一般的には教師ありの方法に比べて、性能が低い。ATR においては教師ありの方法に関して VQ codebook 間の写像構成に基づく方法が研究されてきた [28]。さらにこれを NN で実現する方法が提案されその有効性が示されている [59]。ATR においてはその発展として時間構造を持った音響セグメントの間の写像構成問題に取り組み、恒等写像による初期学習と未知発声者の単語音声による追加学習とを組合せた新たな方式を提案し、その有効性を示している [35]。その NN の構造を図 18に示す。また適応系、認識系の処理の流れを図 19に示す。音素認識結果を表 10に示す。

### 4.4 教師なし構造変換 NN

音声の諸分野において 2 つの集合の要素の対応付けに帰着される問題は数多い。集合の各要素の対応付けに関する何らかの教師信号に基づいた写像構成方法は強力であるが、教師なし話者適応において成功を収めてきた教師信号を自動的に生成し写像を構成する方法に比べて、教師信号を与えなければならないという点で柔軟性に乏しい。NN の重み係数を教師なしで自動的に学習できれば、より多くの学習パターンを提示できることになり、未学習パターンに対する識別性能を向上できる可能性を持つ。文献 [36] では 図 20 に示すような VQ と BP とを組み合わせた学習方法と図 21に示すような階層的な学習方法とを提案している。

## 5 今後の研究課題

今後の Neural Network における研究項目を以下に示す。これらのうち、ATR において大半は既に研究が開始され成果が報告されている。

- Fuzzy 学習
- Phoneme Neural Network Filter [32], [33]
  不動点アルゴリズムを用いた解析 [97]
- 話者適応 [35]
- 構造の学習 (Speaker Adaptation) [36]
- Frequency Block TDNN [37]
- TDNN-LR の高性能化 [38]
- Reccurent Network の評価
- HMM と NN との融合
- Boltzmann Machine[51]
- 不特定話者への対応 [55]
- NN による文法などの記述

文献 [55] において論文 [92] で提案されている多数話者用の NN およびその改良モデルの性能を評価しており、非常に高い性能を得ている。

6

## 6 むすび

ATR におけるニューラルネットワーク研究を概観し、それぞれの研究の関連性などについて述べた。本報告では音声情報処理に応用されている実験的な側面について述べた。上で述べた今後の課題について検討を加え知見を得ると同時に、NN でなくてはできない音声処理技術分野の開拓へと発展させていきたい。

### 謝辞

日頃御指導いただく樽松社長に感謝します。また本資料執筆に当たって有益な助言をいただいた、鹿峨山研究室長をはじめとする音声情報処理研究室の諸氏に感謝します。また有益な討論を下さった ATR 視聴覚機構研究所の片桐主任研究員、岩見田研究員、マクダーモット研究員に感謝します。

## 参考文献


[1] 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, pp.19-24 (1987-12).

[2] 鹿野, 中村, 田村, Waibel, ニューラルネットワークの音声情報処理への応用, 音響学会誌, Vol44, No.10, pp.798-804 (1988).

[3] 田村, 沢井, 中村, 鹿野, ニューラルネットを用いた音声処理, テレビジョン学会誌, Vol.43, No.9 pp.935-943 (1989).

[4] 中村, 鹿野, コネクショニストモデルによる単語列予測の検討, 音学講論, 3-P-8, pp.243-244 (1988-03).

[5] 田村, ワイベル, Neural Network を使った波形入出力による雑音抑圧, 音学講論, 3-P-13, pp.253-254 (1988-03).

[6] P.Haffner, A.Waibel, K.Shikano, Fast Back-Propagation Learning Methods for Neural Networks in Speech, Rec. Fall Meet. Acoust. Soc. Jpn. 2-P-1, pp.203-204 (1988-10).

[7] 中村, 鹿野, ニューラルネットによる N-gram 単語列予測モデルの検討, 音学講論, 2-P-2, pp.205-206 (1988-10).

[8] 沢井, ワイベル, 鹿野, 時間遅れ神経回路網による音節スポッティングの検討, 音学講論, 2-P-11, pp.223-224 (1988-10).

[9] A.Waibel, H.Sawai, K.Shikano, Phoneme Recognition by Modular Construction of Time-Delay Neural Networks, Rec. Fall Meet. Acoust. Soc. Jpn. 2-P-12, pp.225-226 (1988-10).

[10] 田村震一, 波形入出力による雑音抑圧ニューラルネットワークの解析, 音学講論, 2-P-18, pp.237-238 (1988-10).

[11] 船橋, ニューラルネットワークの capability について, MBE88-52, pp.733-740 (1988).

[12] 船橋, 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, Vol.J73-A, No.1, pp.139-145 (1990-01).

[13] P.Haffner, H.Sawai, A.Waibel, K.Sikano, Fast Back-Propagation Learning Methods for Large Phonemic Neural Networks, Rec. Spring Meet. Acoust. Soc. Jpn. 1-6-14, pp.27-28 (1989-03).

[14] 宮武, 沢井, 鹿野, 全音韻を統合した時間遅れ神経回路網 (TDNN) による音韻スポッティング, 音学講論, 2-P-24, pp.277-278 (1989-03).

[15] H.Sawai, M.Miyatake, K.Shikano, Spotting Phonemes by Hierarchical Construction of Time-Delay Neural Networks, Rec. Spring Meet. Acoust. Soc. Jpn., 2-P-25, pp.279-280 (1989-03).

[16] 片桐, マクダーモット, 横田, 自己組織化特徴写像を用いた動的特徴の表現, 音学講論, 2-P-17, pp.263-264 (1989-03).

[17] E.McDermott, S.Katagiri, LVQ2 for Multi-Phoneme Recognition, Rec. Spring Meet. Acoust. Soc. Jpn. 2-P-28, pp.285-286 (1989-03).

[18] 横田, 片桐, マクダーモット, LVQ 音韻認識システムにおける学習の最適化, 音学講論, 2-P-29, pp.287-288 (1989-03).

[19] S.Tamura, An Analysis of a Noice Reduction Neural Network, Proc. ICASSP89, Vol.3, pp.2001-2004 (1989-05).

[20] 南, 宮武, 沢井, 鹿野, TDNN 音韻スポッティングと拡張 LR パーザを用いた文節音声認識, 音響学会講演論文集, 3-1-11, pp.97-98 (1989-10).

[21] 畑崎, 小森, 田中, 川端, 鹿野, スペクトログラムリーディング知識に基づく音韻認識エキスパートシステムにおける音韻識別ニューラルネットワークの融合法の検討, 音響学会講演論文集, 3-1-14, pp.103-104 (1989-10).

[22] 丸山, 中村, 川端, 鹿野, HMM 音韻認識と NETgram を用いた単語音声認識, 音響学会講演論文集, 2-P-8, pp.145-146 (1989-10).

[23] 中村, 田村, 宮武, 沢井, ニューラルネットワークベンチシステムの開発, 音響学会講演論文集, 2-P-25, pp.181-182 (1989-10).

[24] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K.J. Lang, Phoneme Recognition Using Time -Delay Neural Networks, Trans., ASSP-37, No.3, pp.328-339 (MAR. 1989).

[25] 小森, 川端, 鹿野, 畑崎, 音韻認識エキスパートシステムにおける母音認識 TDNN による母音スポッティング, 音響学会講演論文集, 2-P-18, pp.155-156 (1990-03).

[26] 小森, 鹿野, 南, 時間構造を考慮したニューラルネットワークによる音韻認識, 音響学会講演論文集, 2-P-19, pp.157-158 (1990-03).

[27] 小森, 杉山, 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評価, 電子情報通信学会論文誌投稿中.

[28] 中村, 鹿野, ベクトル量子化話者適応の TDNN 音韻認識への適用, 音響学会講演論文集, 2-P-22, pp.163-165 (1990-03).

[29] 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, pp.35-36 (1990-03).

[30] 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, pp.161-162 (1990-03).

[31] 田村, 中村, 波形入出力による雑音抑圧ニューラルネットワークの改良, 音響学会講演論文集, 3-4-18, pp.303-304 (1990-03).

[32] 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, pp.167-168 (1990-03).

[33] 中村, 田村, 嵯峨山, ニューラルネットによる音韻フィルタを用いた母音認識, 音声研究会資料, pp.17-24 (1990-06).

[34] 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, pp.9-16 (1990-06).

[35] 福沢, 他, ニューラルネットワークによる恒等写像を用いた話者適応, 音響学会講演論文集, 1-8-16, pp.31-32 (1990-09).

[36] 杉山, 他, ニューラルネットによる集合間写像の教師なし学習, 音響学会講演論文集, 2-P-10, pp.149-150 (1990-09).

[37] 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, pp.153-154 (1990-09).

[38] 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, pp.151-152 (1990-09).

[39] 大倉, 杉山, 波形入出力による雑音抑圧ニューラルネットの音声認識への応用, 音響学会講演論文集, 1-8-3, pp.5-6 (1990-09).

[40] 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, pp.159-160 (1990-09).

[41] M.Nakamura, K.Maruyama, T.Kawabata, K.Shikano, Neural Network Approach to Word Category Prediction for English Texts, COLING'90, Helsinki (1990-08).

[42] P.Haffner, 他, 音声ニューラルネットワークのためのバックプロパゲーションアルゴリズムの高速化, TR-I-0058 (1988-11).

[43] P.Haffner, ニューラルネットワークにおける高速学習プログラム, TR-I-0059 (1988-11).

[44] 鹿野, 他, ニューラルネットワークの音声情報処理への応用, TR-I-0063 (1988-12).

[45] 中村, 他, ベクトル量子化話者適応の時間遅れ神経回路網による音韻認識への適用, TR-I-0098 (1989-8).

[46] 宮武, 他, 時間遅れ神経回路網 (TDNN) による音韻スポッティングのための効果的学習法, TR-I-0103 (1989-08).

[47] 遠藤, 他, ニューラルネットワークによる予測モデルを用いた音韻認識, TR-I-0107 (1989-8).

[48] 中村, 他, ニューラルネット開発用ワークベンチシステム - ネットワークエディタおよびモニタ機能について -, TR-I-0113 (1989-09).

[49] 田村, フィードフォワードニューラルネットの解釈, TR-I-0116 (1989-10).

[50] 川端, 構成的ニューラルネットによる音声認識, TR-I-0122 (1989-11).

[51] J.C.Dang, 田村, 沢井, Shift-invariant Deterministic Boltzmann Machines for Phoneme Recognition, TR-I-0130 (1989-12).

[52] 丸山, 他, NETgram を用いた HMM 英単語音声認識の改善, TR-I-0133 (1990-02).

[53] 南, 他, TDNN 音韻スポッティングと予測 LR パーザを用いた大語彙単語音声認識, TR-I-0144 (1990-02).

[54] 南, 沢井, TDNN の構造の音韻認識率、シフトインバリアンス性への影響, TR-I-0145 (1989).

[55] 中村, 沢井, TDNN による不特定話者音素認識, TR-I-0178 (1990).

[56] S.Katagiri, C.H.Lee, A New Speech Recognition Algorithm based on HMM and LVQ, 音響学会講演論文集, 2-P-7, pp.143-144 (1990-09).

[57] S.Katagiri, C.H.Lee, B.H.Juang, A Generalized Probabilistic Decent Method, 音響学会講演論文集, 2-P-6, pp.141-142 (1990-09).

[58] 岩見田, 片桐, マクダーモット, LVQ-HMM による不特定話者音韻認識, 音響学会講演論文集, 1-8-18, pp.35-36 (1990-09).

[59] 磯, 麻生川, 吉田, 渡辺, ニューラルネットワークによる話者適応, 音学講論, 1-6-16, pp.31-32 (1989-03).

[60] 磯, ニューラルネットワークによる予測モデルを用いた音声認識, 音声研究会資料, SP89-23, pp.81-87 (1989).

[61] 渡辺隆夫, 音声認識のためのニューラルネットと時間軸正規化の組合せ学習, 音学講論, 2-P-27, pp.283-284 (1989-03).

[62] 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, 信学技報, NC89-103, (1990-3).

[63] 鈴木, 河原, 平均曲率を用いた神経回路網学習法の汎化能力について, 神経回路学会平成2年全国大会講演論文集, O1-4, pp.18 (1990-9).

[64] 金谷, 神経回路網によるパターン認識のベイズ統計的振舞いについて, 通信学会技術報告, PRU89-15, pp.9-16 (1989-6).

[65] 相川, 他, 音素セグメント単位のニューラルネットワークを用いた音声認識, 音声研究会資料, S90-13, pp.33-39 (1990-06).

[66] 高木, 坪香, ニューラルネットを用いた音韻セグメンテーション, 音響学会講演論文集, 2-P-7, pp.215-216 (1989-10).

[67] Esther Levin, Word Recognition using Hidden Control Neural Architecture, Proc. ICASSP90, S8.6, pp.433-436 (1990-04).

[68] N.Morgan and H.Bourland, Continuous Speech Recognition using Multilayer Perceptrons with Hidden Markov Models, Proc. ICASSP90, S8.1, pp.413-416 (1990-04).

[69] Les T. Niles and H.F. Silverman, Combining Hidden Markov Model and Neural Network Classifiers, Proc. ICASSP90, S8.2, pp.417-420 (1990-04).

[70] Minami, etal, On Sensitivity and Robustness of HMM and Neural Network Speech Recognition Algorithms, ICSLP90 (1990-11).

[71] S.J.Cox, J.S.Bridle, Simultaneous Speaker Normalization and Utterance Labelling Using Baysian/Neural Net Techniques, ICASSP90.

[72] J.S.Bridle, Alpha-Nets: A reccurent "neural" network architecture with a Hidden Markov Model interpretation, RSRE Reseach Report, Oct-1989.

[73] J.S.Bridle, Training Stochastic Model Recognition Algorithms as Networks can lead to Maximum Mutual Information Estimation of Parameters, Advances in Neural Information Processing Systems 2, 1990.

[74] S.J.Young, Competitive Training: A Connectionist Approach to the Discriminative Training of Hidden Markov Models, CUED/ F-INFENG/ TR.41 (1990-03).

[75] Mike Chong, etal, Classification and Regression Tree Neural Networks for Automatic Speech Recognition,

[76] R.W.Prager, etal, The Modified Kanerva Model: Results for Real Time Word Recognition, Proc. IEE First International Conference on Artificial Neural Networks (1989-10).

[77] P.C.Woodland, Isolated Word Speech Recognition based on Connectionist Techniques, Br Telecom Technol J. Vol.8, No.2, pp.61-66 (1990-04).

[78] Michael W.H. Chong, etal, Implementation of Neural Networks for Speech Recognition on a Transputer Array, CUED /F-INFENG /TR.8 (1988-03).

[79] A.J.Robinson, etal, Phoneme Recognition from the TIMIT Datebase using Recurrent Error Propagation Networks, CUED /F-INFENG /TR.42 (1990-03).

[80] S.V.B.Aiyer, etal, A Theoretical Invetstigation into the Performance of the Hopfield Model, CUED /F-INFENG /TR.36.

[81] A.J.Robinson, etal, Generalising the Nodes of the Error Propagation Networks, CUED /F-INFENG /TR.25 (1988).

[82] F.Fallside, Connetionist Models in Speech Recognition, A Brief Summary and Bibliography, CUED /F-INFENG /TR.18 (1988-04).

[83] K. Lari, etal, The Estimation of Stochastic Context-Free Grammars using the Inside-Outside Algorithm, Computer Speech and Language, No.4, pp.35-56 (1990).

[84] R.W.Prager, etal, The Modified Kanerva Model for Automatic Speech Recognition, Computer Speech and Language, No.3, pp.61-81 (1989).

[85] R.W.Prager, etal, Boltzmann Machines for Speech Recognition, Computer Speech and Language, No.1, pp.3-27 (1986).

[86] M.Niranjan, F.Fallside, Neural Networks and Radial Basis Functions in Classifying Static Speech Patterns, CUED/F-INFENG/TR 22 (1988).

[87] Dominique G. Beroule, Guided Propagation: Current State of Theory and Applications, Neuro Computing: Algorithms Architecture and Application, (1990).

[88] Dominique G. Beroule, Guided Propagation inside A Topographic Memory, IEEE First International Conference on Neural Networks, IV.469-476 (1987-06).

[89] Dominique G. Beroule, The Never-Ending Learning, NATO ASI Senes, Vol.F41, Neural Computers (1988).

[90] Dominique G. Beroule, The Adaptive, Dynamic and Associative Memory Model: A Possible Future Tool for Vocal Human-Computer Communication, The Structure of Multimodal Dialogue (1989).

[91] A.Waibel, Connectionist Speech and Language Research, Semiannual Progress Report (1989-02).

[92] J.Hampshire and A.Waibel, The Meta-Pi Network: Connectionist Rapid Adaptation for High-Performance Multi-Speaker Phoneme Recognition, Proc. of ICASSP, (1990-04).

[93] A.N.Jain, A.H.Waibel, Robust Connectionist Parsing of Spoken Language, Proc. of ICASSP (1990-04).

[94] M.Franzini, K.F.Lee, A.Waibel, Connectionist Viterbi Training: A New Hybrid Method for continous Speech Recognition, Proc. of ICASSP (1990-04).

[95] J.Tebelskis, A.Waibel, Large Vocabulary recognition Using Linked Predictive Neural Networks, Proc. of ICASSP (1990-04).

[96] U.Bodenhausen, The Tempo-Algorithm: Learning in a Neural Network with Variable Time-Delays, Proc. IJCNN Vol.1, pp.597-600 (1990-01).

[97] 小島, 相補性と不動点, 産業図書 1981.

## A 他研究機関における NN 研究論文一覧

1. NEC
   [59], [60], [61]
2. NTT
   [62], [63], [64], [65]
3. CMU
   [91], [92], [93], [94], [95]
4. Cambridge University
   [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85]
5. RSRE
   [71], [72], [73]
6. Bell
   [67]
7. LIMSI
   [87], [88], [89], [90]

## B NN 関連の ATR Technical Report

音声処理 NN に関する ATR の Technical Report の一覧を載せる。

- haffner-tr-0058:[42]
- haffner-tr-0059:[43]
- shikano-tr-0063:[44]
- nakamura-tr-0098:[45]
- miyatake-tr-0103:[46]
- endo-tr-0107:[47]
- nakamura-tr-0113:[48]
- tamura-tr-0116:[49]
- kawabata-tr-0122:[50]
- dang-tr-0130:[51]
- maruyama-tr-0133:[52]
- minami-tr-0144:[53]
- minami-tr-0145:[54]
- nakamura-tr-0178:[55]

## C 図表

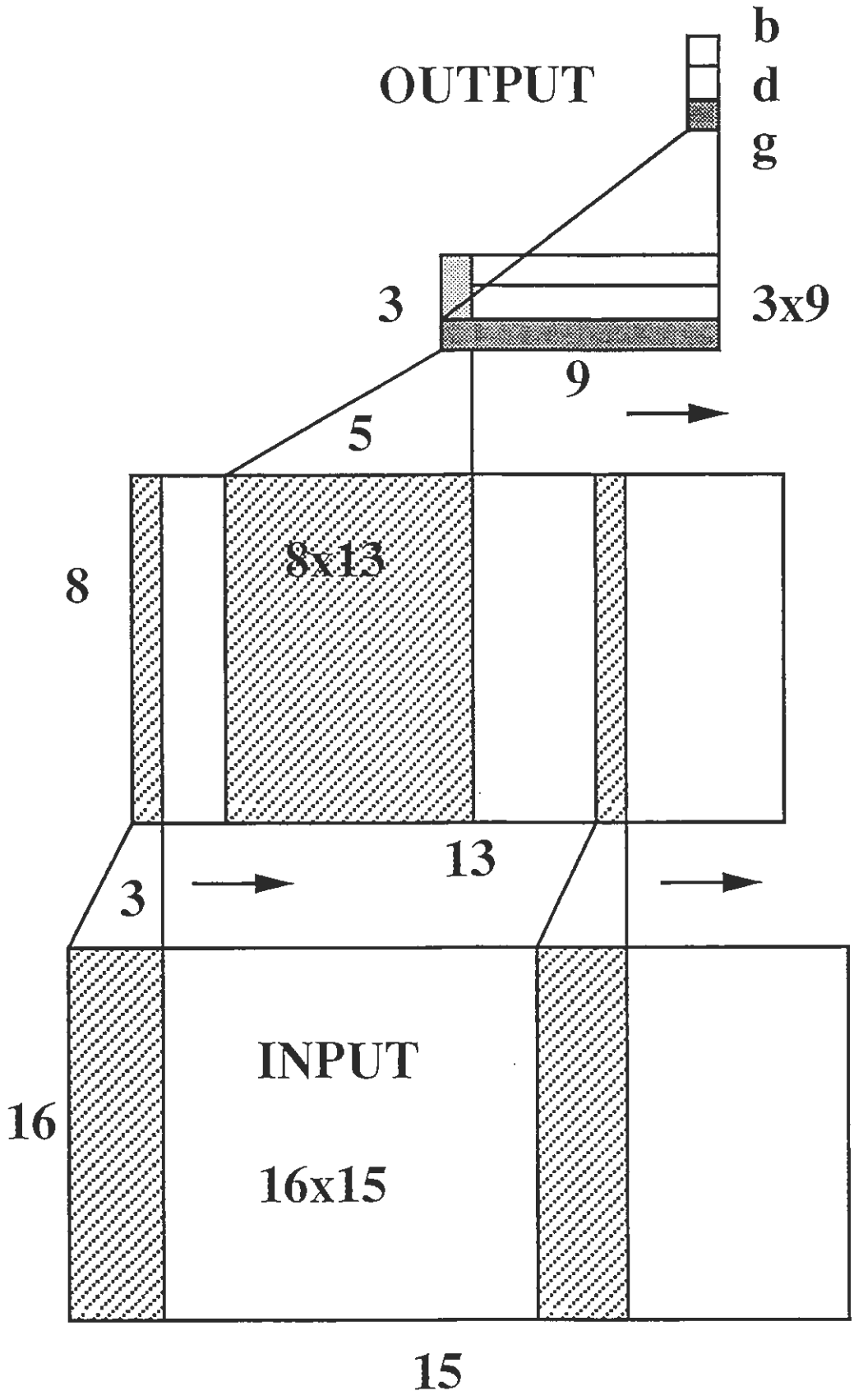


図 1: TDNN の構造

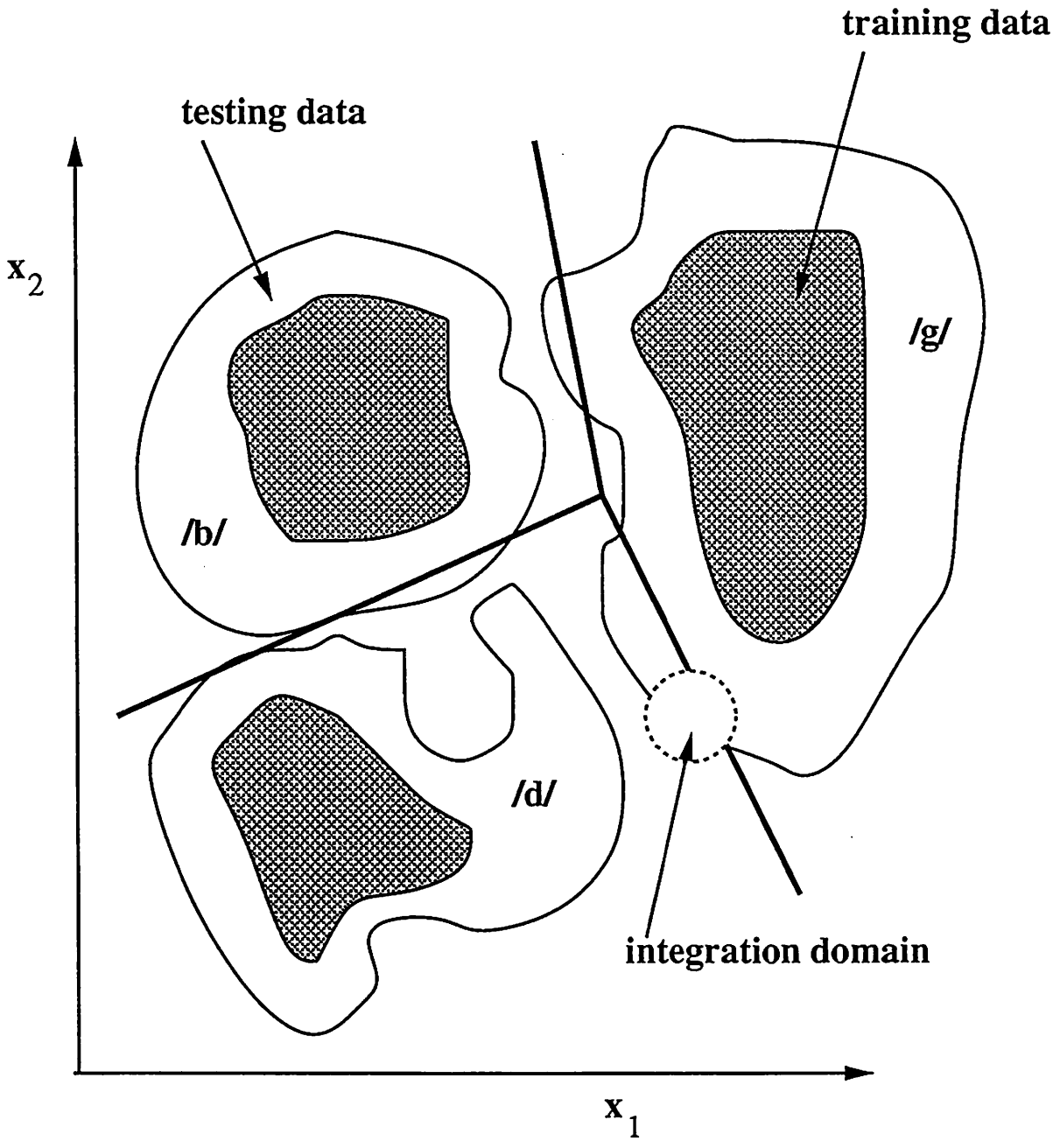


図 2: TDNN 出力値の積分による平滑化法

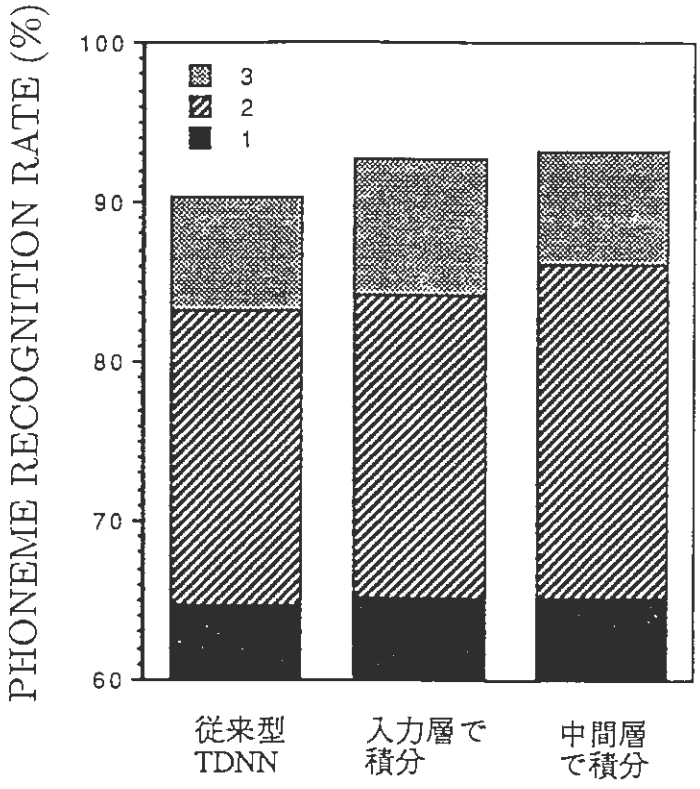


図 3: TDNN 出力値の積分による平滑化法による音素識別率

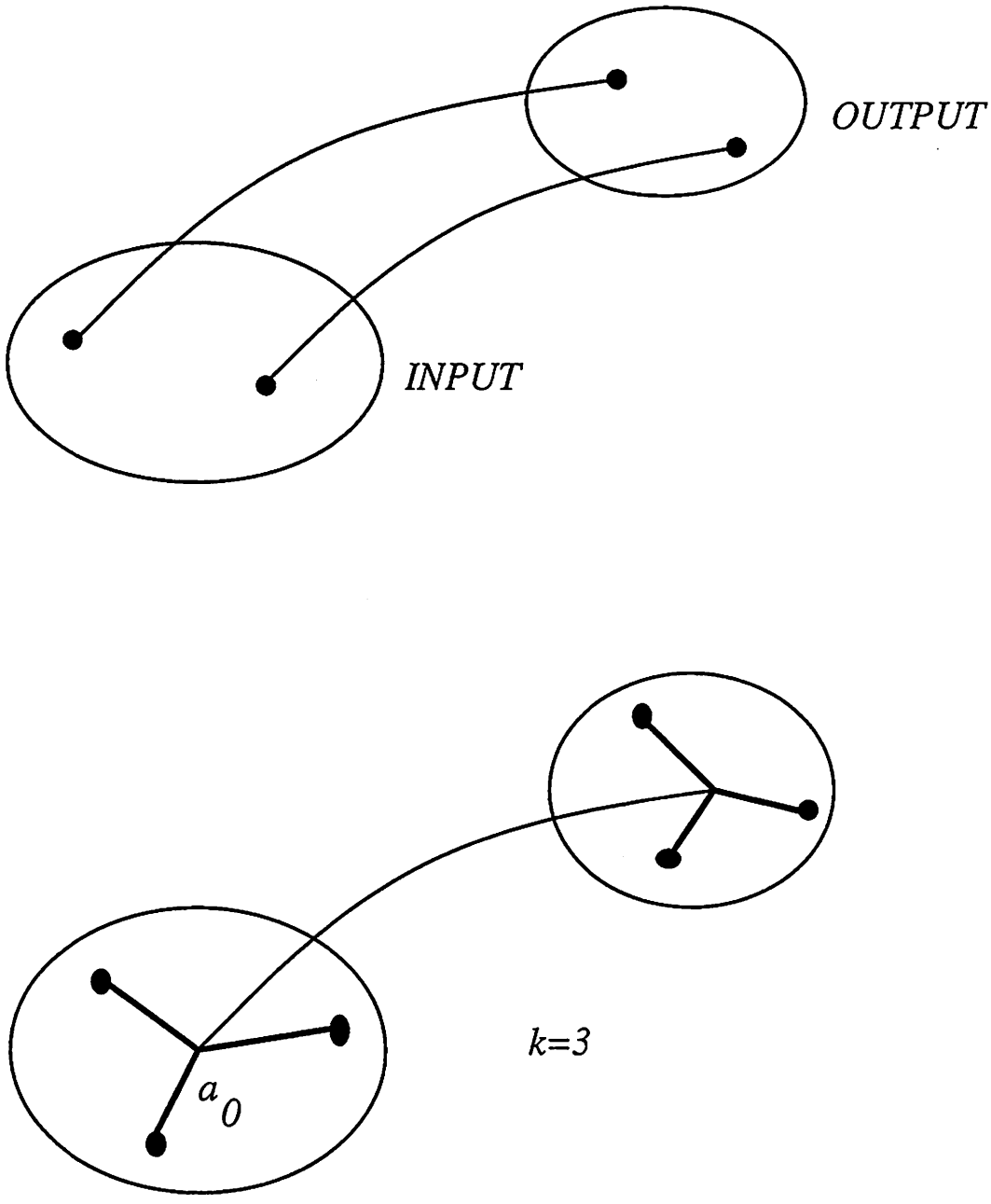


図 4: KNIT 法の概念

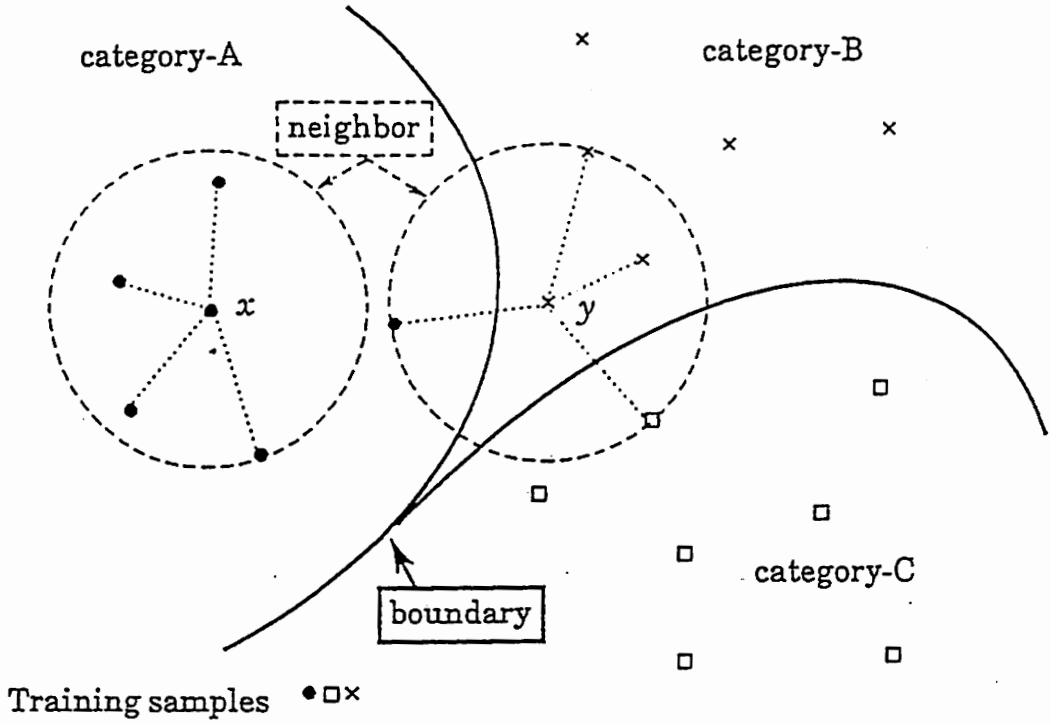
category-A
category-B
neighbor
x
y
boundary
category-C
Training samples

図 5: FuNI 法の概念

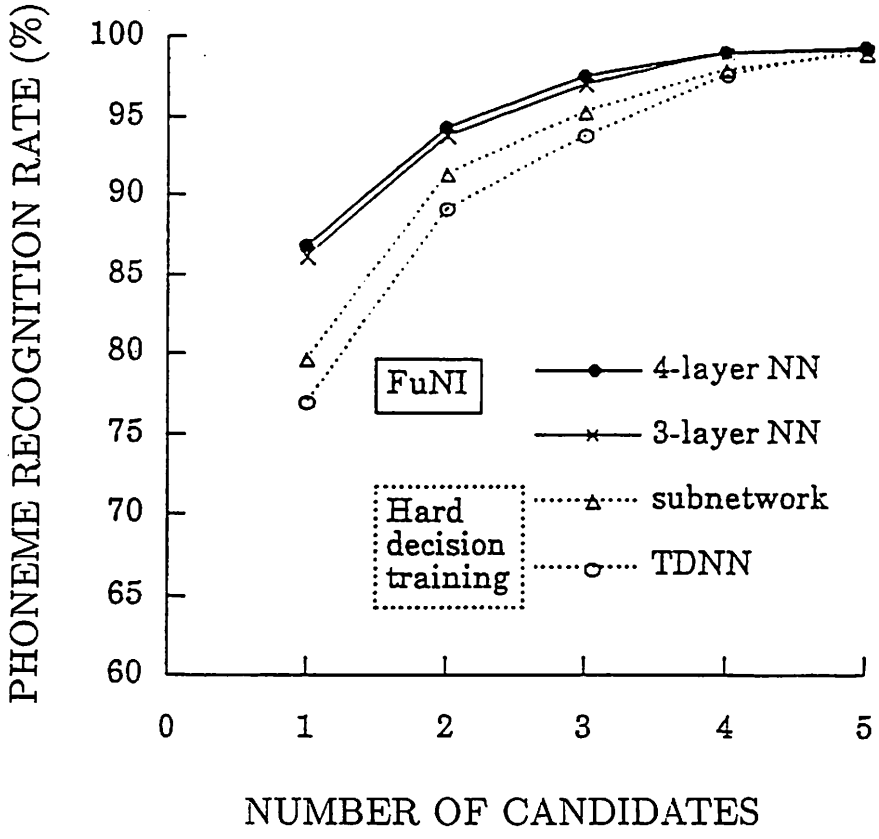


図 6: FuNI 法を用いた音素認識

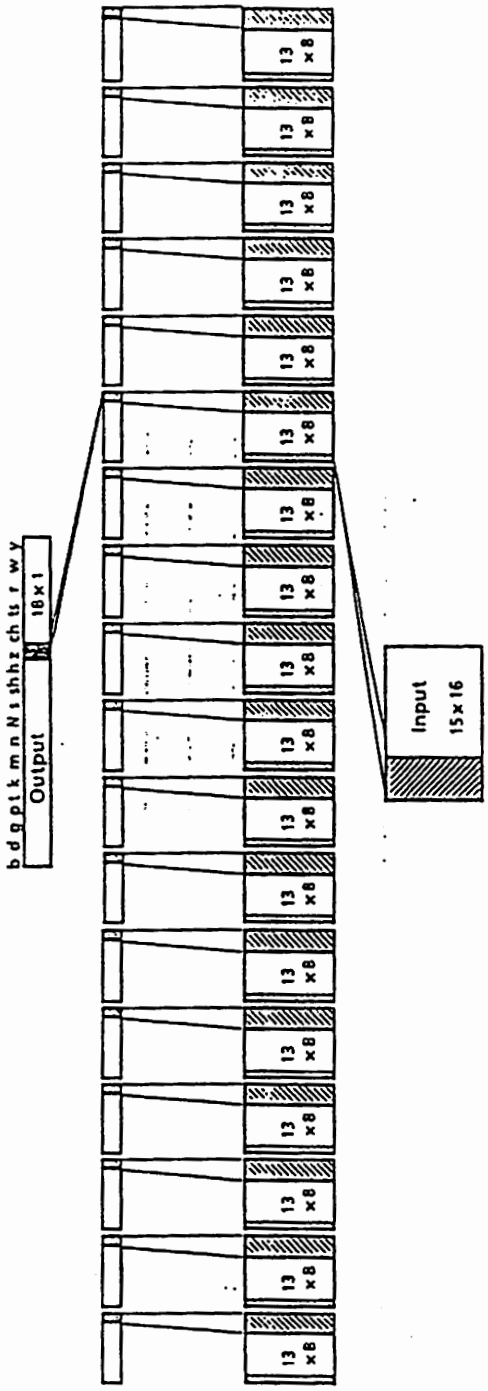


図 7: モジュール構成の比較 (分散型)

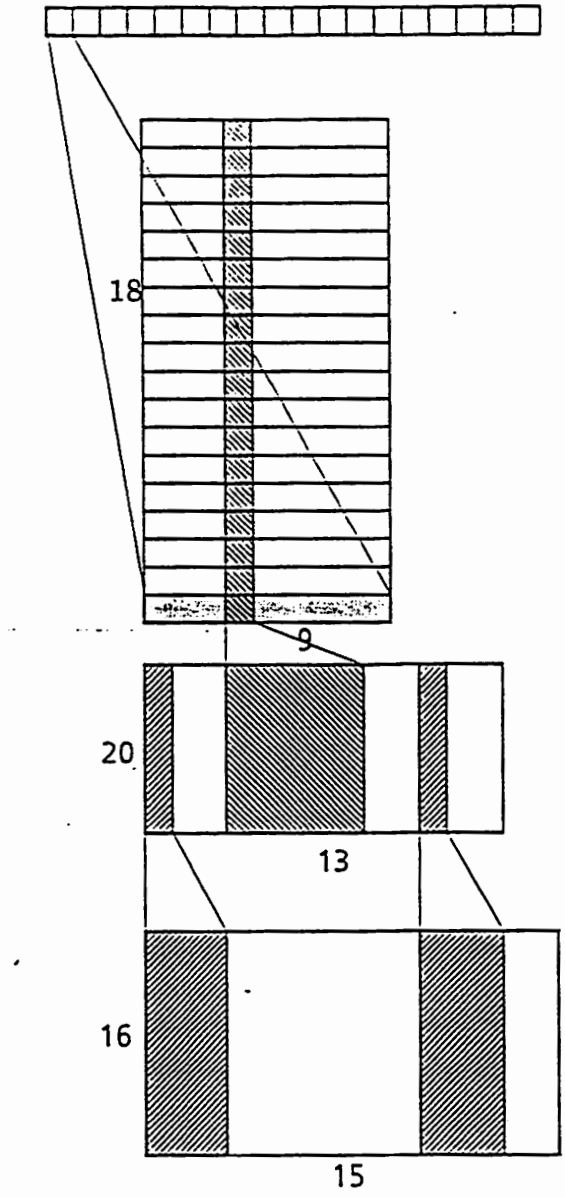


図 8: モジュール構成の比較（集中型）

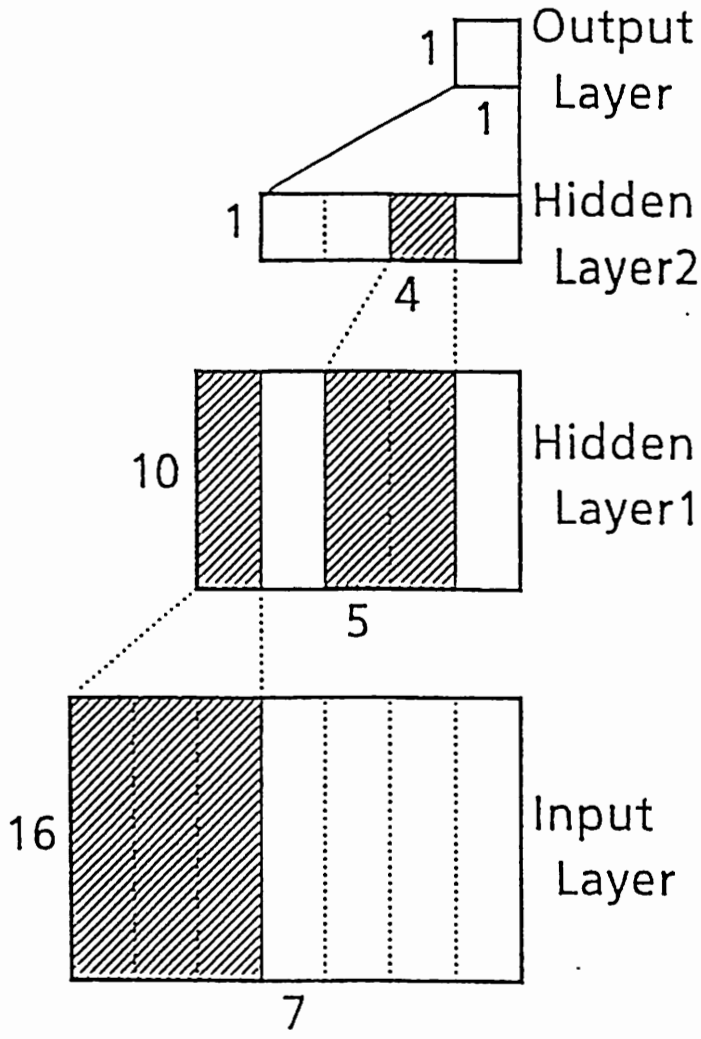


図 9: 対判定型 TDNN の構造

C 図表

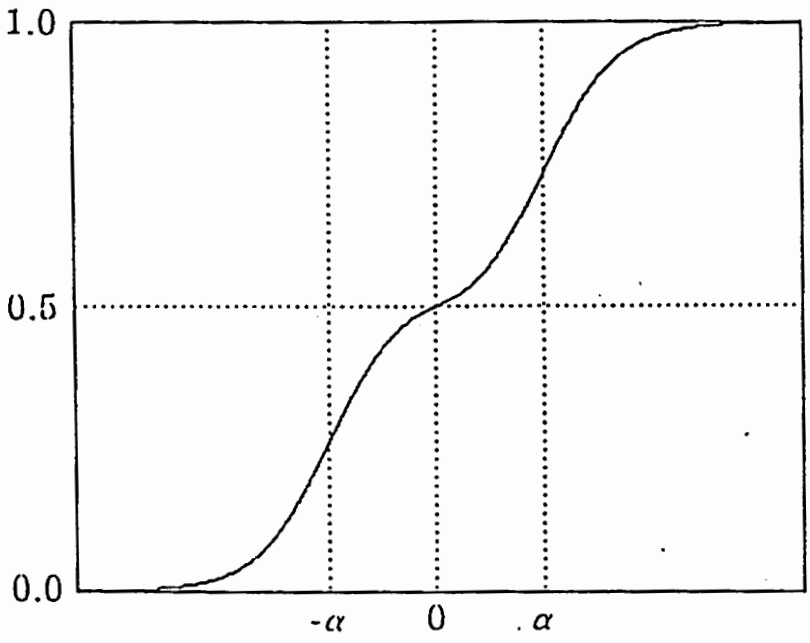


図 10: 出力ユニット内の非線形関数の概形

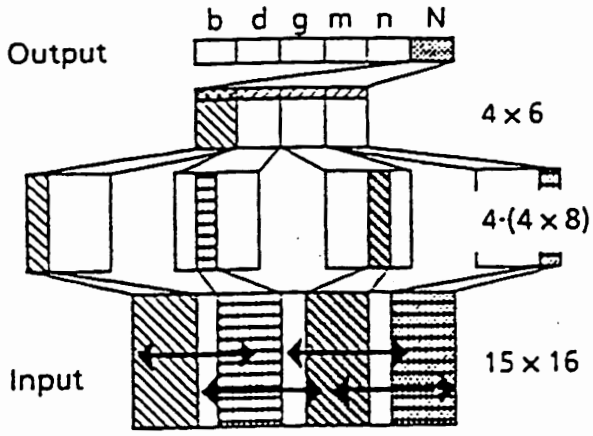


図 11: 時間構造考慮 NN の構造 (TS)

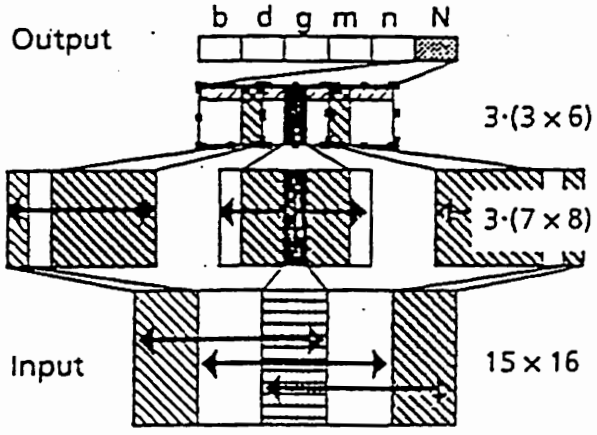


図 12: 時間構造考慮 NN の構造 (STS)

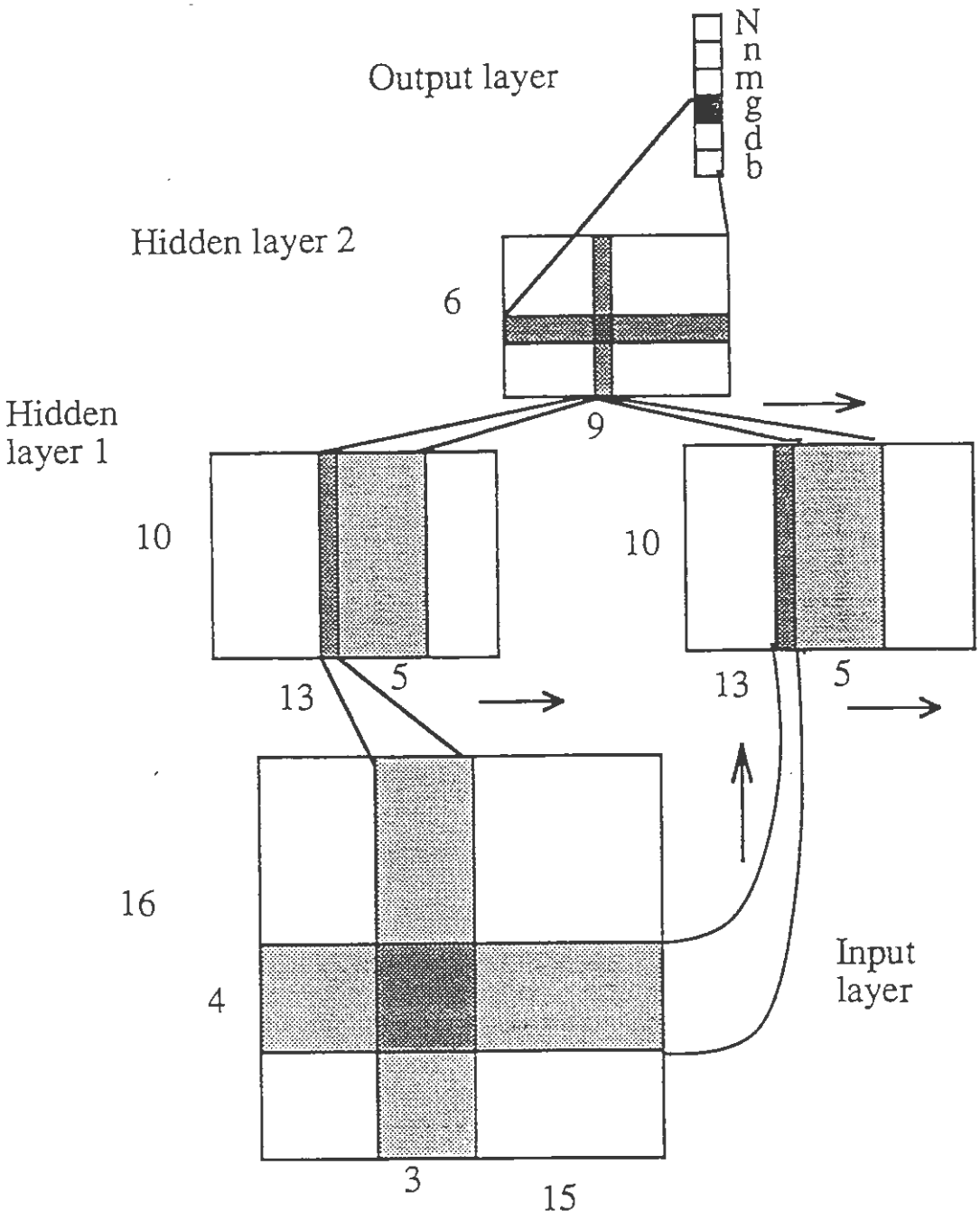


図 13: 時空間 block 統合 NN の構造 1 (TF)

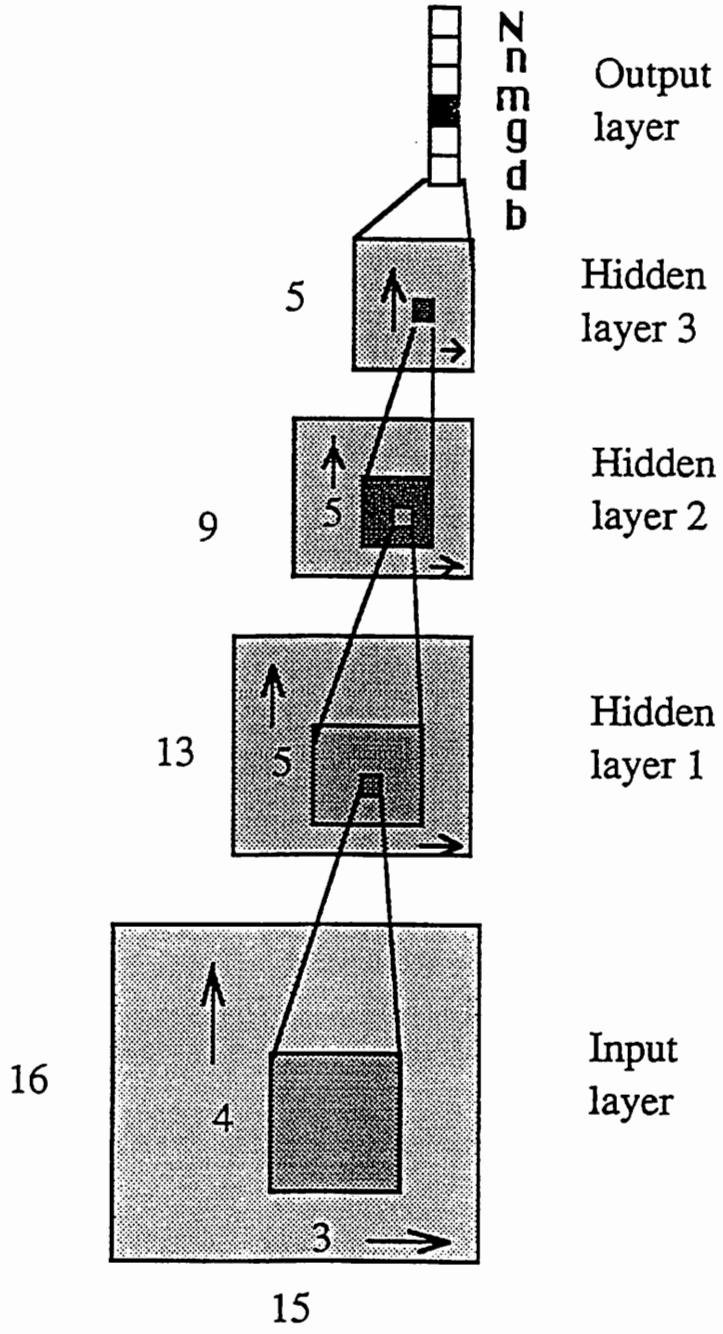


図 14: 時空間 block 統合 NN の構造 2 (BW)

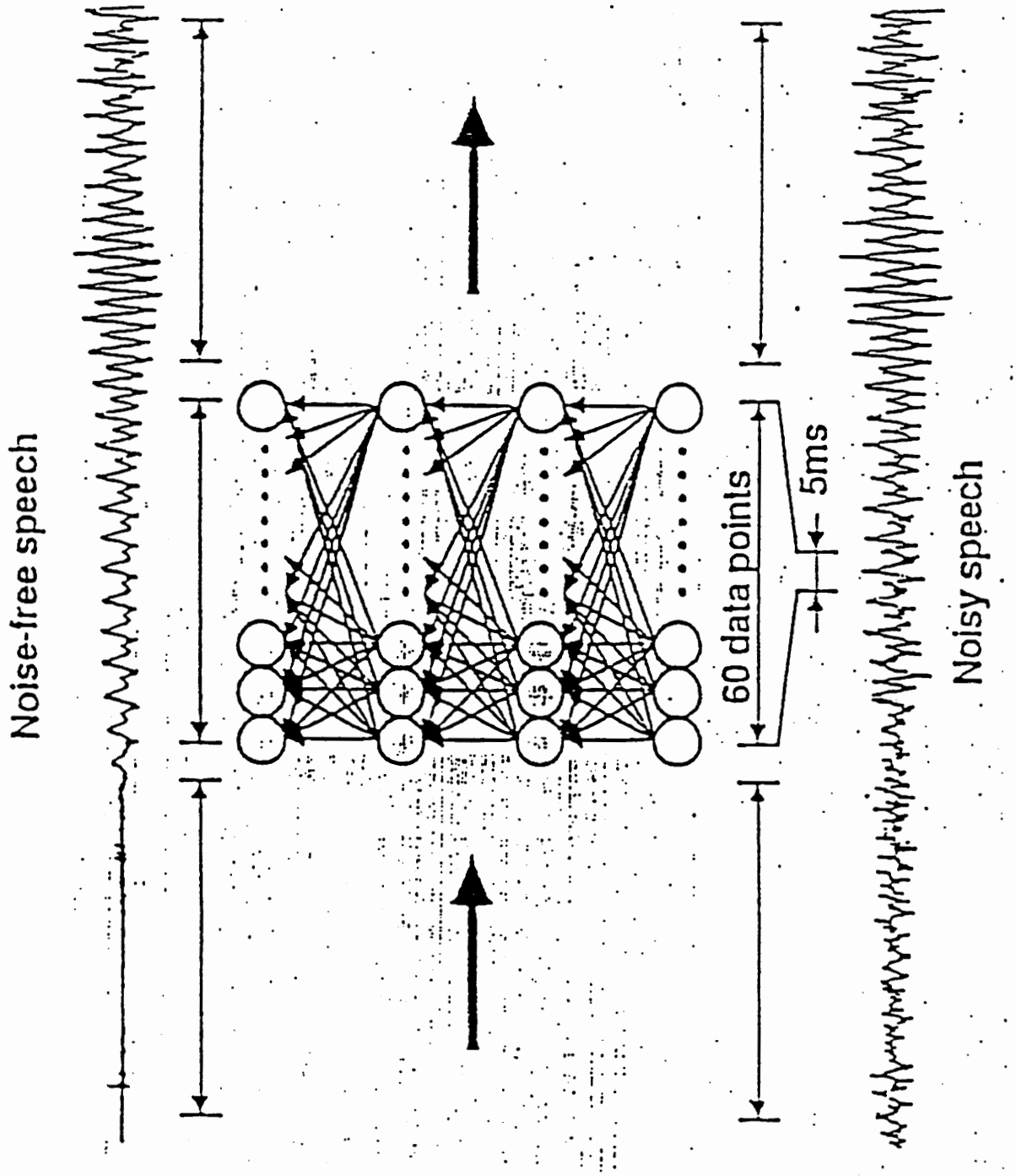


図 15: 雑音抑圧 NN の構造

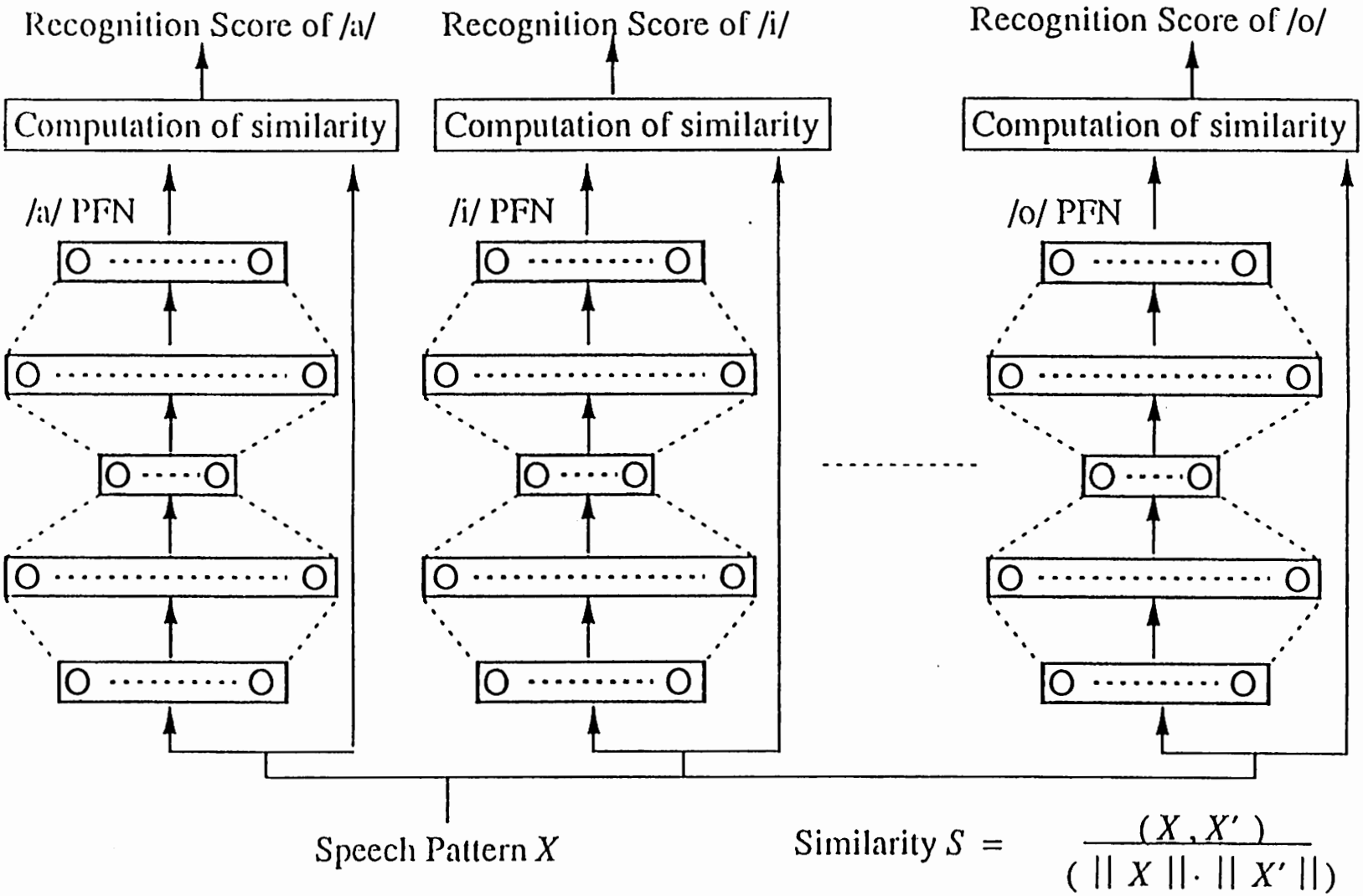


図 16: 音素ニューラルフィルター(PFN)の構造

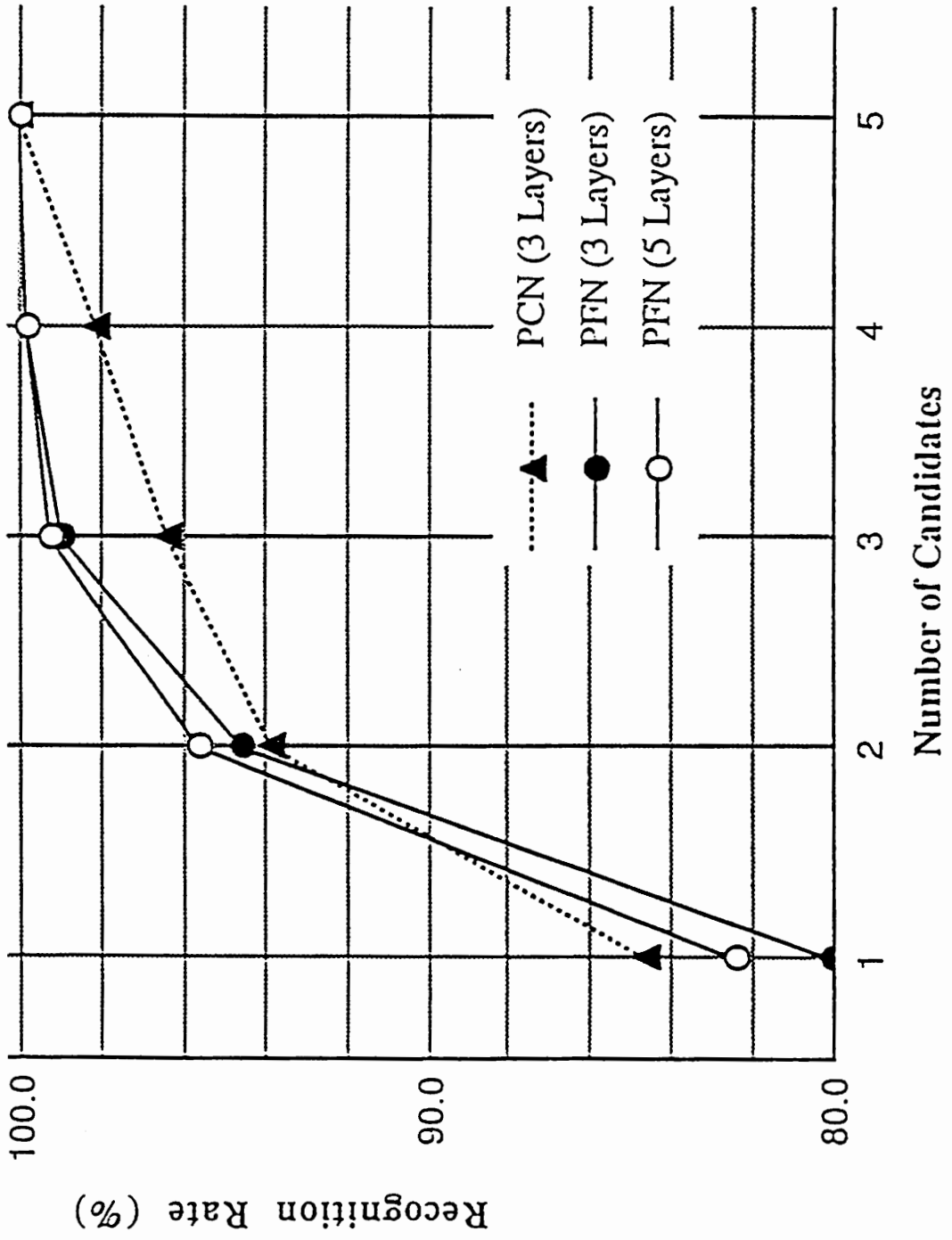


図 17: 音素ニューラルフィルター (PFN) による母音認識率

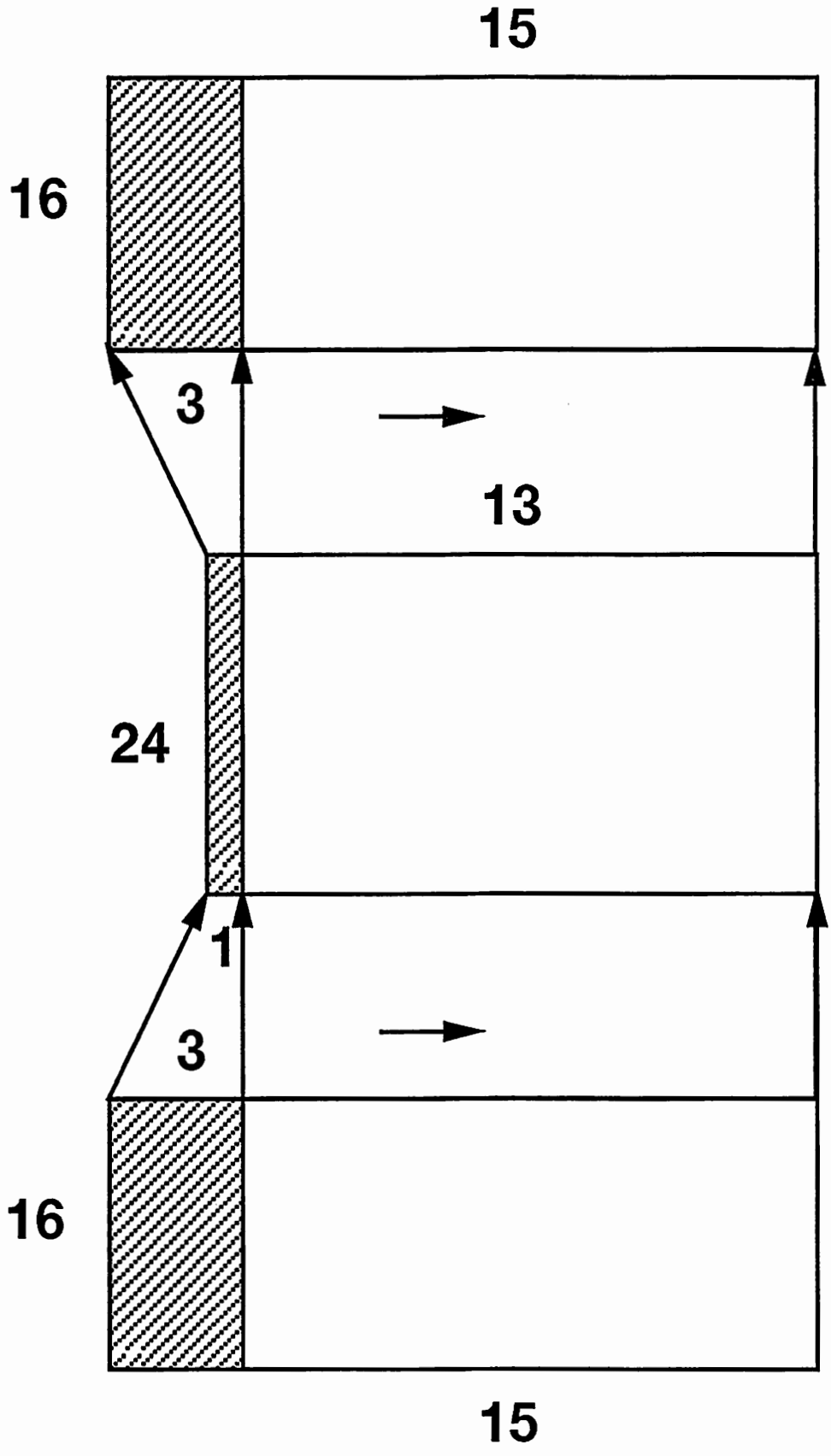


図 18: 話者適応 NN の構造

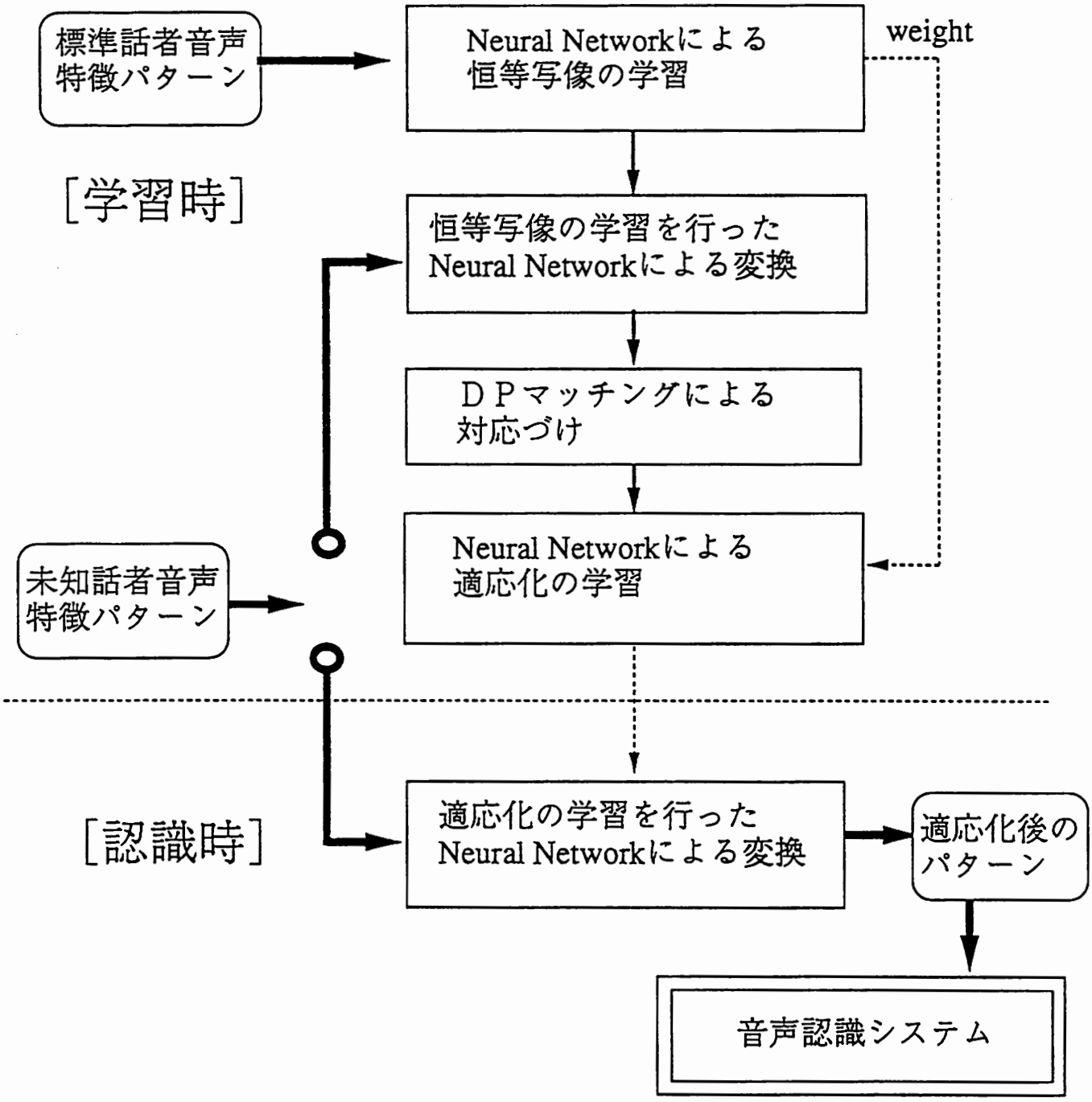


図 19: NN による話者適応化の処理のながれ

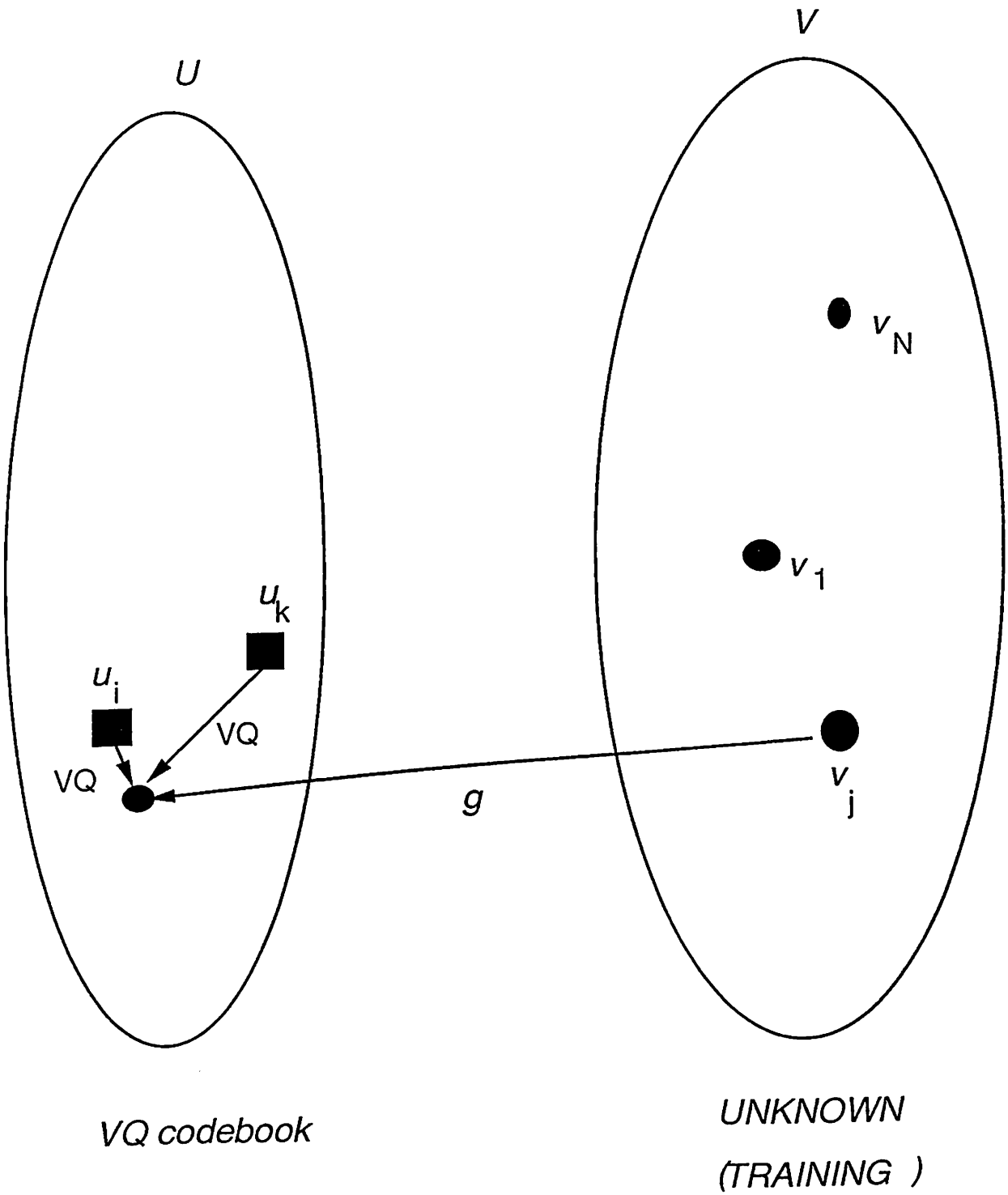


図 20: VQ + BP による学習

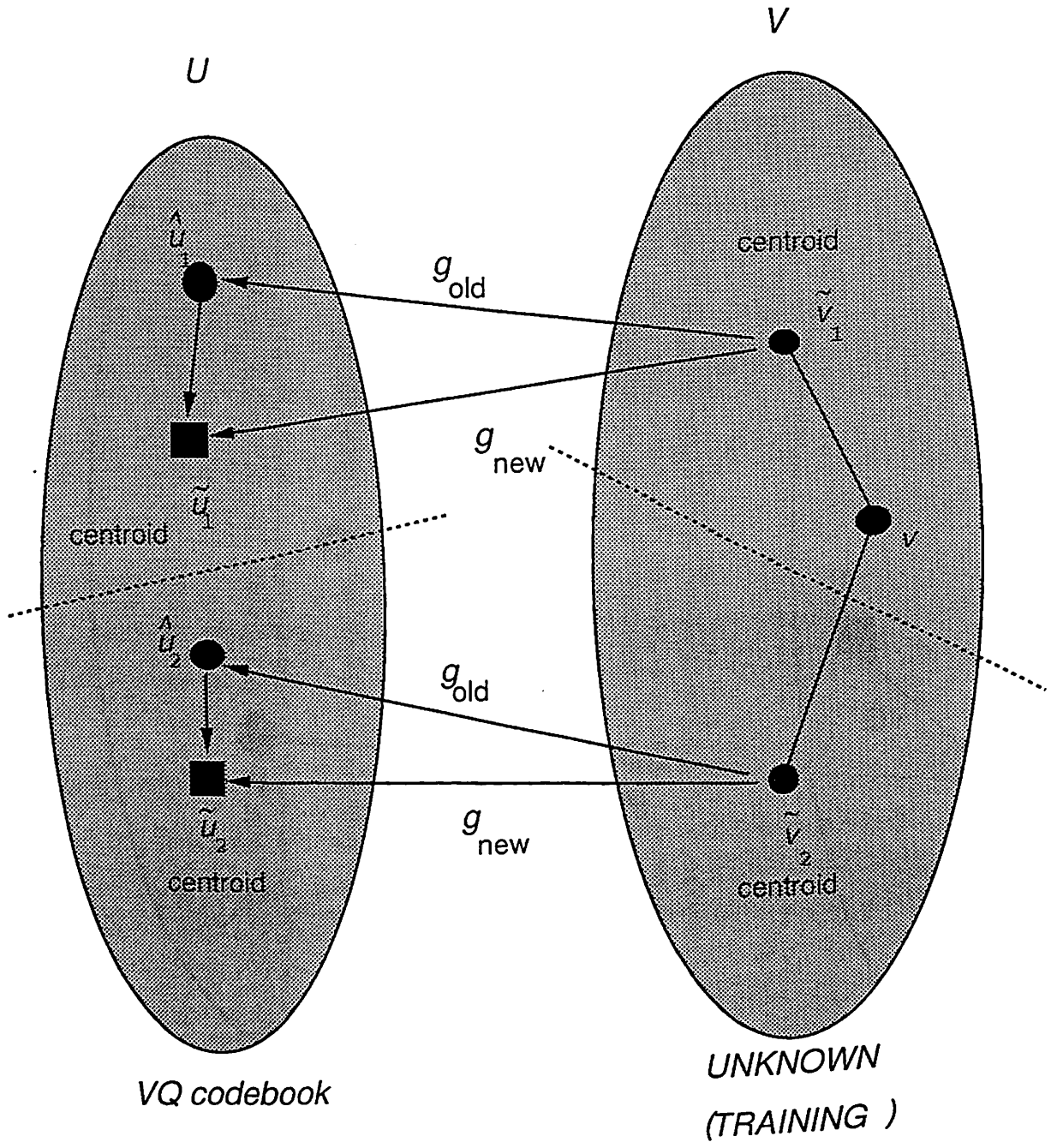
V
U
centroid
$\hat{u}_1$
$g_{old}$
$\tilde{v}_1$
$g_{new}$
$\tilde{u}_1$
centroid
v
$\hat{u}_2$
$g_{old}$
$g_{new}$
$\tilde{u}_2$
centroid
$\tilde{v}_2$
centroid
VQ codebook
UNKNOWN
(TRAINING )

図 21: 階層的な学習