The University of Osaka
Institutional Knowledge Archive

Title	時系列パターン認識機構の神経回路モデルに関する研究
Author(s)	山内, 康一郎
Citation	大阪大学, 1994, 博士論文
Version Type	VoR
URL	https://doi.org/10.11501/3075131
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

時系列パターン認識機構の
神経回路モデルに関する研究

山内康一郎

1994年1月

概要

本論文は，人間の音声処理能力を模倣した，時系列パターン (音声) 処理機構のモデルについて述べたものである．

一般に，時系列パターン認識を行なうには，時間伸縮に柔軟に対応するメカニズムが必要である．これまでにも，この問題を克服するための時系列パターン認識システムが多く提案されてきた．その多くは，入力された時系列パターンから，どの特徴がどのような順序で現れたかという情報のみを検出することによってそのパターンを認識する．すなわち，時系列パターンの各特徴の継続時間長を無視することによってその時間伸縮に対処していた．

その一方で，モールス信号のように，継続時間の中に情報が含まれるような信号も存在する．このような信号を，これら従来のシステムによって識別することは，難しい．しかし人間は，このモールス信号のように，継続時間に情報が含まれるようなパターンであっても，その時間伸縮に影響を受けずに正しく識別できることが知られている．このことから，人間の聴覚系は，入力された時系列パターンの時間構造を壊すことなくその時間伸縮に対応する能力を持っていると考えられる．本論文では，この人間の能力を模倣するため，時系列パターンをその時間構造を保持しながら処理する形式を採り，且つ時間伸縮に対処できるモデルを提案している．

本論文は5章から成る．第1章の序論では，まず，生理学的な側面から見た生体の聴覚系の働きと，心理学的な側面から見た人間の音声処理能力について概観したあと，既に提案されている時系列パターン (音声) 認識システムについて触れる．そして，心理学的な立場から見た人間の音声処理方法は，音声信号の時間構造を保ったままその時間伸縮に対処している可能性が高いのに対して，ほとんどの従来型の時系列パターン認識システムでは，その時間構造を破壊して時間伸縮に対応していることを示す．

第2章では，序論で指摘した従来型システムの問題点を克服する時系列パターン認識モデルを提案する．このモデルは，時系列パターンを，その時間構造を保持しながら処理・認識する形式を採り，且つその時間伸縮に対処する能力を持っている．ここでは，このモデルを神経回路モデルとして構築している．この章では，提案したモデルの説明と，その計算機シミュレーションを示している．また，このモデルと，人間の音声処理方法に関す

る心理学的な知見との比較，および解剖学的に見た聴覚系の構造に関する知見との比較についても述べる．

第3章では，第2章で提案したモデルを構成する神経回路の学習理論を提案している．ここではまず，これまでに提案された神経回路の学習法をいくつか示す．そしてこれらの学習法を用いて時系列パターンを学習させようとすると，時系列パターンのどの部分を学習すべきかを明示的に示さなければならなかったり，あるいは，そうでない場合にも莫大な数の細胞数を必要とするという問題が生ずることを示す．そして，その問題を解決するための新しい学習理論を提案する．この学習法は教師なし学習法で，時系列パターンを提示されるだけで，その中の重要な特徴を自動的に発見し，保持する能力を持つ．

第4章では，第2章で提案した時系列パターン認識モデルを使った音声認識について述べる．ここで提案している音声認識システムはまず，音声スペクトルからいくつかの特徴を抽出する．抽出された特徴データは，先に提案した時系列パターン認識モデルに送られ，認識される．計算機シミュレーションでは，単語音声パターンを使って，本システムの音声パターンの伸縮や，スペクトル形状の変形に対する汎化能力を確かめている．具体的には，このシステムに，男性が普通のスピードで発音した，いくつかの単語音声を学習させた後，女性が様々なスピードで発音した単語音声を提示した．その結果，本システムはこの女性の声を，そのスピードに影響されずに正しく認識した．ここでは小規模な音声認識実験しか行なっていないが，その結果は，このシステムが，音声信号を喋るスピードや話者の違いに影響されることなく正しく認識できるという，高い汎化能力を持つことを示唆している．

目 次

第1章 序論 **1**

1.1 本研究の目的 1

1.2 解剖学的に見た聴覚系の構造 2

1.2.1 聴覚器官 2

1.2.2 蝸牛神経核 5

1.2.3 上オリーブ核 6

1.2.4 下丘 7

1.2.5 内側膝状体 8

1.2.6 大脳聴覚皮質 9

1.3 心理学的に見たヒトの聴覚系 11

1.3.1 音韻知覚 11

1.3.2 単語知覚 14

1.3.3 単語知覚モデル 18

1.4 時系列パターン (音声) 認識研究の背景 21

1.4.1 DP マッチングを用いた手法 22

1.4.2 状態遷移を基礎に置くモデル 23

1.4.3 遅延素子を使用する手法 25

1.5 本論文の主張 29

第2章 時系列パターン認識モデル **31**

2.1 序言 31

2.2 提案するモデルの概要 32

2.3 各認識ブロックの構造 34

2.4 速度制御 39

2.5 システムの出力 41

2.6 計算機シミュレーション 41

2.7 心理学および解剖学的知見との比較 44
2.8 結言 45

第 3 章 時系列パターンの学習 **47**
3.1 序言 47
3.2 従来型学習法 47
3.2.1 教師有り学習法 48
3.2.2 教師なし学習法 50
3.2.3 ネオコグニトロン型学習法 57
3.3 従来型学習法の時系列パターン学習時の問題点 59
3.3.1 ここで前提とするネットワークの構造 59
3.3.2 教師あり学習法の時系列パターン学習時の問題点 61
3.3.3 教師なし学習法の時系列パターン学習時の問題点 62
3.4 改良型学習法 66
3.4.1 改良型学習法の概要 66
3.4.2 $p^t(\nu, k, m)$ の更新 67
3.4.3 $r^t(k)$ の更新 70
3.4.4 $c^t(\nu, k, m)$ の更新 71
3.4.5 計算機シミュレーション 73
3.5 結言 76
3.6 付録 77
3.6.1 比較実験で使用した従来型学習法と改良型学習法のパラメータ . . . 77
3.6.2 式 (31) の解析 77

第 4 章 音声認識へのアプローチ **79**
4.1 序言 79
4.2 特徴抽出部 80
4.2.1 特徴抽出部の概要 80

4.2.2 特徴抽出部を構成する細胞 81
4.2.3 入力部 (Input Block) 82
4.2.4 持続型細胞層 (Sustained Response Layer) 83
4.2.5 CF 抽出部 (CF-Extracting Block) 84
4.2.6 FM 抽出部 (FM-Extracting Block) 85
4.3 認識部 86
4.3.1 U_D層〜U_{C2}層の構成 87
4.3.2 U_{S3},U_{C3}層の構成 91
4.4 学習法 91
4.4.1 興奮性結合の更新 91
4.4.2 S 細胞のパターン選択性の変化 92
4.4.3 S 細胞と V 細胞の結合領域の更新 93
4.5 音声認識実験 94
4.6 結言 96

第 5 章 総括 **99**

謝辞
参考文献
関連発表論文

第1章　序論

1.1　本研究の目的

パターン認識装置を構築するに当たって，柔軟なパターン認識能力を持つ脳の仕組みは，大いに参考となると考えられる．しかし，脳のメカニズムは，そのごく一部を除いてほとんど明らかになっていない．例えば聴覚系については，聴覚系のごく末梢の段階，すなわち鼓膜から，過牛を経て聴覚野に至るまでのメカニズム以外は，ほとんど明らかにされていないのが現状である．

このような中，人間の脳機構を予測する有力な方法の一つに，機能モデルを立て，そのモデルの振舞いと実際の脳の振舞いとを比較して予測したモデルが正当かどうかを検証するという方法がある．本論文では特に，人間の音声認識能力を模した時系列パターン認識機構の神経回路モデルを構築する．

一般に知られているように，人間は音声を聴きとる際，話者や喋るスピードに関わらず，その音声を正しく認識できる．これまでにも，この柔軟な時系列パターン認識能力を模倣する音声認識システムが多く提案されている．これらのシステムでは，時系列パターンの時間軸の非線形伸縮に対処する能力を実現している．

しかし，その多くは，時系列パターンの各特徴の継続時間長を無視することによってその時間伸縮に対応していた．したがって，このようなシステムでは，モールス信号のように，継続時間の中に情報が含まれるような信号を識別することが難しい．これに対して，人間はこのようなパターンであっても，正しく識別できる．そこで本論文では，時系列パターン中の各特徴の継続時間を無視せず且つ時間伸縮に対処できる認識機構の神経回路モデルを構築することを目的とする．

また，このモデルを構築するにあたって，神経回路による時系列パターンの学習法についても新たに考慮する必要があった．従来の神経回路の学習法の多くは，時系列パターンの中のどの部分を記憶するべきか，あるいは，どの部分がどのカテゴリーに属するかを人手によって明示的に示されなければ，うまく学習できないことが多かった．しかし，人間は幼い頃，音素に相当する部分や，語に相当する部分などを一つ一つ教えられること無く，自然に覚えてゆく．そこで，本論文では，時系列パターンを提示されるだけで，その中の重

要な特徴を自動的に獲得する教師なし学習法も併せて構築する.

1.2 解剖学的に見た聴覚系の構造

最初に解剖学的な側面から見た生体の聴覚系の仕組みについて概観する. 図1に蝸牛から大脳聴覚皮質に至るまでの経路を示す. 耳から入った音声はまず, 蝸牛に入る. 蝸牛は音声を周波数分割する役目をする. 蝸牛で周波数分割された音声信号は, 大脳聴覚皮質に至るまでに, いくつかの神経核を経由する. 聴覚神経系には求心性経路 (蝸牛から大脳聴覚皮質に向かう経路) と遠心性経路 (大脳聴覚皮質から蝸牛に向かう経路) の二つがある. 求心性経路は, 与えられた音声を処理する役目をし, 遠心性経路は, 音声信号の特定の部分の感度を高めたりするなどの能動的な処理を行なう役割をするとも言われている. 本章では, この二つの経路のうち, 求心性経路についてのみ概観する.

1.2.1 聴覚器官

空気中を伝わる音は, 聴覚器官によって検出される. 図2(a) に, その聴覚器官全体の構造を示す.

音は, 外耳道を通り, その奥にある鼓膜を振動させる. 鼓膜の振動はつち骨, きぬた骨, あぶみ骨をとおしてインピーダンス変換されて蝸牛に伝えられる. 蝸牛は, 伝えられた鼓膜の振動を, 周波数分割して大脳へ送り出す役目をする. すなわち, 蝸牛は音声信号を音声スペクトルに変換するフィルターバンクの一種と言えよう.

蝸牛は, かたつむりのような形状をした骨で覆われている. この蝸牛をまっすぐに引き延ばしたと仮定して, その断面構造を描くと図2(b) のようになる. 蝸牛の中は, リンパ液でみたされており, その中に基底膜と呼ばれる膜が通っている. 基底膜は, 有毛細胞と呼ばれる細胞が3000個ほど並んで構成されている. つち骨, きぬた骨, あぶみ骨をとおして伝えられた鼓膜の振動は, リンパ液を振動させ, この振動が, 基底膜に進行波を発生させる. この進行波の振幅が最大になる点は, 周波数成分によって異なっており, 周波数成分が低いほど管の先端 (図2(b) の右側) に近くなる. 基底膜を構成する有毛細胞は, その細胞が位置する部分での進行波の振幅に応じて, 電気信号を発生する. これらの有毛細胞の出力は, 3000本程の神経繊維によって次の蝸牛神経核へと送られている.

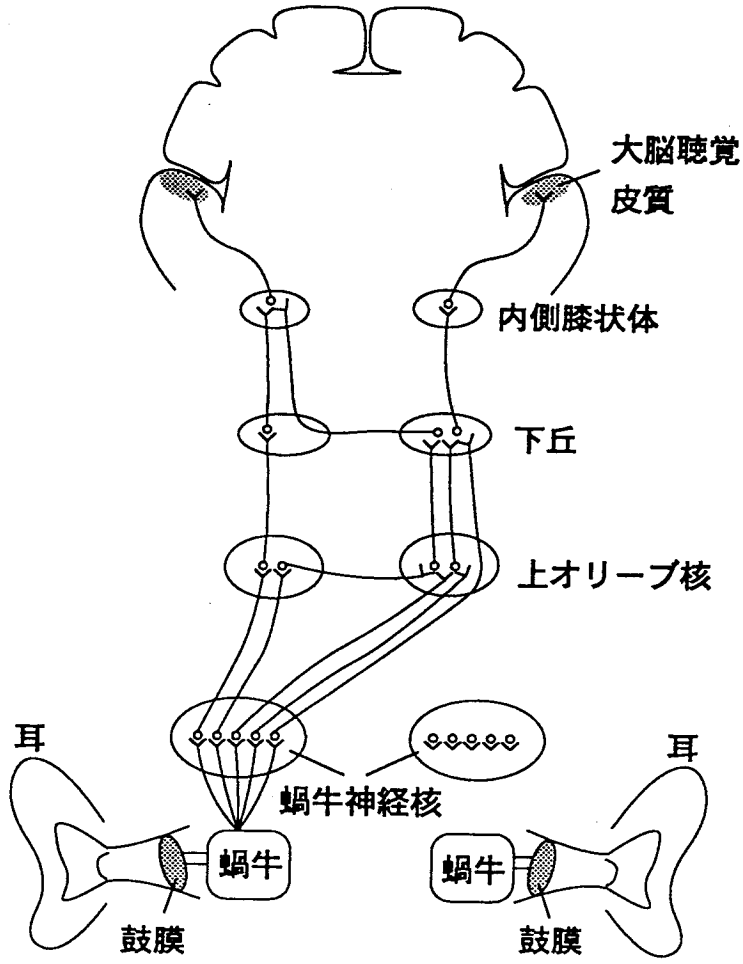


図 1: 耳から大脳聴覚皮質に至るまでの経路 (Morgan [1] を修正)

Békésy (1949) [3] は，この基底膜 (死亡した個体の基底膜) 上の様々な位置における，共振特性を詳しく調べている (図 3参照)．それによると，基底膜上の各位置における共振特性は，さほど鋭いものではなく，広い周波数領域にわたるブロードなものとなっている．

しかし後の研究により，この Békésy の示したデータは，若干修正されている．まず，Johnstone [4] は，生きた個体の基底膜の共振特性は，Békésy が報告した死亡した個体のものよりも鋭いことを報告した．また，Khanna ら [5] は，この基底膜の共振特性の鋭さは，その周波数成分の強さ (音圧) によって適応的に変化することを発見した．具体的には，音圧が小さい時ほど，共振特性が鋭くなり，逆に音圧が大きいほどこの特性はブロードなものとなる．(図 4参照)．

この Johnstone や，Khanna らの知見を基に，平原ら [7] [6] は，適応 Q 型非線形蝸牛フィ

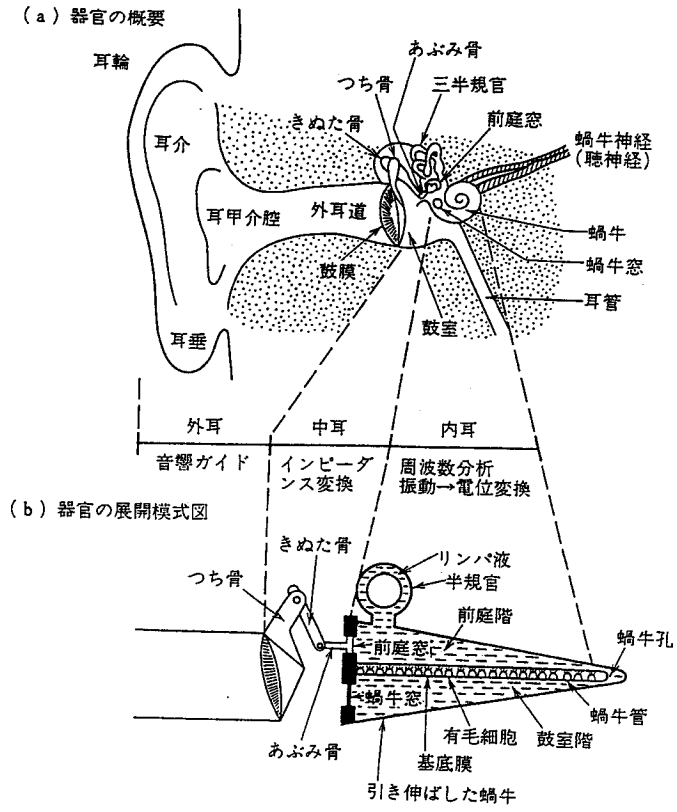


図 2: 耳聴覚器官の構造 (中川他 [2])

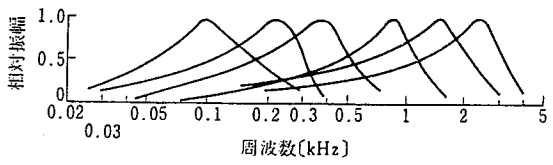


図 3: 基底膜の共振特性 (中川他 [2] を修正)

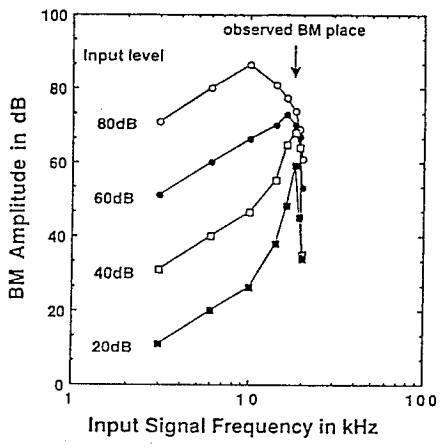


図 4: 基底膜の周波数特性の音圧による変化 (平原 [6])

ルターを提案している．これは，先の基底膜の性質を忠実に模倣するフィルターバンクの一種である．デモンストレーションでは，実際に人間の発声する音声信号を周波数分割させている．それによると，適応Q型非線形蝸牛フィルターは，音声信号中の子音や，周波数変動の大きな部分などの，パワーの小さなスペクトル成分を，明瞭に抽出できることを示している．この適応Q型蝸牛フィルターの詳細については，後の4.2.3節で述べる．

1.2.2 蝸牛神経核

蝸牛から出た周波数分割された音声信号は，最初に蝸牛神経核に入る．蝸牛神経核では，音声スペクトルの各周波数成分毎に，異なった反応を示す細胞がある．これらの細胞の，特徴周波数[1]のトーン信号 (25msec) に対する反応 (発火頻度のヒストグラム) を調べると，図5に示すように，細胞の種類によって異なる形状の出力が得られる．この出力の形によって，これらの細胞を大まかに分類すると，primarylike 型, chopper 型, onset 型, pauser 型, buildup 型の五つに分類できる．

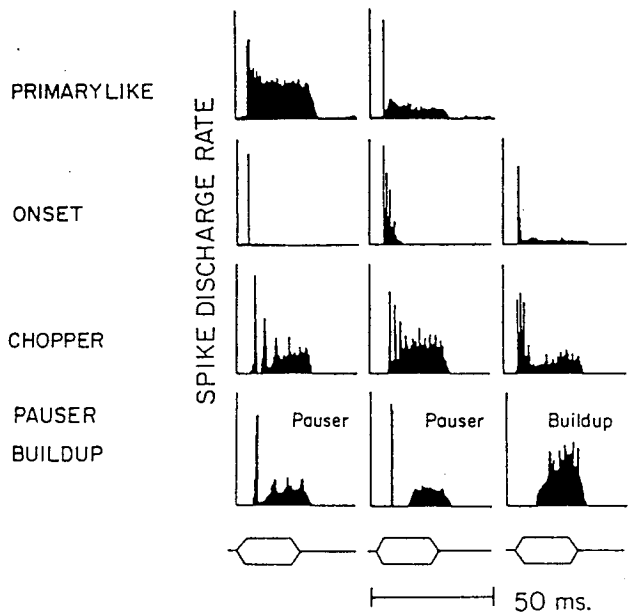


図 5: 蝸牛神経核を構成する細胞の出力 (Morgan[1])

primarylike 型および chopper 型細胞は，おおむね，与えられた入力をほぼ同じ形で出力する性質があるが，これらの細胞の発火頻度は，音声信号のフォルマント周波数に対して

[1]その細胞がもっとも良く反応する周波数.

同期して高くなること (phase-lock) も知られている．すなわち，フォルマントを強調する役目をしていると考えらる．また，primarylike 型細胞の出力は，後述する上オリーブ核に送られ，音源定位に必要な情報を提供している．onset 型及び chopper 型細胞の役割はまだはっきりとは分かっていない．しかし，onset 型の細胞は，ある方向に流れる FM 音に対して非対称な反応を示すことなどから，音声信号の特定の方向に変化する周波数成分を検出している可能性が強い [8]．pauser 型と buildup 型の細胞は，蝸牛神経核よりも上位に位置する上オリーブ核や，下丘からも入力を受けていることが知られている．これらの細胞は，レスポンスを出す帯域幅が，他の細胞よりも広く，ノイズを含んだ音に良く反応する．蝸牛からの出力は，全てこの蝸牛神経核を経由して以降の神経核へ送られることなどから，蝸牛神経核は，以後の処理に必要な特徴を抽出していると考えられる．

1.2.3 上オリーブ核

蝸牛神経核の出力の一部は，次に上オリーブ核に送られる．ここでは，両方の耳からの情報を蝸牛神経核を通して受け取っており，音源定位に役立っていると考えられている [1]．音源定位とは，その音源の方向を検出することで，同一の音が左右の耳に感知される音圧の差および時間差の両方を測定することで実現していると考えられている．音源が正面よりもずれた位置にある場合には，頭の影になった方の耳の受ける音圧がもう一方よりも小さく傾向にある．従って，この音圧の差を検出することによって音源の方位を知ることができる．また同様に，音源が正面よりもずれた位置にある場合には，音源から耳までの距離は，右耳と左耳とでは異なるため，その音源から放たれた音が，左右の耳に到達する時間には，差が現れる．従って，この時間差を測定することでも音の方向を検出することが出来る．

Jeffress(1948) ら [9] は，この時間差を測定する神経回路モデルを提唱した．このモデルは図 6に示すように左右の蝸牛神経核から信号を運ぶ 2 本の Delay-Line と，この Delay-Line から入力を受け取る複数の細胞よって構成される．Delay-Line の入力に近いところでは信号が速く伝わり，遠いところでは遅く伝わる．

図から分かるように上方に配置されている細胞は，一方の Delay-Line の入力に近いとこ

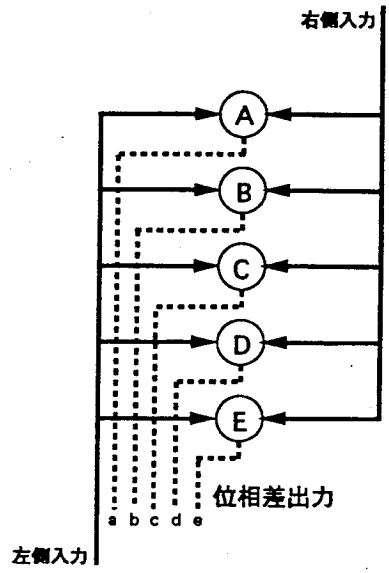


図 6: 音源定位の神経回路モデル (小西 [10])

ろから入力を受け，もう一方からは入力から離れた部分から入力を受けている．逆に，下方に配置されている細胞は，先とは逆の関係が成り立つ．各細胞は，2 本の Delay-Line の両方から同時に信号が入力された時にのみ発火する性質を持つ．したがって各細胞は，同一音声が左右の耳から特定の時間差をもって入力された時にのみ反応することになる．すなわちこれらの細胞は，特定の方向に音源がある時にのみ反応する．

小西ら [11] は，フクロウの上オリーブ核の，層状核において，Jeffress らが予測した時間差計測のための Delay-Line の役目をする細胞の存在を確認している (図 7).

1.2.4 下丘

下丘は，上オリーブ核よりも上位に位置し，上オリーブ核からの出力と蝸牛神経核からの出力の両方が収束している．従って，下丘は，上オリーブ核の出力と蝸牛神経核の出力を統合する役目をしていると考えられている [12]．また，下丘の中には，比較的長い時間長の音に対して選択性を持つ細胞も存在する [1]．例えば，数ヘルツから 1kHz の間で AM 変調を受けた音に対して選択的に反応し，音声のピッチ (基本周波数) を検出するのに役立つと考えられる細胞や，周波数が一定のまま，ある時間以上継続する場合にのみ反応する細胞も見つかっている [13].

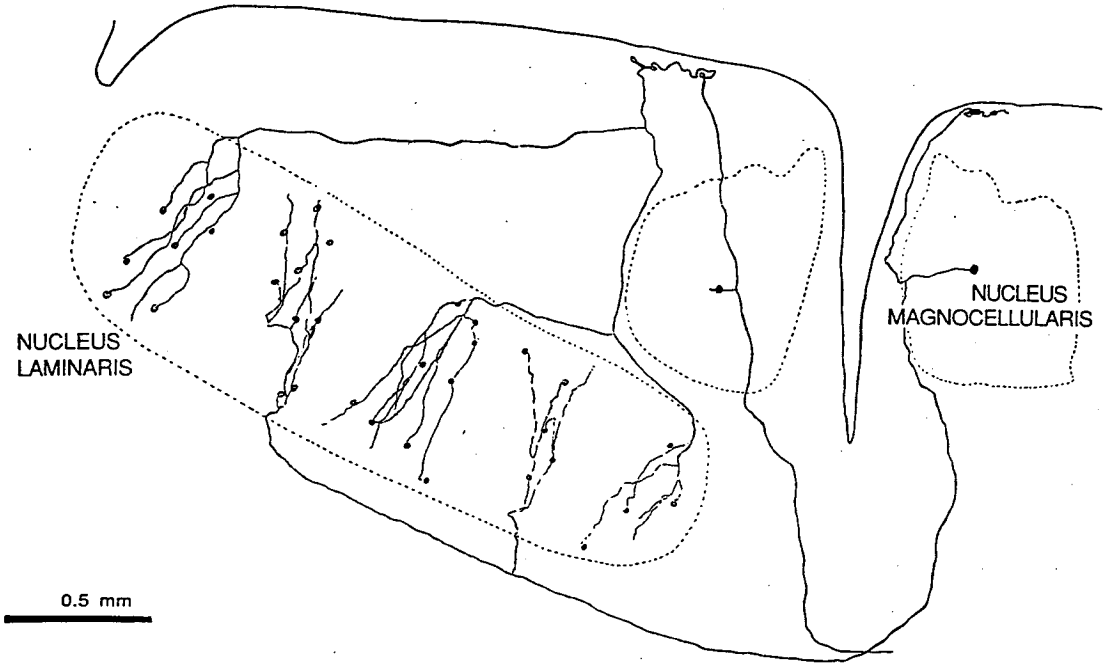


図 7: 層状核 (Nucleus Laminaries) の時間差検出回路
層状核の細胞は左右両側の大細胞核 (Nucleus Magnocellularis) から入力を受ける．黒点は，細胞を表す．この図は，同側と反対側から層状核に入る 2 本の繊維だけを示している．2 本の繊維は，それぞれ，背側と腹側から層状核に入る．繊維の入った所から離れるに従ってインパルスの到着が遅れる．両面間の最大の遅延は，150msec で，フクロウの両耳間の最大時間差と同じである．(小西 [10][11])

1.2.5 内側膝状体

内側膝状体は，腹側領域，内側領域，背側領域に分けられる．この内側膝状体は，蝸牛から大脳聴覚皮質に至る経路の中で一番最後に位置し，それ以前に経由した神経核の出力を大脳聴覚皮質に送り出す役目を果たしている．そのため，内側膝状体には，下位の神経核の出力をそのまま継承する細胞が多く存在する．特に腹側領域はその傾向が強い．この領域には，蝸牛神経核と同じように周波数局在性 (tonotopic organization) が見られる．この領域の細胞は，両方の耳からの入力に反応しるもの，または同側の耳からの入力に抑制されるもの，もう一方の耳からの入力にのみ興奮する細胞に分けられる．また，これらの細胞は，それまでに経由してきたどの神経核の細胞よりも鋭い周波数選択性を示し，この領域で周波数分割が完成すると言われている．さらに，この内側膝状体には，下丘で見られたような，AM 変調を受けた音声に選択的に反応する細胞も存在する．

しかし，内側および腹側領域では先のような傾向は少なく，周波数局在性も見られない[2]．腹側領域の細胞は，周波数に対して緩やかな選択性を持ち，音声信号の onset から数百 msec 遅れて出力を出すなどの性質を持つ．内側領域では，複数の異なった出力を出す細胞が存在する．この領域の細胞は，特定の周波数帯域に選択性を持ち，その音声が入力されるとその直後からレスポンスを出す．この内側領域の細胞は，可塑性を示すことでも知られており，それ以前に経由した神経核の出力だけでカバーできないような滅多に現れない音響特徴を，適応的に獲得しているとも考えられる [1]．

1.2.6 大脳聴覚皮質

内側膝状体の出力は，大脳聴覚皮質に送られる．大脳聴覚皮質は，いくつかの領域に分けることができる [14]．例えば猫の場合，図 8 に示すように，A, A1, A2, DP, P, VP, S, T 等の領域に分けられる．これらの領域の多くは，周波数局在性を持っている [1]．一例とし

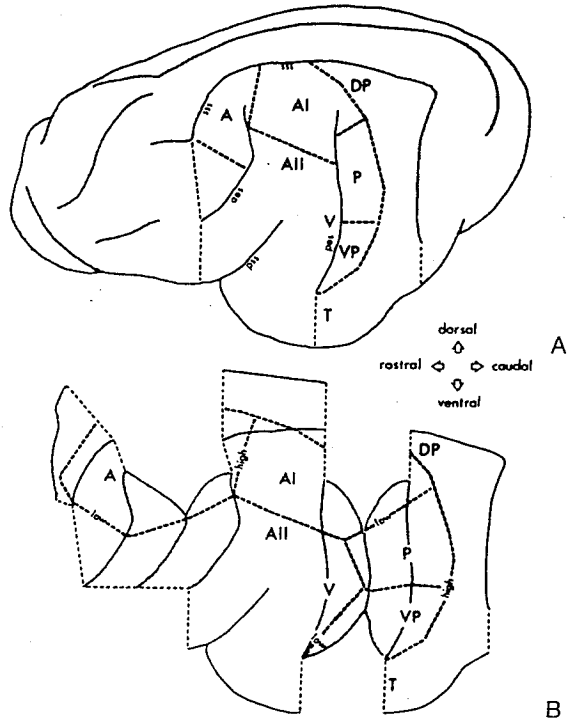


図 8: 猫の大脳聴覚皮質
A:大脳聴覚皮質の位置,B:大脳聴覚皮質を平面状に展開した図 (Donald[14])

[2]これらの細胞が大脳聴覚皮質に投写している部分では周波数局在性が見られる．

て，図 9 に猫の大脳聴覚皮質の A1 領域における細胞の配列を示す．図に示すように，両方の耳からの刺激に対して反応する細胞 (EE で表す) と，同側の耳からの刺激に抑制され，反対側の耳からの刺激に興奮する細胞 (EI で表す) が，特徴周波数毎に交互に帯状につながっている．特にこの A1 野の細胞については，種に関係なく，機能や形態などの点で，他の領域との区別がつき易いという理由から，多くの実験がなされてきた．そして，特定の周波数の AM 音や特定の方向に流れる FM 音 [15]，特定の波形包絡 [16]，帯域雑音 [15] [17]，特定の母音に対して反応する細胞 [18][19] など，複雑な音響特徴に選択的に反応する細胞が数多く発見された．

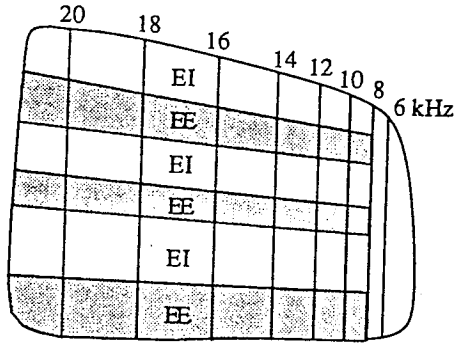


図 9: 猫の A1 野
EI:反対側の耳からの刺激に対して興奮する細胞, EE:両方の耳からの刺激に対して興奮する細胞 (Morgan [1])

A1 野 (あるいは聴覚野全体) の細胞の特徴の一つとして，刺激に対して連続反応を示さず，ON/OFF 反応を示すことが多いことが挙げられている [1]．解剖学的な知見からみても，蝸牛と聴覚皮質の間のいくつかの神経核は，刺激の onset を強調するような出力を出す事が多く，その積み重ねが大脳聴覚皮質の細胞の出力に反映するとしても不自然ではない．これら連続反応を示さない細胞は，変化する音響特徴に選択的に反応するので，音声知覚にとって重要な手がかりとなる音節などの特徴を抽出するのに有効であるとも考えられた [15]．

しかし，丸山ら p[17][18][19] [20] は，ネコの A1 野の細胞のうち，純音に対して連続反応を示す細胞および ON/OFF 反応を示す細胞は，いずれも帯域雑音に対して連続反応を示し，結局 A1 野の 75%もの細胞が特定の刺激に対して連続反応を示すと指摘している．この連続反応は，刺激を 100msec 以上継続して提示したときのみに見られ，通常 80msec 程

度の潜時と，200 から 300msec の漸増反応の後定常状態に落ち着く．すなわち，これまで ON/OFF 反応しか示さないと見られていた細胞は，最適刺激を提示されていなかったために連続反応をしなかった可能性があり，今後のさらに詳しい実験結果が待たれる．

ところで，最近，電位感受性色素を使った光による計測法が取り入れられるようになり，広範囲の大脳聴覚皮質の反応が時間と共にどのように推移するかを RealTime で計測できるようになった [21][22]．これらの実験結果によると，音声信号 (特定の周波数の帯域雑音など) を短時間提示すると，最初にある領域が活性化され，次に別の領域が活性化された．この活性領域の移動は，大脳聴覚皮質が，音声を階層的に処理している可能性を示唆している．

1.3 心理学的に見たヒトの聴覚系

人間が何を手がかりに認識しているのかを知ることは，音声認識装置を構成する上で，重要なヒントになる．そのため，この音声知覚に関しては，心理学的な側面から多くの調査が行なわれてきた．

本節では，まず，種々の心理実験によって明らかになった音韻知覚および単語知覚の手がかりの中から，いくつかを概観する．そして，これらの知見によって考えられた単語知覚モデルについても概観し，このモデルが予想するヒトの聴覚系のメカニズムについて述べることにする．

1.3.1 音韻知覚

音声知覚の手がかりには，音声スペクトルの形状とその時間構造がある．

一般に母音 (vowel) の知覚に関しては，その第 1 および第 2 フォルマント[3] の位置が手がかりとなり，時間構造にはあまり依存しない．図 10 に，母音/a/,/i/,/u/,/e/,/o/をそれぞれ単独に発音した場合の第 1, 第 2 フォルマントの位置を示す．

図中，楕円は，話者によるばらつきを表している．図から分かるように，母音を単独に発音した場合には，各母音の第 1 第 2 フォルマント位置の組み合わせには，明かな違いを

[3]音声信号の周波数成分を表す．この周波数成分の中で，もっとも振幅の大きいものから第一フォルマント, 第二フォルマント,... と数えてゆく．

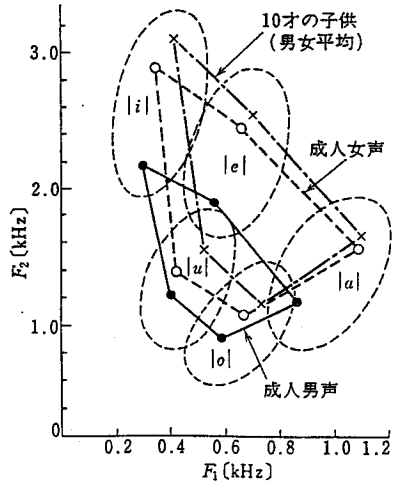


図 10: 日本語の母音フォルマント (中川他 [2])

見いだす事が出来る.

しかし，連続音声中の母音は，その前後に来る音素に応じて変化（調音）し，極めて曖昧である．例えば，/aia/と発音した場合，その/i/の部分における音響的特徴は/e/とほぼ同じか，/i/があったとしても，非常に短い持続時間しかない．しかし人間は，二つの/a/の間に/e/を知覚する事はなく，/i/を知覚する [23]．このような現象は，聴覚処理系の中に，ある種の補正機構が存在し，この補正された特徴が知覚されるために生ずると考えられている．このような考え方を基に，音節中のフォルマント遷移の初期段階における情報から，母音ターゲットを予測するモデル [23]，音声発生機構の物理的制約条件に関する情報を使って母音ターゲットを予測するモデル [24]，そして，調音結合が，前後に来る音素によって決まる事に着目したモデル [25] などが提案されている．

子音 (consonant) の場合は，スペクトル形状と時間情報の両方が知覚の手がかりになっている．特に子音の後に母音が来る音節 (CV 音節) の識別には，時間情報が重要である [26]．例えば，無声破裂音の場合には，破裂部のバーストの後の気音部分の時間長が知覚の手がかりになる．また，有声破裂音の場合には，ONSET 部分のフォルマント遷移の開始位置が重要である．

音節 (sylable) の識別には，フォルマント遷移のスピードが重要な手掛かりになる事がある．例えば，音節/wa/と/ba/の弁別には，フォルマントの遷移スピードが，手掛かりになっている．このフォルマントの遷移スピードは，/ba/の方が/wa/よりも明らかに速い．従っ

て，/ba/のフォルマント遷移部分を時間的に伸ばすと，/wa/に知覚される [27].

この/ba/および/wa/の知覚については，次のような興味深い心理実験結果も報告されている．すなわち，音節/ba/および/wa/の中間程度のフォルマント周波数遷移速度を選び，そのフォルマント遷移速度を一定にしたまま，長い音節と短い音節を作って[4] 被験者に聞かせると，音節が長い場合には，/ba/に知覚される割合が増し，逆に短い場合には，/wa/と知覚される割合が増すという [27] (図 11参照)．これは，音節の時間長から，発話速度が

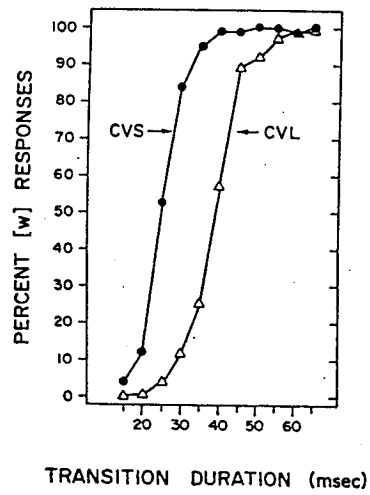


図 11: 音節の時間長と/wa/と知覚される割合との関係
縦軸は/wa/と知覚される割合を表し，横軸はフォルマント遷移時間を表す．[w]–[b] 間の様々なフォルマント遷移時間を選んで，それぞれについて長い音節の場合と短い音節の場合の両方で/wa/と知覚される割合を示している．図中●は時間幅の短い音節を表し，△は時間幅の長い音節を表す (Pisoni [28]).

知覚され，それを手がかりに，聞き手が音声信号の時間長を正規化することによって，生ずる現象と考えられている [27]．ちなみに，これに似た現象が，非言語音でも生ずる事が示され，言語知覚特有の現象ではない事も明らかにされている [28].

実際には，音韻知覚には，以上示してきたものだけではなく，その音韻が構成する単語の知覚にも大きく依存することが知られている．たとえば，単語中の一つの音素を雑音で完全に置き換えても，その雑音部分が，本来あるべき音素に付加雑音が乗ったような形で知覚されるという現象がある [29]．この現象は，音素修復と呼ばれ，雑音部分以外の音素

[4] フォルマント遷移部分の時間長を一定にし，フォルマント遷移が無い部分 (母音部分) の時間長を変化させることによって音節全体の時間長を変化させる．

からその単語が知覚され，その単語知覚結果の情報が音韻知覚へフィードバックされることによって生ずるものと考えられている．また，単語の語頭部分の音韻知覚速度は，語尾部分の音韻知覚速度よりも遅い [30] という現象も，明らかに音韻知覚が単語情報の影響を受けていることを示唆している．すなわち，音韻知覚と単語知覚は，互いに影響を及ぼし合いながら同時並行的に実現されている可能性が高い．

1.3.2 単語知覚

我々が，紙に書かれている英語の文章を読む際には，単語と単語の間にある空白を目印に，各単語を読みとることが出来る．しかし，我々が喋る言葉は通常，単語と単語の間に無音期間を空けることはなく，連続的に発話する．例えば，“I want to eat cheeze toast.” という文章を発話したものを無理矢理書き下すとすれば，“Iwanttoeatcheezetoast.” となろう．つまり，音声信号中の単語と単語の境界は，極めて曖昧なものである [31]．従って，音韻知覚の場合と同様に，単語知覚が何を手がかりに行なわれているのかについては，これまでに多くの心理学的側面からの調査が行われてきた．

当初，単語知覚の諸現象としては，知覚対象となる単語の出現頻度や，その単語の位置する文章の意味情報がその単語の知覚に影響するという結果が報告されていた．たとえば，Miller [32] は，種々の SN 比を持つ雑音下での単語の了解度を計測すると，日常生活でよく利用される単語ほど，了解度が大きい事，また，その単語が文章中にある場合には，その単語が文章と意味的につながるほど明瞭に知覚されることを示している．このうち，意味情報が単語知覚に影響する知見は，単語知覚がその上位の意味知覚との間で互いに相互作用を及ぼし合いながら実現していることを示唆するものである．

このような知見から，人間の単語知覚には，次の 3 つの段階があると予想された [33]．

1. **Access**: 知覚刺激を基に単語候補の活性値[5]を計算する．

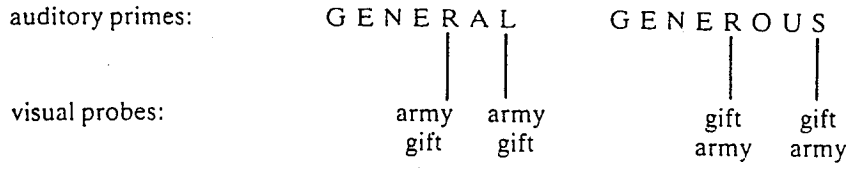
2. **Selection**: *access* された複数の単語候補の中からもっとも適当な候補 (スコアーの高いもの) を一つだけ選び出す．この選択には，意味および文法といった高次の情報が使われる．

[5]その単語候補に対する自信の度合を表す．もし，脳の中に特定の単語に対して反応する細胞があると仮定すればその細胞の出力値とも考えることが出来る．

3. **Integration**: 選択された単語候補を高次の表現へ関係付ける.

研究者の主な関心事は，これら三つの段階が，いつどのように働くのかということであった.

Marslen-Wilson [34] は，クロスモーダルプライミングと呼ばれる心理実験を行ない，人間が単語のどの音素までを聞きとった時にその単語を知覚できるか (Selection がいつ行なわれるか) を調査している．この実験では例えば，"General(陸軍大将)" および "Generous(気前の良い)" の二つの単語を被験者に聞かせながら，視覚刺激として gift/army を被験者に見せ，各々の場合に，視覚刺激が提示されてから，その視覚刺激が何であるかがわかるまでの時間 (Reaction-Time) が視覚刺激を提示する時間位置によってどのように変化するかを測定する．この視覚刺激を提示する時間位置は，この二つの単語の単語を区別出来る時点/r/ (uniqueness point と呼ばれる) 以前と，それ以降の部分である (下図参照)．その結果，



uniqueness point /r/での視覚刺激 gift/army に対する reaction-time には変化がなかったが，uniqueness point 以降の視覚刺激に関しては，"General" の場合には army が，"Generous" の場合には gift がそれぞれ短い reaction-time を示した．すなわち，uniqueness point /r/を境に，どちらか一方の単語の活性値がもう一方よりも大きくなった事を意味している．この結果から彼は，人間の単語知覚のプロセスは単語毎の処理で，逐次的な処理によって単語の uniqueness point を見つけ，これをきっかけに Selection を実現しているとした.

また，単語知覚における Selection は，高次の意味情報の知覚との相互差用によって実現するが，この意味情報の効果がどの時点でどのように現れてくるのかについても調べられている．Zwitserlood [35] は，知覚対象となる単語に対して，文脈的制約条件を付けない文章 (carrier phrase), 弱い意味的構造的制約条件を付ける文章 (neutral context), そして強い意味的，構造的制約条件を付ける文章 (biasing context) を用意し，それぞれの場合において，知覚対象となる単語の各時点における活性値をクロスモーダルプライミング手法を使って調べている．図 12 は，その実験結果を示している．図は，横軸が時間を表し，縦軸が正しい単語候補 (Actual Word) と，それ以外の単語候補 (Competitor) のそれぞれに関係

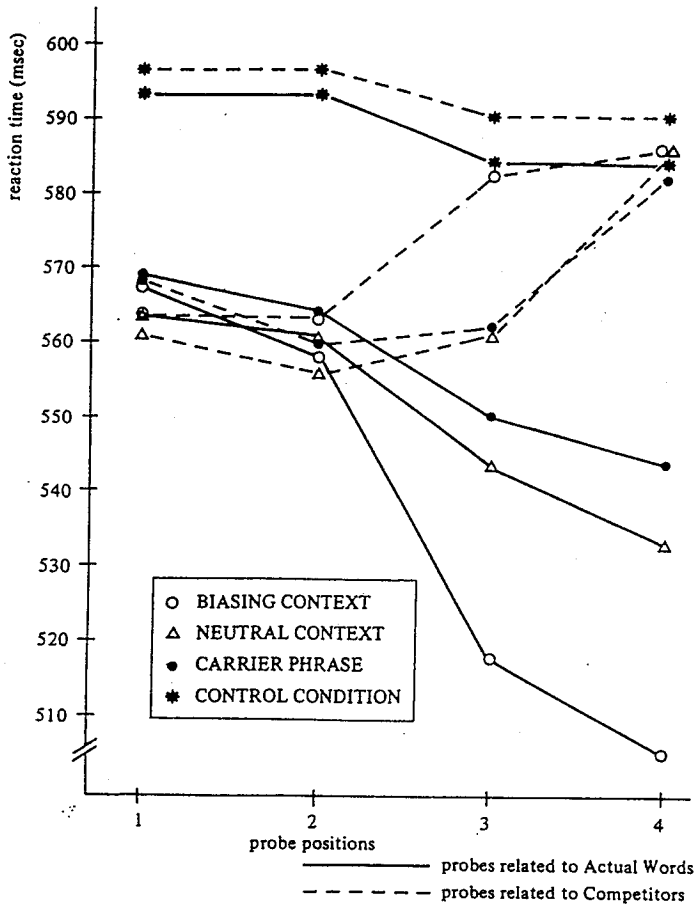


図 12: actual-word と competitor に関連した視覚刺激の reaction-time (Zwiserlood[35]).

する視覚刺激に対する reaction-time を示している．横軸のラベル 3 は，単語の uniqueness point に相当する時点である．この図から，uniqueness point(時点 3) 以前では，文脈効果の大小にかかわらず，正しい単語候補とそれ以外の単語候補の reaction-time にはさほど差が見られないものの，uniqueness point 以降では，文脈効果が大きいほど，両者の間の reaction-time の差が大きい事がわかる．すなわち，Integration と Selection は，uniqueness point 以降に同時に働くことを示している．

しかし，他の単語の一部になり得るような短い単語の場合，先のような uniqueness point が，存在しない場合もあり，必ずしも人間の単語知覚が uniqueness point を手がかりに単語を知覚しているとは考え難いとする知見がある．Grosjean [36] は，“I saw the” に様々な単語とその単語に関連のある文を後続させて作った連続音声を，複数の披験者に聞かせ，その中の I saw the の直後に来る単語がその単語自身と後続する文の各時点でどのように

知覚されるかを被験者に書かせた．図 13 に，その結果の一つを示してある．

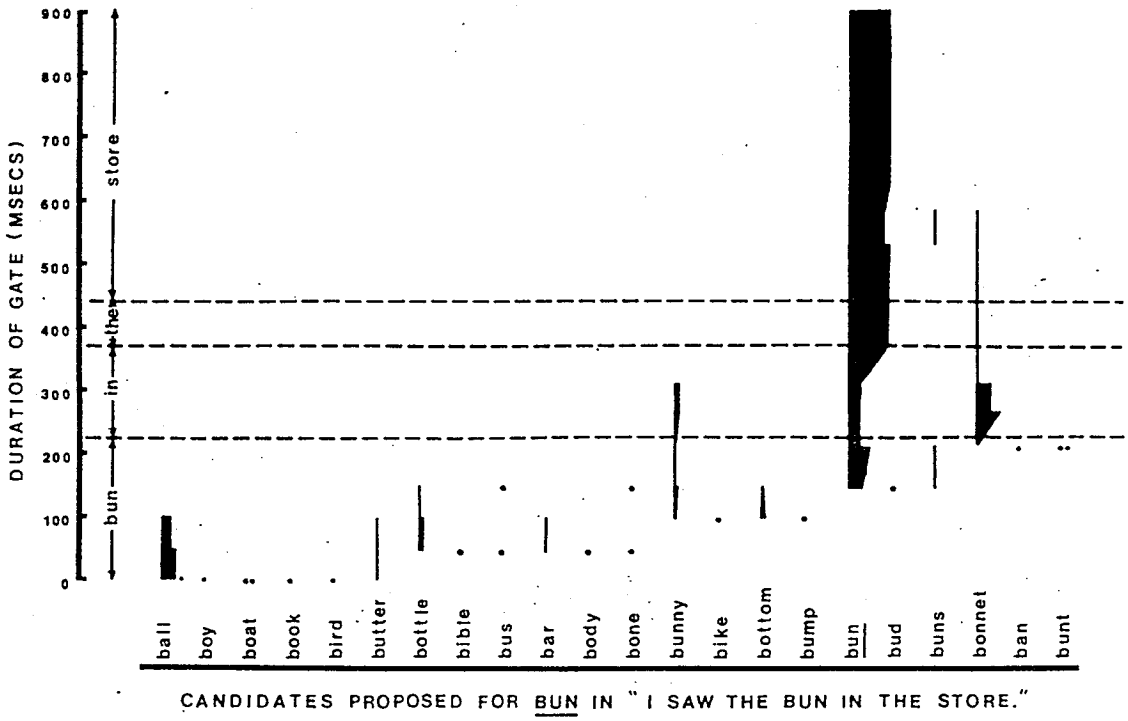


図 13: 単語 "bun" に対する候補の推移 (Grosjean[36]).

この図は，"I saw the *bun* in the store." を聞かせた時の "bun" に対して挙げられた候補とその候補を挙げた披験者の人数を表している．図中，縦軸が時間で横軸が各候補を挙げた人数である．この図から分かるように，*bun* の offset を聞いた段階では，"bun" を候補に挙げている人は少なく，後続する前置詞 "in" を聞いた段階で "bun" を候補に挙げている人が多い．すなわち，聞き取る単語の長さが短い場合には，その単語の offset が過ぎ去った後しばらくして，その単語と，その後に来た単語が同時に知覚されることを示している．このような傾向は，その短い単語が別の単語の一部になり得るような場合に顕著である．この結果から，人間の単語知覚のプロセスは必ずしも逐次的な，あるいは時間軸に沿った左から右への処理，単語毎の処理ではない可能性が強い．

1.3.3 単語知覚モデル

以上のような心理学的な知見を基に，いくつかの単語知覚モデルが提案されている．ここでは，そのなかでも代表的な単語知覚モデル: COHORT [34] および，TRACE [25] に

ついて概観する．

Marslen-Wilson [34] は，**COHORT** [6] と呼ばれる単語知覚モデルを提唱した．このモデルは，人間の単語知覚は，単語毎の逐次的な処理であって，その単語知覚は，単語の uniqueness point まで聞きとった段階で成立するという考え方をもとに構成されている．

このモデルはまず，単語の先頭の音素が入力されると，その音素で始まる単語を候補として挙げ，それ以外の単語は COHORT から外す．そして第二の音素が入力されると，その先頭の音素と第二の音素の組み合わせを持つ単語に候補を絞る．このように次々と音素が入力されるたびにマッチしない単語候補，そして，構文や意味情報にそぐわない単語候補についても COHORT から取り除いて行き，一つの単語候補が残った段階で，その候補を認識結果として出力する．すなわち，COHORT は，前節で示した三つの段階 Acces, Selection, Integration がが互いに影響を及ぼし合いながら単語を知覚するモデルといえる (図 14参照)．前節で述べたように，当初，人間が単語を知覚する時点は，その単語の uniqueness point

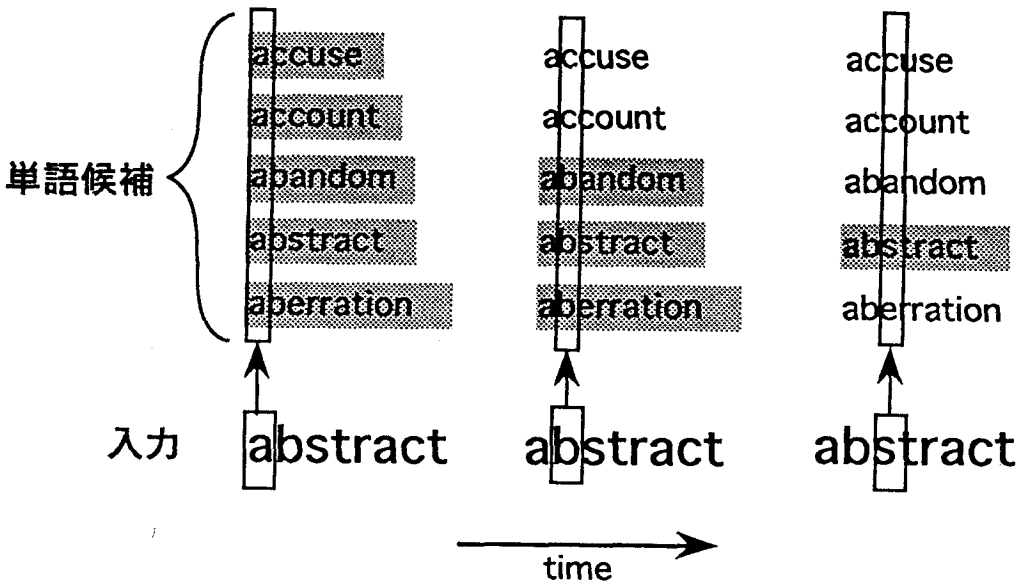


図 14: COHORT

であることが心理学的にも確かめられていたため，この COHORT の仮説は正しいとされていた [34]．

しかし，COHORT には，以下のような致命的とも言える欠点がある [2]．それは，第一

[6] 古代ローマ語で，グループを表す．

に，このモデルの入力としては，極めて精度の高い音声信号が必要であるという点である．すなわち，単語の先頭の音素が雑音などの原因で誤った別の音韻とみなされてしまった場合には，正確な単語がCOHORTから取り除かれてしまい，高次の意味情報をもってしても後から修復できないという事態が発生してしまう．第二の問題点は，このモデルが単語の先頭と終端の音韻が正確に知覚されることを前提にしている点である．これに対して，実際の連続音声における単語境界は極めて曖昧 [31] で，単語の先頭と終端の音韻を，常に正確に検出することは難しい．

Elman と McCelland [25] は，このCOHORTを出発点として，上に挙げたCOHORTの欠点を克服する単語知覚モデル **TRACE** [7] を提案した．TRACEは，先のCOHORTの問題点を克服できるだけでなく，数多くの単語知覚現象を説明できるという点でも優れている [37].

このモデルは音響特徴層，音韻層，単語層の三つの層で構成されている (図15参照)．各層には，閾値，活性度およびresting-valueを持つユニットが並んでいる．隣合う層間のユニット (細胞) は互いに興奮性結合で継っているが，同一層内のユニットは，互いに抑制性の結合で結ばれている．従って，一つのユニットは，そのユニットがつながる他の層の細胞を活性化するように働くと同時に，同一層内にある他のユニットの活性化を妨げるように働く．これにより，ある入力が与えられると，その入力にもっともふさわしいユニットのみが勝ち残り，活性化される．各ユニットの入力には時間幅があり，時間軸方向に並進対称に並べられている．そのため，入力された音声信号は，時間軸方向に空間展開して処理される．同一層内のユニット同士は，抑制性結合でつながれているが，この抑制性結合は同一時間内の処理に関係するするユニット同士だけである．

このような構造を持つことにより，TRACEは，実際の音声信号のように，音素間，あるいは語と語の間には無音期間があることが少なく，その境界が曖昧 [31] であっても，その中から単語を認識することができる．すなわち，各単語ユニットは時間方向に並進対称に並んでいるため，あらゆる部分が単語の語頭あるいは語尾であると仮定して処理してゆくことになり，各時点においてもっとも確からしい単語候補を立てることが可能である．

また，TRACEは，単語の語頭音や途中の音素が抜け落ちたり正しく識別されなかった

[7] 足跡，痕跡の意．

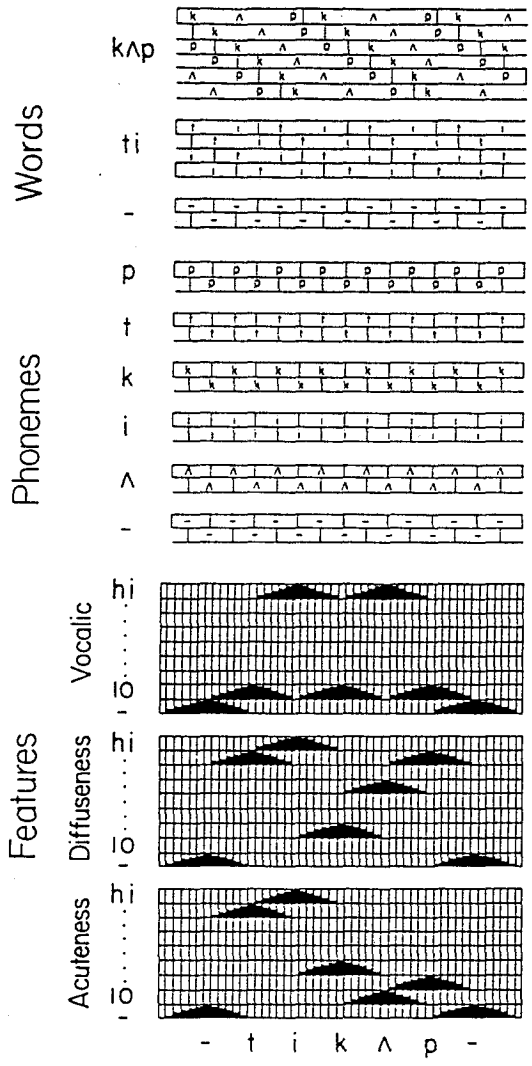


図 15: TRACE (Elman[25]).

りした場合でも，それを修復するという音素修復 [29] に関する現象が説明できる．すなわち，正しく識別できた他の音韻を手がかりにして，それに対応する単語ユニットを活性化させ，その単語ユニットからのフィードバック信号により，抜け落ちた音素に対応するユニットを活性化させることができる．

さらに，長い単語の一部になり得るような短い単語が入力された場合には，その単語の offset が過ぎ去ってからしばらくしてその単語が知覚されるという現象 [36] も，TRACE は説明できる．その短い単語の offset までを入力された段階では，その単語に対応するユニットとそれを含む長い単語に対応するユニットの両方が活性値を持つ．しかし，その後，後続する音素を入力されることにより，その長い単語に対応するユニットが，後続する音素

で構成される別の単語ユニットに抑制されることによって，活性値が下がり，結果としてその短い単語に対応するユニットが勝ち残る．

以上述べたように，TRACE は，多くの単語知覚現象を説明できる．従って，このモデルに似た構造が，人間の聴覚系の中にも存在する可能性が高いといえよう．

しかし，この TRACE にも，次のような問題点が残されている [37]．まず，TRACE では，各ユニットの時間幅があらかじめ固定されているため，時間方向の変化への対処が困難であり，発声速度の変化に対処できない．また，個人差によるスペクトル形状の変動を吸収できないことや，学習則が示されていないことも問題である．

1.4 時系列パターン (音声) 認識研究の背景

前章で述べてきたような正理学的あるいは心理学的な側面からの聴覚系の構造予測の一方で，時系列パターン認識，音声認識のための必要論的な立場からの研究も数多くなされてきた．

一般に柔軟な時系列パターン認識能力を得るためには，次の二つの機構が必要である．

- 入力パターンの非線形な時間伸縮に影響されない機構．

  一般に音声信号の時間伸縮は，母音について顕著で，子音についてはそれほど大きな伸縮を受けない [26]．従って，音声信号は全体として非線形に伸縮する傾向があり，非線形伸縮に対応できる機構が必要である．

- 各単位時間毎に現れるパターンの変形に影響されない機構．

  特に連続音声中の母音は，調音結合すると大きく変形するので，この変動に対処できる機構が必要である．

これまでに，この二つの機構を実現するために，多くのモデルが提案されている．これらのモデルを，時間軸方向の処理方法という観点から大きく分けると，次の 3 つに分類される．

① 動的計画法を用いた時空間パターンマッチング法

② HiddenMarkov 法など，状態遷移を基礎に置く方法

③ 遅延素子を使用する方法．すなわち，遅延素子を使って時空間パターンを空間パターンに展開し，空間パターン認識手法を使って認識する方法．

以降の節ではこれらを順に説明して行くことにする．

1.4.1 DP マッチングを用いた手法

DP マッチングを用いると，非線形に伸縮した時系列パターンと，あらかじめ用意された標準パターンとをうまくマッチングすることができる [38]．この手法はまず，入力された時系列パターンが，標準パターンの一つにもっとも近くなるように，その入力パターンを局所的に時間伸縮させる (図 16参照)．そして，両パターン間の距離を測定する．このような距離測定を，全ての標準パターンとの間で行ない，もっとも距離が短かった標準パターンを認識結果とする．

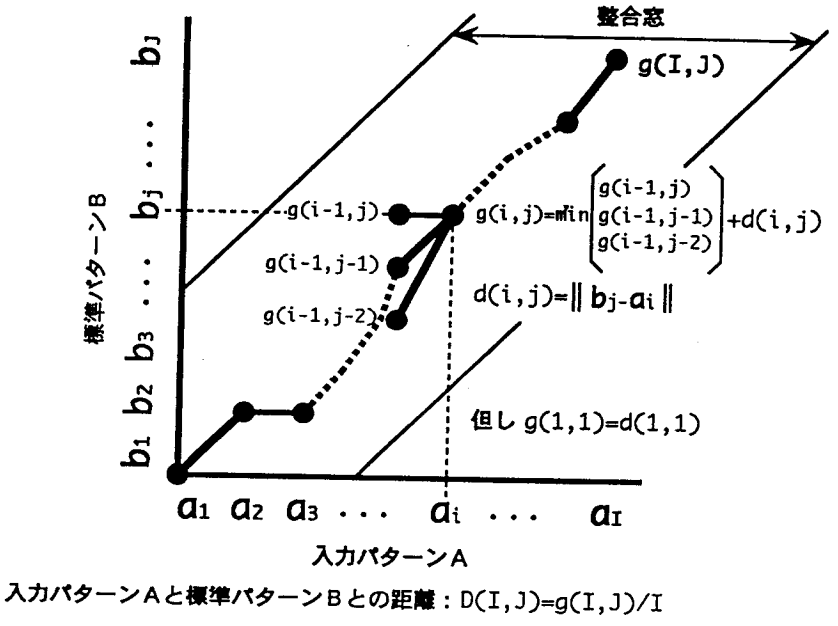


図 16: 動的計画法を用いた時間軸の適応正規化の一例 (中川他 [2] を修正)

この手法は，時間軸方向の伸縮に対しては対処できるものの，音声スペクトルパターンの変形に対して弱いという欠点を持つ．しかし，最近この手法を，人工ニューラルネットワークのの構造の中に組み込む試みがなされ，パターンの変動と時間軸の伸縮の両方に対処しようとする試みがなされている [39] [40]．

1.4.2 状態遷移を基礎に置くモデル

状態遷移に基礎をおくモデル [34] [41][42] [43] は，時系列パターンをその各特徴の入力された順序に基づいて識別する．1.3.3節で説明した単語知覚モデル COHORT [34] は，この典型である．従って，これらのモデルは，時間軸の非線形伸縮にほとんど影響を受けない．

一般に，これらのモデルの振舞いは，そのシステムの内部状態によって説明することが出来る．すなわち，これらのモデルでは，一時刻前のシステムの状態と，現在与えられている入力とから，現在のシステムの状態が決定される．そして，入力が終了した時点での状態が認識結果となる．

しかし，この状態遷移モデルには，次のような問題点が指摘されている [37]．例えば，単語認識への応用を考えた場合，単語中の音素の一つが抜け落ちていたり，あるいは単語中の音素が，一つでも誤って識別されると，それ以降の状態遷移は，本来のものとは全く別のものになってしまう．

このような，入力パターンの変動に付随する問題を許容する手法として，Hidden Markov Model (HMM) が知られている [44]．HMM は，時系列パターンがある確率モデルから発生したものという仮定に基づいて構成されている．すなわち，HMM は，いくつかの状態とその状態間の遷移確率，一つの状態遷移に付随しておきるシンボル (1 フレーム毎のパターン) の発生確率によって記述された確率モデルである (図 17)．一度の状態遷移には，あらゆる種類のシンボルの発生が許されているが，シンボルの種類によって発生確率の大小が異なっている．また，一つの状態からは，複数の状態への遷移が許されるが，どの状態へ遷移するかによって，遷移確率の大小が異なっている．

これを時系列パターン認識に使用する際には，認識対象となるカテゴリーそれぞれについて，状態遷移図よって記述された確率モデルを用意する．入力が与えられると，各カテゴリー毎に，その確率モデルの中から，入力パターンと同じシンボル系列が発生する確率を計算する．そして，もっとも高い発生確率を示した確率モデルのカテゴリーが，認識結果となる．すなわち HMM は，入力パターンの変動をシンボルの発生確率という形で許容し，時間軸の伸縮に対しては，状態遷移確率の形で許容するモデルと見る事ができる．

ちなみに，この手法によって小型の確率モデルをいくつか構築しておくと，それらを複

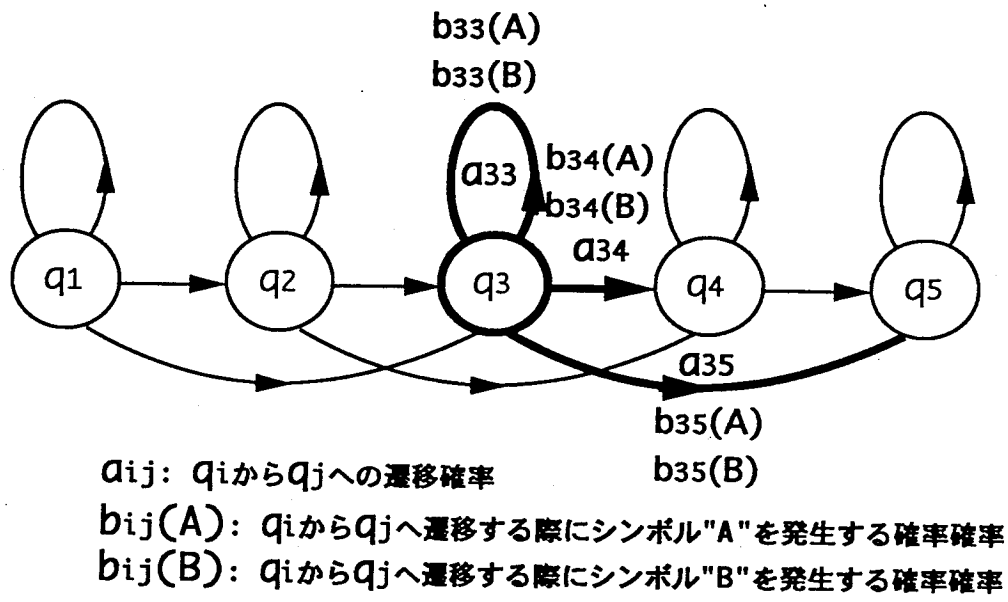


図 17: HMM における状態遷移とシンボルの発生
A または B の 2 種類のシンボルの発生を記述する状態遷移図の例 (中川他 [2] を修正)

数接続するだけで，容易に大きな確率モデルへ拡張することが出来る．例えば，音韻毎の確率モデルを構成しておくと，それをつなぎ合わせるだけで，単語などの確率モデルを構築できる [45]．この手軽さが要因で，現在音声認識にもっとも多く使われており，不特定話者を対象とした，連続音声認識 [46] も実現している．また，この HMM と同等の振る舞いをニューラルネットワークの構造の中に組み込む試み [47] [48] も行なわれている．

ところで，音声認識には，各シンボル (スペクトルパターン) の継続時間長も重要な手がかりの一つであることは，1.3.1節でも述べた．これに対して，この HMM では，各シンボルの継続時間を陽に確率モデルに組み込むことができない．最近，この点が問題視されるようになり，これに対する解決法が考案されるようになった．例えば，有木 [45] は，従来の HMM が，1 フレームのパターンの発生確率のみを使って，確率モデルを構築していたことに対して，1 フレームのパターンの発生確率と，その継続確率とをペアにして確率モデルを構築する方法を提案している．これにより，英語音声に対する音韻認識率が従来の HMM に比べて 8%程度向上することを示している．

1.4.3 遅延素子を使用する手法

多次元の時系列パターンをあらかじめ遅延素子を使って時空間パターンを空間パターンに展開してから，空間パターン認識手法を使って認識する方法が多くある．1.3.3 で示した単語値各モデル TRACE [25] はその典型であろう．前節の状態遷移モデルが，基本的にパターンの順序のみを検出していたのに対して，これらのモデルではその時間構造も検出することが可能である．しかしこの場合，時間伸縮への対処方法が問題となる．

当初，福島 [49] は，遅延素子を使って時系列パターンを空間展開し，そのパターンを自己想起型の連想メモリーの入力として使用した (図 18参照)．遅延素子の入力端には，時系列パターンが入力されると同時に連想メモリーの各出力も入力される．計算機シミュレー

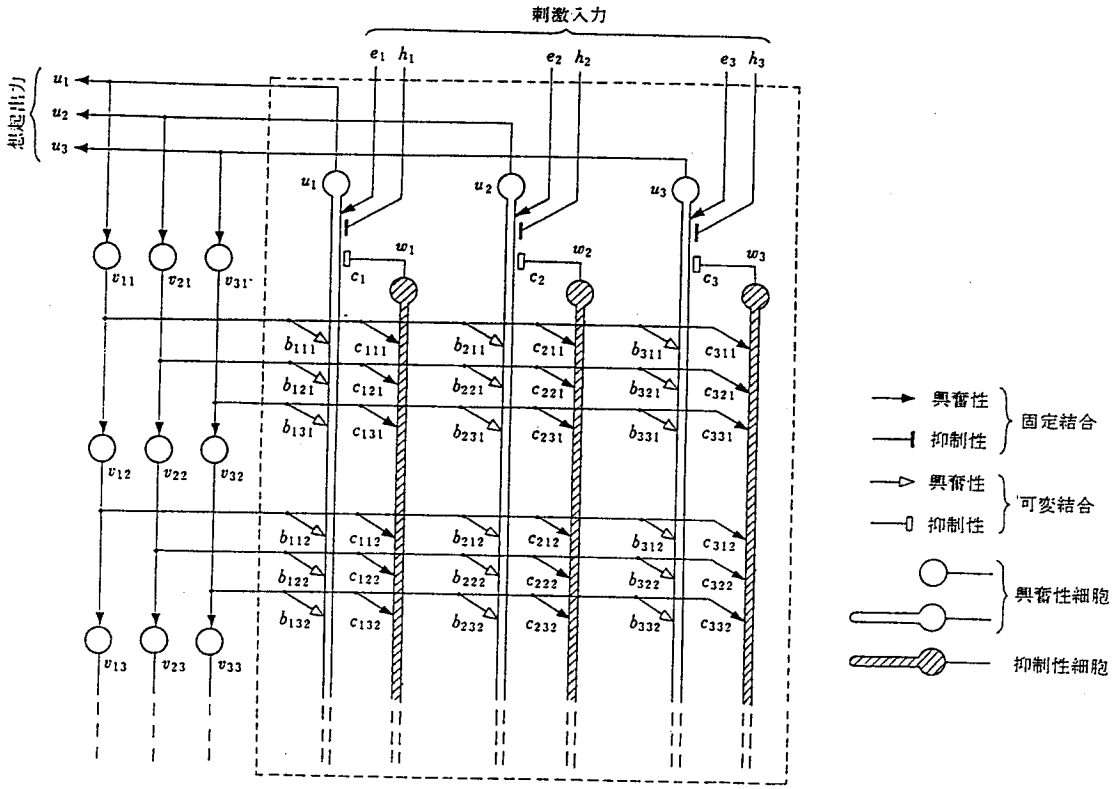


図 18: 時空間パターンを想起する連想メモリー (福島 [49])

ションでは，電光掲示版にながれるような文字列を想起するように (例: 'ABCDEF...') 学習させた後，ノイズが載った時系列パターンを一部 (文字列の一部) だけ提示した．すると，このモデルは，その時空間パターン全体 (文字列全体) を正しく想起した．しかし，このモデルは，時間軸方向の伸縮に対応する能力は持ち合わせていない．

Tank,Hopfield [50] は，先の福島のモデルに似た構造を持ち，時間伸縮に対応できるネットワークを提案している．先のモデルとの主な違いは，時間伸縮に対応できるようにするために，遅延素子上を流れる時空間パターンに対して時間軸方向のボカシを施していることである (図 19)．彼らは，このネットワークにいくつかの都市名をコンテキストで提示

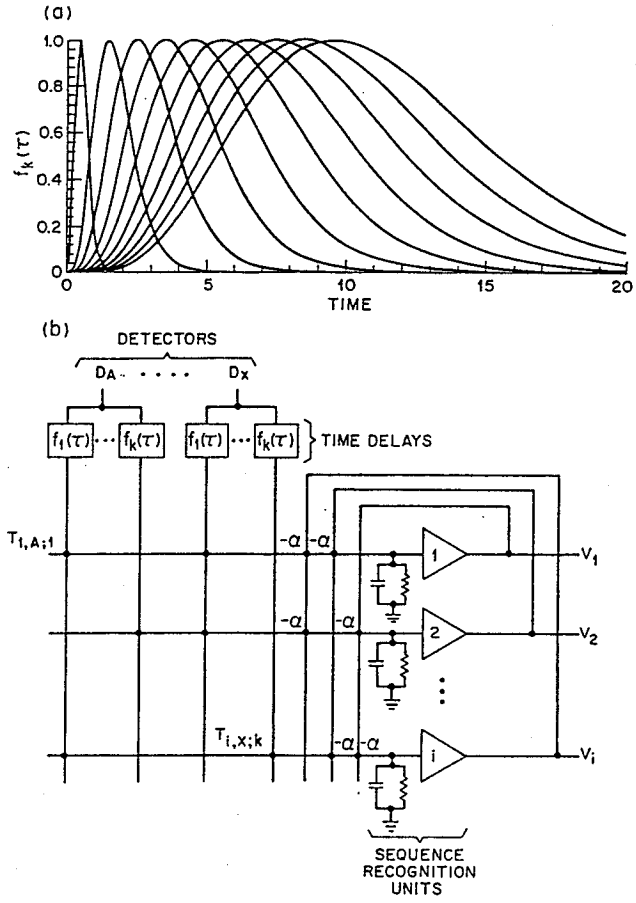


図 19: Tank, Hopfield の提案したモデル (Tank,Hopfield[50])

し[8]，それを想起できるように各ユニットの入力結合の強度を決定した．その後，様々なスピードで都市名を提示し (例:速いパターンを *NEWMEXICO* とすれば，遅いパターンは *NNNEEEWWWMMMEEEXXXIIICCCOOO* となる)，それを想起させる課題を行なった．その結果，このネットワークは時間伸縮に対して対処しつつその都市名を正しく想起した．しかし，時間伸縮に対処できる反面，コンテキストの順序が少々入れ替わってもそれを区

[8]図 19 に示されているように，ネットワークの入力部には，アルファベット 'A'〜'Z' に対する detector がそれぞれ用意されている．従って例えば，都市名 *WASHINGTON* に対応する時系列パターンを提示する場合には，最初に 'W' に対する detector に信号を送り，次に 'A' に対する detector に信号を送り... というふうに入力する．

別することができなくなるという問題が生じている (例:*AWSHINGTON* と入力されても，*WASHINGTON* と同じカテゴリーであるとみなされてしまう).

伊藤，福島は [51][52]，図形パターンの位置づれや変形に強力に対処できる視覚パターン認識モデル "ネオコグニトロン" [53] を利用し，時系列パターン認識への応用を試みた (図20参照). このモデルは，聴覚系の細胞の反応特性を参考にし，一次遅れ特性を持つ細胞を

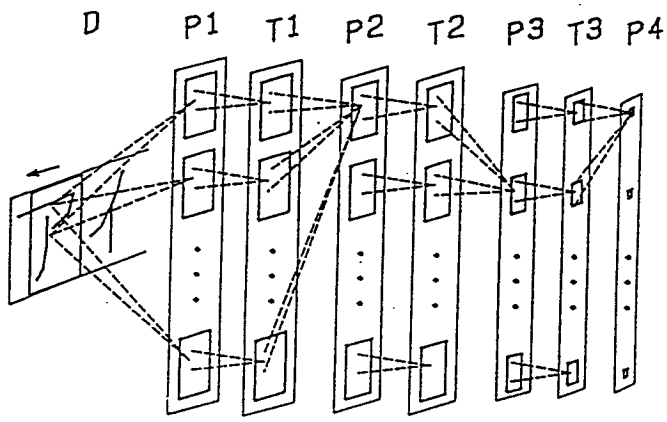


図 20: ネオコグニトロンを使用した時系列パターン認識 (伊藤 [51])

時空間パターンの時間方向へのボカシとして使用している. 計算機シミュレーションでは，音声スペクトルに見立てた電光掲示版に流れるような文字パターンを提示し，流れてくる各文字を認識するように学習させた. その後，崩れた形状の文字で時間方向に伸縮を受けたパターンであっても正しく認識した.

Waibel [54] は，Time Delay Neural Network (TDNN) と呼ばれる Time-Shift-Invariant な階層ネットワークモデルを提案している. このモデルが，通常の階層ネットワークと違う点は，各層の出力が遅延素子によって空間パターンに展開され，それが次の層の入力となる点である (図21参照). 計算機シミュレーションでは，このモデルに，極めて似通った音響特徴を持つために従来の機械認識では，区別をすることが難しいとされていた音韻/b,d,g/を識別するように教師有り学習法:Back-Propagation アルゴリズムによって学習させている. その結果，このネットワークは，誤り率 1.5%でこれらを識別し HMM(誤り率 6.5%) よりも高い能力があることを示した.

また，McDermott [55] は，Shift-Tolarant LVQ を使って音韻認識を行なわせた. このモデルは，Kohonen が提案した Phonetic Typewriter [56] を発展させたものである. Shift-

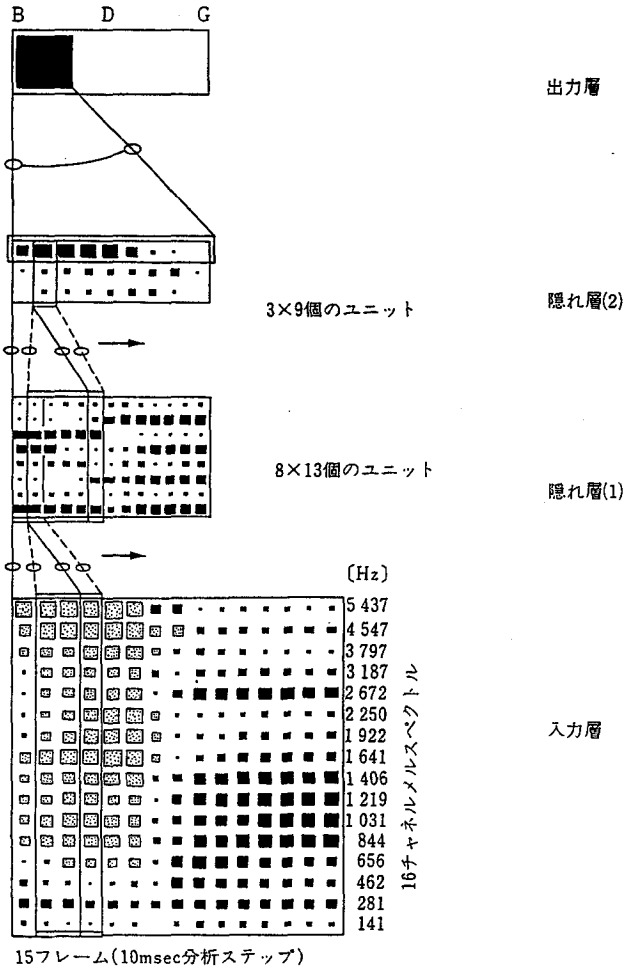


図 21: TDNN (鹿野他 [2])

Tolarant LVQ では，図 22 で示すように，音声スペクトルを遅延素子で空間展開し，それを 2 層のネットワークで認識する．この第一層目はベクトル量子化[9] を行うネットワークになっており，遅延素子で空間展開された音声スペクトルパターンを分割する役目をする．第二層目では，第一層目で分割されたパターンを各カテゴリー毎に束ね，どのカテゴリーに属する参照ベクトルが入力パターンに最も近いかを判定する．学習では，このベクトル量子化器の参照ベクトルを求めている．計算機シミュレーションでは，/b,d,g/の識別において，TDNN を上回る識別率を示した．

[9]ベクトル量子化は，大量のデータを小数のコードブック (参照ベクトル) によって，表現する，情報圧縮の一手法である．具体的には，一つの参照ベクトルで，そのベクトルに近い複数のパターンを表現することによって小数の参照ベクトルで大量のデータを記述する．

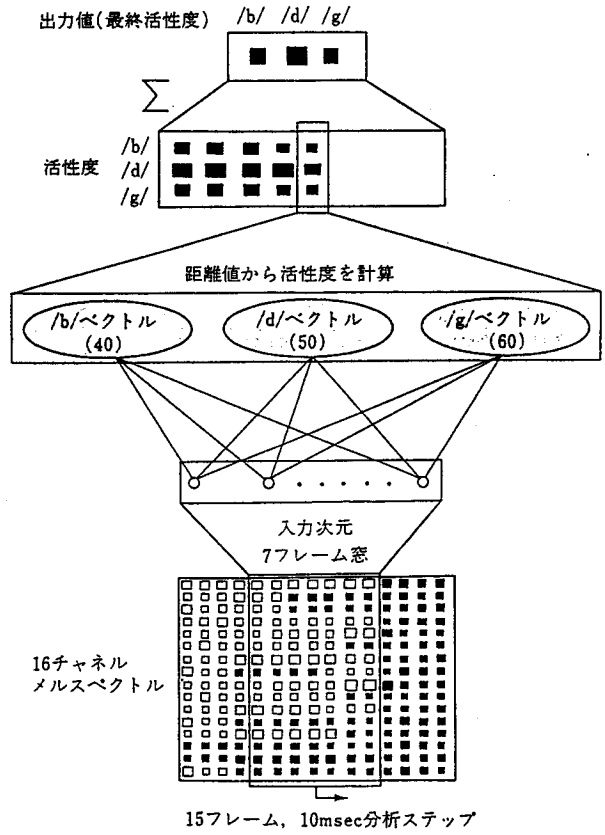


図 22: Shift-Tolarant-LVQ の構造 (鹿野他 [2])

1.5 本論文の主張

1.4節で概観した，時系列パターン認識の神経回路モデルを含めた音声認識システムのうち，動的計画法および状態遷移を基礎に置くシステムは，比較的少ない計算量で，時間軸の非線形伸縮に対処できる．従ってこれらのシステムは，音声認識装置のアルゴリズムとして多用されている．しかし，これらの多くは音声信号中の各特徴の継続時間長を無視することによって先の能力を実現している．

このことによって例えば，モールス信号の認識においては，トーン信号の相対的な時間長の中に情報が含まれるが，これらのシステムでは，時間長を捕らえることが難しく，識別が困難であるという問題も生ずる．これに対して，人間の場合には，このモールス信号が時間伸縮を受けたものであっても正しく識別できる．

1.3節で触れたように，単語知覚に関する諸現象から類推される人間の音声処理方式は，時間軸に沿った左から右への逐次的な処理とは考え難く [36]，単語知覚モデル TRACE [25]

のように音声信号を時空間パターンに展開しながら処理する形式となる．実際に，このような処理形式を採れば，時間長の情報を検出しながら認識することが出来るため，先のモールス信号であっても識別できよう．また，音響パラメータ時系列を専門家が視察することにより，連続音声を音素単位に分割する，いわゆるラベリングを行なう際に，専門家らが分割の手がかりにしている知識を音声データベース化し，自動ラベリングを実現する試みがある [57][58]．このラベリングを行なう専門家は，音素中のイベントに関する継続時間長の情報も一つの手がかりにして音素を識別することが知られている [58][45]．すなわち，音声信号の識別には，各特徴の継続時間長が重要な手がかりになっている可能性が高い．

しかし，逆に音声信号を時空間パターンに展開しながら処理する形式を採ると，時間軸方向の伸縮にどのように対処するかが問題となる．TDNN [54]，Shift-Toralant-LVQ [55] は，いずれも音声スペクトルを時空間パターンに展開して処理する形式を採っている．これらのシステムは，音韻認識において HMM よりも高い認識率を示した．しかし両システム共に，時間伸縮や変形スペクトルに対処するための汎化能力を得るために，変形パターンを含めた多くのサンプルパターンを学習する必要がある．従って，これらのモデルでは，学習パターンで教えられた以外の時間伸縮には対処し難く，単語などの激しい時間伸縮を受ける時系列パターンの識別には必ずしも向いていないと思われる．

以上のような観点から，従来の音声認識システムを人間の音声処理の機能モデルとして見た場合には，まだ不十分な点が多くあると言えよう．人間の柔軟な音声処理をさらに良く説明する機能モデルとしては，次の 2 点を同時に満足する必要があろう．

1. 音声信号を，その時間構造を破壊することなく処理し，音声信号中の各特徴の継続時間長に含まれる情報を検出できる．

2. 時間伸縮に対処できる．

第2章　時系列パターン認識モデル

2.1　序言

前章で指摘したように，人間は音声信号をその時間構造を保ったまま処理し，認識している可能性が強い．しかも，話すスピード (時間軸の伸縮) に対して，ほとんど影響を受けることが無い．このことは，我々がモールス信号を，その時間伸縮の影響を受けずに正しく認識できることからも容易に推測できよう．従って，人間の音声処理方式をうまく説明できる時系列パターン認識モデルとは，時系列パターンの時間構造を保ったまま処理する方式を採りながら，その時間軸方向の伸縮に対処できるモデルと考えられる．しかし，時間構造を保つことと，時間軸方向の伸縮を許容することは一見相反する要素である．

それでは，人間は如何にしてこの相反する要素を両立させているのだろうか．1.3.1節で述べたように，これまでに人間の音韻単語知覚の手がかりを発見するための多くの心理学的な調査が行なわれてきた．その中で，音節/wa/と/ba/の識別の手がかりについては，先の両立方法の一端をかいま見ることができる．これを再度述べると，この音節/wa/と/ba/の識別には，そのフォルマントの遷移速度が重要な手がかりになり，/ba/の方が/wa/よりも速く遷移する [27]．そして，この/ba/および/wa/のフォルマント周波数の遷移速度を人為的に一定にした場合，音節全体の時間長を短くした場合には，/wa/に知覚される割合が増し，逆に，時間長を長くすると/ba/と知覚される割合が増す (図11参照)．

Miller [27] はこの現象を，音節の時間長から発話速度が知覚され，それを手がかりに音声信号の時間長を正規化することによって生ずると解釈した．すなわち，人間は音声パターンの時間軸を正規化するという処理を施すことによって，その入力パターンの時間構造を破壊せず且つ時間伸縮に対処している可能性が高い．

本章では，このような知見に基づき，時系列パターンのスピードを検出し，そのスピードによって入力パターンを正規化しながら認識する時系列パターン認識の機能モデルを提案する [59][60][61]．ただし，現時点では，聴覚系の構造についての解剖学的知見はその末梢系を除いて乏しく，その構造を忠実に再現することは困難である．従ってここでは，その機能モデルを工学的に構成している．

しかし2.7節では，解剖学的な知見から，生体の中にもこのモデルの構造の一部に似た部

分が存在することも述べる．

2.2 提案するモデルの概要

このモデルの基本構造を図 23 に示す．このモデルは，一つのメインブロックと，二つ

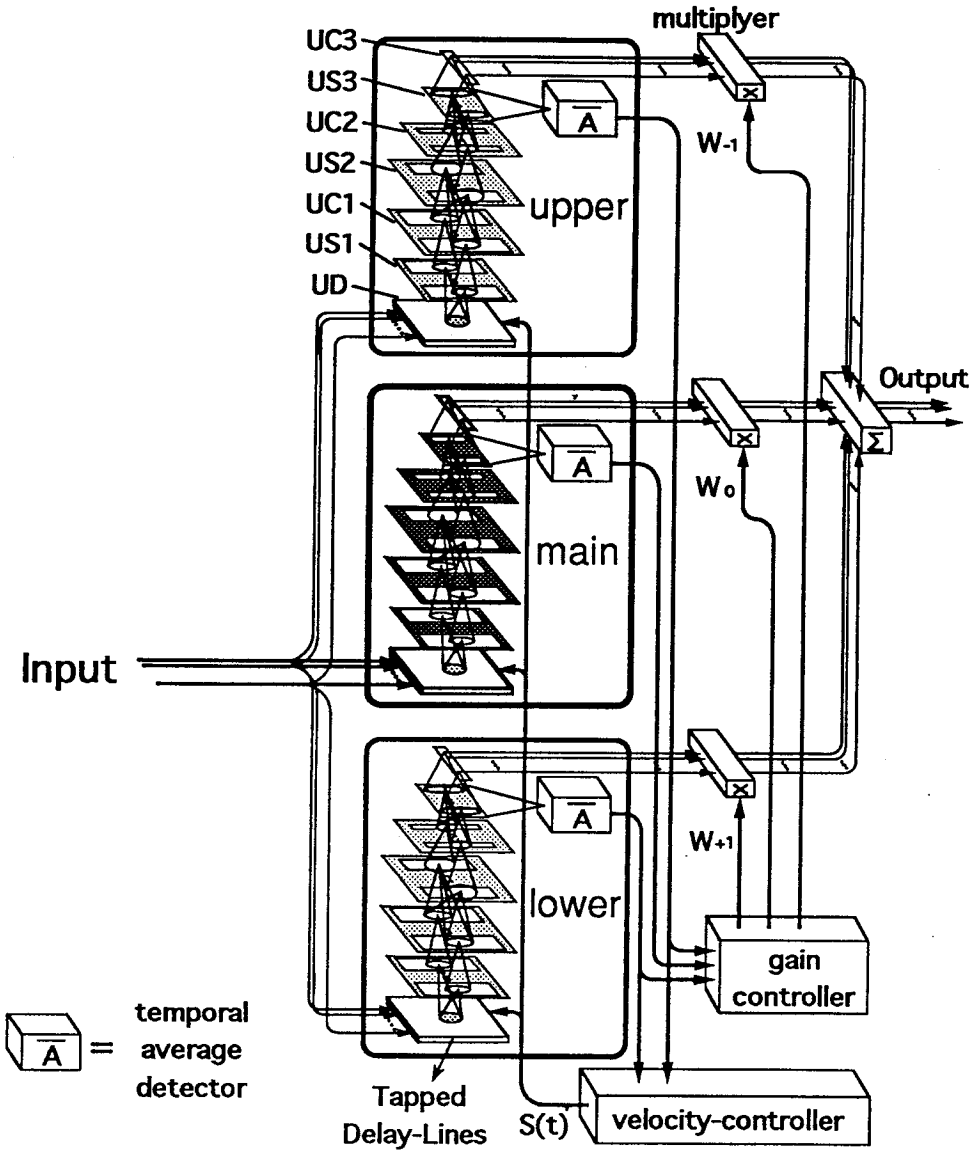


図 23: 提案する時系列パターン認識モデルの構造

のサブブロックから構成される．メインブロックは，時系列パターンを認識する役目をし，二つのサブブロックは，時系列パターンのスピード (伸縮率) を検出する役目をする．各ブロックは，速度可変のタップ付き Delay-Line を入力層として持つネオコグニトロン状の構

造[53] [51][52] を持っている[10]．ネオコグニトロンは，もともと，視覚パターン認識を行なう階層型神経回路モデルとして提案されたものである．ここでは，このモデルを，Delay-Line によって展開された時空間パターンの認識に使用する．この三つのブロックはいずれも同じ構造をしているが，Delay-Line のパターン伝搬スピードの比が異なっている．すなわち，上下のサブブロック及び中央のメインブロックの Delay-Line のスピードは，それぞれ速，遅，中に設定されている．

時系列パターンの認識に先だって，標準時系列パターンを学習させておく．この学習は，主にメインブロックで行なわれ，上下のサブブロックの入力結合の強度は，メインブロックと同じになるように更新されるとする．

これら三つのブロックの Delay-Line のスピードは，可変だが，そのスピードの比は常に一定に保たれている．この Delay-Line のスピードは入力パターンの伸縮率に合わせて，図 23 の *Velocity-Controller* が三つのブロック共に連動して制御する．もし，入力されたパターンが，縮んだパターンである場合には，Delay-Line のスピードを上昇させる．逆に入力されたパターンが伸びたパターンである場合には，Delay-Line のスピードを下降させる．すなわち，常にメインブロックからの出力が最大になるように Delay-Line の速度を制御する．

入力パターンの伸縮率は，上下のサブブロックの平均活性値 (図 23の$\bar{A}$) の大小を比較することによって検出される．例えば，縮んだパターンが入力された場合には，上側のサブブロックが，下側のサブブロックよりも大きな出力を出す．そして，Delay-Line の速度の制御は，この両側のサブブロックの出力が同じになるように行なわれる．こうすることにより，常にメインブロックからの出力が最大になるように，Delay-Line の速度を制御することになる．

システムの出力は，主にメインブロックから出力される．しかし，実際には，Delay-Line の速度が最適値に達していない場合には，メインブロックからの出力が得られない場合がある．そこで，システムの出力として，3 個のブロックの出力の重み付き平均を採用している。各ブロックの重みは，そのブロックの平均活性値の大小に合わせて制御される．すなわち，平均活性値が大きなブロックは重みも大きくなるが，平均活性値が小さなブロッ

[10]ただし，伊藤，福島のモデル [51][52] に含まれる 1 次遅れを持つ素子は，使わない．

クは重みも小さくなる．この重みの制御は図 23 の *Gain-Controller* によって行なわれる．

次節から，このモデルを数式表現によって述べる．なお，このモデルで使用する学習法は，従来のネオコグニトロン型の学習法を時系列パターン学習用に改良したものである．この改良型学習法の意義および説明の詳細については，次の第 3 章に委ね，本章では 2.6節でその概要だけを述べるにとどめる．

2.3 各認識ブロックの構造

各ブロックは，図 24 に示すように，delay-line の層 (U_D層と呼ぶ) の上に，U_S 層と呼ぶ細胞層と U_C 層と呼ぶ細胞層が交互に並んだネオコグニトロン状の構造 [53] を持つ．

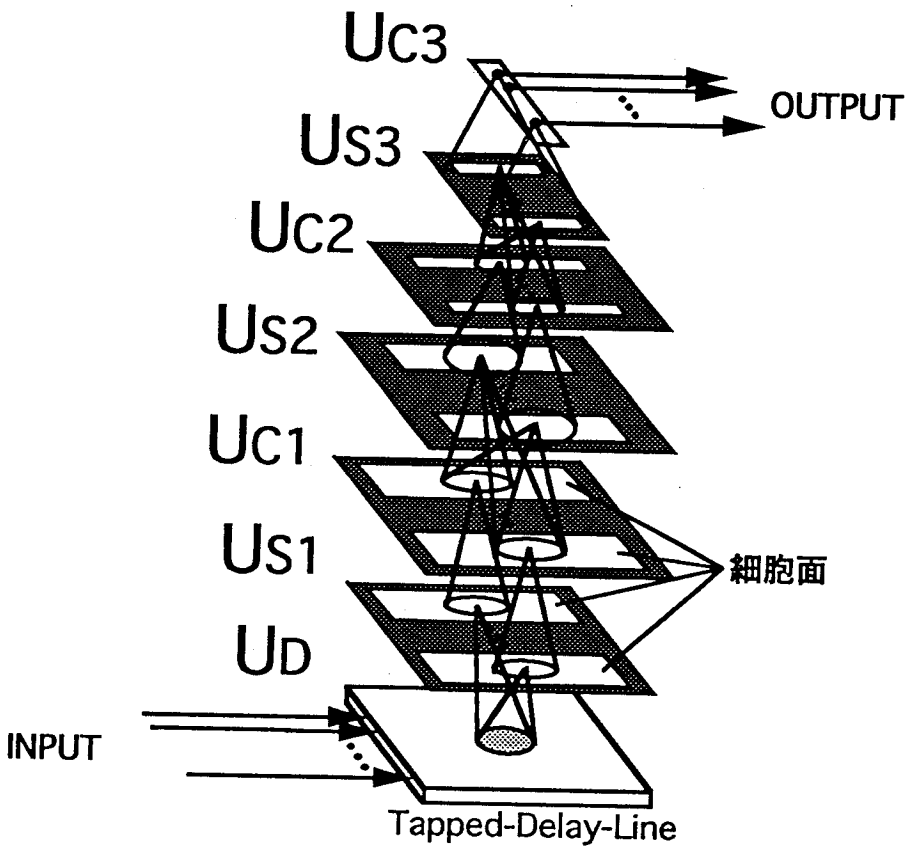


図 24: 1 ブロックの構造

入力された時系列パターンはまず，delay-line の層 (U_D層) に入力される．U_D層は，Delay-Line が複数並んだ構造をしている．この Delay-Line は，多次元の時系列パターンを空間パターンに展開する役目を果たす．

図 24 の中央のブロック (メインブロック) から数えてξ 番目 $(-1 \leq \xi \leq 1)$ のブロックの, n_x 番目の Delay-Line 上の n_y 番目の遅延素子の, 時刻 t における出力 $u_D^{\xi}(n_x, n_y, t)$ は, 次式で与えられる.

$$u_D^{\xi}(n_x, n_y, t) = u_D^{\xi}\left(n_x,\ n_y - 1,\ t - \frac{k^{\xi} \cdot \tau_0}{S(t)}\right) \quad (\xi = -1, 0, 1) \tag{1}$$

ここに, $\tau_0(< 1)$ Delay-Line 上をのパターンの伝搬する速度を決定する定数である. k は定数で, 後に説明する計算機シミュレーションでは $k = 1.5$ としている. つまりメインブロック $(\xi = 0)$ の Delay-Line は上側のサブブロック $(\xi = -1)$ の Delay-Line よりも速く, 下側のサブブロック $(\xi = +1)$ の Delay-Line はメインブロックの Delay-Line よりも速い. $S(t)$ は, Delay-Line のスピードの制御を行うための係数 (以後, 速度制御係数と呼ぶ) で, 標準スピードからの倍率を表わし, 全てのネットワークについて共通である. 但し, $S(t)$ の変化は, 遅延素子の遅延時間よりも十分に遅いと仮定している. また, $S(t)$ の初期値は 1 とする. すなわち, 認識を開始する時刻を $t = 0$ とすれば, $S(0) = 1$ である.

Delay-Line を構成する各遅延素子の出力は, そのすぐ上の U_S層に送られる. U_S層は, 複数の平面状に並べられた S 細胞と呼ばれる特徴抽出細胞の集団 (細胞面と呼ぶ) で構成される (図 25).

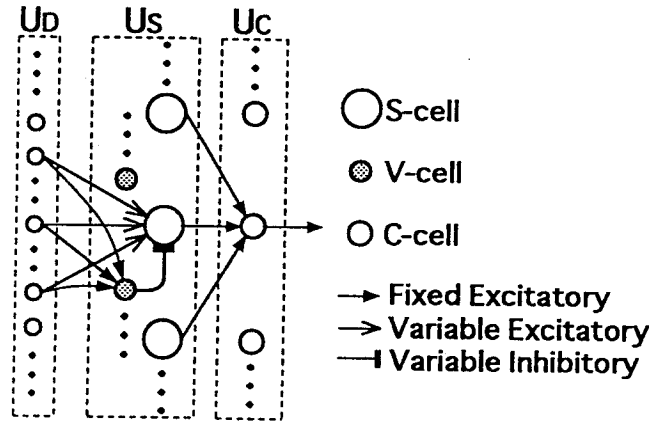


図 25: 異なる層間の結合

同一細胞面上の各 S 細胞は, それぞれその 1 つ前の層の異なった位置の小領域の細胞から, 同一の空間分布の興奮性結合を介して信号を受け取ると共に, その S 細胞と同じ位置にある V 細胞と呼ばれる細胞からの, 抑制性の信号を受け取っている. V 細胞は, 対応す

るS細胞と同じ領域からの信号を興奮性結合を介して受け取り，その前層の結合領域にある細胞の出力の平均値を出力する．S細胞は，この前層からの興奮性信号と，V細胞からの抑制性信号の両方を受け取ることによって，前層の結合領域中の細胞の特定の発火パターンに選択的に反応する能力を持つ．すなわち，S細胞は，前層の局所的な特徴抽出を行なう．U_S層の出力はそのすぐ上のU_C層に送られる．U_C層は，C細胞と呼ばれる細胞が並んだ細胞面で構成される．同一細胞面上の各C細胞は，そのすぐ下のU_S層の特定の細胞面上のそれぞれ異なった位置における小領域のS細胞から，興奮性固定結合を介して信号を受け取っている．したがって各C細胞は，出力を出しているS細胞の位置が空間的に少しずれていても反応を出し続ける性質を持つ．つまり，U_S層の出力の空間的な位置ずれを少し許容する役目をする．このU_S層とU_C層の組が複数段重なることにより，最終段のU_S層では，最初の層 (U_D層) 上の空間パターンの位置ずれや変形にあまり影響されずに正しく特徴を抽出できる．また，最終段のU_C層のC細胞がネットワークの最終的な認識出力を出す認識細胞であり，この層の各細胞面に存在する細胞数は1個である．後述する計算機シミュレーションでは，簡単のため各ネットワークは，U_S層とU_C層の組を2段だけ持つとした．

数学的には，S細胞，C細胞，V細胞の出力はそれぞれ以下のように表される．以降では，$\boldsymbol{n}$を，その細胞の結合領域の中心の座標を表すものとし，kを細胞面のインデックス，tを時間とする．例えば第ξブロックのl段目に位置するS細胞の出力は，$u^{\xi}_{Sl}(\boldsymbol{n},k,t)$で表され，次式で表現される．

$$u^{\xi}_{Sl}(\boldsymbol{n},k,t) = r^{\xi}_{l}(k)\cdot\varphi\left[\frac{1+\sum_{\kappa=1}^{K_{Cl-1}}\sum_{\boldsymbol{\nu}\in A_l} a_l(\boldsymbol{\nu},\kappa,k)\cdot u^{\xi}_{Cl-1}(\boldsymbol{n}+\boldsymbol{\nu},\kappa,t)}{1+\dfrac{r^{\xi}_{l}(k)}{1+r^{\xi}_{l}(k)}\cdot b_l(k)\cdot u^{\xi}_{Vl}(\boldsymbol{n},k,t)}-1\right], \tag{2}$$

ここに$\varphi[\,]$は半波整流特性を示す関数で$\varphi[x]=\max(x,0)$で定義される．

式(2)の$a_l(\boldsymbol{\nu},\kappa,k)(\geq 0)$は，前層 ($U_{Cl-1}$層) の位置$\boldsymbol{n}+\boldsymbol{\nu}$のC細胞からつながるS細胞の興奮性可変入力結合の強度を表している．ここに$\boldsymbol{\nu}$は，そのS細胞の結合領域の中心座標からの相対位置を表す．A_lは，U_{Sl}層のS細胞の結合領域を表している．K_{Cl-1}は，U_{Cl-1}層の細胞面の数を表す．

この興奮性可変結合の強度 $a_l(\boldsymbol{\nu},\kappa,k)(\geq 0)$ は，次式で表すことができる．

$$a_l(\boldsymbol{\nu},\kappa,k) = \{c_l(\boldsymbol{\nu},\kappa,k)\}^2 \cdot p_l(\boldsymbol{\nu},\kappa,k), \tag{3}$$

ここに $p_l(\boldsymbol{\nu},\kappa,k)(\geq 0)$, $c_l(\boldsymbol{\nu},\kappa,k)(\geq 0)$ 可変パラメータである．$p_l(\boldsymbol{\nu},\kappa,k)$ は，S 細胞の興奮性可変結合の強度を決定する．$c_l(\boldsymbol{\nu},\kappa,k)$ は，S 細胞の結合領域の各場所からの信号に対する感度を表すパラメータであり，この変数も後述する学習手続きによって変化する．なお，興奮性入力結合の強度 $a_l(\boldsymbol{\nu},\kappa,k)$ は，第 k 番目の細胞面上の全ての S 細胞が共有する．従って，$a_l(\boldsymbol{\nu},\kappa,k)$ には，S 細胞の位置を表す記号$\boldsymbol{n}$ は，含まれない．また，この入力結合の強度については，全てのブロックについて同じであるとしているので，ブロックを表す記号ξも省略されている．

$b_l(k)(>0)$ は，V 細胞から，その S 細胞への抑制性可変結合の強度を表している．その値は，S 細胞の興奮性入力結合の強度に応じて次のような値をとる．すなわち，

$$b_l(k) = \sqrt{\sum_{\kappa=1}^{K_{Cl-1}} \sum_{\boldsymbol{\nu}\in A_l} \{c_l(\boldsymbol{\nu},\kappa,k)\}^2 \cdot \{p_l(\boldsymbol{\nu},\kappa,k)\}^2}. \tag{4}$$

式 (4) 中，$c_l(\boldsymbol{\nu},\kappa,k)$, $p_l(\boldsymbol{\nu},\kappa,k)$ は，式 (3) のものと同じである．

$r_l^{\xi}(k)$ は，S 細胞のパターン選択性を決定する正のパラメータである．もし，このパラメータ $r_l^{\xi}(k)$ が大きければ，その S 細胞のパターン選択性は高まり，逆に小さければ，パターン選択性は低くなる．

このパラメータ $r_l^{\xi}(k)$ は後述の第 3 章で示すように，学習期間中には，$b_l(k)$ の強度に応じて変化するが，学習が終了した時点で固定される．

ちなみに，この S 細胞のパターン選択性が低くなると，その S 細胞が出力を出すパターンのスピード (伸縮率) の範囲が広くなる．この性質を利用し，上下の二つのサブブロックのパラメータ $r_l^{\xi}(k)$ を，真中のメインブロックよりも小さく設定して，上下二つのサブブロックがメインブロックよりもブロードなスピード選択性を持つようにしている．すなわち，

$$r_l^{-1}(k) = r_l^{+1}(k) = \rho_l \cdot r_l^{0}(k). \tag{5}$$

ここにρ_l は，定数で $(0 < \rho_l \leq 1)$ である．これによって，全てのブロックが同一のダイナミックレンジを持つ場合に比べて，広いスピードの変化に対処することができる (図 26 参照)．

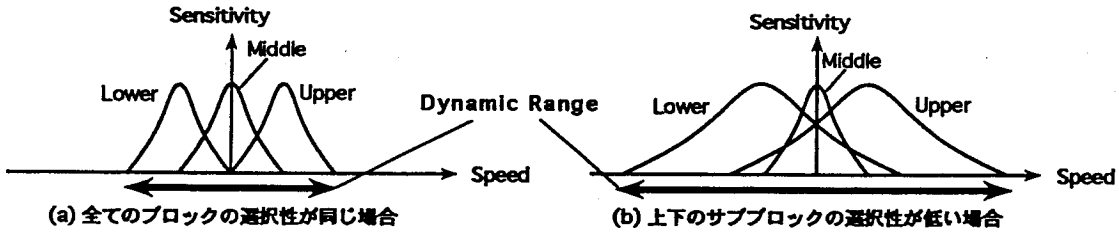


図 26: 異なるダイナミックレンジを持つブロックの効果

$u^{\xi}_{Vl}(\boldsymbol{n},k,t)$ は，第 k 細胞面上にある V 細胞の時刻 t における出力を表す．V 細胞は，S 細胞と同じ細胞群から信号を受け取り，その細胞群の出力の 2 乗平均の値に等しい出力を出す．すなわち，

$$u^{\xi}_{Vl}(\boldsymbol{n},k,t)=\sqrt{\sum_{\kappa=1}^{K_{Cl-1}}\sum_{\boldsymbol{\nu}\in A_l}\{c_l(\boldsymbol{\nu},\kappa,k)\}^2\cdot\{u^{\xi}_{Cl-1}(\boldsymbol{n}+\boldsymbol{\nu},\kappa,t)\}^2}\ , \tag{6}$$

で表せる．この式の $c^t(\nu,k,m)$ は，式 (3) で使われたものと同じもので，遅延素子から V 細胞への時刻 t における興奮性可変入力結合の強度でもある．ここに $u^{\xi}_{Cl-1}(\boldsymbol{n}+\boldsymbol{\nu},\kappa,t)$ は，U_{Cl-1}層の第 k 番目の細胞面の C 細胞の出力を表しており，次式で表される．

$$u^{\xi}_{Cl}(\boldsymbol{n},k,t)=\psi\left[\sum_{\boldsymbol{\nu}\in A_l}d_l(\boldsymbol{\nu})\cdot u^{\xi}_{Sl}(\boldsymbol{n}+\boldsymbol{\nu},k,t)\right]. \tag{7}$$

$\psi[\ ]$ は，C 細胞の出力関数で，

$$\psi[x]=\frac{\varphi[x]}{1+\varphi[x].} \tag{8}$$

である．ここに $d_l(\boldsymbol{\nu})$ は，C 細胞の興奮性固定結合の強度を表しており，$|\boldsymbol{\nu}|$ が増化するにつれて減少する．後述する計算機シミュレーションでは，$d_l(\boldsymbol{\nu})$ を，その結合領域の範囲内 $(\boldsymbol{\nu}\in A_l)$ では，2 次元のガウス関数となるように定めた．

式 (7) は，$u^{\xi}_{Cl-1}(\boldsymbol{n}+\boldsymbol{\nu},\kappa,t)$ を $u^{\xi}_{D}(\boldsymbol{n}+\boldsymbol{\nu},t)$ で置き換え，K_{Cl-1} を 1 で置き換えれば，$(l=1)$ でも成立する．

最も上位に位置する C 細胞は，ブロックの認識結果を出力する．この C 細胞の出力を $u^{\xi}_{C2}(k,t)$ で表すことにする．この最上位の C 細胞の数は，一細胞面当たり 1 個であるため，結合領域の中心位置を表すパラメータ $\boldsymbol{n}$ は，省略されている．これらの C 細胞の出力値は，入力パターンのスピード (伸縮率) に応じて，変化する．すなわち，このパターンのスピードが標準パターンともっとも近くなった時にこれらの C 細胞は最大の出力値を出す．

2.4 速度制御

速度制御部は，三つのブロックの Delay-Line の信号伝搬速度の比を保ったまま全ての Delay-Line の速度を連動制御する．この速度制御部は，各ブロックの平均活性値 (図 23の $\bar{A}$) を比較し，真中のメインブロックの平均活性値を最大にするためのスピードの変化量を決定する．この平均活性値は，各ブロックの *temporal average detector* が算出する．

実際には，この平均活性値$\bar{A}$は，最上位段の U_C層の出力から計算されるのではなく，最上位段の U_S層の S 細胞の出力から計算される．最上位段の S 細胞の出力値は，入力パターンのスピードに依存するのに対して，C 細胞の出力値は，S 細胞の出力値のみならず，出力を出した S 細胞の数にも依存する．そこで，入力パターンのスピードに関する情報が正確に反映されている最上位段の U_S層の出力から，平均活性値を算出している．この入力パターンのスピードに関する情報をより正確に抽出するため，*temporal average detector* は，最上位段の U_S 層から，最大出力を出す S 細胞を選択し，その出力値を使って，平均活性値を計算する．

この *temporal average detector* は，漏洩のある積分器 (leaky integrator) によって構成されている．いま，$\bar{A}_\xi(t)$ を第ξ番目のブロックの時刻 t における *temporal average detector* の出力とする．この出力$\bar{A}_\xi(t)$ は，次の式によって変化する．

$$\frac{\tau}{S(t)} \cdot \frac{d\bar{A}_\xi(t)}{dt} = \max_{k,\boldsymbol{n}} \left[u^{\xi}_{SL}(k, \boldsymbol{n}, t) \right] - \bar{A}_\xi(t), \tag{9}$$

ただし，L は最上位段を表し，後述する計算機シミュレーションでは，$L = 2$ としている．ここに$\tau/S(t)$ は，積分器の時定数である．この時定数は，Delay-Line の伝搬スピードに応じて変化する．そして，正定数τ は，時定数$\tau/S(t)$ が常に各時系列パターン (例えば単語など) の時間長の 2〜3 倍程度となるように設定されている．

後述するように，もし，時系列パターンが速ければ，velocity-controller は，$S(t)$ の値を大きくし，遅ければ小さくする．従って，時系列パターンのスピードが途中で変化したとしても，時定数$\tau/S(t)$ は各時系列パターンの 2〜3 倍程度に保たれる．

velocity-controller は，$S(t)$ の値を次の式 (10) で表される$\Delta S(t)$ だけ変化させる．

$$\Delta S(t) = \alpha \cdot S(t) \cdot \left\{ \frac{\bar{A}_{-1}(t) - \bar{A}_{+1}(t)}{\bar{A}_{-1}(t) + \bar{A}_{+1}(t) + \varepsilon} \right\} \cdot T(t), \tag{10}$$

ここにα $(0 < \alpha < 1)$ は，$S(t)$ の変化スピードを決定する定数である．

関数$T(t) = 1$ *or* 0 は，不応期間をもつ細胞 (タイミング細胞と呼ぶ) の出力を表現している．このタイミング細胞は，最上位段のS細胞の少なくとも1個でも出力を出している間は，連続的にトリガーされる．しかし，一度$T(t)$ が1に活性化されると，最上位段のS細胞が出力を出していたとしても，しばらくの間0となる．すなわち，この細胞は，定期的に出力を出す．velocity-controller は，このタイミング細胞が発火した時だけ働くようになっており，不連続な制御を行なう．これは，このシステムがDelay-Lineを含んでいるために，フィードバックループに時間遅れが含まれるからである．すなわち，不連続な制御を行なうことによってこの時間遅れを相殺するようにしている．

$S(t)$ は基本的に，式(10) の分子$\bar{A}_{-1}(t) - \bar{A}_{+1}(t)$ に比例した値だけ変化する．この値は，$\bar{A}_0$ を最大にする値 $(= S_d)$ と，実際の値 $S(t)$ との差に比例した値となる (図27の点線).

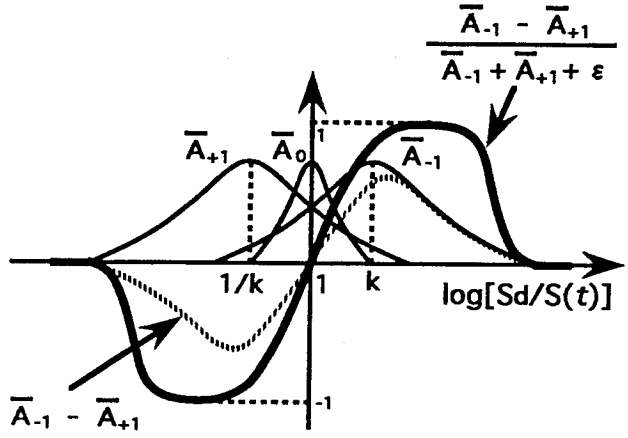


図 27: $\bar{A}_\xi(t)$ と $S_d(t)/S(t)$ との関係

しかし，もし$S(t)$ の変化が，$\bar{A}_{-1}(t) - \bar{A}_{+1}(t)$ だけに依存して決められるとすると，図27の点線に示されるように，$S_d > k \cdot S(t)$ あるいは $S_d < S(t)/k$ となった時に $|\Delta S(t)|$ の値が小さ過ぎて，速やかな制御を実現することができない．そこで，式(10) で示されるように，$\{\bar{A}_{-1}(t) - \bar{A}_{+1}(t)\}/\{\bar{A}_{-1}(t) + \bar{A}_{+1}(t) + \varepsilon\}$ を$\bar{A}_{-1}(t) - \bar{A}_{+1}(t)$ の代わりに使用した．ここに，$\varepsilon \ll 1$ である．この値は，図27の実曲線に示されるように，$S_d > k \cdot S(t)$ あるいは $S_d < S(t)/k$ の時には値 1.0 または -1.0 に飽和する．

ただし，$\Delta S(t) < 0$ の時に，$|\Delta S(t)|$ の値が大きくなり過ぎて，$S(t)$ が負にならないよ

うにするため$\Delta S(t)$ は，$S(t)$ に比例するようになっている．

2.5 システムの出力

$u(k,t)$ をシステムの最終出力とすると，

$$u(k,t) = \sum_{\xi} W^{\xi}(t) \cdot u^{\xi}_{CL}(k,t), \tag{11}$$

である．

第ξ番目のブロックの重み $W^{\xi}(t)$ は，*gain-controller* によって次式のように制御される．

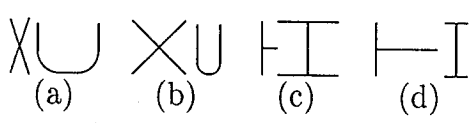
$$W^{\xi}(t) = \frac{\bar{A}_{\xi}(t)}{\sum_{\xi=-1}^{+1} \bar{A}_{\xi}(t).} \tag{12}$$

2.6 計算機シミュレーション

以上述べてきた時系列パターン認識モデルの動作を確認するため，計算機シミュレーションを行なった．

ここでは，このモデルの基本動作を確認するために，図 23 に示したモデルを少し簡略化してシミュレーションを行なっている．まず，式 (5) のρ_l の値は 1 とし，三つのブロック全てが同じダイナミックレンジを持つとしている．各ブロックの U_S層および U_C層の組は 2 段とし，各層のサイズを表 1 のように選んだ．また，システムの最終出力としては，三つのブロックの重みつき平均を計算する代わりに，平均活性値が最大となるブロックの最上位段 (U_{C2}層) の出力を採用するようにしている．

認識対象となるパターンとして，後の第 4 章では音声スペクトルを扱っている．しかし本節では，このモデルが各特徴の継続時間の比を検出し識別する能力を持つことを明確に示すために，以下に示すような電光掲示版に流れるような文字パターンを使用する．このパターンは，音声スペクトルに見立てた多次元の時系列パターンで，下に示すような 4 つのパターンセットを用意している．



Layer	Number of Cell-Planes	Size of Cell-Plane (f-axis) × (τ-axis)	Sizes of Connecting Regions of Cells (f-axis) × (τ-axis)
U_D	1	15 × 25	———
U_{S1}	10	15 × 25	5 × 5
U_{C1}	10	17 × 27	5 × 5
U_{S2}	10	3 × 9	15 × 18
U_{C2}	10	1 × 1	3 × 9

表 1: 各ブロックのサイズ．f-axis は，空間方向を表し，τ-axis は，時間方向を表している．

この 4 パターン (a)(b)(c)(d) は，それぞれ 90° 回転させた 2 種類の文字によって構成されている．この 4 パターンの各文字の長さの比はそれぞれ異なっている．

認識実験に先だって，速度制御を行なわせずに，図 23 の中央にあるメインブロックの Delay-Line に先の 4 パターン (a)(b)(c)(d) を標準スピードで提示しながら，学習させた．この時系列パターンの学習には，従来型のネオコグニトロン型学習法 [62] を，時系列パターン学習用に改良したもの [63][64] を使用した．従来型の学習法で時系列パターンを学習させると，同一の特徴であっても時間毎に少しずつシフトしたパターンを別の特徴とみなして学習するため，膨大な数の S 細胞が必要になるという問題が生ずる．改良型学習法では，このような問題を解決するため，時系列パターン中の重要な特徴のみを記憶するようにして，冗長な特徴を学習しないようにする工夫がなされている．この重要な特徴の発見は自動的に行なわれる．実際に学習させる際には，記憶させたい時系列パターンをランダムな順序で繰り返し提示するだけでよい．例えば，上のパターンを学習させる場合には，図 28 で示すように (a)(b)(c)(d) をランダムな順序で並べたものを提示する．計算機シミュレーションでは，(a)(b)(c)(d) をランダムに 300 セット並べたパターンを一通り提示した．この学習手法の詳細については，次の第 3 章で詳細する[11]．

学習終了後のメインブロックの学習パターンに対する各層の反応例を図 29 に示す．図から，U_{S1}, U_{C1}層には線分の各傾きに選択的に反応する細胞，U_{S2}, U_{C2}層は上の 4 つのパター

[11]ただし，本節のシミュレーションでの学習法は，第 3 章で後述する学習法で行なわれているようなパラメータ $c_l(\nu, \kappa, k)$ の更新については省略している．また後述するパラメータ σ_{rl} は学習期間中一定値に固定している．

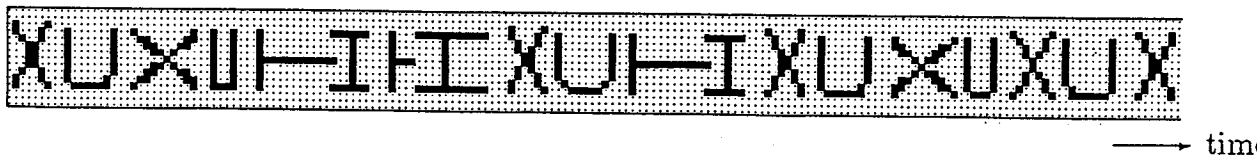


図 28: 学習パターンの一例

ンそれぞれに反応する細胞ができあがっていることが分かる．すなわち本モデルは，学習途中に，線分や個々のカテゴリーに対応するパターンなど重要な情報を自動的に発見し，記憶したことを意味している．

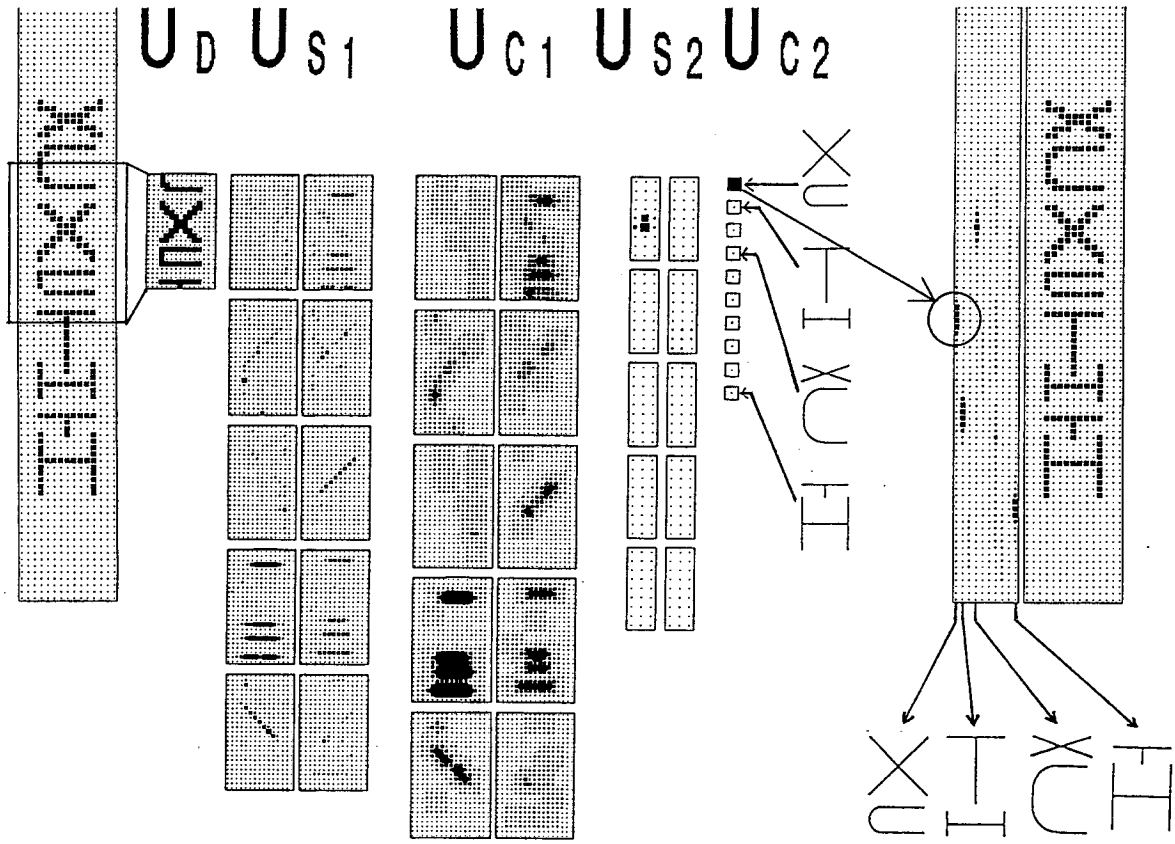


図 29: メインブロックの学習パターンに対する反応例

学習終了後，メインブロックの各 S 細胞の入力結合の強度は，上下のサブブロックにコピーした．そして，上の 4 種類の時系列パターンが時間方向に伸縮し，且つ空間方向に変形したようなパターンをシステムの入力に提示し，速度制御を行なわせながら認識させた．図 30 は，その認識結果を示している．図は，メインブロックの Delay-Line のレスポンスと，システムの最終出力を示している．この図から，本システムが伸縮したパターンを徐々に標

準パターンに近付けながら認識していることが分かる[12]．また，各カテゴリー (a)(b)(c)(d) に対する最終出力は，図 29 とほぼ同じであることが分かる．

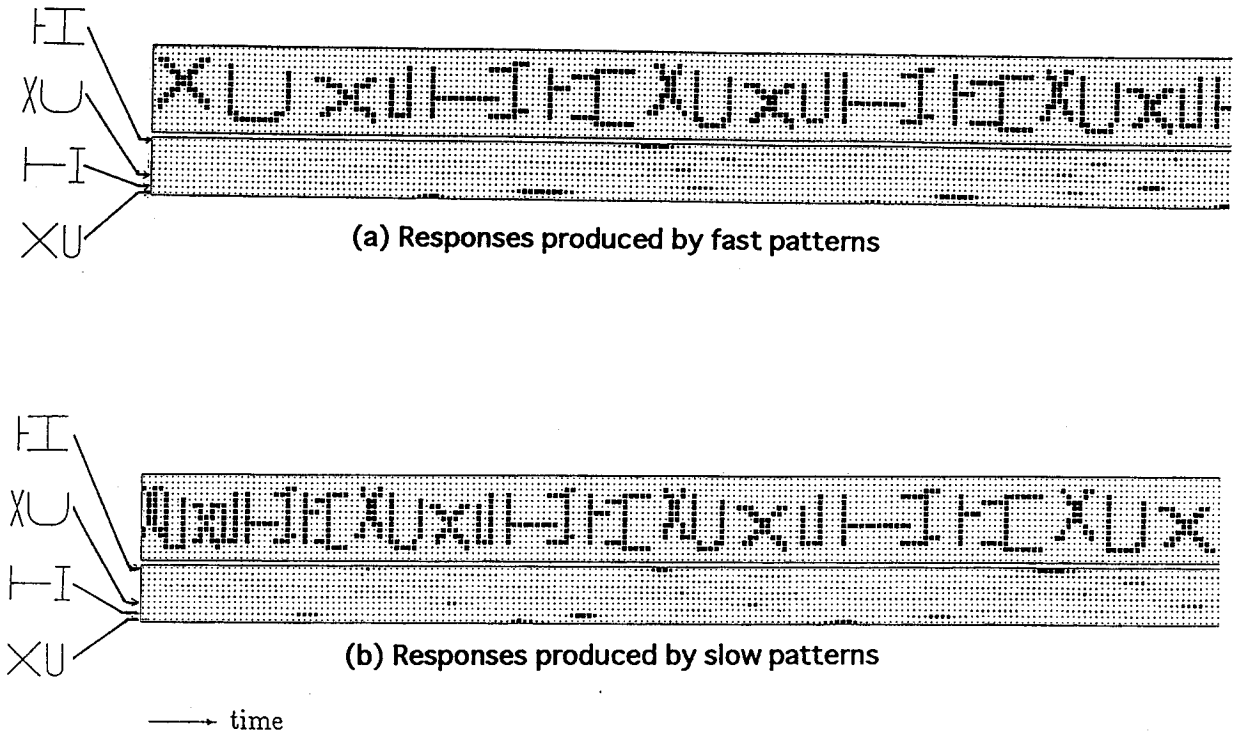


図 30: 本システムの伸縮・変形パターンに対する反応例

2.7 心理学および解剖学的知見との比較

ここで提案した時系列パターン認識モデルは，時系列パターンのスピードを検出し，そのスピードに合わせて Delay-Line のスピードを制御することによって，入力パターンの時間長を標準パターンに近付けながら認識するものである．従って，入力された時系列パターンの各特徴の継続時間長の比を破壊することなくそれを検出し，認識結果に反映することが可能である．このような手続きは，実際に人間の音節知覚にもその一端を見ることができる．例えば 1.3.1節で示したように，Miller ら [27] は，音節/wa/,/ba/識別境界がその音節の時間長によって変化することを発見している (図 11参照)．この知見は，人間が一つの音節全体の継続時間を手がかりに，音声信号の伸縮率を検出し，それを基に音声信号の時間長を正規化して認識していることを強く示唆するものである．

[12]図 30 は，実際には，入力パターンをその時間軸のスケールを Delay-Line のスピードに合わせて変えて表示している．この時間軸のスケールは値 $1/S(t)$ に比例するようにしている．

また，本モデルにおける速度制御のもう一つの意義として，少ない細胞数でも大きな伸縮を受けた音声信号の認識を可能にしている点が挙げられよう．このような機構は，不均一網膜 [65][13]と，眼球運動の関係に見られるような，少数の細胞を有効に利用するための工夫が生体にあることを考えれば不自然ではないと考えられる．

ところで，ここで提案したモデルの構造が，実際に生体にあるかどうかは，明らかではない．しかし，その一部の要素については，解剖学的にその存在を示唆する報告がいくつかある．

例えばDelay-Lineについては，フクロウの層状核において，その存在が確認されている [11] (1.2.3節参照)．このDelay-Lineは同一音が両耳から検出される時間差を検出する役目をし，音源定位に必要な情報を提供するためにあると考えられている．従って，生体が実際に，Delay-Lineを使って情報処理を行なっている可能性は高いと考えられる．

また，サルの大脳視覚皮質の17野では，見ようとしている物体が目の焦点からどれほど離れているか(奥行き)を検出している細胞群が発見されている [66]．これらの細胞は，焦点から離れたものに反応するもの，近いものに反応するもの，焦点にあったものに反応するものの3種類に分けることが出来る．これらの細胞は，両目の角度を調節して焦点を合わせる役目をしていると考えられている．従って，聴覚系においても，音声信号のスピードを検出するために，異なるスピードに選択性を持つ細胞が存在している可能性があると言えよう．同様に，ここで提案したモデルは，三つの異なるスピードに選択性を持つブロックで構成されている．この構造は，先にも述べたように，入力パターンのスピードを検出するのに有効である．

2.8 結言

時系列パターンを，その時間構造を保持したまま処理し，且つその時間伸縮に対応できる時系列パターン認識モデルを提案した．このシステムは，入力された時系列パターンの伸縮率を検出し，それに合わせてDelay-Lineのパターン伝搬速度を制御し，入力されたパターンの時間長を正規化しながら認識する．時系列パターンの伸縮率は，特定のスピードに選択性を持つブロックを二つ用意し，それらの出力の平均値の差から算出される．計算

[13]網膜の中心部分での視細胞の分布密度は高いが，周辺部分での分布密度は低い．

機シミュレーションでは，同じ文字セットで構成されていながら，その継続時間の比の異なるパターンセット (例：短い 'T' の後に長い 'H' が来るパターンと長い 'T' の後に短い 'H' が来るパターン) が時間軸方向に伸縮を受けていても正しく区別できることを示した．すなわちこの結果は，このモデルが，時系列パターンの各特徴の継続時間長の比の違いを，学習・弁別出来ることを示唆するものである．

このように，入力されたパターンの時間長を正規化しながら認識するという手続きは，Miller [27] らが発見した人間の音節知覚に付随する現象 (1.3.1節参照) にも裏付けられるように，実際に人間の音声処理の中でも行なわれている可能性が高い．また，本モデルの構造の一部と類似な構成が，生体の中にも見受けられる．

すなわち，ここで提案したモデルは，心理学的な知見や，解剖学的な知見から見て，生体の時系列パターン認識プロセスをうまく模倣していると考えられる．

第 3 章　時系列パターンの学習

3.1　序言

一般に，物事を覚えたり理解することを，学習と呼ぶ．この学習は，神経細胞同士をつなぐ入力結合の強度が，変化することによって実現すると考えられている．しかし，この学習が，どのようなメカニズムによって行なわれているかについては，生理学的にほとんど分かっていないのが現状である．従って，これまでに提案された神経回路の学習法のほとんどは，人工ニューラルネットワークにパターンを記憶させるための必要論的な立場から構築された．

これらの学習の目的は，一般に，次々に与えられる多数のパターンを有限個の神経細胞とその細胞間の結合強度の中に，何らかの形で集約して記憶させることにある．この集約方法如何によって，そのネットワークの記憶容量や汎化能力[14] が決定される．特に，時系列パターンの学習には，時々刻々と入力される膨大なパターンをどのように集約するかが，大きな問題となってくる．

本章では，これまでに既に提案されているいくつかの学習法を概観した後，これらの学習法を時系列パターン学習に適用した場合の問題点を指摘し，それを解決する改良型学習法を提案する [63][64].

3.2　従来型学習法

これまでに提案された学習方法は，大きく分けて二つある．一つは，ネットワークに，記憶させたいパターンとネットワークの望ましい出力 (そのパターンの属性とも言える) を共に与え，それらを基に記憶させるものである．これを教師あり学習と呼ぶ．もう一つは，記憶させたいパターンを単に提示するだけで，パターンの属性は一切与えずにネットワークに記憶させる方法である．これを教師なし学習と呼ぶ．本節では，この二つの学習法を概観する．また後の議論の準備として，従来からネオコグニトロンに採用されてきた学習法については，教師なし学習法の一つではあるが，別の節を設けて詳しく説明する．

[14]例えば，ネットワークに文字を学習させた場合，学習後に学習サンプルの中に含まれていなかったような変形した文字を提示された時に，それを正しく認識できる能力を言う．

3.2.1 教師あり学習法

教師あり学習法の多くは，ネットワークに，記憶させたいパターンとそのパターンの属性を明示的に与えて記憶させるものである (図 31)．これは，出力細胞が，特定の入力に対

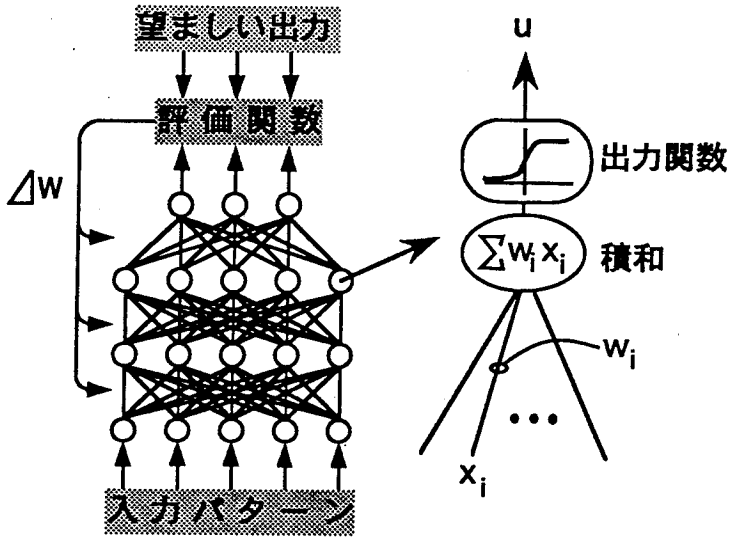


図 31: 教師あり学習の例

して，望ましい出力値を出すように，細胞間の一本一本の入力結合の強度を選択するという最適化問題の一つとして考えることが出来る．しかし，この最適化問題は，現実的には近似的にしか解くことが出来ない．

例えば，この最適化問題を，一つの細胞に収束する入力結合の強度の決定であると簡単化してみよう．さらに，入力結合 1 本の取り得る強度が量子化された m 値を取ると簡単化すれば，全体で n 本ある入力結合の強度の選択方法は，m^n 通りとなる．すなわち，本来ならば，この m^n 通りの組合せを総当たりして，その中からもっとも望ましい細胞の出力を発生させる入力結合強度の組み合わせを選択すれば良い．しかし，通常一つの細胞に収束する結合の数 n は数千〜数万であることと，入力結合の強度が連続値 (m は十分大きい) を取ることを考えれば，m^nは極めて膨大な数となる．つまり，このような総当たり戦略は現実的ではない．従って，これまでに提案された教師有り学習法は，上の最適化問題を近似的に解く方法となっている．

具体的には，記憶させたいパターンベクトルの組$x_1, x_2, \cdots, x_p$ と，それらのパターンに

対する望ましいネットワークの出力ベクトルの組 (あるいはそのパターンの属性とも言える) $\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_p$ ，そしてネットワーク内のパラメータ (例えば，細胞間の多数ある入力結合の強度) をひと塊にしたベクトル$\boldsymbol{W}$ を使って，評価関数 $L[\boldsymbol{W}]$ を定義する．すなわち，

$$L[\boldsymbol{W}] = \sum_{i}^{p} l[\boldsymbol{x}_i, \boldsymbol{d}_i, \boldsymbol{W}]. \tag{13}$$

関数 $l[\boldsymbol{x}_i, \boldsymbol{d}_i, \boldsymbol{W}]$ は，例えば，望ましいネットワークの出力$\boldsymbol{d}_i$ と実際のネットワークの出力の差を表すように定義される．

$$l[\boldsymbol{x}_i, \boldsymbol{d}_i, \boldsymbol{W}] = \| \boldsymbol{d}_i - \boldsymbol{f}[\boldsymbol{W}, \boldsymbol{x}_i] \|^2. \tag{14}$$

ここに，$\boldsymbol{f}[\boldsymbol{W}, \boldsymbol{x}_i]$ は，ネットワークのパラメータ (細胞間の入力結合強度) のベクトル$\boldsymbol{W}$ と入力パターンベクトル$\boldsymbol{x}_i$ によって記述されたネットワークの出力 (ベクトル) を表す．

そして，入力結合の強度$\boldsymbol{W}$ は，関数 $L[\boldsymbol{W}]$ を極小化する方向に少しずつ変化させる．例えば，$\boldsymbol{W}$ の変化方向を，次式で示すように，関数 $L[\boldsymbol{W}]$ の点$\boldsymbol{W}$ における最も勾配の急な方向として選ぶ．すなわち，

$$\Delta \boldsymbol{W} = -\alpha \cdot \frac{dL[\boldsymbol{W}]}{d\boldsymbol{W}} \tag{15}$$

ここにα は，$\boldsymbol{W}$ の変化速度を表す定数で，$0 < \alpha < 1$ である[15]．この式 15 は，$L(\boldsymbol{W})$ の極小点で (右辺がゼロとなる$\boldsymbol{W}$で) 平衡状態に落ち着く．

ただし，図 32 で示すように，極小点は必ずしも関数 $L[\boldsymbol{W}]$ が最小となる点とは限らない．この学習結果の善し悪しは，$\boldsymbol{W}$ の初期値$\boldsymbol{W}_0$ (すなわち，学習開始時の$\boldsymbol{W}$) に依存する．通常，この関数 $L[\boldsymbol{W}]$ の形状は，学習前から予測できないものであり，$\boldsymbol{W}_0$ はランダムに選択されることが多い．

実際には，式 (15) を実行する代わりに，一つのパターンが提示される度に以下の式で表される量だけ入力結合の強度を変化させることが多い．

$$\Delta \boldsymbol{W} = -\alpha \cdot \frac{dl[\boldsymbol{x}_i, \boldsymbol{d}_i, \boldsymbol{W}]}{d\boldsymbol{W}} \tag{16}$$

この式を，パラメータα の値を十分に小さくして $(0 < \alpha \ll 1)$ 実行すれば，式 (13) で定義される評価関数を近似的に極小化することになる．

[15] この式 (15) が成立するには，各細胞の出力関数が連続で，微分可能であることが必要である．一例として，図 31の細胞の出力関数に示すような単調増化関数が挙げられる．

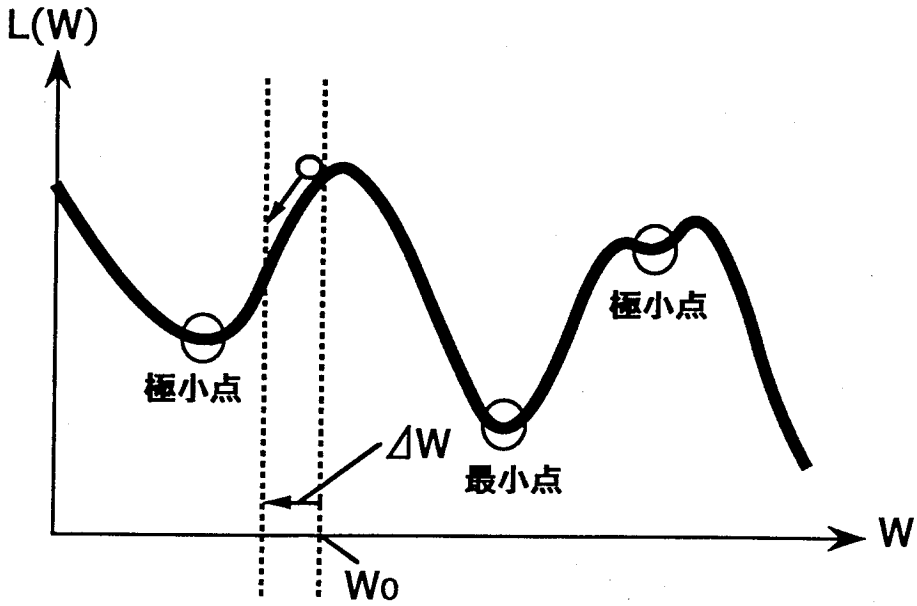


図 32: 入力結合の変化量の決定
W が一次元の場合の関数 $L(W)$ の形状の一例と，W の変化の方向とを表す．横軸が W を表し縦軸が $L(W)$ である．

また本論文では詳しく述べないが，対象とするネットワークが，図 31ので表すようなフィードフォワード型の階層ネットワークの場合，各層の細胞の出力はその一つ前の層の出力とその細胞の持つ入力結合の強度の積和によって書き表される．従って，このようなネットワークの中間層の入力結合強度の変化量を式 (15) によって計算すると，その一つ上の層の望ましい出力と，実際の出力の差 (誤差) を集約した形で記述される．つまり，各入力結合強度の変化量は，先の '誤差' を，一段づつ前の層へ伝搬させるような形で求めて行く形になる．従って，この手法を back-propagation 法と呼ぶことが多い [67].

3.2.2 教師なし学習法

教師なし学習法は当初，von der Malsburg [68] ，福島 [69][53] ，甘利 [70] ，Kohonen [71] らによって提案された．これらのモデルはいずれも，視覚野にある単純型細胞 (方位選択性を持つ細胞) や複雑型細胞などの自己組織的形成のモデルである．これらの学習法は，モデルによって少しづつ異なるものの，図 33で示すような形のネットワークを使っておおよそ次のように説明することが出来る．

これらのモデルは，ベクトル量子化[16] のための，コードブック (参照ベクトル) の決定手法の一つとして位置付けることができる．

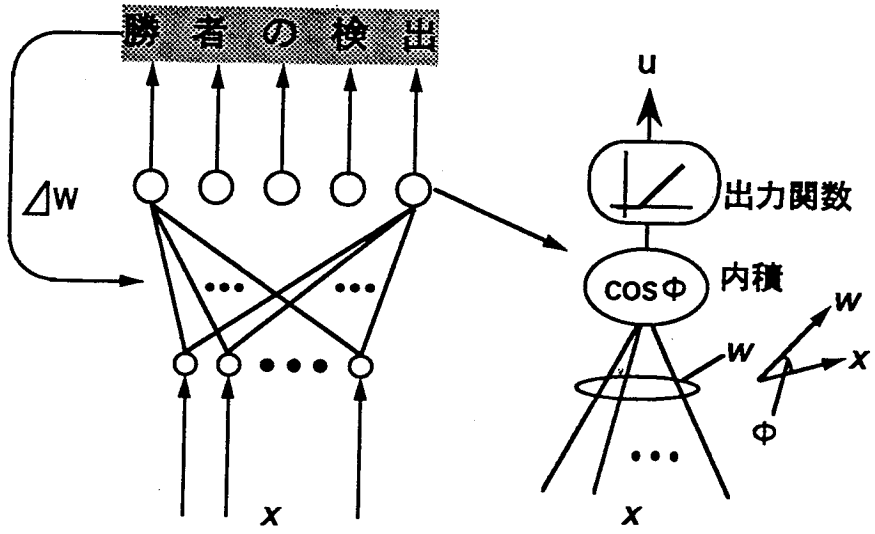


図 33: 教師なし学習の例

すなわち，一つの細胞は，入力 $\boldsymbol{x}_k$ (k はパターンの種類を表すインデックス) をその細胞の入力結合強度のベクトル $\boldsymbol{W}_j$ を通して受け取る．ネットワークの第 j 番目の細胞の活性値は次で表される．

$$u_j = \varphi[\cos\phi - \theta_j]\,. \tag{17}$$

ここに， $\varphi[x]$ は単調増加関数で，例えば $\varphi[x] = \max(x, 0)$ のように選ばれる．また， ϕ は，入力ベクトル $\boldsymbol{x}_k$ と入力結合強度のベクトル $\boldsymbol{W}_j$ のなす角度を表している．すなわち，

$$\cos\phi = \frac{(\boldsymbol{W}_j \cdot \boldsymbol{x}_k)}{\|\boldsymbol{W}_j\| \cdot \|\boldsymbol{x}_k\|} \tag{18}$$

である．式 (17) 中の θ_j (≥ 0) は，その細胞の閾値を表している．この閾値は 0 に設定することが多い．

学習は次のように行なわれる．まず， $\boldsymbol{W}_j$ の初期値 (あるいは初期方向) はランダム選んでおく．そして，パターンがネットワークに提示される度に，式 (17) 表される活性値 u_j

[16]ベクトル量子化は，大量のデータを小数のコードブック (参照ベクトル) によって，表現する，情報圧縮の一手法である．具体的には，一つの参照ベクトルで，そのベクトルに近い複数のパターンを表現することによって小数の参照ベクトルで大量のデータを記述する．

がもっとも大きくなった細胞(勝者)が，その入力結合の強度のベクトル $\boldsymbol{W}_j$ を変化させる[17]．その変化量は，提示されているパターンベクトル $\boldsymbol{x}_k$ に少しづつ近付ける形で行なわれる．すなわち，$\boldsymbol{W}_j$ の変化量は，

$$\Delta \boldsymbol{W}_j = \alpha \cdot (\boldsymbol{x}_k - \boldsymbol{W}_j) \cdot \delta(j,k) \tag{19}$$

である．ここに $\delta(j,k)$ は，パターン $\boldsymbol{x}_k$ がネットワークに提示された時に，第 j 細胞が勝者となった場合に値 1 をとり，そうでない場合には値 0 をとる関数である．α (<1) は，$\boldsymbol{W}_j$ の変化量を決定する正定数である．第 j 細胞が勝者となる領域は，その細胞の入力結合のベクトル $\boldsymbol{W}_j$ と他の細胞の入力結合のベクトル $\boldsymbol{W}_i$ $(i \neq j)$ によって決定される．従って関数 $\delta(j,k)$ は，本来ならば全ての細胞の入力結合のベクトルも変数に含まねばならないが，ここでは簡単のために単に $\delta(j,k)$ と表すことにする．

このような更新を，パターンが提示される毎に繰り返すことによって，各細胞の入力結合の強度のベクトル $\boldsymbol{W}_j$ は，次のようなベクトル $\overline{\boldsymbol{W}}_j$ に収束する．

$$\overline{\boldsymbol{W}}_j = \frac{\sum_k p_k \cdot \boldsymbol{x}_k \cdot \delta(j,k)}{\sum_k p_k \cdot \delta(j,k)} \tag{20}$$

ここに p_k は，パターン $\boldsymbol{x}_k$ が提示される確率(あるいは出現確率)を表す．すなわち，$\overline{\boldsymbol{W}}_j$ は，第 j 細胞が勝者となることが出来る範囲のパターンの平均値に落ちつくことになる．これによって，もし，第 j 細胞が勝者となることが出来る範囲の中で特に高い頻度で現れるパターンがあったとすると，$\overline{\boldsymbol{W}}_j$ はその高頻度で現れるパターンに近付いていくことになる．従って一般に，各細胞の入力結合 $\boldsymbol{W}_j$ は，高い頻度で現れるパターンの周囲に多く分布する傾向にある．これを，提示されるパターンが一次元の場合を一例に示そう．パターン x_k が提示される確率が図 34のように表される場合，各細胞の入力結合の強度が収束する強度 $\overline{W}_j$ は，図 34の縦実線で表すように収束する．すなわち，提示される確率の高いパターンに対応する $\overline{W}_j$ は，密に分布するが，提示される確率の低いパターンに対応する $\overline{W}_j$ は，疎に分布する．従って，この教師なし学習法は，提示されるパターン $\boldsymbol{x}_k$ の確率分布を近似的に入力結合の強度の分布に映し出す能力を持っている．

[17]Kohonen のモデル等では，連続的な特徴地図を得るために，勝者とその近傍にある細胞の両方が入力結合の更新を受けるが，ここでは簡単のためにこのような機構は考えない．

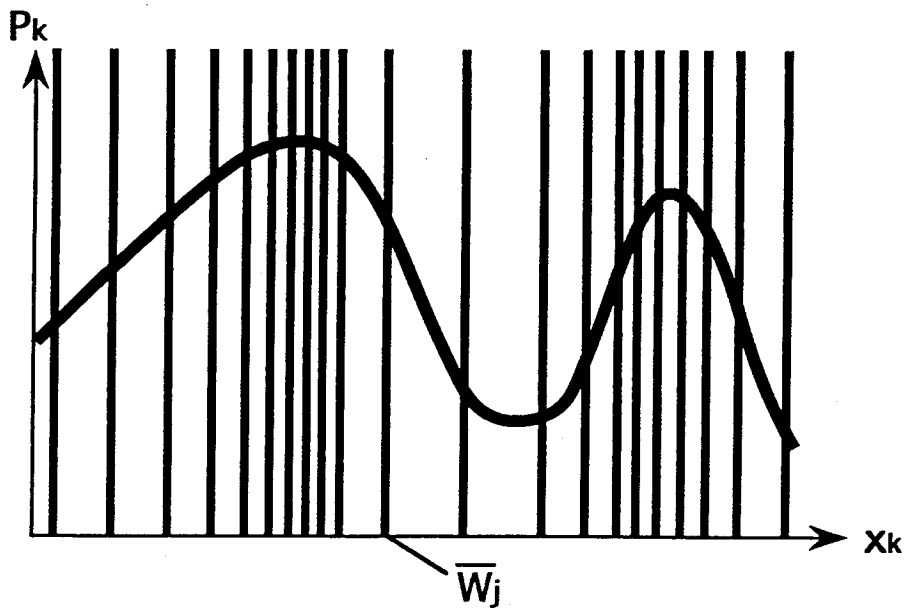


図 34: 提示されるパターンの出現確率と$\overline{W}_j$の分布
曲線が確率分布を表し，縦線の各位置が$\overline{W}_j$を表す．

Kohonen [56] は，この性質を利用して，提示パターン x_k として連続音声信号を周波数分割したスペクトルパターンをネットワークに次々に提示した (各細胞は 1 フレーム分の時間窓をもってこの音声スペクトルを観測する)．その結果，出来上がった $\overline{W}_j$ の多くは，音声信号中の母音のスペクトルを表していた．これは，連続音声においては，スペクトルがパターンが定常になる部分は，母音であるため，母音のスペクトルパターンの出現頻度が高いことを意味している．

このような学習法では，各パターン x_k の出現頻度が途中で変化するといったような，環境の変化が起きた場合に，$\overline{W}_j$ の分布もそれに合わせて適応的に変化する性質を持つ (図35参照)．これは，環境に適応するという意味で興味深いが，その一方で，過去に記憶した情報が失われるという問題も生ずる．例えば，図 35 の破線で表されるパターンを表現する細胞は，最初は 1 番であったが，環境の変化に伴い，2 番 3 番と変化していることが分かる．

しかし，実際に我々は，環境が変わったとしても，過去に記憶した事柄は忘れないものである．例えば，日本で生まれ育った人が，アメリカに移住した場合には，その人は，日本語を忘れずに，英語も喋れるようになる．すなわち，より人間に近い学習法としては，過

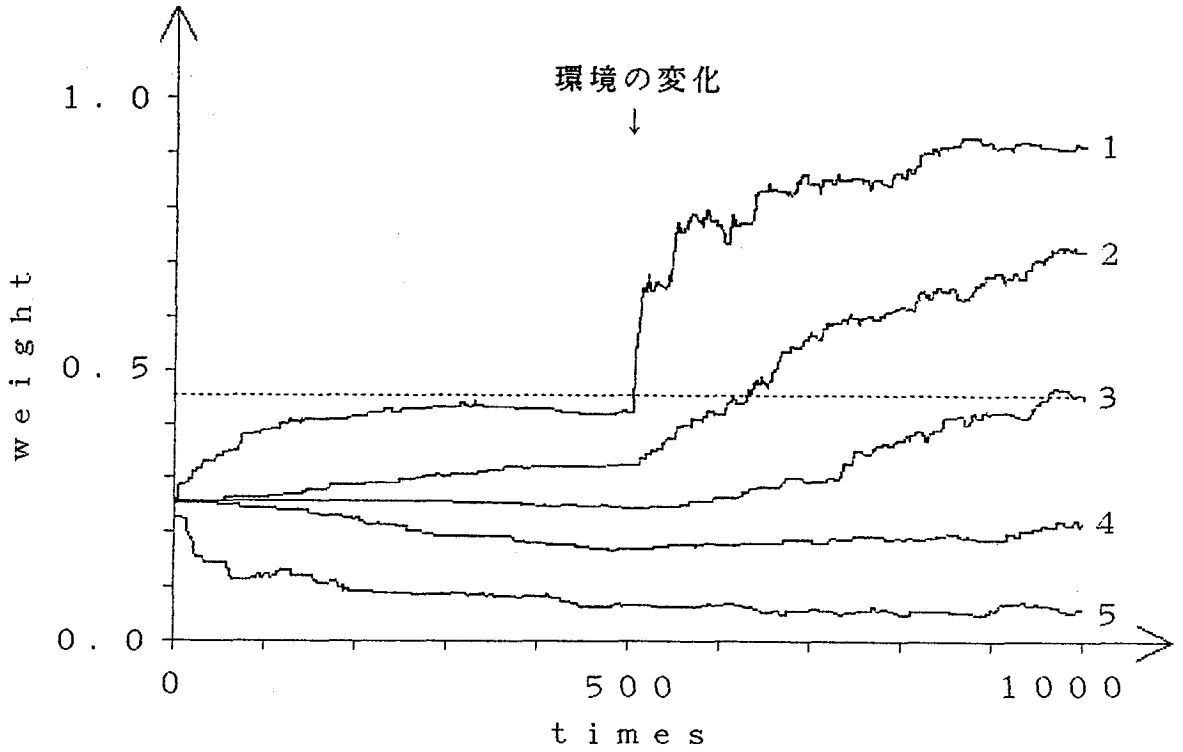


図 35: Kohonen のモデルの環境の変化に対する応答
縦軸は w_j の大きさを表し，横軸が学習ステップを表す．提示パターンは一次元で，最初の 500 パターンは，0〜0.5 までの一様乱数を提示し，次の 500 パターンは，0〜1.0 までの一様乱数を提示した．式 (17) の閾値 θ_j は 0 としている．

去に記憶した重要な情報は，保持しつつも，新たな環境に適応できる能力を持つべきであろう．

そこで筆者は以前，新しいパターンの追加学習が可能な教師なし学習法を実現するネットワークを提案した [72][73]．このネットワークの構造を図 36 に示す．このネットワークは，F 細胞，T 細胞，D 細胞，N 細胞の 4 種類の細胞で構成されている．T 細胞と D 細胞は 1 対 1 に対応している．この中の，D 細胞の結合強度が入力パターンを学習する．この学習法では，各 D 細胞の出力は，基本的に式 (17) と同様であるが，閾値が $\theta_j > 0$ にセットされている．即ち，一つの D 細胞は，その結合強度のベクトル $\boldsymbol{W}_j$ から離れたパターンに対しては，全く出力を出さない[18]．

このネットワークの学習は二つのフェーズを持つ．即ち，パターンが提示された時に，

[18] N 細胞は，2.3節で説明した V 細胞と同じ役割を果たす．すなわち，与えられたパターンの 2 乗平均値を出力し，D 細胞を分流的に抑制する．これにより D 細胞は，式 (17) で表されるような特性を得ている．しかし，簡単のため，ここでは N 細胞については，詳しく触れない

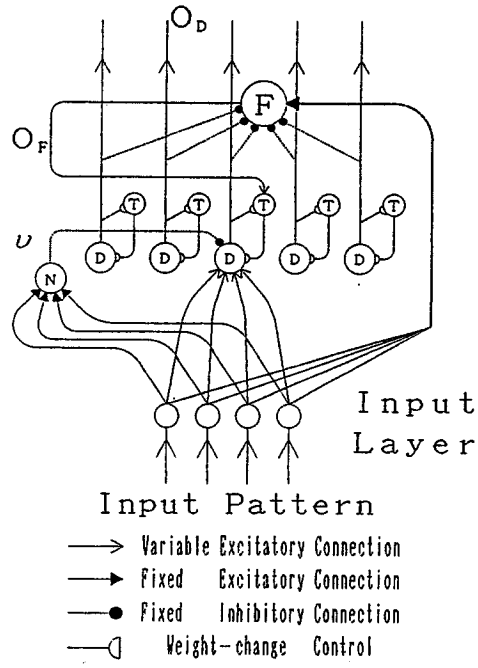


図 36: 追加学習可能な自己組織化機構 (山内 [72])

ネットワークを構成するどの細胞からも出力が出なかった場合と，一つの細胞が最大出力を出す場合である．

前者の場合には，図 36 中，F 細胞が出力を出す．この F 細胞が出力を出すと，T 細胞と名付けられた細胞の一つが，出力を出す．この T 細胞の出力は，その相手の D 細胞に送られる．この信号が来ると，D 細胞は，提示されているパターンベクトル$\boldsymbol{x}_k$ をその結合強度のベクトル$\boldsymbol{W}_j$ に写しとる．この時，先の T 細胞は，相手の D 細胞に信号を送ったことを，記憶し，再び F 細胞が出力を出した時に，相手の D 細胞に信号を送らないようにしている．一度，提示パターンをを入力結合に写し取った D 細胞は，それ以降，再び写し取ったパターン$\boldsymbol{x}_k$，あるいはそれに似たパターンが提示されると，それに対して活性出力を出す．

後者の学習は，提示されたパターンに対してある D 細胞が他に比べて最も大きな反応出力を出した場合に生ずる．この場合，その D 細胞は，$\boldsymbol{W}_j$ を，提示されているパターンに近付くように少しづつ更新する．この変化量は，式 (19) と同様である．ただし，$\boldsymbol{W}_j$ の初期値は，長さ 0 のベクトルに設定されている．

このようにすることによって，途中で環境が変化し，それまでに無かったような全く新た

なパターンの分布が現れたとしても，まだパターンを入力結合に写し取っていないＤ細胞が，新しいパターンに対処することになる[19]．従って，先の図 35 で示したような，過去の記憶を失うという事態は生じない (図 37参照)．だが，このようにして次々と新しいパター

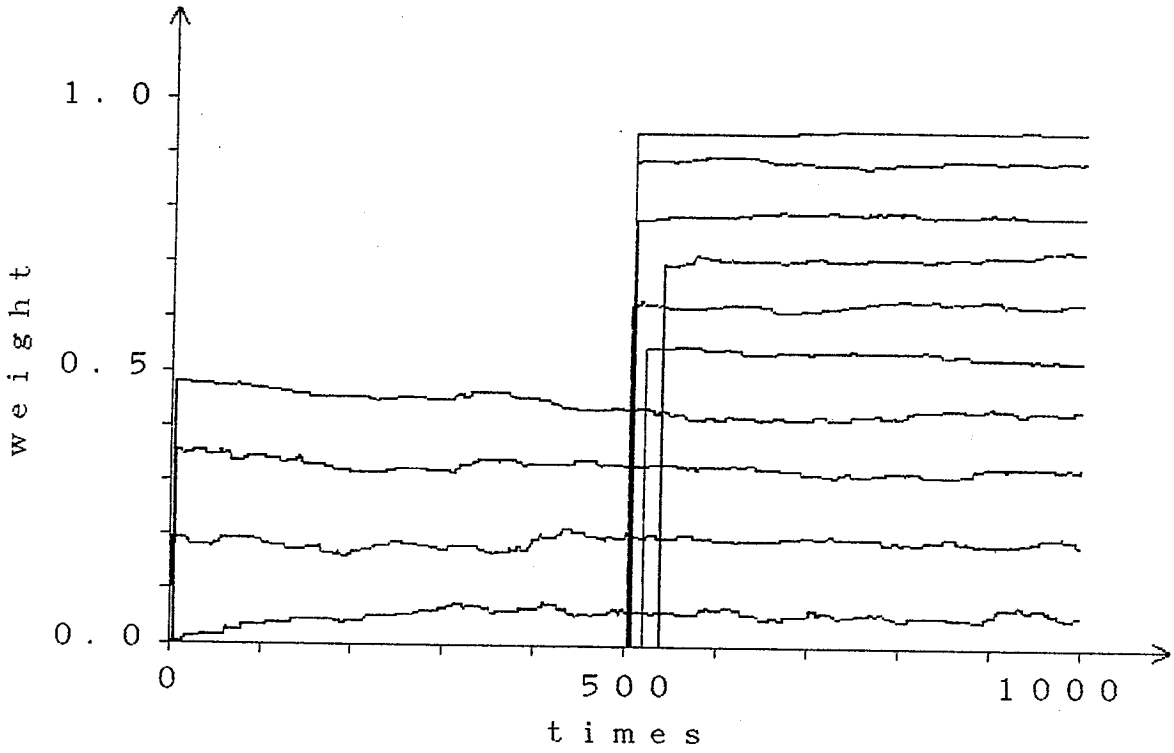


図 37: 追加学習可能なモデルの環境の変化に対する応答提示パターンは一次元で，最初の 500 パターンは，0〜0.5 までの一様乱数を提示し，次の 500 パターンは，0〜1.0 までの一様乱数を提示した (山内 [72])．

ンを記憶していくと仮定すると，膨大な数のＤ細胞が必要となる．これは，現実的ではない．そこで，Ｔ細胞は，相手のＤ細胞の出力を出す頻度が極端に低い場合には，Ｆ細胞が出力を出した際に，再度そのＤ細胞に対して信号を送り込み，新たなパターンを記憶させるようにしている．すなわち，必要のない記憶は，淘汰される．

ちなみに，次節で説明する，従来からネオコグニトロンに採用されていた学習法 [62] は，この追加学習可能なモデルに似た性質を持つ．すなわち，ネオコグニトロンの一個のＳ細胞は，式 (17) における閾値が，何も学習していない初期状態では，$\theta_j = 0$ であるが，あ

[19]閾値が$\theta_j > 0$ にセットされているため，既にパターンを学習したＤ細胞が，記憶しているパターンから離れたパターンを提示された際には，出力を出さない．従ってこの場合に，既にパターンを学習したＤ細胞が，入力結合の更新を受けることは無い．

るパターンを学習すると$\theta_j > 0$に変化することにほぼ等しい．従って，環境が変化しても，先のモデルと同様に，何もパターンを記憶していないS細胞が新たなパターンを記憶するために，過去の情報に影響を与えることは少ない．しかし，一度パターンを記憶したS細胞は，以後そのパターンを忘却することはほとんどないため[20]，次々と新たなパターンを提示し続けるとすれば，膨大な数のS細胞が必要となる．この問題は，後の3.3節で詳しく取り上げることにする．

3.2.3 ネオコグニトロン型学習法

福島ら [62] が提案したネオコグニトロンは，視覚パターン認識機構を説明する神経回路モデルである．このモデルは，図23で示したネットワークのブロック一つ分の構造とほぼ同じ構造をしている．そこでここでは，2.3節で定義した1個のブロックを使って，ネオコグニトロン型学習法を説明していくことにする．

2.3節でも述べたように，このブロックは，U_S層とU_C層の組が何段か重なって構成されている．この中で入力結合の更新を受ける細胞は，U_S層を構成するS細胞である．S細胞の出力は，式(2) (3) (4) (6) で定義されている．ここでは，第l段目のU_{Sl}層を構成するS細胞の学習に着目して，説明していくことにする．なお，ここではブロックを表す記号ξは省略する．

このネオコグニトロンの学習法は，前節で述べた教師なし学習法の一つとして位置付けることができる．この学習法と，他の学習法の主な違いは，入力結合の更新を行なう細胞の決定方法である．この違いは，ネオコグニトロンの構造に関係している．

2.3節でも述べたように，ネオコグニトロンのU_S層は，同一の入力結合強度を持つS細胞を二次元状に並べた面 (細胞面と呼ぶ) が何枚か集まって構成されている．同一細胞面上のS細胞は，それぞれ前段のU_{Cl-1}層上の，異なる位置からの入力を受けている．

学習は，この細胞面上から，学習を行なう細胞の代表者 (学習における核) を選び，その細胞面上の核以外の細胞は，その核となった細胞と同一の入力結合強度の変更を受けるようになっている．すなわち，同一細胞面上にあるS細胞は全て同じ入力結合の強度分布を

[20] 入力パターンの出現頻度が，十分ゆっくりと変化する場合には，それに追従して$\overline{\boldsymbol{W}}_j$が移動するので，過去の表現を失うこともある．

持つように学習する．の核となる細胞は，以下ようにして選ばれる．

ここで，U_{Cl-1}層上のある小さな領域 (これを競合領域と呼ぶ) に結合領域の中心を持つ一群のS細胞を，超コラムと呼ぶとにする．一つの超コラムの中には，全ての細胞面の細胞が含まれていて，それらの結合領域はほぼ一致している．

今，入力層 (図23では，U_D層に対応する) に学習パターンが提示されると，各超コラム内で最大出力を出したS細胞が，ひとまず核になる細胞の候補として選ばれる．この候補が，もし自分の属する細胞面上で唯一の候補である場合には無条件で核として選ばれる．しかし，同一細胞面上に2個以上の候補が選ばれた場合には，そのなかで，最大の出力を出している候補が核として選ばれる．

このようにして，U_{Sl} 層上の第 k 細胞面上のS細胞の一つが核として選ばれると，その細胞面上の全てのS細胞の興奮性入力結合の強度を決定する変数 $p_l(\boldsymbol{\nu}, \kappa, k)$ (式 (3) 参照) は，その核に選ばれたS細胞が受け取る入力の強度に比例した量だけ強化される．すなわち，$p_l(\boldsymbol{\nu}, \kappa, k)$ は，次式であらわされる量だけ強化される．

$$\Delta p_l(\boldsymbol{\nu}, \kappa, k) = q \cdot u_{Cl-1}(\boldsymbol{\nu} + \hat{\boldsymbol{n}}, \kappa) \tag{21}$$

ここに q は，強化の速度を決める正の定数である．

なお，この変数 $p_l(\boldsymbol{\nu}, \kappa, k)$ は，その細胞免状の全てのS細胞について共通なので，核として選ばれたS細胞の位置を表す記号 $\boldsymbol{n}$ は含まれていない．

ただし，$p_l(\boldsymbol{\nu}, \kappa, k)$ の初期値は十分に小さな正の値を持つと共に，その細胞面上のS細胞のパターン選択性を決定する変数 $r_l(k)$ の初期値についてもゼロである．これによって，どのようなパターンが提示された時でも，その細胞面上のS細胞から，何らかの出力が出るようにし，その細胞面から核が選ばれる可能性が出るようにしている．この変数 $r_l(k)$ は，一旦その細胞面の中から核が選ばれると，正定数 r_l に固定される ($r_l(k) = r_l$) [21]．

[21] オリジナルのネオコグニトロン [62] では，$r_l(k)$ は正定数で，S細胞の抑制性結合の強度 $b_l(k)$ の初期値がゼロであった．これに対して，2.3節の式 (4) の定義では，$b_l(k)$ は，興奮性結合の強度 (常にゼロ以上) に比例するので，常にゼロ以上の強度を持つ．従って，このオリジナルの学習法を2.3節の定義に合わせるとすれば，$r_l(k)$ は，そのS細胞に属する細胞面から核が選ばれないうちはゼロとすることに等しい．

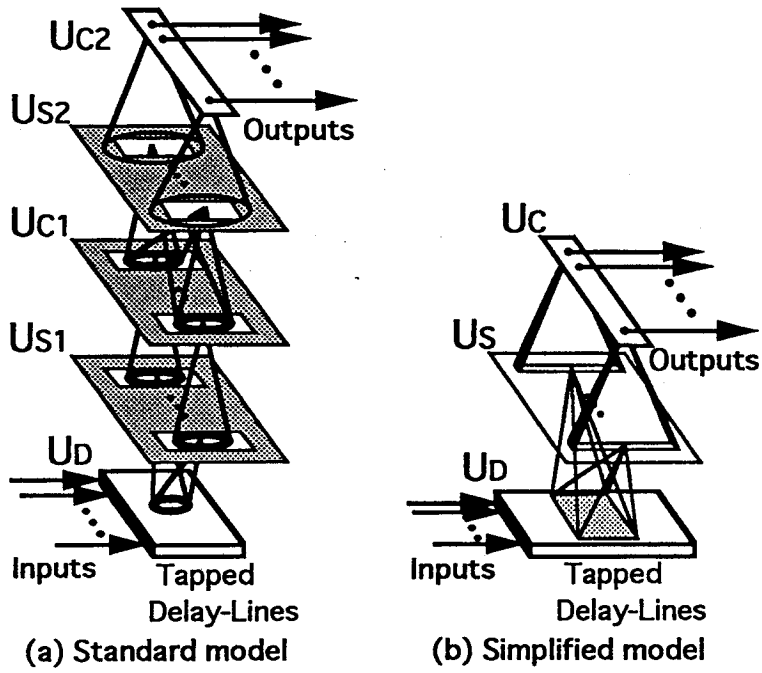


図 38: 時系列パターンを認識する階層ネットワークの構造

3.3 従来型学習法の時系列パターン学習時の問題点

前節では，これまでに提案された教師あり学習法，教師なし学習法，および従来からネオコグニトロンに採用されてきた学習法について，概観してきた．本節では，これらの学習法を，時系列パターンの学習に利用しようとした場合の問題点について述べることにする．

3.3.1 ここで前提とするネットワークの構造

従来の学習法の問題点について議論する前に，ここで扱うネットワークの構造について説明する．前提とするネットワークの構造は，基本的に，先に第2章で示した時系列パターン認識モデルの1ブロック分に相当する．即ち，時系列パターンを Delay-Line 上に空間展開し，それを基に認識を行なうネットワークを考える．このネットワークの構造を再度，図 38(a) に示す．ただし，以後の議論を簡潔にするために，Delay-Line (U_D層) は，伝搬速度が固定されているとする．すなわち，m 番目の *delay line* の時刻 t における入力パターンを，$u_p^t(m)$ $(m = 1, 2, \cdots, M)$ とおくと，その *delay line* 上の n 番目の遅延素子の，時刻

tにおける出力 $u_D^t(n,m)$ は，次式で与えられる．

$$u_D^t(n,m) = u_p^{t-n}(m) \tag{22}$$

また，後述する教師なし学習法の問題点については，主に従来からネオコグニトロンに採用されていた学習法について取り上げる．ここは，より考察し易くするために，扱う時空間パターンを単純化して，空間方向について位置づれが無く，時間軸方向にも伸縮がかかっていないような入力を扱うことにする．

そのため，この教師なし学習法の議論においては，図 38(a) のネットワークを簡略化して，U_S 層と U_C 層の組を 1 段だけ持つ図 38(b) のようなネットワークで考えていくことにする．(変形パターンを認識させるには，図 38(b) に示すネットワークを図 38(a) に示すネットワークと同様の構造に拡張すればよい [62]．従って，学習法についてのみ考察するならば，図 38(b) に示すネットワークで十分である)．図 38(b) に示すネットワークでは，U_S 層の各細胞面上の各 S 細胞の入力結合は，U_D層の時間軸方向については，一定範囲にある細胞とだけ結合しているが，空間軸方向には全結合している．そして S 細胞の入力結合は，時間軸方向にのみ並進対称性をもっている．また，U_C 層の C 細胞の数は 1 細胞面あたり 1 個に選ばれており，各 C 細胞は，U_S 層の特定の細胞面上にある全ての S 細胞から，信号を受け取っている．

ここで，後述するネオコグニトロン型学習法の議論を簡潔にするため，第 2 章で示した各層の数式表現を，図 38(b) で示すネットワーク用に簡略化しておこう．

第 k 番目の細胞面に含まれ，受容野の中心位置 (時間軸方向) n の S 細胞の時刻 t における出力 $u_S^t(n,k)$ は，

$$u_S^t(n,k) = r^t(k) \cdot \varphi\left[\frac{1 + \sum_m \sum_{|\nu| \le N} a^t(\nu,k,m) \cdot u_D^t(n+\nu,m)}{1 + \frac{r^t(k)}{1+r^t(k)} \cdot b^t(k) \cdot u_V^t(n,k)} - 1\right] \tag{23}$$

で表せる．N は，1 個の S 細胞が入力を受ける範囲の広さを表わす．$a^t(\nu,k,m)$ (≥ 0) は，U_D 層の遅延素子から S 細胞への時刻 t における，興奮性可変入力結合の強度を表し，

$$a^t(\nu,k,m) = \{c^t(\nu,k,m)\}^2 \cdot p^t(\nu,k,m) \tag{24}$$

である．ここに $p^t(\nu,k,m)$ (≥ 0) は，S 細胞の興奮性可変入力結合の内部変数で，学習時には，この変数を更新する．$c^t(\nu,k,m)$ (≥ 0) は，受容野内の各場所からの信号に対する感度を調節する係数である．

V 細胞から S 細胞への抑制性可変入力結合の時刻 t における強度 $b^t(k)$ (>0) は，

$$b^t(k)=\sqrt{\sum_m \sum_{|\nu|\leq N} \{c^t(\nu,k,m)\}^2\{p^t(\nu,k,m)\}^2} \tag{25}$$

で表せる．

第 k 番目の細胞面上にある，抑制性細胞 (V 細胞) の出力 $u_V^t(n,k)$ は，

$$u_V^t(n,k)=\sqrt{\sum_m \sum_{|\nu|\leq N} \{c^t(\nu,k,m)\}^2\{u_D^t(n+\nu,m)\}^2} \tag{26}$$

である．この式の $c^t(\nu,k,m)$ は，式 (24) で使われたものと同じもので，遅延素子から V 細胞への時刻 t における興奮性可変入力結合の強度でもある[22]．

U_S 層の出力はその上の U_C 層に送られる．この U_C 層上の第 k 番目の C 細胞の出力を $u_C^t(k)$ とおくと次式で表わされる．

$$u_C^t(k)=\psi\left[\sum_n d(n)\cdot u_S^t(n,k)\right] \tag{27}$$

ここに，$\psi[\,]$ は飽和特性を表す関数で，$\psi[x]=\varphi[x]/(1+\varphi[x])$ である．式 (27) の $\sum_n$ は，U_S 層上の第 k 番目の細胞面上にある全ての S 細胞に対して総和をとることを意味している．$d(n)$ (≥ 0) は S 細胞からの興奮性固定入力結合の強度を表わし，S 細胞面の中心位置 $n=0$ で最大値をとり，$|n|$ が大きくなるほど小さくなるような正の値に設定されている．

3.3.2 教師あり学習法の時系列パターン学習時の問題点

教師あり学習法によって時系列パターンを学習させる試みの一つとして，1.4.3節で触れた TDNN [54] が挙げられる．TDNN は，図 38(a) で示したネットワークと類似の構造を持っている (図 21参照)．

この教師あり学習法については，学習させるパターンを人間が明示的に指定しなければならない．この，TDNN を学習させる場合には，あらかじめ，記憶させたい時系列パター

[22]式 (24)(25)(26) の $\{c^t(\nu,k,m)\}^2$ が，従来のネオコグニトロン [62] の $c(\nu,m)$ に対応する．

ン (音声スペクトル) を人間が切りだし，作成する必要がある．そして，学習時には切り出した一つの学習サンプルを入力層 (図 38(a) の U_D層に相当する) に展開してから，各層の細胞の入力結合の強度を更新させている．すなわち，時系列パターンのどのタイミングで，学習するかを指示するのと同等である．従って教師あり学習法を使う場合には，予めその時系列パターンのどの部分が重要なのかを人間が知っておかなければならない．

しかし，この時系列パターンが未知のものである場合には，どの時点からどの時点までが重要なのかが分からないため，教師あり学習法を使うことは難しいと考えられる．

また，教師あり学習法を，人間の音声信号の学習法として見た場合にも，次のような不自然な点が挙げられる．すなわち，一般に良く知られているように，人間は幼い頃，誰からも教えられること無く，周囲の人の会話を聞いているだけで自然に言葉を覚えて行く．従って，人間が教師あり学習法のみによって，時系列パターンを学習しているとは考え難い．

3.3.3 教師なし学習法の時系列パターン学習時の問題点

一方，教師なし学習法は，パターンを提示されるだけで，そのパターンの確率分布を近似的にネットワーク内に取り込むことが出来る．これによって，例えば音声スペクトルパターン中の母音部分などのような重要な特徴を自動的に学習できるできることも知られている [56]．しかし，これらの学習法では，予め提示するパターンの環境が固定されている必要があり，次々と変化する実際の環境においては，せっかく獲得した重要な記憶が失われてしまうという欠点がある．これを解決する目的で提案された，追加学習可能なネットワーク [72] は，過去に記憶した重要な情報を忘れることなく，新たな環境に適応できる．また，このような性質は，従来から，ネオコグニトロンに採用されていた学習法にも見ることが出来る．しかし，以降に示すように，過去に記憶した情報を保持する形式を採る学習方式では，逆に，次々と新たなパターンが提示されるような時系列パターンの学習に利用しようとすると，膨大な数の細胞数が必要になるという問題が生ずる．本節では，対象とする従来型学習法として，従来からネオコグニトロンに採用されてきた学習法を選び，この問題点について，詳しく論ずることにする．なお，本節では，従来型のネオコグニトロンの学習法を，単に従来型学習法と呼ぶことにする．

この問題点を議論する前に，図 38(b) に示すネットワークに従来のネオコグニトロン型学習法 [62] を適用した場合の学習手続きをまとめておこう．

図 38(b) の U_D 層にパターン u_p が入力されたときに，U_S 層において，第$\hat{k}$ 細胞面上の位置$\hat{n}$ にある S 細胞の時刻 t における出力 $u_S^t(\hat{n},\hat{k})$ が，その競合領域 (Ωで表す) にある他の S 細胞よりも大きいとき，すなわち

$$u_S^t(\hat{n},\hat{k}) = \max_{|\nu|\leq\Omega,k} u_S^t(\hat{n}+\nu,k) \tag{28}$$

を満たす時には，その S 細胞をひとまず，学習における '核' [62] となる細胞の候補として選ぶ．この S 細胞が，第$\hat{k}$ 番目の細胞面上での唯一の候補であった場合には，その S 細胞は，無条件に '核' になる細胞となる．しかし，その細胞面内に 2 個以上の候補が選ばれたときには，それらの中で最も大きな出力で反応している候補を，核とする．この '核' として選ばれた S 細胞の興奮性可変入力結合の内部変数を，次の値だけ更新する．

$$\Delta p^t(\nu,\hat{k},m) = q\cdot u_D^t(\hat{n}+\nu,m) \tag{29}$$

ここに，q は入力結合の強化速度を決める正定数である．ただし，式 (29) は，核として選ばれた S 細胞の入力結合だけでなく，その S 細胞が属する細胞面のすべての S 細胞の入力結合も自動的に同じ値に更新することを意味している．

この，従来型学習法の場合には，$c^t(\nu,k,m)$ は変化させず，kによらない値に固定していた．すなわち，

$$c^t(\nu,k,m) = c_0(\nu) \tag{30}$$

としていた．ここに $c_0(\nu)$ は $|\nu|$ が大きくなるほど小さくなるような正の値をとる関数である．

また，$p^t(\nu,k,m)$ は，初期状態では，十分小さな正の値を持つ．パラメータ $r^t(k)$ は，第 k 細胞面から核が選ばれないうちは $r^t(k)=0$ とし，一旦核が選ばれると，$r^t(k)=r \quad (r>0)$ とする[23]．

[23] $b^t(k)$ については，オリジナルのネオコグニトロン [62] の場合には，初期値を 0 とし，その細胞面から '核' に選ばれる細胞が現れる毎に $u_V^t(n,k)$ に比例した値を加えるようにしていた．また $r^t(k)$ も正定数 r に固定されていた．しかし，ここでの定義のように，$r^t(k)$ を，その細胞面から核が選ばれないうちはゼロとし，一旦核が選ばれると正定数 r に固定するように定義を改めれば，式 (25) が成立すると考えても差し支えない．

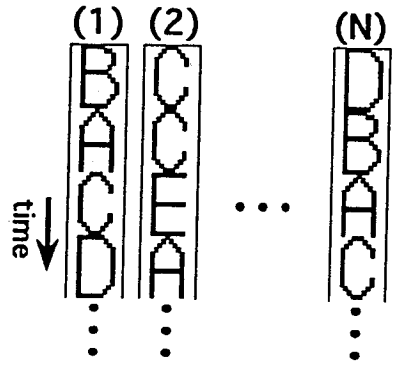


図 39: 入力パターンの例

さて，一般に認識対象となる刺激パターンはいくつかのカテゴリー (特徴) のパターンの組み合わせによって構成されると考えられる．例えば，人間がしゃべる言葉は，いくつかの語の組み合わせで構成される．そこで，後の議論を容易にするために，基底膜を通して得られる，周波数分割された人間の音声信号 (基底膜上に現れる時空間パターン) を個々のカテゴリーが明確に区別できるようなパターンに抽象化し，電光掲示板に映し出される文字系列 (例えば A,B,C,D,... のような形状のパターンを並べたもの) のようなパターンを考える．そして，各語に相当する部分が各文字であるとすれば，基底膜から得られる時空間パターンは，図 39 に示すように，文字パターンが時間軸方向に近似的にランダムな順序で並べられたものと考えることができる．そこでこのような時空間パターンを，図 38(b) で示すネットワークの U_D 層に入力し，従来型の学習法を使って学習させてみた．

U_D 層に未学習の時空間パターンが入力されると，最初に U_D 層の入力端 (図 38(b) に示す U_D層の左端) の近くにある細胞が反応出力を出す．すると，この反応パターンが現れた領域から入力を受け取る U_S 層上の S 細胞 (図 38(b) に示す U_S層の左端に位置する S 細胞) の一つが核として選ばれ，その反応パターンを学習する．一旦このようにして一つの S 細胞が学習すると，同一細胞面上の S 細胞の入力結合には，並進対称性があるので，学習したパターンが *delay line* 上を伝搬する間，U_S層上の左端 (入力端) から順に右に位置する S 細胞が，その S 細胞を中心とした競合領域の中で最大の出力を出し続ける．この結果，一度学習したパターンが，入力側から幅 Ω (競合領域の幅の 1/2) を伝搬する間は，ネットワークは新たな特徴を学習しない．つまり，ネットワークは，図 40 で示すように，競合領域の大きさ Ω に一致する時間間隔で未学習の細胞面に新しい特徴を記憶させるように働く．

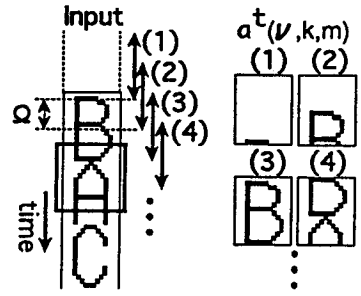


図 40: 従来型学習法の学習のタイミング

しかし，このように一定期間毎に新しい特徴を学習したとしても，次のような問題が起こる．簡単のため，基底膜から得られる時空間パターンが，i 種類の文字で構成されるとし，その時間長がすべて，S 細胞の受容野の時間軸方向の長さとほぼ一致すると仮定する．すると，S 細胞の学習する特徴は，1 個の文字全体を含む場合と，A の後半と B の前半のような 2 文字の境界を含む場合が生じる (ここでは，空白も 1 つの文字として扱い，空白と文字の一部を含む特徴も，2 文字の境界を含む特徴と呼ぶ)．このうち，空白だけで構成される特徴を除いて，2 文字の境界を含む特徴の組み合わせを考えると，前半に来る文字の種類と後半に来る文字の組み合わせだけを考えても $(i+1)^2-1$ 通りある．そして，これに時間的なずれが生ずると，同一の組み合わせでも異なる特徴となるので，さらに多数の組み合わせがあることが分かる．つまり，2 文字の境界を含む特徴だけでも極めて多数の特徴を学習することになる．従って，このような方法で多くのカテゴリーを含む実際の時空間パターンを学習させてゆくには，十分多数の S 細胞を用意しなければならない．これは，有限個の素子で構成される現実のシステム (生物を含む) には，実際的な方法とは言い難い．

このような問題を解決するには，単独の文字全体を含むパターンを 1 特徴として学習し，文字と文字の境界を含むパターンは，学習しないようにできれば良いと考えられる．このようにできれば，図 39 に示すような時空間パターンの特徴抽出には，i 種類の文字を認識する i 個の S 細胞があれば十分である[24]．個々の文字を正確に学習するための，もっとも単純な方法としては，学習させるべき各文字の前後に全くパターンが存在しない期間を設

[24]単語を学習する場合には，複数の文字を含む特徴を獲得する必要があるが，これは，高次の層に広い受容野を持つ S 細胞があって，複数の短い文字の集まりを 1 まとまりのカテゴリーとして学習すればよい．しかし，ここではこの問題は扱わない．

定することが考えられる [52]．すなわち，パターンとパターンの間の切れ目が一定期間続いたときに反応する細胞を仮定して，その細胞の出力があったときに，U_D 層上を伝搬している活性パターン (最後に入力された文字) を学習させればよい．しかし，実際の連続音声などを学習させる場合には，語と語の間に無音期間があることは少なく，さらにそれらの境界が不明確 [31] であるので，このような方法で個々のカテゴリーを正確に学習することは難しいと思われる．そこで我々は，語と語の間に無音期間がほとんどないような入力刺激であっても，それを提示するだけで，個々の語を正確に学習する改良型学習法を提案する．

3.4 改良型学習法

3.4.1 改良型学習法の概要

幼い子供が，親の音声を聞いているだけで，誰からも教えられること無く自然と言葉を覚えて行くことは，一般に良く知られている．これは，人間が，音声信号の中のどこからどこまでが一まとまりの言葉かという情報を一切与えられなくても，これらの言葉を正確に学習する機構を持っていることを意味している．では，なぜこのようなことが可能なのだろうか．

ここで，3.2 節と同様に音声信号を，文字がランダムな順序で並べられたパターンとみなし，特徴抽出細胞群が各文字 1 個分の時間幅とほぼ同じ時間幅で入力を受け取っている場合を考えよう．すると先の議論から，これらの細胞群が観測する特徴には，1 文字を単独に含む特徴と，2 文字の境界を含む特徴がある．

さて，特定の 1 文字を単独に含む特徴は，その文字が入力されれば必ず観測されるが，特定の 2 文字の境界を含む特徴については，その 2 文字が順に入力されなければ観測されないことに着目してみよう．各文字はランダムな順序で入力されると仮定すると，特定の 2 文字が同じ順序で入力される率は，特定の 1 文字が入力される率よりもかなり低くなると考えられる．従って，特定の 1 文字を単独に含む特徴の出現頻度は，特定の 2 文字の境界を含む特徴の出現頻度よりも十分に大きいと考えることができる．つまり，人間は，出現頻度の大きい特徴のみ，一まとまりの言葉を表す重要な特徴として，選択的に学習し，自

然に言葉を覚えていくのではないだろうか.

このような観点から，われわれは，従来の学習法を改良し，各S細胞は，学習した特徴の出現頻度がある値よりも大きい場合には，その特徴を重要な特徴とみなして保持し，それ以外の場合には学習した特徴を忘却して，再び別の特徴を学習するように働く改良型学習法を提案する．この学習法では，先のように各S細胞の受容野が，学習する特徴と同じサイズであるという仮定は必要なく，受容野よりも小さなサイズの特徴でも学習が可能である．そしてこの場合には，S細胞の受容野のうち，その小さなサイズの特徴の外側の部分(すなわち受容野の周辺部)からの入力信号に対する感度 $c^t(\nu,k,m)$ を減少させる．その結果，その部分からの入力結合の値すなわち $a^t(\nu,k,m)(=\{c^t(\nu,k,m)\}^2p^t(\nu,k,m))$ は減少する．つまり，その特徴のサイズに合わせてS細胞の受容野を狭くさせることに等しい．これにより，学習した小さな特徴の周囲に来るパターンが，S細胞を必要以上に活性化したり抑制したりしないようしている．

次節から，この改良型学習法を数式表現で説明するが，本論文では，理解し易くするために，まず $p^t(\nu,k,m)$ の更新手続きについて説明し，その後に，$c^t(\nu,k,m)$ の更新手続きを説明する．

3.4.2 $p^t(\nu,k,m)$ の更新

改良型学習法では，S細胞の入力結合の強度を，そのS細胞が属する細胞面が核として選ばれる頻度に比例するように変化させることを目的としている．そして，その入力結合の強度がある一定値よりも大きければ，そのS細胞には，学習した特徴を保持させるが，逆に小さければその特徴を忘却させ，再び新たな特徴を学習させる．これを実現するために，S細胞の興奮性可変入力結合の内部変数 $p^t(\nu,k,m)$ は，そのS細胞が核に選ばれた場合には増加させるが，逆に核に選ばれなかった場合には減少させるようにしている．すなわち $p^t(\nu,k,m)$ の変化量は，

$$\Delta p^t(\nu,k,m) = -\alpha\cdot f[b^t(k)]\cdot(p^t(\nu,k,m)-p_0)+\delta_{k\hat{k}}\cdot q[u_V^t(\hat{n},k)]\cdot u_D^t(\hat{n}+\nu,m) \qquad (31)$$

で表せる．ここに $\hat{k}$ は，核として選ばれた細胞面を表し，$\delta_{k\hat{k}}$ は，Kronecker のデルタを表す．

式 (31) の第一項は忘却項で $p^t(\nu,k,m)$ を減少させる項である．この中の $f[x]$ $(0 < f[x] \leq 1)$ は忘却率を決定する関数で，$x \to 0$ のとき $f[x] \to 1$ となり，$x \to \infty$ のとき $f[x] \to \varepsilon$ $(0 < \varepsilon \ll 1)$ となる単調減少関数である．後述のシミュレーションでは $f[x]$ を次式のように定めた．

$$f[x] = 1 - \frac{1-\varepsilon}{1+\exp[-\zeta(x-\theta)]} \tag{32}$$

ここにθ (>0) は，xがどの程度の大きさになったときに $f[x]$ が小さくなるかを決定する正定数で，$\zeta(\gg 0)$ は，忘却率の変化の緩やかさを決定する正定数である．すなわち，$b^t(k) > \theta$ のときには，忘却率は大きくなり，逆に $b^t(k) < \theta$ のときには，忘却率は小さくなる．また，α $(0 < \alpha < 1)$ は，忘却率を決定する比例定数である．p_0 (>0) は，十分に小さな正定数で，第 k 番目の細胞面の S 細胞が十分に長い時間，核として選ばれない場合に，$p^t(\nu,k,m)$ が収束する値である．つまり，十分に長い時間，パターンが提示されない場合には，全ての S 細胞の入力結合が，忘却項によって減衰して十分小さくなるが，0 にまで減衰してしまうと，後に入力が与えられても，S 細胞から出力が出ないため，核が選出できなくなってしまう．そこで，S 細胞の入力結合を常に 0 以上の値に保ち，入力が与えられたときには S 細胞から 0 以上の出力が出るようにして，常に核を選出できるようにしている．

式 (31) の第二項は，$p^t(\nu,k,m)$ を強化する項である．この中の $q[x]$ は結合強度の強化速度を決定する関数で，x が十分に小さいときには正定数となり，x が大きいときには $q[x] \propto 1/x$ となる．後述のシミュレーションでは $q[x]$ を次式のように定めた．

$$q[x] = \frac{Q}{\sigma_q + x} \tag{33}$$

ここに，σ_q (>0)，Q (>0) は定数である．したがって，式 (31) の第二項は，$u_V^t(\hat{n},k) \gg \sigma_q$ のときには，受容野に提示されている入力の強度を $u_V^t(\hat{n},k)$ で正規化したものに比例する．その結果，$p^t(\nu,k,m)$ のノルム (ベクトルとみなしたときの長さ) は，式 (31) が平衡状態に落ち着いたときには，$u_D^t(\hat{n}+\nu,m)$ のノルムにはよらず，第 k 番目の細胞面から核が選ばれる頻度 ($\delta_{k\hat{k}}$ が 1 となる頻度) にのみ依存した値となる (付録 3.6.2節参照)．したがって，式 (25) で与えられる $b^t(k)$ も，第 k 番目の細胞面から核が選ばれる頻度にのみ依存した値となる．

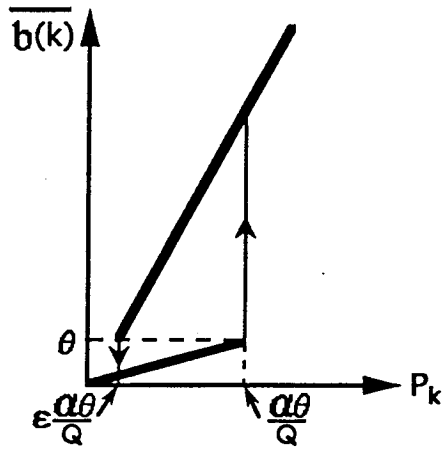


図 41: $\overline{b(k)}$と P_kの関係

忘却率 $\alpha \cdot f[b^t(k)]$ は，$b^t(k)$ の値によって非線形に変化する．したがって，$p^t(\nu, k, m)$ のノルム及び，$b^t(k)$ の値は，短時間平均をとってみると，その細胞面から核が選ばれる頻度 (P_kとおく) に応じて，図 41 のようにヒステリシスのある変化特性を示す．図は，$b^t(k)$ の短時間平均$\overline{b(k)}$ と，その細胞面から核が選ばれる頻度 P_k との関係の概要を示したものであるが，$p^t(\nu, k, m)$ についても同様の特性がみられる．ただしこの図では，式 (32) のパラメータς は十分に大きいと仮定している．入力結合の強度のこのような非線形的変化は，次の 2 つの効果をもたらす．

① ある S 細胞が，出現頻度の小さい特徴を学習した場合には，$\delta_{k\hat{k}} = 1$ となる頻度が低いので，その細胞の興奮性可変入力結合の内部変数 $p^t(\nu, k, m)$ の値も大きな値にはならない．したがって，式 (24) および式 (23) で与えられる S 細胞の出力もあまり大きくならない．一方，高頻度で現れる特徴を学習した場合には，$\delta_{k\hat{k}} = 1$ となる頻度が高いので，$p^t(\nu, k, m)$ の値は十分大きくなる．したがって，その特徴が来たときには，S 細胞は大出力を出す．

② ある S 細胞が高頻度で核に選ばれて，一旦 $b^t(k)$ の値が閾値θ を超すと，式 (32) から分かるように $p^t(\nu, k, m)$ の忘却率 $\alpha \cdot f[b^t(k)]$ は十分小さくなる．したがって，後に環境が変わって核に選ばれる頻度がある程度小さくなったとしても $p^t(\nu, k, m)$ は大きく保たれたままほとんど減衰しない．すなわち，S 細胞は一度学習した重要な特徴

を忘却し難くなる．

3.4.3 $r^t(k)$ の更新

ある S 細胞が学習した特徴の出現頻度が小さい場合，前述のようにその S 細胞の興奮性可変入力結合の内部変数 $p^t(\nu,k,m)$ および抑制性可変入力結合の強度 $b^t(k)$ は十分に小さくなり，学習した特徴を急速に忘却する．学習した特徴を忘却した S 細胞には，再び別の特徴を学習させることによって再利用させる．これを実現するために，V 細胞からの抑制性入力の効果を決定する変数 $r^t(k)$ を，抑制性入力結合の強度 $b^t(k)$ に応じて変化させ，$b^t(k) \to 0$ のときには，$r^t(k) \to 0$ となるようにする ($r^t(k) \to 0$ となると，第 k 番目の細胞面上の S 細胞の特徴選択性が小さくなる)．すなわち，入力結合の強度が十分小さくなった S 細胞は，その特徴選択性が小さくなるので，以前学習した特徴以外の特徴に対しても出力を出すようになる．つまり，学習した特徴を忘却した S 細胞は，再度別の特徴を学習する可能性が高まる．シミュレーションでは $r^t(k)$ を次式で定めた．

$$r^t(k) = R \cdot \frac{b^t(k)}{\sigma_r^t + b^t(k)} \tag{34}$$

ここに，R (>0) は，$r^t(k)$ の上限を設定する定数で，$b^t(k)$ が十分に大きくなると $r^t(k)$ は Rに飽和する．また，σ_r^t (>0) は，$b^t(k)$ がどの程度の強度になったときに，$r^t(k)$ が飽和するかを決定する変数である．

σ_r^t を十分に小さな値にすると，$b^t(k)$ が少し増加しただけでも，S 細胞の特徴選択性が高まるので，S 細胞は一旦ある特徴を学習すると，以後 (その特徴の出現頻度が極端に低下して，忘却が起こる場合を除けば)，その特徴とほぼ同一のパターンが現れたときにのみ反応する．逆にσ_r^t を大きな値にすると，$b^t(k)$ が少々増加しても，S 細胞の特徴選択性は大きくならないので，S 細胞は一旦学習した特徴の一部だけを含むようなパターンが現れても反応する．したがって，σ_r^t を小さな値にすると，S 細胞は受容野のサイズとほぼ同じサイズの特徴を学習する傾向が強まり，σ_r^t を大きな値にすると，S 細胞は受容野のサイズよりも小さなサイズの特徴であっても学習する傾向が強まる．

しかし，学習開始時からσ_r^t の値を大きな値にすると，次のような不都合が生ずる．σ_r^t が大きな値を持つときには，特徴選択性が低いので，S 細胞は学習したパターンの一部分の

みが含まれるようなパターンに対しても反応する．したがって，あるS細胞が，その受容野とほぼ同じサイズの特徴を，少しずれた状態(その特徴の一部が受容野外にはみだした状態)で学習してしまうと，そのS細胞は，学習した特徴がそのずれた位置に来たときにしか反応しなくなる．つまり，S細胞は受容野とほぼ同じサイズの特徴であっても，その特徴と一致したものではなく，その一部が受容野外にはみ出し，ずれたものを保持してしまうことがある．そこでシミュレーションでは，σ_r^t を初期状態では小さな値にし，学習開始時からの時間経過と共に徐々に大きな値に変化させるようにした(σ_r^t の時間変化の式は，付録に示してある)．このようにパラメータを変化させることによって，S細胞は学習の初期段階では受容野とほぼ同程度のサイズの特徴を学習し，時間の経過と共に，徐々に受容野よりも小さなサイズの特徴を学習するようになる．

3.4.4 $c^t(\nu,k,m)$ の更新

第 k 番目の細胞面のS細胞が，その受容野よりも小さなサイズの特徴に反応するようになった場合，そのS細胞は，核として選ばれる度に，その小さな特徴とその周囲にあるパターンとを学習する．このとき，第 k 番目の細胞面上のS細胞とV細胞の興奮性入力結合のうち，その小さなサイズの特徴の周辺部からの信号に対する感度 $c^t(\nu,k,m)$ を下げる．つまり，その小さな特徴のサイズに合わせて，そのS細胞およびV細胞の受容野を狭くする．この $c^t(\nu,k,m)$ の更新には，S細胞が学習するパターンの，次のような性質を利用している．すなわち，各特徴の前後に来る特徴の種類がランダムなので，1つのS細胞が学習するパターンは，そのS細胞を反応させる小さな特徴以外は，学習する毎に異なる可能性が高い．

そこで，第 k 番目の細胞面上のS細胞が核として選ばれると，$p^t(\nu,k,m)$ と，提示されている入力 $u_D^t(\hat{n}+\nu,m)$ とを比較して，その中で一定の割合以上の差がある部分の $c^t(\nu,k,m)$ を減少させるようにする．逆に第 k 番目の細胞面上のS細胞が核として選ばれないときは，$c^t(\nu,k,m)$ を少しづつその初期値 $c_0(\nu)$ $(\geq c^t(\nu,k,m) > 0)$ に回復させる．すなわち，

$c^t(\nu,k,m)$ の変化量は，次式で表せる．

$$\Delta c^t(\nu,k,m) = \alpha' \cdot f[b^t(k)] \cdot (c_0(\nu) - c^t(\nu,k,m)) - \delta_{k\hat{k}} \cdot q' \cdot c^t(\nu,k,m) \cdot E\left[\tilde{u}^t(\hat{n},\nu,k,m), \tilde{p}^t(\nu,k,m)\right] \tag{35}$$

この式の第一項が回復項で，$c^t(\nu,k,m)$ を初期値に回復させる方向に働く．この中の $f[x]$ は，回復率を決定する関数で，式 (31) で使用した式 (32) と同じ単調減少関数である．つまり，回復率は，式 (31) の忘却率の大きさと連動して変化し，$b^t(k) < \theta$ のときには回復率は大きくなり $b^t(k) > \theta$ のときには回復率は小さくなる．α' $(0 < \alpha' < 1)$ は，回復率を決定する比例定数である．この回復項により，第 k 番目の細胞面が十分に長い時間核として選ばれない場合は $c^t(\nu,k,m)$ は，再利用に備えてその初期値 $c_0(\nu)$ に回復する．

式 (35) の第二項は，$c^t(\nu,k,m)$ を減少させる項である．この中の $E[x,y]$ は，x と y の間に，一定の割合 ($\theta'(>1)$ で表す) 以上の差がある場合に正となる関数で，シミュレーションでは，次式を使用した．

$$E[x,y] = \varphi[x - \theta' y] + \varphi[y - \theta' x] \tag{36}$$

ここに，$\varphi[\ \ ]$ は式 (23) で使用した関数と同じで，半波整流特性を表す関数である．すなわち $E[x,y]$ は，$x > \theta' y$ のとき，あるいは $x < y/\theta'$ のときのみ正の値を取りそれ以外は 0 である．$\tilde{p}^t(\nu,k,m)$ は，$p^t(\nu,k,m)$ を $c^t(\nu,k,m)$ で重み付けしたものを正規化した強度で，$b^t(k)$ を使って次式で表される．

$$\tilde{p}^t(\nu,k,m) = \frac{c^t(\nu,k,m) \cdot p^t(\nu,k,m)}{b^t(k)} \tag{37}$$

同様に$\tilde{u}^t(\hat{n},\nu,k,m)$ は，$u_D^t(\hat{n}+\nu,m)$ を $c^t(\nu,k,m)$ で重み付けしたものを正規化した強度で，$u_V^t(\hat{n},k,m)$ を使って次式で表される．

$$\tilde{u}^t(\hat{n},\nu,k,m) = \frac{c^t(\nu,k,m) \cdot u_D^t(\hat{n}+\nu,m)}{u_V^t(\hat{n},k,m)} \tag{38}$$

また，q' (>0) は，$c^t(\nu,k,m)$ の減少速度を決める正定数である．つまり式 (35) は，正規化された $\{c^t(\nu,k,m) \cdot p^t(\nu,k,m)\}$ と，正規化された $\{c^t(\nu,k,m) \cdot u_D^t(\hat{n}+\nu,m)\}$ の間に，割合θ' 以上の誤差がある箇所は，不必要な部分とみなして $c^t(\nu,k,m)$ を減少させる．この場合，回復項の効果が十分に小さいとすれば，$c^t(\nu,k,m) \to 0$ となる．

3.4.5 計算機シミュレーション

図38(b) で示されるネットワークに，従来型と改良型のそれぞれの学習法を用いて，同一のパターンを学習させ，それぞれの場合に出来上がったS細胞の入力結合を比較する．1個のS細胞が入力を受け取る広さは，15(空間方向) × 19(= $2N+1$，時間方向) とした．U_S 層の細胞面の数は，従来型については300面，改良型については10面としている (改良型の場合に，300面の細胞面を用意しても，その大部分が，未使用のまま残ってしまうので，10面だけ用意した)．この他，シミュレーションで使用したネットワークの細胞数及び，パラメータについては，付録で詳しく述べる．学習パターンは，A,B,C,D,E,F の6個の文字とスペースを含む7個のパターン (画素は1,0の二値をとる) を縦方向に前述の図39のように，ランダムに2400個並べたパターン (文字1800個とスペース600個) を使用した．ただし，様々な時間長の文字が混在するテストパターンを使用できるようにするため，各文字の時間長を A,D が10，B,E が13，C,F が17，スペースが8とした．従来型および，改良型学習法を採用した場合に，得られた入力結合の強度をそれぞれ図42，と図43に示す．図は，結合の強度を黒塗りの四角形の大きさで表している．両図の，ラベル 'a' はS細胞の興奮性入力結合の強度 $a^t(\nu,k,m)$，ラベル 'b' はS細胞の抑制性入力結合の強度 $b^t(k)$ を表す．また，図43には，参考のために $a^t(\nu,k,m)$ の内部変数 $A^t(\nu,k,m)$ (ラベル 'A')，及び $\{c^t(\nu,k,m)\}^2$ (ラベル 'C') も並べて表示してある．ただし従来型の場合には，300面の細胞面すべてについて表示することはできないので，図42ではその一部のみを示した．

従来型学習法の場合，ほとんどの細胞面が文字と文字の境界を含む特徴を学習している．また，図には一部しか示されていないが，300面の細胞面すべてが，何らかの特徴を学習した．

一方，改良型学習法の場合，10面の細胞面のうち，結合強度が大きいのは6面だけで，それ以外の細胞面の結合強度は十分に小さくなっている．そして，この6面の細胞面は，全て単独の文字を正確に学習している．さらに，この6面の中で，短い文字 (A,B,D,E) を学習した細胞面の 'a' $(= \{c^t(\nu,k,m)\}^2 \cdot A^t(\nu,k,m))$ は，学習した短い文字パターンのみを正確に表している．ちなみに，この短い文字パターンを学習した細胞面の $a^t(\nu,k,m)$ の内部変数を見ると，$A^t(\nu,k,m)$ は，その短い文字パターンを表す部分の前後に，他の文字の

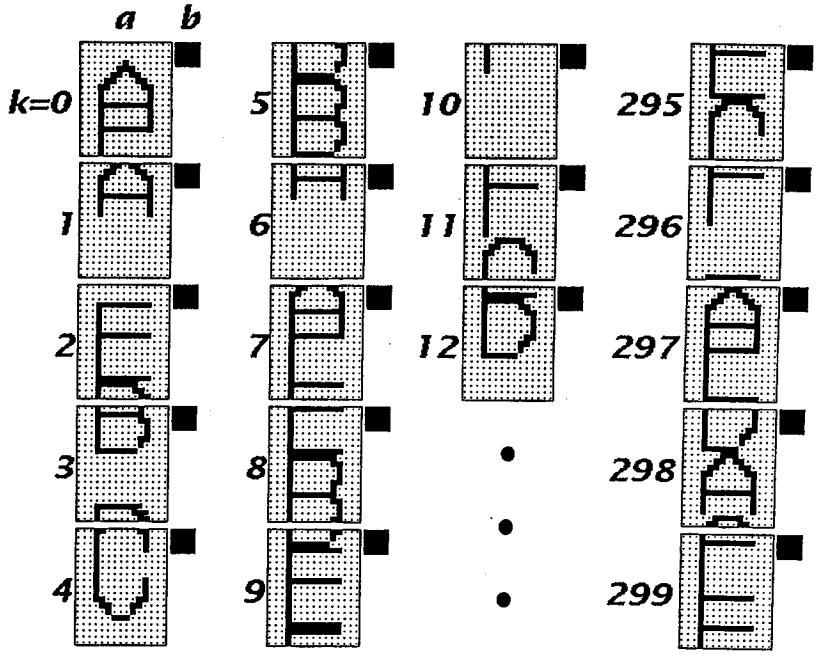


図 42: 従来型学習法によって決定された結合強度
ほとんどの細胞面が文字と文字の境界を含む特徴を学習している．また，一部しか示していないが，300 面の細胞面すべてが，何らかの特徴を学習した．

影響で，大きな値を持つ部分があるが，その部分では，$c^t(\nu,k,m)$ が，ほぼ 0 となっているので，$a^t(\nu,k,m)$ の値も十分に小さくなっている．

改良型学習法で学習させた後に，ネットワークに学習パターンを入力したときの最終段 (U_C層) の C 細胞の出力 $u_C^t(k)$ を図 44 に示す．2.2 節でも述べたように，このネットワークでは，U_C 層の第 k 番目の C 細胞は，U_S 層の第 k 番目の細胞面上にある全ての S 細胞の出力を，興奮性入力結合を介して受け取っている．したがって，第 k 番目の細胞面上にある S 細胞が，1 個でも出力を出せば，第 k 番目の C 細胞は出力を出す．図 44 は，横軸を時間軸，縦軸を出力の大きさとして，上段に入力パターン，下段に最終段の 10 個の C 細胞の出力の時間変化を示している．下段の認識出力の部分では，上から順に $k=0,1,2,\ldots$ 番目の C 細胞の出力を示している．また，その右端に矢印と共に記してある文字は，その C 細胞が選択的に反応する文字パターンを表す．図 43 と図 44 を比較すると，入力結合の強

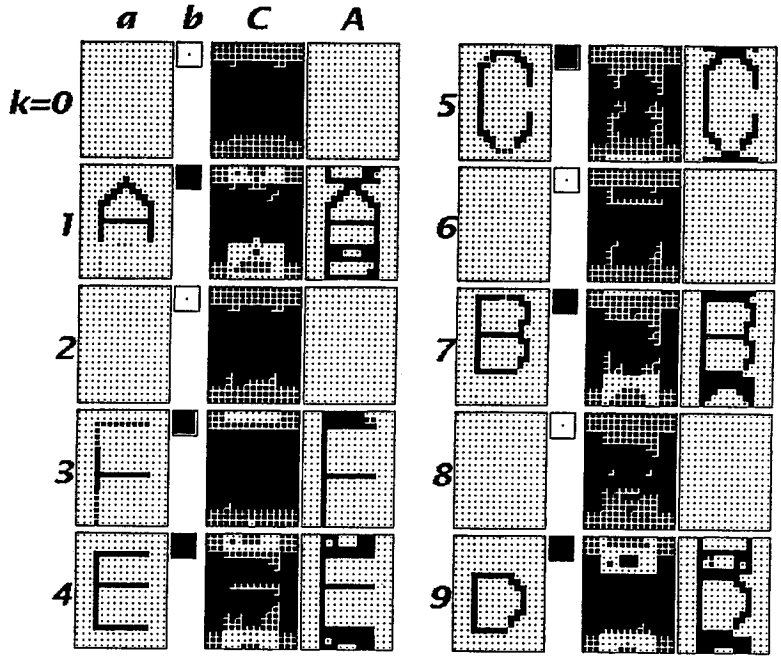


図 43: 改良型学習法によって決定された結合強度
10 面の細胞面のうち，結合強度が大きいのは 6 面だけで，それ以外の細胞面の結合強度は十分に小さくなっている．そして，この 6 面の細胞面は，全て単独の文字を正確に学習している．さらに，この 6 面の中で，短い文字 (A,B,D,E) を学習した細胞面の 'a' は，学習した短い文字パターンのみを正確に表している．

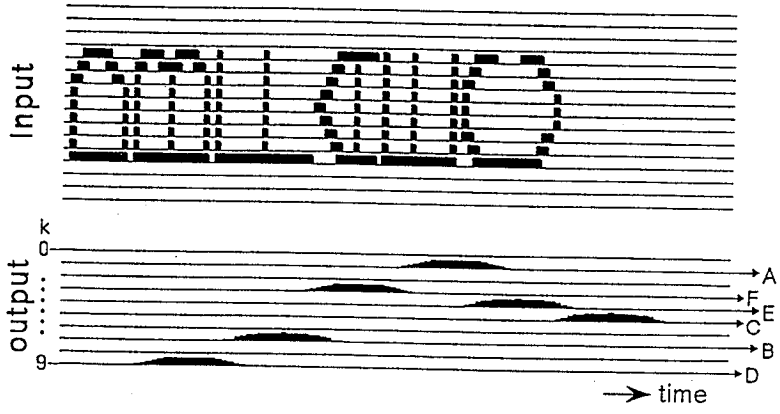


図 44: ネットワークの出力

度が大きな細胞面のS細胞のみが，特定の文字に対して選択的に出力を出していることが分かる．一方，入力結合の強度が小さな細胞面のS細胞からは，出力が出ていないことが分かる(実際には表示できないほど小さな出力が出ている)．つまりネットワークは，必要な特徴のみを正確に検出している．

3.5 結言

従来型の学習法の問題点を指摘し，この問題点を克服する改良型学習法を提案した．この学習法は，ネオコグニトロン型学習法を改善したもので，高い頻度で現れる重要な特徴のみを選択的に保持する能力を持つ．この学習法は，頻繁に現れる特徴を学習したS細胞には，その特徴を保持させ，逆に，低い頻度でしか現れない特徴を学習したS細胞には，その特徴を忘却させる．そして，一旦学習した特徴を忘却したS細胞には，再び別の特徴を学習させることによって，そのS細胞を再利用する．これにより，少数の細胞を用意するだけで，時系列パターンの中の重要な特徴を学習させることが可能である．また，S細胞が，その受容野よりも小さなサイズの特徴を学習する場合，受容野のサイズがその特徴のサイズに合わせて自動的に狭くなるので，S細胞は，その小さな特徴の周囲に来るパターンから影響を受けないようになる．

この改良型学習法を適用した小規模のネオコグニトロンを使って，時系列パターンの中の重要な特徴のみを正確に学習できることをシミュレーションによって確認した．

ただし，現段階では，学習させるパターンの性質に合わせて，人間が各パラメータを調節する必要があるという問題がある．例えば，式(32)のθは，入力の中に現れる特徴の出現頻度を知った上で，人間が調節する必要がある．また，パラメータσ_r^tを学習開始時からの時間経過と共に徐々に変化させているので，追加学習が難しいという問題もある．例えば，学習開始の時点から十分に時間が経過した時に，S細胞に，その受容野とほぼ同じサイズの未学習の特徴を提示すると，S細胞はこれをうまく学習できないことがある(4.3節参照)．今後，如何にして未知のパターンを与えるだけで各パラメータを自動的に決定させるか，また如何にして追加学習できるようにするかが課題である[25]．

[25]パラメータσ_r^tについては，S細胞の受容野の広さと時系列パターン中の各特徴のサイズに極端な差が無い場合には，学習期間中，一定値に固定していても差し支えない．

このように二，三の課題が残されているが，この改良型学習法によって，ネットワークは，実際に我々人間が受ける知覚刺激とほぼ同様のパターンを，継続的に提示されるだけで，自動的にその中の重要な情報を発見し，認識できるようになった．

3.6 付録

3.6.1 比較実験で使用した従来型学習法と改良型学習法のパラメータ

シミュレーションで使用したネットワークの細胞数は，U_D層が15(空間方向) × 40(時間方向)，U_S層が1(空間方向) × 21(時間方向)，U_C層が1(空間方向) × 1(時間方向) である．改良型学習法のパラメータは $R = 15$, $Q = 0.075$, $\sigma_q = 0.05$, $\alpha = 0.15$, $\theta = 0.48$, $\zeta = 50$, $\theta' = 1.5, q' = 0.3, \alpha' = 0.002, A_0 = 0.0001$ とし，σ_r^t は，$\sigma_r^t = 1/\{0.3 + 5.7\exp(-0.00017\,t)\}$ で決定した．比較のための従来型学習法のパラメータは $r = 15$, $q = 0.5$ とした．両学習手続きにおいて，U_S 層の競合領域の広さ Ω は，時間方向に 10 とした．また，$p^t(\nu, k, m)$ の初期値は [0, 0.0002] の一様乱数とした．$c_0(\nu)$ は，$\{c_0(\nu)\}^2$ の形状が，$|\nu| \leq N\ (= 10)$ の範囲内では標準偏差が 17 の 1 次元ガウス関数に，その外側は 0 となるようにした．ただし，$\sum_m \sum_{|\nu| \leq N} \{c_0(\nu)\}^2 = 1$ となるように定めた．C 細胞の入力結合の強度 $d(n)$ は $|\,n\,| \leq 10$ で，標準偏差 5 の 1 次元ガウス関数になるように定めた．

3.6.2 式 (31) の解析

式 (31) が平衡状態に落ち着いたときの入力結合 $b^t(k)$ の平均値$\overline{b(k)}$ を求める．ただし，$c^t(\nu, k, m)$ の変化が，$p^t(\nu, k, m)$ の変化に比べて十分にゆっくりとしているものとする．ここでは理解し易くするために，式 (31) をベクトル空間上で考える．そこで，$\{c^t(\nu, m, k) \cdot p^t(\nu, k, m)\}$, $\{c^t(\nu, m, k) \cdot u_D^t(\hat{n} + \nu, m)\}$ $(|\nu| \leq A;\ m = 1, 2, \cdots, M)$ をそれぞれベクトル$\tilde{\mathbf{p}}^t(k)$, $\tilde{\mathbf{u}}^t(k)$ で表す．すると式 (25) (26) より $u_V^t(n, k) = \|\tilde{\mathbf{u}}^t(k)\|$, $b^t(k) = \|\tilde{\mathbf{p}}^t(k)\|$ となる．

ここで，入力の強度が十分に大きく $u_V^t(\hat{n}, k) \gg \sigma_q$ となる場合を考える．すると式 (33) から $q[u_V^t(\hat{n}, k)] \simeq Q/u_V^t(\hat{n}, k)$ となる．このとき，式 (31) の両辺に $c^t(\nu, m, k)$ を乗じ，

$p_b(\nu, m)$ を十分に小さいものとして無視すれば，次の近似式を得る．

$$\Delta\tilde{\mathbf{p}}^t(k) \cong -\alpha \cdot f[b^t(k)] \cdot \tilde{\mathbf{p}}^t(k) + \delta_{k\hat{k}} \cdot Q \cdot \frac{\tilde{\mathbf{u}}^t(k)}{\|\tilde{\mathbf{u}}^t(k)\|} \tag{39}$$

つまり，$\tilde{\mathbf{p}}^t(k)$ を強化する項は$\tilde{\mathbf{u}}^t(k)$ を正規化したベクトルとなる．ここで，式 (39) の実行によって $b^t(k)(= \|\tilde{\mathbf{p}}^t(k)\|)$ が受ける変化量を求めると，

$$\Delta b^t(k) \cong -\alpha \cdot f[b^t(k)] \cdot b^t(k) + \delta_{k\hat{k}} \cdot Q \cdot \lambda^t(k) \tag{40}$$

となる．ただし，$\lambda^t(k)$ $(0 < \lambda^t(k) \leq 1)$ は，

$$\lambda^t(k) = \frac{\|\tilde{\mathbf{p}}^t(k) + Q \cdot \hat{\mathbf{u}}^t(k)\|}{\|\tilde{\mathbf{p}}^t(k)\| + Q} \tag{41}$$

である．ここに$\hat{\mathbf{u}}^t(k)$ は$\tilde{\mathbf{u}}^t(k)$ を正規化したベクトルを表す．この$\lambda^t(k)$ は，$\tilde{\mathbf{u}}^t(k)$ と$\tilde{\mathbf{p}}^t(k)$ がほぼ平行である場合には$\lambda^t(k) \simeq 1$ となる．簡単のため，忘却率を決定する関数を正定数として $f[x] = \rho$ $(\rho < 1)$ としよう．すると，第 k 番目の細胞面が核に選ばれる頻度が P_k のとき，式 (39) が平衡状態に落ち着いたときには，式 (40) も平衡状態となるので，

$$-\alpha \cdot \rho \cdot \overline{b(k)} + P_k \cdot Q \cdot \overline{\lambda(k)} = 0 \tag{42}$$

となる．ここに$\overline{\lambda(k)}$ は$\lambda^t(k)$ の平均を表す．ここで$\overline{\lambda(k)} \simeq 1$ と近似すると，式 (42) から，

$$\overline{b(k)} \cong \frac{Q}{\alpha \cdot \rho} \cdot P_k \tag{43}$$

となる．つまり$\overline{b(k)}$ は，入力の強度によらず，第 k 番目の細胞面が核に選ばれる頻度にのみ依存する．

第4章　音声認識へのアプローチ

4.1　序言

本章では，第2章で提案した時系列パターン認識モデルを使って，音声信号を学習・認識するシステムを構築する [74][75] [76] [77]．そして計算機シミュレーションでは，本システムの音声信号の時間軸方向の伸縮や，スペクトルの変形に対する汎化能力を調べるために，一人の男性が普通のスピードで発音したいくつかの単語をシステムに学習させた後，女性が様々なスピードで発音した単語を提示し，正しく認識できるかどうかを確認する．

この音声認識システムを構築するに当たって，ここでは，人間の聴覚系が，大きく二つのモジュールに分けられると仮定している (図45参照)．

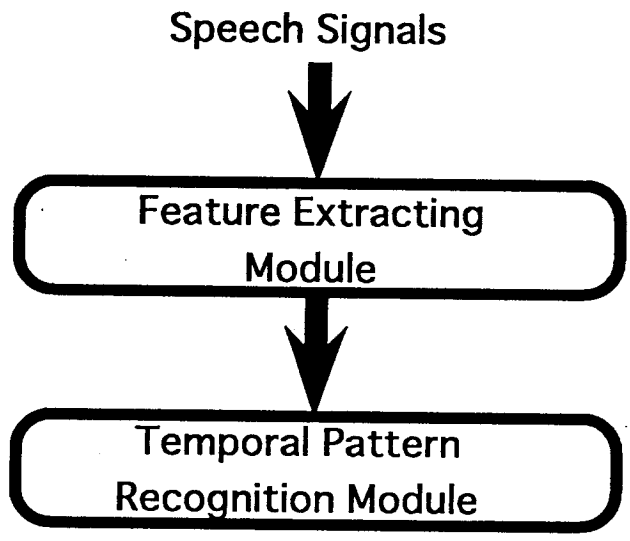


図 45: システムの概要

一つは，音声信号を周波数分割し，その中からいくつかの特徴を抽出する "特徴抽出モジュール" である．この部分は，1.2節で概観した，解剖学的に見た聴覚系の構造のうち，蝸牛から，大脳聴覚野に至るまでの経路に相当する．この経路上にあるいくつかの神経核の働きは，上オリーブ核の音源定位を除けば，主に音声信号の周波数分割が目的と見られている．また，これらの神経核を構成する細胞の中には，音声信号中の特定の特徴に選択的に反応する細胞も見い出されており，特徴抽出が行なわれているとも考えられる．

もう一方のモジュールは，特徴抽出モジュールの出力を入力とし，いくつかの特徴で表現された音声信号を認識する“認識モジュール”である．このモジュールは，大脳聴覚皮質に対応すると仮定している．ここでは，このモジュールを，第2章で提案した時系列パターン認識モデルによって構成する．

4.2 特徴抽出部

4.2.1 特徴抽出部の概要

特徴抽出部では，音声信号を音声スペクトルに変換し，その音声スペクトルから，周波数一定成分 (Constant-Frequency)，周波数上昇 (Frequency-Ascending), および下降成分 (Frequency-Descending) の抽出を行なう．以後表現を簡潔にするため，周波数一定成分を CF, 周波数上昇成分と周波数下降成分をそれぞれ FM-A, FM-D で表す．この特徴抽出部は，伊藤，福島 [78] のモデルをベースとしているが，彼らのモデルよりも構造が簡単である．

この特徴抽出部は，入力ブロック (input block), CF 抽出ブロック (CF-extracting block), FM 抽出ブロック (FM-extracting block) の三つのブロックで構成される (図 46参照).

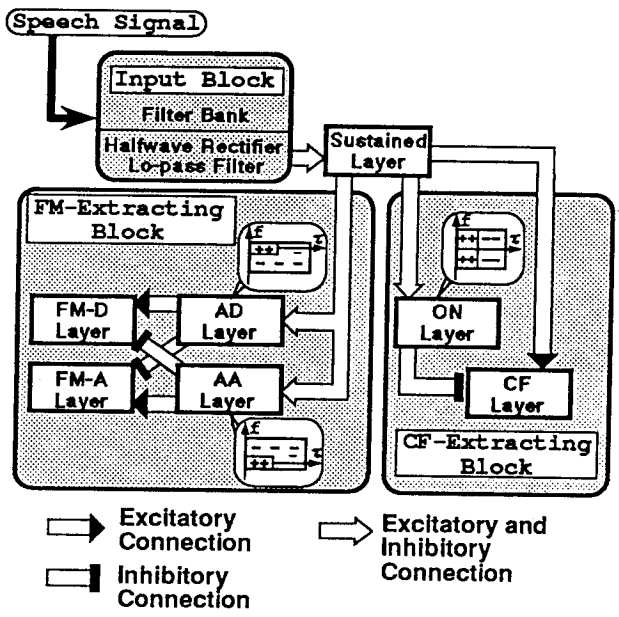


図 46: 特徴抽出部の構造の概要

入力ブロックは，いわば蝸牛の有毛細胞に相当する部分で，フィルターバンクと，半波整流部によって構成される．この入力ブロックの出力は，持続型細胞層を通る．この部分

では，各周波数成分に対して連続反応する細胞で構成されている．これらの細胞はそれぞれ他の周波数成分に反応する細胞と，抑制性の結合で結ばれており，各周波数成分の出力の強度を，近似的に正規化する役目をする．持続型細胞層の出力は，CF 抽出ブロックと FM 抽出ブロックの両方に送られる．

この持続型細胞層，CF 抽出ブロックおよび FM 抽出ブロックは，いずれも異なった特徴周波数[26] を持った細胞が周波数軸方向に並んだ細胞層によって構成されている．これらの細胞の特徴周波数は，Bark scale(1 から 16 Bark) で配置されている．さらに，同一細胞層の細胞は，特徴周波数の違いを除けば，全て同じ特徴選択性を示す．特徴抽出モジュールは，このような細胞層を複数組み合わせて構成されている．

4.2.2 特徴抽出部を構成する細胞

細胞層を構成する細胞は，実際の聴覚神経細胞を参考にしている．すなわち，これらの細胞の入力および出力は，その細胞が反応する周波数成分の強度に比例している．図 47 は，この細胞の構造を示している．

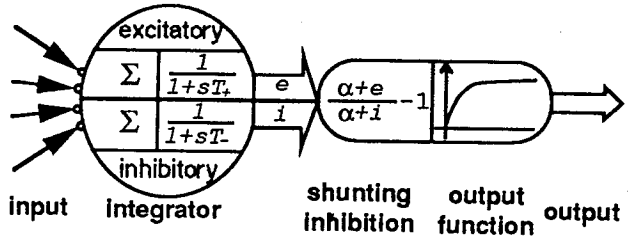


図 47: 聴覚神経細胞のモデル
T_+, T_- は，積分器の時定数を表す．α は，出力関数の飽和特性を決定するパラメータである．

図に示すように，この細胞は，興奮性と抑制性の入力結合をある時定数を持った積分器を通して受け取っている．このような結合と積分器を通して，この細胞は，入力信号を周波数軸と時間軸の両方から積分していることに等しい．抑制性の結合を通して入力される信号は，細胞の出力を分流的に抑制する．この細胞の出力関数は，飽和関数である．この細胞の，時空間特性は，この興奮性と抑制性の結合，および積分器の時定数を調節するこ

[26]特徴周波数とは，その細胞がもっとも良く反応する周波数を指す．

とで決定することが出来る．例えば，興奮性と抑制性の結合の分布が同じで，積分器の時定数が，興奮性の方が抑制性のものよりも小さく設定されている場合には，on-type の特性を持つ細胞となる．これを図 48 に示す．この図の左側は，興奮性と抑制性の両方の結合

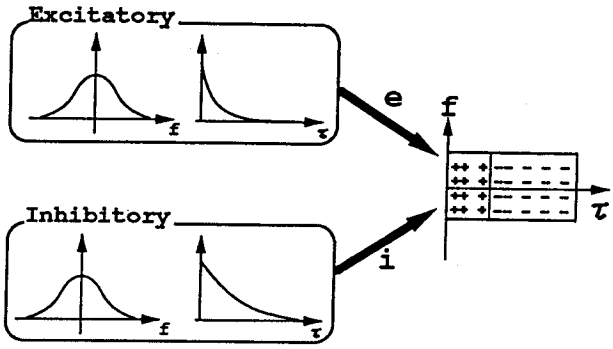


図 48: 細胞のパターン選択性の一例

の分布と，積分器の時定数を示している．図の右側は，左側に示す結合の分布と時定数によって決定される細胞の時空間パターン選択性を示している．図で使われている記号 'f' と 'τ' は，それぞれ周波数と時間を表している．

4.2.3 入力部 (Input Block)

入力部は，基底膜のモデルとなっており，入力音声信号を周波数分割する役目をしている．ここではこれを，平原ら [7] が提案した "適応 Q 型蝸牛フィルター" を基に構成している．このフィルターの特徴は，各周波数成分の音圧に応じて，適応的にフィルターの Q を変えながら周波数分割を行なうことである．これは，実際の基底膜の各位置における共振特性から，参考にされたものである．基底膜の各位置における共振特性は，1.2.1節の図 4 で示したように，音圧が小さい時ほど，共振特性が鋭くなり，逆に音圧が大きいほどこの特性はブロードなものとなる．これは，次のような効果をもたらす．例えば，人間の発音する音声信号の中で，周波数の変化の激しい部分 (FM 成分や子音等) は，音声知覚のための重要な情報が多く含まれている．しかし通常，このような部分の音圧は低い事が多い．そこで，このような音声信号を，上に示すような特性を持つ有毛細胞で周波数分割すれば，このような部分の情報を明確に抽出することができる．

平原らは，この適応 Q 型フィルターを図 49 で示すように，直列につないだノッチフィル

ターで構成している．図に示すように，各ノッチフィルターの出力は，バンドパスフィル

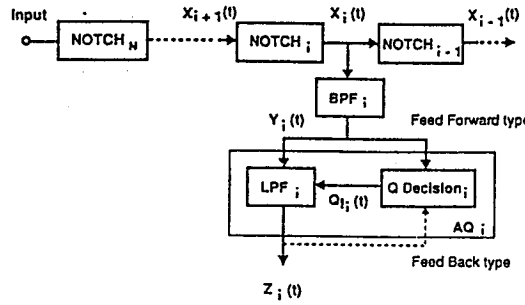


図 49: 適応 Q 型蝸牛フィルター (平原 [6])

ターに送られる．そして，各バンドパスフィルターの出力は，Q 決定回路，および，Q 可変ローパスフィルターに送られる．Q 決定回路は，Q 可変ローパスフィルターの Q を制御し，バンドパスフィルターの出力強度が小さい時には，Q を大きくし，逆に出力強度の大きい時には Q を小さくする．この適応 Q 型蝸牛フィルターの最終出力は，各 Q 可変ローパスフィルターの出力となる．

4.2.4 持続型細胞層 (Sustained Response Layer)

持続型細胞層は，入力パターンの各周波数成分に対して，持続的に出力を出す性質を持つ細胞によって構成されている．この層の出力は，以下に示すような機構によって，入力ブロックの出力を時間毎に近似的に正規化したものとなっている．これによって，音声信号中に含まれる背景雑音を抑制する効果がある．

このようにするのは，持続型細胞層に後続する下丘の細胞に相当する CF ブロック，および蝸牛神経核の onset 型の細胞のような特性を示す FM 抽出ブロックが，正常に動作するようにするために必要となる．

後述するように，これらのブロックを構成する細胞には，周波数成分が変化した場合に反応出力を出す細胞が存在している．例えば，CF ブロックには，各周波数成分の立ち上がり部分に反応する細胞が存在する．従って，このような細胞が持続型細胞層を介さずに，入力ブロックの出力を受け取っているとすると，背景雑音の成分にも反応してしまうことが多い．これが，そのブロックの出力に悪影響を及ぼすことがある．

そこで，CF およびFM 抽出ブロックの前に，持続型細胞層を配置し，この背景雑音を抑制している．先にも述べたように，これは，音声スペクトルの各周波数成分の強度を，時間毎に正規化することによって，行なわれる．すなわち，この正規化によって，背景雑音の振幅が音声信号の振幅よりも小さい時には，フォルマント成分が強調され，背景雑音の成分が抑制される．

この正規化は次のような構造によって実現されている．すなわち，持続型細胞層を構成する各細胞は，入力部から，興奮性と抑制性の両方の信号を同じ時定数の積分器を通して受け取っている．抑制性入力結合の分布は，興奮性のものよりも十分に広く，各細胞は，ほぼ全周波数成分からの抑制性信号を受け取っている．これにより，各細胞の出力は，その細胞が興奮性結合で信号を受け取っている周波数成分を，全周波数成分の出力で近似的に正規化したものになる[27]．

4.2.5 CF 抽出部 (CF-Extracting Block)

CF 抽出ブロックは，継続する一定周波数成分を抽出する．例えばこれは，母音の各フォルマントに相当する．このブロックはON 層と CF 層で構成される (図 46参照)．

ON 層の各細胞は，持続型細胞層から，興奮性，抑制性の両方の結合を通して信号を受け取っている．一つの細胞は，特定の周波数成分の ONSET に対して反応するが，その周波数成分が一定期間以上持続した場合には，その出力は減衰する．この細胞の興奮性および抑制性の分布は共に同じだが，時定数は，興奮性の方が抑制性よりも小さい (図 48参照)．

ON 層の各細胞は，各周波数成分の ONSET に対して反応する．この ONSET 成分は，周波数の変化する成分も含まれる．持続型細胞層は，入力音声の周波数成分をほぼそのまま反映するので，その出力は，周波数一定成分と周波数の変化する成分の両方を含んでいる．CF 層は，ON 層の出力を使って，持続型細胞層の出力の中に含まれる，周波数の変化する部分を抑制している．

CF 層を構成する各細胞は，興奮性結合を持続型細胞層からのみ受け取り，抑制性結合を，ON 層のみから，受け取っている．すなわち，CF 層は，長く持続する周波数一定成分にのみ反応する．

[27]抑制性信号は，細胞の出力を分流型に抑制する．

4.2.6 FM 抽出部 (FM-Extracting Block)

FM 抽出部は，周波数上昇成分と周波数下降成分を抽出する．

このFM抽出部には，二つの非対称層がある．一つは，AA (非対称上昇型層：asymmetric-ascending) 層で，もう一つは，AD (非対称下降型層：asymmetric-descending) 層である．AA 層の振舞いは，先の ON 層の振舞いに似ているが，周波数下降成分に反応しない点が異なっている．すなわち，AA 層は，周波数上昇成分と，周波数一定成分，および周波数下降成分の ONSET に反応する．同様に AD 層の振舞いについても，周波数下降成分と，周波数一定成分，および周波数上昇成分の ONSET に反応する．

AA 層を構成する細胞 (AA 細胞と呼ぶ) と，AD 層を構成する細胞 (AD 細胞と呼ぶ) は，共に持続型細胞層から，興奮性および抑制性の結合を受けている．AA および AD 細胞の興奮性結合の結合領域は，抑制性結合の結合領域よりも狭い．そして，これらの細胞の抑制性結合の分布は，非対称である．例えば，AA 細胞に収束する抑制性結合の分布は，高周波数側に広がっている．積分器の時定数は，AA，AD 細胞ともに，興奮性の方が，抑制性よりも小さくとられている．

例えば，周波数の下降する音を，フィルターバンクに入力したとすると，AA 細胞は，最初に，持続型細胞の高い特徴周波数の細胞からの抑制性信号を受け取り，その後，低い特徴周波数の細胞から興奮性信号を受け取る．この時，抑制性の効果は，興奮性の効果よりも大きいため，AA 細胞は，周波数下降の音には反応しない．

AA 細胞と AD 細胞は，周波数上昇，および下降成分だけでなく，それ以外の ONSET 成分に反応する．これを，周波数上昇，下降成分のみを取り出すため，FM-D(周波数上昇) および FM-D(周波数下降) 層は，AA，AD 層の両方の出力を使って，周波数上昇下降成分以外の成分を抑えるようにしている．具体的には，FM-A，FM-D 層の細胞は，同一の特徴周波数の AA 細胞と AD 細胞が，同時に反応する成分を抑制するようにしている．これを実現するために，FM 層では，AA，AD 層のいずれか一方の層から興奮性結合を受け取り，もう一方から抑制性結合を受け取っている．

例えば FM-A 層の細胞は，AA 細胞からの興奮性結合を受け取ると共に，その AD 細胞から，抑制性結合を受け取っている．他の ONSET 成分を完全に抑制するため，抑制性結

合の結合領域は，興奮性結合の結合領域よりも広く設定されている．これにより，FM-A 層からは，純粋に，周波数上昇成分のみが抽出される．

同様に，FM-D 層の細胞は，AA 細胞からの抑制性結合と，AD 細胞からの興奮性結合を受け取っている．

4.3 認識部

特徴抽出部の出力は，認識部に送られ，認識される (図 50)．認識部 [74][75] の構造は，第 2 章で提案した時系列パターン認識モデルとほぼ同じである．

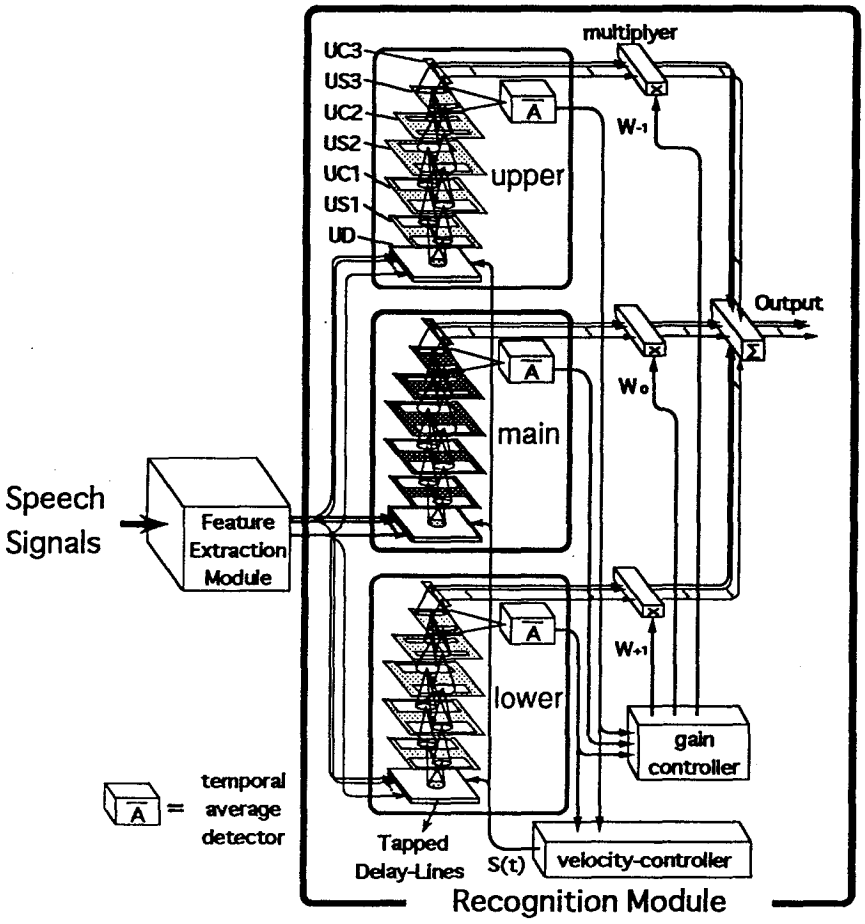


図 50: 認識部の構造

ただし，先のモデルを音声認識に適用するには，三つある認識ブロックのネットワークサイズ (特に各細胞の結合領域等) の選択に，注意が必要である．

これは，音声スペクトルパターンが，第 2 章で認識対象としていた文字パターンのように，区別のつき易いものではなく，違った音韻でも，非常に似通ったパターンが多いから

である．すなわち，各認識ブロックの，変形に対する許容能力が大き過ぎると，逆に区別しなければならない音韻セットが区別できなくなってしまう．

そこで，本節では図 50 で示す時系列パターン認識モデルのうち，認識ブロック一つ分の構成方法を中心に説明していくことにする．

ここでは，一ブロックは，U_S，U_C層の組を 3 段持つとしている．各段の役割は，1 段目で，異なるフォルマント形状を抽出し，2 段目で音韻識別を行なう．そして，最上位段である 3 段目は，単語を認識する．次の節からは，このブロック一つ分の構造を，音韻識別を行なう U_D層〜U_{C2}層の構成と，単語識別を行なう U_{S3}〜U_{C3}層の構成に分けて説明していく．

4.3.1 U_D層〜U_{C2}層の構成

特徴抽出部の出力は，CF，FM 上昇，FM 下降の三つに分かれる．これら三つは，それぞれ U_D層に並列に入力される．図 51 は，特徴抽出部から，一つの認識ブロックへの結合を示したものである．図から分かるように，CF，FM 上昇，FM 下降の三つの特徴はそれ

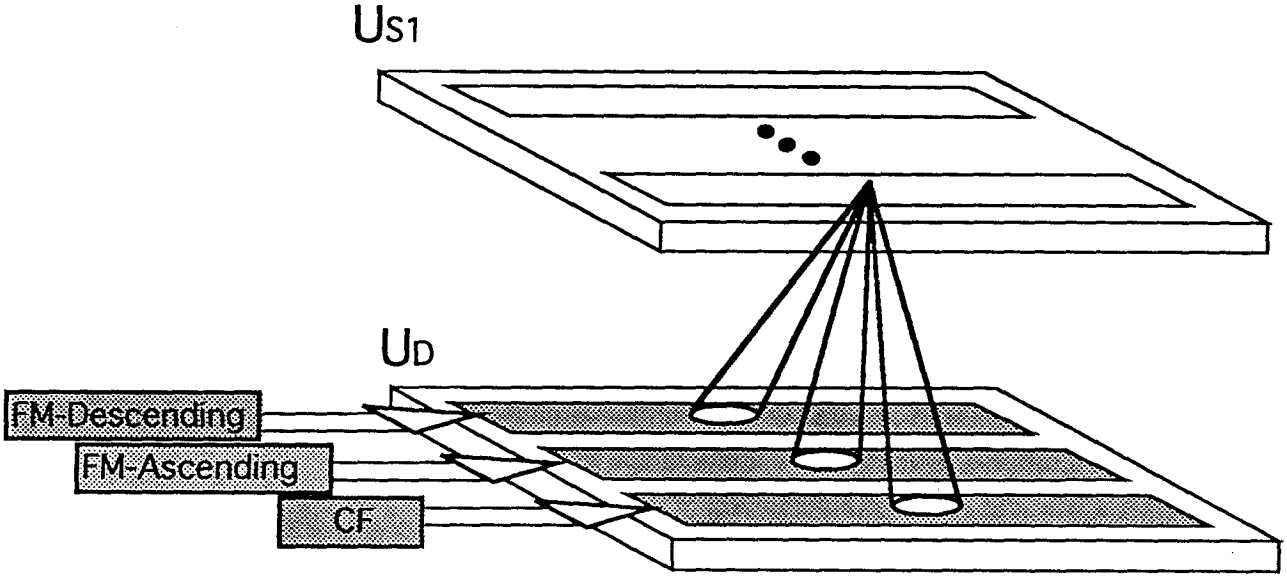


図 51: 特徴抽出部と各認識ブロックとの結合関係

ぞれ U_D層上の異なる面に入力される．この三つの面は，U_S，U_C層における細胞面に相当

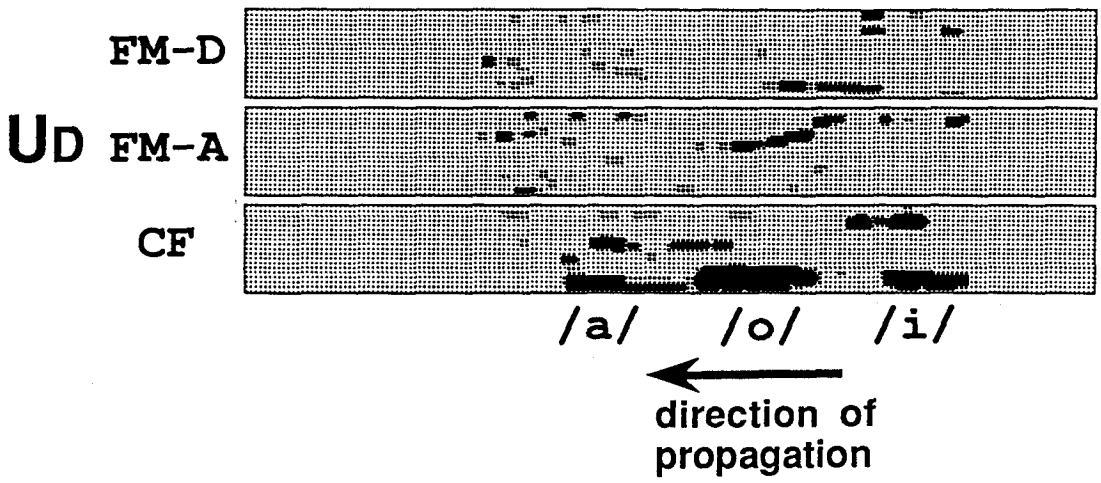


図 52: U_D層上に展開された時空間パターンの例

するものである[28].

U_{S1}層の一つの S 細胞は，この U_D層上の三つの面からの信号を同時に受け取る．ただし，一つの S 細胞の各面との結合領域は，全て同じ時間領域であり，且つ同じ周波数領域である．

ここで，特徴抽出部の出力が，U_D層上に，どのように空間展開されるかを説明しよう．CF 抽出部の出力は，音声信号の母音部分のフォルマントの時空間パターンに相当する．そして，FM 上昇，FM 下降抽出部の出力については，音節の周波数変化部分の時空間パターンに相当する．これらの出力が，U_D層上に空間展開されると，図 52 のようになる．しかし，このような時空間パターンから，異なる母音や音節を識別するには，かなり細かな差異を手がかりにする必要がある．これを，母音の識別を例にとって考えてみよう．

母音の識別には，周波数一定成分 (フォルマント) の周波数位置の違いが，その識別の手がかりとなる．そこで，周波数一定成分を抽出する CF 抽出部の出力が U_D層上で展開されたパターンを考える．図 53 は，日本語の母音/i,e,a,o,u/を入力した場合の U_D層上での CF 部分の時空間パターンの摸式図である．図から分かるように，各母音は，フォルマント位置は互いに似通っている場合が多い．特に，/o/と/u/のフォルマント位置の違いはごくわずかである．

[28]本来ならば，式 (1) を，複数の面を持つように表現を書き換えるべきであるが，ここでは省略する．

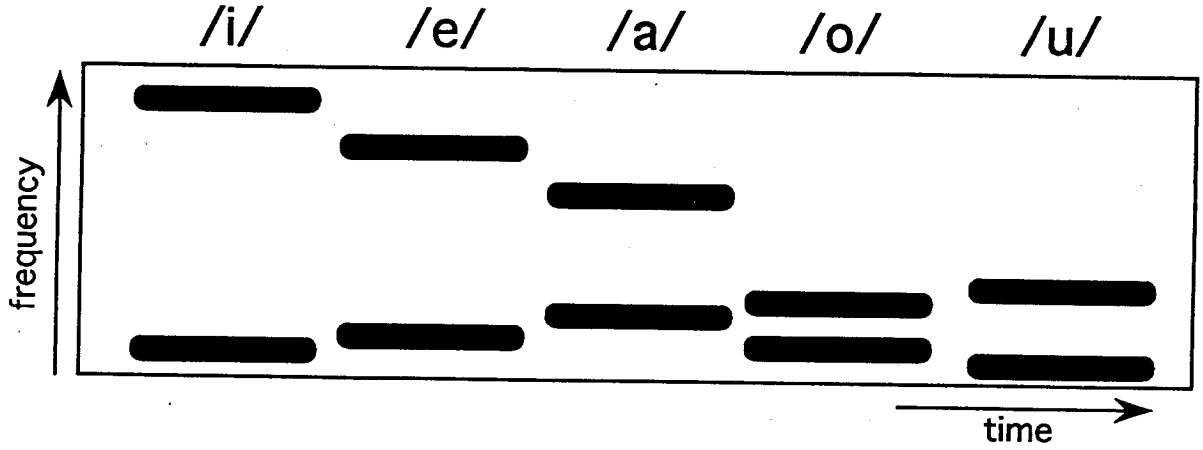


図 53: 母音/i,e,a,o,u/に対するフォルマント位置

一方，ネオコグニトロン型の構造を持つネットワーク特徴として，入力パターンの (U_D 層上のパターン) の位置ずれや，変形に対して影響されることなく，それを認識する点が挙げられる．この能力は，U_C層の S 細胞が，同一の特徴を異なった位置から抽出し，U_C層を構成する C 細胞が，U_S層の細胞の出力の位置ずれを少しづつ許容するというプロセスを，複数段にわたって繰り返すことによって実現されている．

従って，ネオコグニトロン型の構造を持つブロックによって先のような母音の弁別を可能にするためには，各ブロックの位置ずれ (特に周波数方向) の許容能力を抑えるように各細胞の入力結合の結合領域を選択する必要がある．

そこで，後に示す計算機シミュレーションでは，U_D層〜U_{C2}層の各層の細胞の結合領域 (周波数軸方向) を次の図 54 のように決定した．図から分かるように，U_{S1}層の一つの S 細胞は，U_D層の周波数方向の約 1/3 の領域からの入力を受け取るようになっている．また，U_{C1}層の C 細胞は，前段の S 細胞の出力の位置づれを許容する役目をするが，許容し過ぎないように，比較的狭い結合領域を持っている．U_{S2}層の細胞は，前段のほぼ全域の細胞からの信号を受け取っている．

このような結合によって，U_{S1} 層の 1 個の S 細胞は，母音/o/,/u/などのように，第一フォルマント (周波数の低い方) と第二フォルマント (周波数の高い方) が接近して存在する場合には，この第一第二フォルマントの組を観測することになる．また，/a/,/i/などのように第一第二フォルマントが離れている場合には，各々のフォルマントを別々に観測することになる．

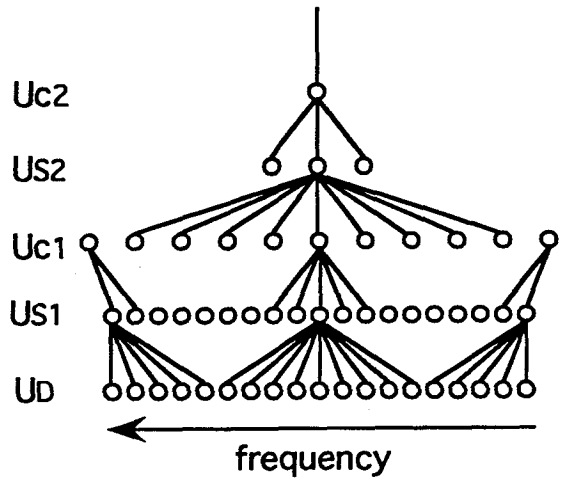


図 54: U_D層〜U_{C2}層までの各層の細胞の周波数軸方向の結合領域

従って，学習終了後のU_{S1}層では，例えば，母音/o/と/u/のように，第一第二フォルマントが接近し，かつ似通った母音の場合でも，第一第二フォルマントの相対位置の違いによって，U_{S1}層上のそれぞれ別の細胞面のS細胞が反応し，後の区別の手がかりを与えると期待される．また，/a/,/i/のように第一第二フォルマントが離れた母音が入力された場合には，個々のフォルマントに対して，U_{S1} 層上の異なる周波数位置のS細胞が反応する．つまり，このS細胞の出力する位置の違いが，後の区別の手がかりを与える[29]．

U_{S2}層は前層のほぼ全域からの入力を受けている．従って，U_{S1}層の，出力を出す細胞面の違いと，個々のS細胞の出力位置の違いとを検出し，それを基に母音を識別する．例えば，母音/a/と/i/に対するU_{S1}層上の反応は，主に第二フォルマントに対応するS細胞の出力位置が異なる．U_{S2}層では，この反応位置の差を手がかりに/a/と/i/を識別することになる．

U_{C2}層のC細胞は，U_{S2}層の全域から入力を受け，周波数軸方向への並びを持たない(時間軸方向には並んでいる)．音韻識別はこのU_{C2}層で完了する．U_{C2}層のC細胞の，時間軸方向の結合領域の長さは，標準的なスピードで喋った単語の一つの音韻の時間長とほぼ同一の長さに選ばれている．

[29]母音が入力されるとこれらのS細胞は連続反応を示すことになる．これは，大脳聴覚皮質の細胞の性質として，入力のonsetやoffsetに反応すると言う知見[1]と矛盾する．しかし，1.2.6節で述べたように，丸山ら[17][18][19] [20]は，ネコのA1野の細胞のうち，75%もの細胞が特定の刺激に対して連続反応を示すと指摘している．従って，これらのS細胞のように連続反応を示す細胞を仮定しても不自然ではないかもしれない．

4.3.2 U_{S3},U_{C3}層の構成

U_{S3}層，およびU_{C3}層では，単語の識別が行なわれる．これらの層の細胞は，周波数軸方向への並びはなく，全て時間軸方向へのみ並べられている．U_{S3}層の一つのS細胞は，前段のU_{C2}層の出力を，一定の時間軸方向の幅をもって入力を受け取っている．この時間軸方向の幅は，記憶させたい単語の時間長よりも長く設定されている[30]．U_{C3}層の1個のC細胞は，U_{S3}層の特定の細胞面上の全てのS細胞からの出力を受け取っており，単語認識結果を出力する．このU_{C3}層の一つの細胞面におけるC細胞の数は1個である．

4.4 学習法

学習法については，3.4節で提案したものとほぼ同一である．ここでは，これを第2章で提案した時系列パターン認識モデルに合うように表現し直すことにする．ただし，ここではこの学習法の意味については詳しく触れず，数式表現のみを示すことにする．この学習法の意味等の詳細については，第3章を参照されたい．

先にも述べたように，学習は主にメインブロック ($\xi = 0$) で行なわれ，両側のサブブロックの入力結合については，メインブロックの入力結合と同じ強度を持つように更新されるとする．以後，メインブロックの各細胞の，出力やパラメータを表す際には，$\xi = 0$ として表すことにする．例えば，メインブロックのS細胞の出力は，$u^{\xi}_{Sl}(\boldsymbol{n}, k)$ を単に，$u^{0}_{Sl}(\boldsymbol{n}, k)$ と表す．また，この入力結合の強度については，全てのブロックについて同じであるとしているので，ブロックを表す記号は含まれていない．

また，$a_l(\boldsymbol{\nu}, \kappa, k), p_l(\boldsymbol{\nu}, \kappa, k), c_l(\boldsymbol{\nu}, \kappa, k), b(k), r^{\xi}_l(k)$ をそれぞれ，$a_l(\boldsymbol{\nu}, \kappa, k, t), p_l(\boldsymbol{\nu}, \kappa, k, t), c_l(\boldsymbol{\nu}, \kappa, k, t), b(k, t), r^{\xi}_l(k, t)$ で表現する．ここに t は，学習ステップを表す．

4.4.1 興奮性結合の更新

表現を簡潔にするため，$u^{0}_{Sl}(\boldsymbol{n}, k)$ をメインブロックの第 l 段，第 k 細胞面の位置 $\boldsymbol{n}$ のS細胞とする．

2.3節でも述べたように，$u^{0}_{Sl}(\boldsymbol{n}, k)$ の興奮性可変結合の強度 $a_l(\boldsymbol{\nu}, \kappa, k)(\geq 0)$ は，$a_l(\boldsymbol{\nu}, \kappa, k) =$

[30]学習終了後には，この結合領域は，単語の時間長とほぼ等しくなる．

$\{c_l(\boldsymbol{\nu},\kappa,k)\}^2 \cdot p_l(\boldsymbol{\nu},\kappa,k)$ で表される．ここに $p_l(\boldsymbol{\nu},\kappa,k)(\geq 0)$, $c_l(\boldsymbol{\nu},\kappa,k)(\geq 0)$ は，興奮性結合の強度を決める変数である．この $p_l(\boldsymbol{\nu},\kappa,k,t)$ は，学習期間中，次のように更新される．

この変数の初期値は，十分に小さな正の値を持つとする $(0 < p_l(\boldsymbol{\nu},\kappa,k,0) \ll 1)$.

$$\Delta p_l(\boldsymbol{\nu},\kappa,k,t) = -\alpha_l \cdot f_l[b_l(k,t)] \cdot (p_l(\boldsymbol{\nu},\kappa,k,t) - p_0) + \delta_l(\boldsymbol{n},k,t) \cdot q_l[u_V^0(\boldsymbol{n},k,t)] \cdot u_{Cl-1}^0(\boldsymbol{\nu}+\hat{\boldsymbol{n}},\kappa,t) \tag{44}$$

ここに，$\alpha_l(<1)$，$p_0 \ll 1$ は，共に正定数である．

式 (44) の第 1 項は，減衰項である．この項の中で，関数 $f_l[x]$ は，減衰率を決定する関数で，次で表される．

$$f_l[x] = 1 - \frac{1-\varepsilon}{1+\exp[-\zeta(x-\theta_l)]} \tag{45}$$

ここに$\varepsilon(\ll 1)$ である．ζ は，関数 $f_l[x]$ が，x がどの程度の値になったときに飽和するかを決定する正定数である．

式 (44) の第二項は，強化項である．この項の関数$\delta_l(\boldsymbol{n},k,t)$ は，細胞 $u_{Sl}^0(\boldsymbol{n},k)$ が時刻 t にシードセル (学習における核) に選ばれた時には 1 を取り，そうでなければ 0 を取る関数である．

関数 $q_l[x]$ は，強化スピードを表す関数で，次で表される．

$$q_l[x] = \frac{Q_l}{\sigma_{ql}+x,} \tag{46}$$

ここに Q_l は，正定数で，強化スピードを決定する．σ_{ql} は，関数 $q_l[x]$ の飽和度を決定する正定数である．

4.4.2 S 細胞のパターン選択性の変化

変数 $r_l^0(k,t)$ は，メインブロックの S 細胞のパターン選択性を決定する変数である．この変数は，その抑制性可変結合の強度 $b_l(k,t)$ に応じて変化し，$b_l(k,t)$ が小さいときには，$r_l^0(k,t)$ も小さく，$b_l(k,t)$ が大きいときには，$r_l^0(k,t)$ も大きくなる．すなわち，

$$r_l^0(k,t) = R_l \cdot \frac{b_l(k,t)}{\sigma_{rl}+b_l(k,t),} \tag{47}$$

ここにσ_{rl} は，$r_l^0(k,t)$ の飽和度を決定する変数である．3.4.3節でも述べたように，この変数は通常学習の初期段階では大きな値に設定しておき，学習が進行するに従って，徐々に小さな値にした方が望ましい．これは，様々なサイズのパターンを学習する際に必要となる．ただし，後の計算機シミュレーションでは，この変数を一定値に固定して学習させている[31]．

4.4.3 S 細胞と V 細胞の結合領域の更新

S 細胞と V 細胞の結合領域は，その S 細胞が抽出する特徴のサイズに合わせて小さくする (その特徴が S 細胞の元の結合領域よりも小さい場合)．この結合領域の更新は変数 $c_l(\boldsymbol{\nu},\kappa,k,t)$ の値を減少させる事によって，行われる．具体的には，その S 細胞の興奮性可変結合の値と提示されているパターンとの間に，ある一定値以上の差がある場合に $c_l(\boldsymbol{\nu},\kappa,k,t)$ を減少させる．すなわち，

$$\Delta c_l(\boldsymbol{\nu},\kappa,k,t) = \alpha_l' \cdot f_l[b_l(k,t)] \cdot (c_l'(\boldsymbol{\nu}) - c_l(\boldsymbol{\nu},\kappa,k,t))$$
$$-\delta(k,\boldsymbol{n},t) \cdot q_l' \cdot c_l(\boldsymbol{\nu},\kappa,k,t) \cdot E\left[\tilde{u}_{Cl-1}^0(\boldsymbol{\nu},\hat{\boldsymbol{n}},\kappa,t), \tilde{p}_l(\boldsymbol{\nu},\kappa,k,t)\right]. \tag{48}$$

である．この式の第一項は，回復項で，$u_{Sl}^0(\boldsymbol{n},k)$ がシードセルに選ばれなかった際には $c_l(\boldsymbol{\nu},\kappa,k,t)$ を，後の再利用に備えて，その初期値 $c_l'(\boldsymbol{\nu})(\geq c_l(\boldsymbol{\nu},\kappa,k,t))$ に回復させる．この初期値 $c_l'(\boldsymbol{\nu})$ は，$|\boldsymbol{\nu}|$ が大きくなるに従って小さくなる．第一項の$\alpha_l'(<1)$ は回復率を決定する正定数である．関数 $f_l[\ ]$ は回復率を決定する関数で，式 (45) と同一である．

式 (48) の第二項は，$c_l(\boldsymbol{\nu},\kappa,k,t)$ を減衰させる項である．この項の中の $E(x,y)$ は，x と y の間に一定の割合 ($\theta' > 1$) 以上の差がある場合に正の値を取り，それ以外ならば 0 を取る関数で次式で表される．

$$E[x,y] = \varphi[x - \theta' \cdot y] + \varphi[y - \theta' \cdot x]. \tag{49}$$

また$\tilde{u}_{Cl-1}^0(\boldsymbol{\nu},\hat{\boldsymbol{n}},\kappa,t)$ は，$c_l(\boldsymbol{\nu},\kappa,k,t) \times u_{Cl-1}^0(\boldsymbol{\nu}+\hat{\boldsymbol{n}},\kappa,t)$ を正規化した値を表し，次式で表現される．

$$\tilde{u}_{Cl-1}^0(\boldsymbol{\nu},\hat{\boldsymbol{n}},\kappa,t) = \frac{c_l(\boldsymbol{\nu},\kappa,k,t) \cdot u_{Cl-1}^0(\boldsymbol{\nu}+\hat{\boldsymbol{n}},\kappa,t)}{u_{Vl}^0(\hat{\boldsymbol{n}},k,t)} \tag{50}$$

[31] U_S層の S 細胞の結合領域の広さを，学習するパターンに比べて極端に大きくしない限りは，パラメータ σ_r^t は固定していても差し支えない

同様に$\tilde{p}_l(\boldsymbol{\nu},\kappa,k,t)$ は，$c_l(\boldsymbol{\nu},\kappa,k,t) \times p_l(\boldsymbol{\nu},\kappa,k,t)$ を正規化した値を表し，次の式で表される．

$$\tilde{p}_l(\boldsymbol{\nu},\kappa,k,t) = \frac{c_l(\boldsymbol{\nu},\kappa,k,t) \cdot p_l(\boldsymbol{\nu},\kappa,k,t)}{b_l(k,t)}. \tag{51}$$

これらは式 (44) (50) の中の $u^0_{Cl-1}(\hat{\boldsymbol{n}}+\boldsymbol{\nu},\kappa,t)$ を $u^0_D(\hat{\boldsymbol{n}}+\boldsymbol{\nu},t)$ で置き換えれば，$l=1$ の場合にも成立する．

4.5 音声認識実験

表2に，このシミュレーションで使用した各認識ブロックのネットワークサイズを示す．

Layer	Number of Cell-Planes	Size of Cell-Plane (f-axis) × (τ-axis)	Sizes of Connecting Regions of Cells (f-axis) × (τ-axis)
U_D	3	19 × 181	——
U_{S1}	10	19 × 181	9 × 5
U_{C1}	10	11 × 93	5 × 11
U_{S2}	15	3 × 93	9 × 5
U_{C2}	15	1 × 49	5 × 11
U_{S3}	1	1 × 25	1 × 25
U_{C3}	1	1 × 1	1 × 25

表 2: 各認識ブロックのネットワークサイズ

日本語の単語「あおい」,「みどり」,「きいろ」(それぞれ．/aoi/,/midori/,/kiiro/で表す) を，男性が普通のスピードで一度だけ発音したもののみを学習に使用した．この音声は，テープレコーダに録音し，SUN-SPARC-STATION-2 において，8kHz のサンプリング周波数で A/D 変換した．

これを先に 4.2節で説明した入力部によって，1〜16Bark の周波数範囲 (75Hz〜3200Hz) で，音声スペクトルに変換し，特徴抽出部には 1 フレームの時間長を 1msec として入力されるようにした．認識部の入力 (即ち特徴抽出部の出力) では，1 フレームの時間長を 10msec とした．学習期間中は，このようにして作成した 3 個の単語の音声データを無音期間と共

にランダムな順序でシステムに提示した．ただし，主に学習を行なうのはメインブロックで，上下のサブブロックはメインブロックと同じ結合強度を持つとしている．この学習期間中には，速度制御は行なわれず，U_D層のパターン伝搬スピードはその初期値に固定されている．ただし簡単のため，式(47)のパラメータσ_{rl}は，学習期間中は一定値に固定した[32]．

認識期間中は，各ブロックの変形に対する許容度を高めるため，パラメータ$r^{\xi}(k)$は学習期間中の値よりも小さく設定した(式(47)のR_lを小さくした)．図55に学習パターン/aoi/に対するメインブロックの各U_D層とU_C層の出力例を示した．この図は，単語/aoi/が入力された直後の各層の出力を表している．この図から分かるように，U_D層の出力は特徴抽出ブロックの出力を空間展開したものになっており，音声信号の周波数一定成分(CF)，周波数上昇成分(FM-A)，周波数下降成分(FM-D)の三つの特徴が抽出されているのが分かる．U_{C1}層では，個々のフォルマントに強く反応する細胞や複数のフォルマントが合わさったものに強く反応するもの，及びU_D層上のFM上昇下降成分に反応する細胞も見られる．例えば，図55のU_{C1}層の上から一番目の細胞面は，母音/a,o,i/の個々のフォルマントに強く反応している．これに対して，U_{C1}層の上から四番目の細胞面は，主に母音/o/のように，第一第二フォルマントの組に反応するが，学習時よりもパラメータ$r^{\xi}(k)$を小さくしているために，/a/,/i/の個々のフォルマントにも反応している．U_{C2}層では，母音/a/,/o/,/i/に反応する細胞や，音節/oi/に反応する細胞もある[33]．そしてU_{C3}層では，単語/aoi/に対応する細胞が反応していることが分かる．この結果は，本システムが学習期間中に，音声信号中の母音や音節そして単語などの重要な情報を検出する能力を自動的に獲得していったことを示している．

次に，速度制御を行なわせながら，先の三つの単語を女性が様々なスピードで喋った音声信号をシステムに提示した(学習パターンは，男性が普通のスピードで喋った単語であった)．図56は，本システムの学習に使用したパターンと，その後に続く変形パターンの両方に対する出力を示している．この図は，メインブロックのU_D層上のパターンのうち，CF部分のみを示している[34]．この図から，速度制御部は，各単語の時間長が学習させた単語

[32] U_S層のS細胞の結合領域の広さが，学習するパターンに比べて極端に大きくしない限りは，パラメータσ_{rl}は固定していても差し支えない

[33] 単語/aoi/,/midori/,/kiiro/を構成する母音は/a/,/o/,/i/のみなので，ここではこれ以外の母音について反応する細胞面は出来上がっていない．

[34] 実際には，図56は，特徴抽出部の出力をその時間軸のスケールをDelay-Lineのスピードに合わせて変

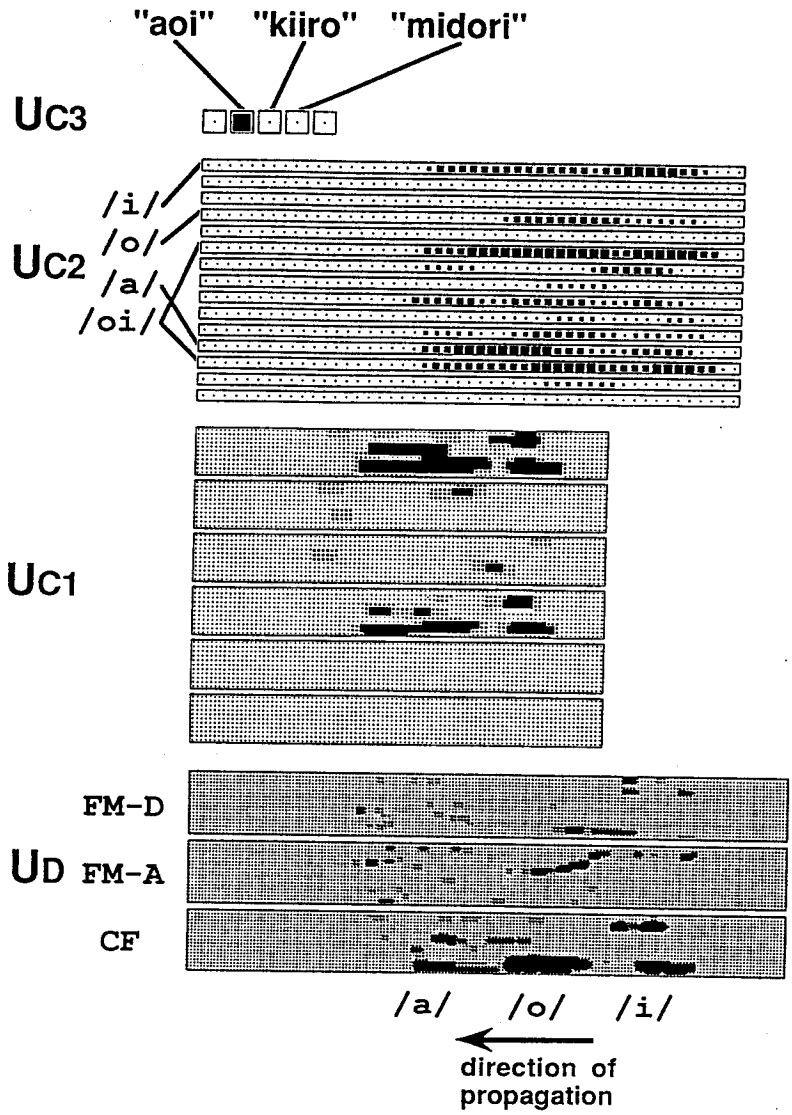


図 55: メインブロックの学習パターンに対する反応例

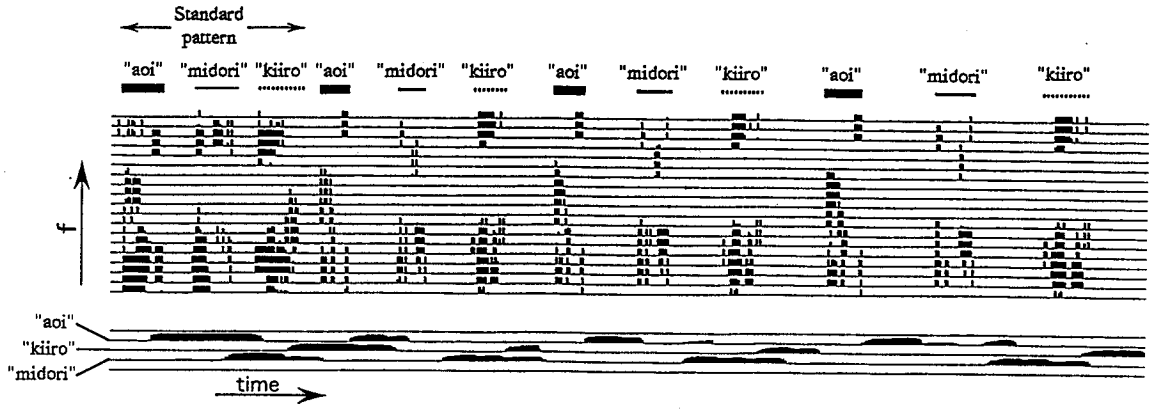
の時間長に近付くように，徐々に Delay-Line のスピードを調節していることが分かる．さらに，その認識出力は，ほぼ正確なものとなっている．

4.6 結言

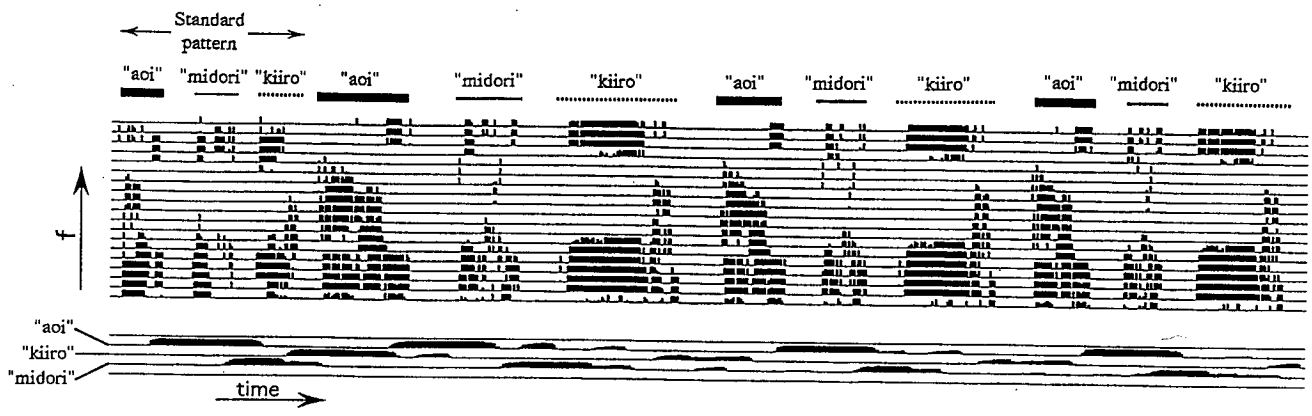
特徴抽出部と，速度可変 Delay-Line を使用した時系列パターン認識部で構成された音声認識システムを構築した．このシステムは，次の二つの仮定の基に構成されている．

一つは，蝸牛から，大脳聴覚皮質に至るまでの経路には，様々な音響特徴音響特徴に反

えて表示している．この時間軸のスケールは値 $1/S(t)$ に比例するようにしている．



(a) Responses produced by fast patterns



(b) Responses produced by slow patterns

図 56: 女性の声で様々なスピードで喋った音声に対するシステムの出力
(a) と (b) 共に，図の上方はメインブロックの Delay-Line の出力のうち CF(周波数一定) 成分を表し，下はそれに対応するシステムの出力を表す．各図で，Standard pattern に後続する単語の時間長が，徐々に Standard pattern における単語の時間長に近付いていることが分かる．例えば (b) の Standard pattern の直後にある/aoi/の時間長は，Standard pattern のものよりもかなり長いが，後の方では短くなっている．

応する細胞があるという解剖学的知見から来ている．たとえば，蝸牛神経核には，周波数の特定の方向への変化に反応する細胞が存在する．また，上丘には周波数一定音に反応する細胞も見つかっている．すなわち，大脳聴覚皮質では，これら抽出された特徴から音声信号を認識している可能性が高い．このような知見から，人間の音声認識プロセスには，

特徴抽出プロセスと，抽出された特徴データからの認識プロセスがあると仮定した．したがって，ここで構成したシステムも，特徴抽出部と認識部の二つのモジュールで構成されている．

第二の仮定は，人間の時系列パターンの認識プロセスでは，時系列パターン中の各特徴の継続時間長を無視することなくその時間伸縮に対処しているということである．このことは，既に1.3節と第2章でも述べた．これは例えば，モールス信号のようにトーン信号の継続時間長に重要な情報が含まれる場合でも，我々はそれらを，そのスピードに関わらず正確に識別できることから，容易に推測できよう．このシステムの認識部は，2.6節で示したように，同じ文字セットで構成されていながら，その継続時間の比の異なるパターンセット (例：短い 'T' の後に長い 'H' が来るパターンと長い 'T' の後に短い 'H' が来るパターン) が時間軸方向に伸縮を受けていても正しく区別できる．また，各認識ブロックが，ネオコグニトロンに似た構造を持つため，学習時に提示していなかったような変形パターンであっても正しく認識できる特徴を持っている．

ここでの計算機シミュレーションでは，小規模な単語音声認識しか行なっていないが，その結果は，このシステムが，不特定話者に対する音声をその伸縮に影響されることなく識別できるという高い汎化能力を持つことを示唆するものである．

以上のことから，ここで提案したシステムは，人間のいくつかの音声知覚現象を説明できる作業仮説となっている可能性があると共に，音声認識手法としての有効な手段になると考えられる．

第5章　総括

自動音声認識システムを構築する立場から，これまでにも多くの音声認識アルゴリズムおよび，および神経回路モデルが提案された．これらの多くは，音声信号の時間伸縮に対処するために，主に音声信号中の各特徴の現れる順序のみを検出する事によって，音声信号を認識している．すなわち，継続時間長を無視することによって，その時間伸縮に対処していた．この順序検出は，いずれも時間軸に沿った逐次的な処理によって実現されている．

一方，人間の音声知覚に関してはこれまでにも，心理学的側面からの調査が多く行われてきた．これらの調査結果の多くは，人間の音声処理が，時間軸に沿った逐次的な処理ではないことを示唆している．例えば，連続音声中の単語知覚に関しては，次のような現象が発見されている．すなわち，短い単語の多くは，その単語の語尾を聞き取った段階では，その単語を知覚されず，しばらくして，その単語の後に来る単語を聞き取った段階で，先の単語と，その後に来た単語が同時に知覚されるという．このような，単語知覚に付随する現象は，Elman らが提案した単語知覚モデル TRACE によって，比較的良く説明できることが知られている．彼らのモデルは，音声スペクトル (実際には疑似特徴) を，時間軸方向に空間展開してから処理する形式を採っている．すなわち，音声信号の時間構造を保持したまま処理し，音声信号中の各特徴の継続時間長を無視しない．このモデルでは，各単語の時間長も含めたテンプレートが時間軸方向に単位時間毎に並べられ，音声信号のあらゆる時点が単語の始点あるいは終点であるという仮定の基にマッチングが行われる．しかし，この各単語に対するテンプレートの時間長が，固定されているため，音声信号の時間伸縮に対処することが難しいという問題点が指摘されている．すなわち，音声知覚のより良い作業仮説として，音声信号を時間構造を保ったまま処理し，且つその時間伸縮に対処するモデルが望まれていた．

本論文では，この人間の音声知覚に関する心理学的知見に基づいた，音声認識モデルについて研究した結果を述べた．第2章では，音声認識の前提となる，時系列パターン認識モデルを提案した．第3章では，先の時系列パターン認識モデルの学習手法を提案した．そして第4章では，このモデルを実際の音声認識に応用した例を示した．以降では，これらの結果を総括し，このモデルの寄与および課題を述べる．

第2章では，人間の音声知覚に関する心理学的知見に基づいた，時系列パターン認識モデルを提案した．参考にした心理学的知見は，音節/wa/と/ba/の識別の手がかりについてである．この音節/wa/と/ba/の識別には，そのフォルマントの遷移速度が重要な手がかりになり，/ba/の方が/wa/よりも速く遷移する．そして，この/ba/および/wa/のフォルマント周波数の遷移速度を人為的に一定にした場合，音節全体の時間長を短くした場合には/wa/に知覚される割合が増し，逆に，時間長を長くすると/ba/と知覚される割合が増す．これは人間が，音節の時間長から発話速度を知覚し，それを手がかりに音節の時間長を正規化しながら認識している可能性を示唆している．このような知見に基づき，時系列パターンのスピードを検出し，そのスピードによって入力パターンを正規化しながら認識する時系列パターン認識モデルを提案した．このモデルは，一つのメインブロックと，二つのサブブロックから構成される．メインブロックは，時系列パターンを認識する役目をし，二つのサブブロックは，時系列パターンのスピード (伸縮率) を検出する役目をする．各ブロックは，速度可変のタップ付き Delay-Line を入力層として持つ階層ネットワークで構成されている．ここでは，この階層ネットワークを，Delay-Line によって展開された時空間パターンの認識に使用する．この三つのブロックはいずれも同じ構造をしているが，Delay-Line のパターン伝搬スピードの比が異なっている．すなわち，上下のサブブロック及び中央のメインブロックの Delay-Line のスピードは，それぞれ速，遅，中，に設定されている．時系列パターンの伸縮率は，上下のサブブロックの出力の差を手がかりに計測される．入力パターンの伸縮率が計測されると，システムは，全ての Delay-Line の速度を連動制御し，メインブロックの Delay-Line 上を伝搬するパターンが標準パターンに近付くように正規化しながら認識する．計算機シミュレーションでは，音声信号を模倣した電光掲示板を流れる文字パターンのような時系列パターンを使ってその動作を確認した．この時系列パターンは，いずれも二つの文字パターンで構成されるが，各文字の時間長の比が異なっている．例えば，長い 'T' の後に短い 'H' が来るもの，短い 'T' の後に長い 'H' が来るものなどである．これらのパターンを第3章で提案した学習手続きを使って各ブロックに学習させた後，これらのパターンが時間軸方向に伸びたり縮んだりしたパターンを提示した．すると，このモデルはその時間伸縮に影響されることなく「長い 'T' の後に短い 'H' が来るもの」と

「短い 'T' の後に長い 'H' が来るもの」を区別した．しかし，このシステムの速度制御は，時系列パターンが入力され始めた初期の段階ではその効果が現れない．これは，時系列パターンの一部が Delay-Line 上に展開され，上下のサブブロックのどちらかがそれを認識した段階ではじめて伸縮率が検出されるために，制御が遅れるためである．従って，速度が次々に急峻に変化する時系列パターンが入力されると，それに追従することが難しくなる．実際には，システムの出力として，三つのブロックの出力の重みつき平均を採用しているので，このような場合においても，正しい認識出力は得られる．しかし，より遅れ時間の少ない速度制御法を考案する必要があろう．

第 3 章では，先の時系列パターン認識モデルの学習法を提案している．これまでにも，多くの神経回路における学習法が提案されていたが，時系列パターンの学習にはまだ課題が残されているものが多い．例えば，教師あり学習法を採用するとすれば，時系列パターンのどこからどこまでを一つのカテゴリーと認めさせるかを，人間が明示的に与える必要がある．これに対して，幼い子供が，親の音声を聞いているだけで，誰からも教えられること無く自然と言葉を覚えて行くことは，一般に良く知られている．従って，人間の時系列パターン学習は，教師あり学習だけで説明できるとは考え難い．これに対して，教師なし学習は，明示的に記憶するべきカテゴリーを示されなくとも，与えられた入力パターンの確率分布を近似的に獲得することによって，記憶するべきカテゴリーを自動的に発見できる点が興味深い．しかし，これまでの教師なし学習は，環境が静的な場合にはうまく働くものの，次々に変化する環境については，過去の記憶を必ずしもうまく保持できるとは言えなかった．そこでここでは，福島らが提案したネオコグニトロン型学習法を改良し，この問題を解決する学習法にした．この学習法は，一度記憶したパターンの出現確率が，ある一定値以上の場合にはその記憶を重要とみなして保持するが，そうでなければ，それを忘却し，新たな別のパターンを記憶することが出来る．このようにして一度重要とみなされ，保持された記憶については，忘却し難くなっており，後に環境が変わって記憶した事柄の出現確率が小さくなったとしても，その記憶を保持できる．これは例えば，日本で生まれ育った人が，アメリカで生活するようになると，日本語を忘れることなく英語を話すことが出来るようになるという事実の説明になろう．さらに，この学習法では，各細胞の受容

野のサイズをその細胞が抽出するべき特徴に合わせて適応的に変化させ，その前後に来るパターンに影響を受けることを極力避けられるような工夫がなされている．ただし，本学習法は，現段階では，いくつかのパラメータを人間が設定しなければならない．たとえば，出現確率がどれほど以上ならば重要なパターンとみなすかを決定するパラメータは，今のところ人間が設定しなければならない．これらのパラメータを如何にして自動的に設定するかが今後の課題である．

第4章では，第3章で提案した時系列パターン認識モデルを使って，音声信号を学習・認識するシステムを構築した．この音声認識システムを構築するにあたって，人間の聴覚系が二つのブロックに分かれると仮定している．すなわち，蝸牛から大脳聴覚野に至るまでの経路に相当する部分と，大脳聴覚野に相当する部分である．ここでは，前者を特徴抽出部とし，後者を認識部とした．特徴抽出部ではまず，蝸牛と同様に音声信号を周波数分割してスペクトルパターンに変換する．次にそのスペクトルパターンから，周波数一定成分，周波数上昇成分，周波数下降成分の三種類の特徴を抽出する．これは，蝸牛神経核において，周波数上昇，下降成分に反応する細胞が発見されていること，および下丘において，周波数一定成分に反応する細胞が発見されていることを参考にしている．認識部は，特徴抽出部で抽出されたデータから音声を認識する．計算機シミュレーションでは，男性が普通のスピードで発音したいくつかの単語をシステムに学習させた後，女性が様々なスピードで発音した単語をシステムに提示した．その結果このシステムは，音声信号の伸縮や，話者の違いに影響されること無く，それらを正しく認識した．ここでの計算機シミュレーションでは，小規模な単語音声認識しか行なっていないが，その結果は，このシステムが，不特定話者に対する音声をその伸縮に影響されることなく識別できるという高い汎化能力を持つことを示唆するものである．ただし現段階では，特徴抽出部において，子音に含まれるバースト音などに対応するような特徴を抽出する部分を省略しているため，音韻情報として，子音を抽出することが出来ない．従って，区別の付かない単語セットがあるものと考えられる．これについては，今後特徴抽出部を拡張することによって解決できよう．またこのモデルでは，異なる単語に対応するS細胞の間の相互抑制結合や，単語から音韻への遠心性経路などが用意されていないため，現段階では人間の単語知覚に関するいくつか

の諸現象を説明できない．だが，これらを解決するには，実際に遠心性経路や単語間の抑制性結合を用意すれば容易に解決できるものと考えられる．

以上，本論文では，人間の音声知覚に付随する現象を参考にした時系列パターン認識モデルについて述べた．ここで提案したモデルは，人間の音声知覚における時間方向の処理に主眼を置いて構築したものである．このモデルは，音声信号をその時間構造を破壊せずに処理し，且つその時間伸縮に対処するという人間の音声知覚プロセスを説明できる．ここでの計算機シミュレーションでは，小規模な音声認識実験しか行なっていないが，その結果は，このシステムが音声信号の伸縮や話者の違いに影響されること無く，それらを正しく認識できるという，高い汎化能力を持つことを示唆している．

この一方で，本モデルには，次のような課題が残されている．それは，このモデルでは，その Delay-Line が保持できる時間長よりも長いパターンは扱うことができないという点である．例えば，一つの文章に相当するような長い音声信号については，第 4 章で扱ったモデルでは扱えない．これを解決するもっとも簡単な方法として，このモデルの Delay-Line を，文章一つ分程度の時間長に引き伸ばすことが挙げられる．しかし，これだけ長い Delay-Line が生体の中に存在するか否かは定かではない．実際に生体の中に発見された Delay-Line が保持する時間長は，ごく短いものであった (1.2節の図 7参照)．これを解決する手段として，単語よりも高次の意味情報については，状態遷移モデルで記述する事が考えられる．例えば，Hidden Markov Model のように扱うとすれば，単語が一つの状態遷移に付随して発生するシンボルという形になろう．

人間の音声知覚の本来の目的は，話者の感情を，音声信号から再構成することと考えられる．この話者の感情に関しては，音声信号から得られるコンテキストだけに含まれるものではなく，抑揚など，さらに低次の特徴の中に含まれるものでもある (例えば，コンテキストが同じでも喋り方を変えただけで相手への意志の伝わり方がずいぶんと違ってくるものである)．今後このシステムが，単語と，音韻，そして音響特徴との相互作用によって音声を認識するシステムに発展し，認識している単語 (コンテキスト) と，音声スペクトルとの対応関係を把握することが可能となることを望む．すなわち，これによって，ノイズや他の音が混ざったような音声信号から，単語などのコンテキストに関する情報のみならず，

その単語に対応するスペクトルパターンを明確に切り出すことができ，それを基に，話者の感情を検出することが可能になるものと思われる．

謝辞

この研究を進めるに当たって，多く御指導と助言を頂いた，大阪大学基礎工学部生物工学科，福島邦彦教授に深謝する．また，本研究について多くの議論，助言を頂いた，大阪大学基礎工学部生物工学科，倉田耕治講師，および岡田真人助手にも深謝する．

また，本研究の共同研究者である，大阪大学基礎工学部生物工学科博士前期課程在学中の福田　愛氏には，第4章の特徴抽出部の改良，構築，およびデータの取り込みを中心に担当して頂いた．彼の協力無くしては，本研究が成り立たなかった事を強調しておきたい．

最後に，研究生活全般にわたってバックアップして頂いた生物工学科の職員および福島研究室の皆さんに深謝する．

参考文献

[1] D. P. Morgan and C. L. Scofield. *Neural Networks and Speech Processing*, pp. 9–40. Kluwer Academic Publishers, London, 1991.

[2] 甘利俊一, 中川聖一, 鹿野清宏, 東倉洋一（編）. 音声・聴覚と神経回路網モデル. オーム社, 1990.

[3] G Von Békésy. The vibration of the cochlear partition in anatomical preparation and in models of the inner ear. *Journal of the Acoustical Society of America*, Vol. 21, pp. 233–245, 1949.

[4] B. M. Johnstone, K. J. Taylor, and A. J. Boyle. Mechanics of guinea pig cochlea. *Journal of the Acoustical Society of America*, Vol. 47, No. 2, pp. 504–509, 1970.

[5] Khanna S. M. and et al. Basilar membrane tuning in the cat cochlear. *Sciense*, Vol. 190, pp. 1218–1221, 1982.

[6] 平原達也. 適応Q型非線形蝸牛フィルタ. 日本音響学会誌, Vol. 47, No. 5, pp. 327–335, 1991.

[7] Tatsuya Hirahara and Takashi Komakine. A computational cochlea nonlinear preprocessing model with adaptive q circuits. *International Conference on Acoustics, Speech, and Signal Processing'89*, Vol. S-10a, No. 8, pp. 496–499, 5 1989.

[8] N. B. Cant and D. K. Morest. The structural basis for stimulus coding in the cochlear nucleus of the cat. In C. Berlin, editor, *Hearing Science*, pp. 371–421. Collegi-Hill Press, San Diego, CA, 1984.

[9] L. A. Jeffress. A place theory of sound localization. *Journal of Comparative and Physiological Psychology.*, Vol. 41, pp. 35–39, 1948.

[10] 小西正一. フクロウの音限定位の脳機構. 科学, Vol. 60, No. 1, pp. 18–28, 1990.

[11] C. E. Carr and M. Konishi. Axonal delay lines for time measurement in the owl's brainstem. *Proceedings of the National Academy of Science of the United State of America*, Vol. 85, pp. 8311–8315, November 1988.

[12] Robert H. Helfert, Celleen R. Snead, and Richard A.Altschuler. The ascending auditory pathways. In Richard A.Altschuler, Richard P. Bobbin, Ben M. Clopton, and Douglas W. Hoffman, editors, *Neurobiology of Hearing The Central Auditory System*, pp. 1–25. Raven Press, New York, 1991.

[13] 大串健吾. 聴覚系の情報処理機構のモデル. 電子情報通信学会論文誌, Vol. (D)54-C, No. 4, pp. 332–339, 4 月 1971.

[14] Donald Wong. Cellular organization of the cat's auditory cortex. In Richard A.Altschuler, Richard P. Bobbin, Ben M. Clopton, and Douglas W. Hoffman, editors, *Neurobiology of Hearing The Central Auditory System*, pp. 367–387. Raven Press, New York, 1991.

[15] Richard A.Altschuler, Richard P. Bobbin, Ben M. Clopton, and Douglas W. Hoffman, editors. *Neurobiology of Hearing The Central Auditory System.* Raven Press, New York, 1991.

[16] 古川原誠, 丸山直滋. ネコ皮質聴ニューロンの鋸波状AM音に対する波形包絡選択性. 新潟医学会雑誌, 第 105 巻, No. 3, pp. 162–176, 3 1992.

[17] 斉藤勝則, 丸山直滋. ネコ聴覚領の純音ユニット及び帯域雑音ユニットについて. 新潟医学会雑誌, 第 96 巻, No. 8, pp. 515–531, 8 1981.

[18] 丸山直滋. 言語音と環境音の識別機序. 日本生理学会誌, Vol. 52, pp. 135–146, 1990.

[19] 工藤雅治, 丸山直滋. ネコ聴覚領のホルマント識別ニューロンについて. 新潟医学会雑誌, 第 96 巻, No. 5, pp. 311–326, 5 1982.

[20] 新沢秀範, 丸山直滋. ネコ聴覚野の帯域雑音受容ニューロンの反応成立要因について. 新潟医学会雑誌, 第 104 巻, No. 9, pp. 763–776, 9 1990.

[21] 福西宏有. 電位感受性色素を用いた脳の聴覚野の観測. 日本音響学会誌, Vol. 48, No. 5, pp. 313–319, 1992.

[22] Ikuo Taniguchi, Junsei Horikawa, Toshio Moriyama, and Masahiro Nasu. Spatio-temporal pattern of frequency representation in the auditory cortex of guinea pigs. *Neuroscience Letters*, No. 146, pp. 37–40, 1992.

[23] 赤木正人, 古井貞熙. 音声知覚における母音ターゲット予測機構モデル化. 電子情報通信学会論文誌, Vol. J69-A, No. 10, pp. 1277–1285, 1986.

[24] 北田宏, 中野馨. 運動理論に基づく音韻知覚神経回路モデル. 電子情報通信学会技術報告, NC91–22, 1991.

[25] J. L. McClelland and J. L. Elman. Interactive processes in speech perception: The trace model. In D. E. Rumelhart, McClelland J. L., and PDP Research Group, editors, *Parallel Distributed Processing*, pp. 58–121. Bradford Books, 2 edition, 1988.

[26] 片桐滋, 東倉洋一, 古井貞熙. 単音節知覚における時間情報の役割. 日本音響学会誌, Vol. 42, No. 2, pp. 97–105, 1986.

[27] J. L. Miller and A. M. Liberman. Some effects of later-occuring infromation on the perception of stop consonant and semivowel. *Perception & Psychophysics.*, Vol. 25, pp. 457–465, 1979.

[28] D. B. Pisoni, T. D. Carrell, and S. J. Gans. Perception of duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics.*, Vol. 34, pp. 314–322, 1983.

[29] A. G. Samuel. Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology : General*, Vol. 110, pp. 474–494, 1981.

[30] P. E. Rubin, M. T. Turvey, and P. van Gelder. Initial phonemes are detected faster in sopken words than spoken nonwords. *Perception & Psychophysics.*, Vol. 19, pp. 394–398, 1976.

[31] L. Nakatani and K. Dukes. Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, Vol. 62, No. 3, pp. 714–719, 1977.

[32] G. A. Miller, G. A. Heise, and W. Lichten. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology : General*, Vol. 41, pp. 329–335, 1951.

[33] R. A. Cole and J. Jakimik. A model of speech perception. In R. A. Cole, editor, *Perception and Production of Fluent Speech.* LEA, Inc, Hillsdale, NJ, 1980.

[34] W. D. Marslen-Wilson and L. K. Tyler. The temporal structure of spoken language understanding. *Cognition*, Vol. 8, pp. 1–71, 1980.

[35] P.Zwitserlood. The locus of the effects of sentential semantic context in spoken-word processing. *Cognition*, Vol. 32, pp. 25–64, 1989.

[36] F. Grosjean. The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics.*, Vol. 38, No. 4, pp. 299–310, 1985.

[37] 天野成昭. 単語知覚モデルの研究動向. 日本音響学会誌, Vol. 48, No. 1, pp. 20–25, 1992.

[38] V. R. Viswanathan, A. L. Higgins, and W. H. Russel. Design of a robust baseband lpc coder for speech transmission over 9.6 kb/s noisy channels. *IEEE TRANSACTIONS ON COMMUNICATIONS.*, Vol. COM–30, No. 4, pp. 663–673, 1982.

[39] 迫江博昭, 磯健一. ダイナミックニューラルネットワークの提案—神経回路網と dp マッチングに基づく新しい音声認識. 電子情報通信学会論文誌, Vol. J71-D, No. 7, pp. 1341–1344, 7 月 1988.

[40] Kiyoaki Aikawa. Phoneme recognition using time-warping neural networks. *The Journal of the Acoustical Society of Japan (E)*, Vol. 13, No. 6, pp. 395–402, November 1992.

[41] 二見亮弘, 星宮望. 相互結合型神経回路網の 状態遷移に基礎をおく時系列パターン認識の神経回路モデル. 電子情報通信学会論文誌, Vol. J71-D, No. 10, pp. 2181–2190, October 1988.

[42] 宮本弘之, 福島邦彦. 階層神経回路モデルによる時系列パターンの認識. 電子情報通信学会技術報告, NC89-83, 1990.

[43] S. Kurogi. Speech recognition by an artificial neural network using findings on the auditory system. *Biological Cybernetics*, Vol. 64, No. 3, pp. 243–249, 1991.

[44] 中川聖一. 音声認識における時系列パターン照合アルゴリズムの展開. 人工知能学会誌, Vol. 3, No. 4, pp. 414–423, 1988.

[45] 有木康雄. HMMを用いた英語の音素認識における継続時間長の効果. 電子情報通信学会論文誌, Vol. J75-D-II, No. 12, pp. 1993–2001, 12 月 1992.

[46] Y. Takebayashi, H. Tsuboi, H. Kanazawa, Y. Sadamoto, H. Hashimoto, and H. Hashimoto. A real-time speech dialogue system using spontaneous speech understanding. *IEICE TRANSACTIONS on Information and Systems*, Vol. E76-D, No. 1, pp. 112–120, JANUARY 1993.

[47] R. P. Lippmann and B. Gold. Neural-net classifiers useful for speech recognition. *Proceedings International Conference on Neural Networks*, Vol. 4, pp. 417–425, 1987.

[48] 中川聖一, 早川勲. シーケンシャルニューラルネットワークを用いた音声認識. 電子情報通信学会論文誌, Vol. J74-D-II, No. 9, pp. 1174–1183, 9 月 1991.

[49] K. Fukushima. A model of associative memory in the brain. *Kybernetik*, Vol. 12, No. 2, pp. 58–63, 1973.

[50] D. W. Tank and J. J. Hopfield. Neural computation by concentrating information in time. *Proceedings of the National Academy of Science of the United State of America*, Vol. 84, pp. 1896–1900, April 1987.

[51] 伊藤崇之, 福島邦彦. 神経回路網モデルによる時空間パターンの認識. 電子情報通信学会技術報告, MBE86–131, 1986.

[52] 伊藤崇之. 階層型神経回路モデルによる時空間パターンの認識. システム制御情報学会論文誌, Vol. 4, No. 1, pp. 21–27, 1991.

[53] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, Vol. 36, pp. 193–202, 1980.

[54] Alex Waibel. Modular construction of time-delay neural networks for speech recognition. *Neural Computation*, Vol. 1, No. 1, pp. 39–46, 1989.

[55] Erik McDermott, Hitoshi Iwamida, Shigeru Katagiri, and Yoh'ichi Tohkura. Shift-tolerant lvq and hybrid lvq-hmm for phoneme recognition. ATR 研究報告書, pp. 425–438, 1990.

[56] T. Kohonen. The "neural" phonetic typewriter. *IEEE COMPUTER*, Vol. 21, No. 3, pp. 11–22, March 1988.

[57] 溝口理一郎, 角所収. 知識工学を応用した音声認識. 日本音響学会誌, Vol. 42, No. 12, pp. 942–947, 1986.

[58] 荒井和博, 鬼山康人, 野村康雄, 山下洋一, 北橋忠宏, 溝口理一郎. 知識処理に基づく音声自動ラベリングシステム. 電子情報通信学会論文誌, Vol. J74-D-II, No. 2, pp. 130–141, 12 月 1991.

[59] 山内康一郎, 福島邦彦. 速度制御を利用した時系列パターン認識モデル. 神経回路学会第 2 回全国大会講演論文集, p. 124, 12 月 1991.

[60] 山内康一郎, 福島邦彦. 速度制御を用いた時系列パターン認識モデル. 電子情報通信学会技術報告, NC91-161, 3 月 1992.

[61] K. Yamauchi and K. Fukushima. Temporal pattern recognition model using velocity-controlled delay-lines. In *Proceedings of the 2nd International Conference on Fuzzy Logic and Neural Networks (IIZUKA'92)*, Vol. 2, pp. 771–774, July 1992.

[62] K. Fukushima and S. Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, Vol. 15, No. 6, pp. 455–469, 1982.

[63] 山内康一郎, 福島邦彦. 時系列パターン認識のためのネオコグニトロン型学習法の改善. 電子情報通信学会論文誌, Vol. J76-D-II, No. 10, pp. 2223–2232, 10 月 1993.

[64] Kunihiko Fukushima, Masato Okada, Kouichirou Yamauchi, Michihiro Ohono, and Kazuhito Hiroshige. Neocognitron with non-uniform receptive fields. In Stan Gielen and Bert Kappen, editors, *ICANN'93 Proceedings of the International Conference on Artificial Neural Networks*, pp. 994–997. Springer-Verlag, September 1993.

[65] J. J. Koenderink and J. van Doorn A. Visual detection of spatial contrast: Influence of location in the visual field, target extent and illuminance level. *Biological Cybernetics*, Vol. 30, No. 3, pp. 157–167, September 1978.

[66] Gian F. Poggio. Cortical mechanisms of binocular vision in the rhesus monkey. In O. Pompeiano and C. Ajmone Marsan, editors, *Brain Mechanisms and Perceptual Awareness.*, pp. 53–66. Raven Press, New York, 1981.

[67] D. E. Rumelhart, G. H. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, McClelland J. L., and PDP Research Group, editors, *Parallel Distributed Processing*, chapter 8, pp. 318–362. Bradford Books, 2 edition, 1986.

[68] von der Malsburg Chr. self-organization of orientation sensitive cells in the striata cortex. *Kybernetik*, Vol. 14, pp. 85–100, 1973.

[69] 福島邦彦. コグニトロンのパターン分離能力の向上. 電子情報通信学会論文誌, Vol. 62–A, No. 10, pp. 650–657, 1979.

[70] S. Amari and A. Takeuchi. Mathematical theory on formation of category detecting nerve cells. *Biological Cybernetics*, Vol. 29, pp. 127–136, 1978.

[71] T. Kohonen. Self-organized formation of topogically correct feature maps. *Biological Cybernetics*, Vol. 43, pp. 59–69, 1982.

[72] 山内康一郎, 神保孝志, 梅野正義. 新奇パターンを学習するための自己組織化機構. 電子情報通信学会論文誌, Vol. J74-D-II, No. 4, pp. 474–481, 4 月 1991.

[73] K. Yamauchi, T. Jimbo, and M. Umeno. Self-organizing architecture for learning novel patterns. *SYSTEMS and COMPUTERS in JAPAN (SCRIPTA TECHNICA, INC.)*, Vol. 23, No. 4, pp. 36–45, June 1992.

[74] 山内康一郎, 福田愛, 福島邦彦. 聴覚神経系を模した特徴抽出機構と速度制御機構を用いた音声認識. 電子情報通信学会技術報告, NC92-106, 3 月 1993.

[75] K. Yamauchi, M. Fukuda, and K. Fukushima. A speech recognition system consisting of auditory feature extracting cells and velocity-controled delay-lines — part ii. recognition module. In *International Joint Conference on Neural Networks, IJCNN'93 Nagoya Japan*, Vol. 1, pp. 259–262, October 1993.

[76] M. Fukuda, K. Yamauchi, and K. Fukushima. A speech recognition system consisting of auditory feature extracting cells and velocity-controlled delay-lines — part i. feature extracting module. In *International Joint Conference on Neural Networks, IJCNN'93 Nagoya Japan*, Vol. 1, pp. 255–258, October 1993.

[77] K. Yamauchi, M. Fukuda, and K. Fukushima. Speed invariant speech recognition using variable velocity delay-lines. submitted to *Neural Networks*, 1993.

[78] Takayuki Ito and Kunihiko Fukushima. A neural network model extracting features from speech signals (in japanese). *IEICE Trans, D-II*, Vol. J70-D, No. 2, pp. 451–462, 1987.

関連発表論文

修士論文

1. "ニューロンモデルを用いた文字認識に関する研究", 名古屋工業大学工学部電気情報工学科, 修士論文, 平成 3 年 2 月.

学術論文

1. 山内康一郎, 神保孝志, 梅野正義: "新奇パターンを学習するための自己組織化機構", 電子情報通信学会論文誌, **J74-D-II**[4], pp. 474–481, 4 月 (1991).

2. K. Yamauchi, T. Jimbo, M. Umeno: "Self-Organizing Architecture for Learning Novel Patterns", *SYSTEMS and COMPUTERS in JAPAN (SCRIPTA TECHNICA, INC.)*, **23**[4], pp. 36–45, June(1992).

3. 山内 康一郎, 福島 邦彦: "時系列パターン認識のためのネオコグニトロン型学習法の改善", 電子情報通信学会論文誌, **J76-D-II**[10], pp. 2223–2232, 10 月 (1993).

4. K. Yamauchi, M. Fukuda, K. Fukushima: "Speed Invariant Speech Recognition Using Variable Velocity Delay-Lines", *Neural Networks*, 投稿中, (1993).

国際会議論文

1. K. Yamauchi, K. Fukushima:, "Temporal pattern recognition model using velocity-controlled delay-lines", In *Proceedings of the 2nd International Conference on Fuzzy Logic and Neural Networks,Iizuka,Japan*, pp. 771–774, July(1992).

2. Kunihiko Fukushima, Masato Okada, Kouichirou Yamauchi, Michihiro Ohono, and Kazuhito Hiroshige:, "Neocognitron with non-uniform receptive fields.", In Stan Gielen and Bert Kappen, editors, *ICANN'93 Proceedings of the International Conference on Artificial Neural Networks*, pp. 994–997. Springer-Verlag, September(1993).

3. K. Yamauchi, M. Fukuda, K. Fukushima:, "A Speech Recognition System Consisting of Auditory Feature Extracting Cells and Velocity-Controled Delay-lines — Part II. Recognition Module", In *International Joint Conference on Neural Networks, IJCNN'93 Nagoya Japan*, pp. 259–262, October(1993).

4. M. Fukuda, K. Yamauchi, K. Fukushima:, "A Speech Recognition System Consisting of Auditory Feature Extracting Cells and Velocity-Controlled Delay-lines — Part I. Feature Extracting Module", In *International Joint Conference on Neural Networks, IJCNN'93 Nagoya Japan*, pp. 255–258, October(1993).

学会発表

1. 山内 康一郎, 神保 孝志, 梅野 正義:, "ニューロンモデルによる新しい自己組織化アルゴリズム", In 第 *37* 回応用物理学関係連合講演会講演予稿集, p. 749, March(1990).

2. 山内 康一郎, 神保 孝志, 梅野 正義:, "新奇パターンを学習する自己組織化機構とそのネオコグニトロンへの応用", In 電気関係学会東海支部連合大会講演論文集, p. 564, October(1990).

3. 山内 康一郎, 福島 邦彦: "速度制御を利用した時系列パターン認識モデル", 神経回路学会第 *2* 回全国大会講演論文集, 124, 12 月 (1991).

研究会発表

1. 山内康一郎, 神保孝志, 梅野正義: "新奇パターンを学習するための自己組織化機構とそのネオコグニトロンへの応用", **PRU 90**-126, pp. 9–14, 2 月 (1991).

2. 山内康一郎, 福島邦彦: "速度制御を用いた時系列パターン認識モデル", **NC91**-161, pp. 221–228, 3 月 (1992).

3. 山内康一郎, 福田 愛, 福島 邦彦: "聴覚神経系を模した特徴抽出機構と速度制御機構を用いた音声認識", **NC92**-106, pp. 89–96, 3 月 (1993).