■シンポジウム 読み書きへの学際的アプローチ

読み書きへの工学的アプローチ
——ニューラルネットによる試み——

筧 一彦*

要旨：人間の読み書きの機能のうち特に「読み上げ」の機能を取り上げ，その工学的実現法について解説した。工学的実現法としては従来からの辞書による単語知識，文法規則，文字から音素への変換規則，韻律の生成規則といった言語・音声の研究の結果を用いる方法を示し，単なる「読み上げ」だけでも多くの知識や規則が必要であることを示した。次にニューラルネットを用いて，文字から音素への変換を教師あり学習させる試みについて紹介し，従来の工学的実現法との相違，人間の学習との類似性について述べた。人間も同時相互制約条件下の課題に対して並列・分散処理を行なっていると考えられ，ニューラルネットはそれを考察する構成的方法の一つとなっている。 **神経心理学，6；82～89**

Key Words：読み，規則合成音声，ニューラルネット，教師あり学習，並列分散処理
reading, speech systhesis by rule, neural network, supervised learning, PDP

I はじめに

計算機の出現によって，人間の知的機能を計算機で実現する試みが継続して行なわれて来ており，この研究分野は人工知能（ＡＩ）と呼ばれている。一旦計算機によってある機能が実現されてしまうと，それに関する研究は人工知能とは呼ばれなくなる傾向がある。したがって，上記のＡＩの定義は必ずしも実態に即していない面があるが，一貫して研究されてきた広義のＡＩ分野には人間の読む，聞く，話す，書くという能力がある。

本来“読む・聞く”ということはその結果として意味を理解することであり“話す・書く”ということは伝達したい意図を言語音声や文章で表すことである。“聞く”ということに関しては音声をすべて正しく transcript するという，いわゆる音声タイプライタ的立場をとらずに音声の意味理解を目指すという研究も行なわれている。“読む”ということに関しては，読み上げる，すなわち漢字かなまじりのテキストを入力して，朗読音声を出力するいわゆるテキストからの音声合成システムがある。このようなシステムでは一般に意味理解に関する解析はほとんど行なわれず，係り受けの解析程度までである。意図を達成するために話し言葉にせよ書き言葉にせよ文章を生成するということは，人間と対話できる機械を作るという夢を実現する上で重要な部分であり，現状ではごく基礎的な研究が行なわれている。しかし限定的な場合を除き，その工学的実現はかなり将来のこととなろう。

かな表記に対応する記号列を与え，かな・漢字変換を行なわせるような日本語ワープロの機能を想定した場合にも，真の意味処理なしには十分なかな・漢字変換はできないが，一度使われた変換を変換候補の上位に持ってくる，あるいは，作られる文章の対象領域が決まっている

1990年2月17日受理
Engineering Approach to Realization of Reading Function——An application of a neural network——
*ＮＴＴ基礎研究所情報科学研究部，Kazuhiko Kakehi：NTT Basic Research Laboratories

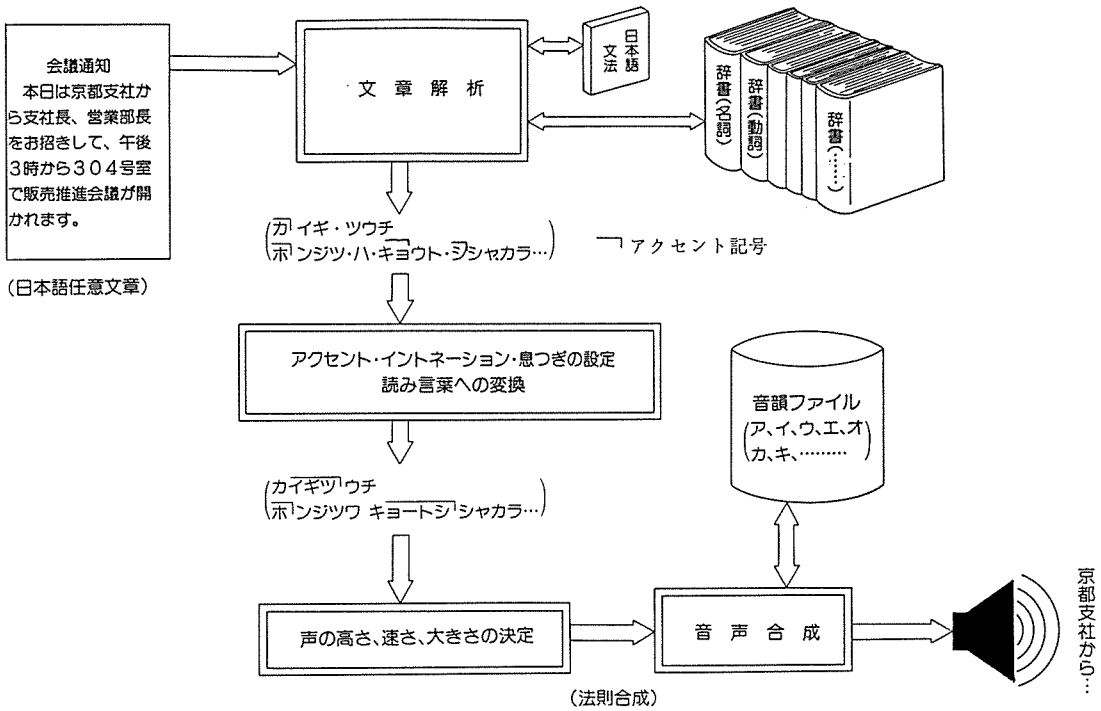


図1　テキストからの音声合成システム

ような場合，専用の用語を登録するなどの手段によって実用的には相当有効な変換が可能である。

工学的な実現は上記に述べたように必ずしも完全なものでなくとも，低次の処理で有効なものがあればそれが使われるようになるものである。工学的に考えられている処理の階層と人間の内部におけるそれは相当異なるものと考えられるが，本稿では“読み上げる”という機械について従来から行なわれている研究を紹介し，それによって読み上げるということにはどのような知識や規則が必要であるか，それらは通常どのようにして抽出されているかについて明らかにする。次いでそれらの機能の一部をニューラルネットにより実現しようとした例を述べ，学習がいかに行なわれるか，また従来の方法との対比を行なう。終りに人間の学習との類似性などについても紹介する。

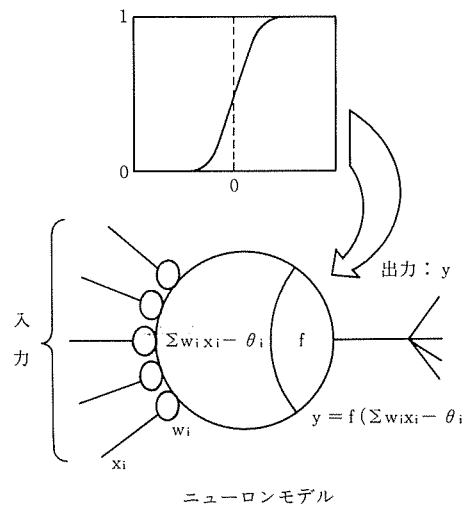
II　読み上げの工学的実現法

図1に現在工学的に実現されている日本語のテキスト音声合成システムの概略を示す（佐藤，匂坂，小暮，嵯峨山，1983)。入力が印刷文字であれば，これを認識する（読み取る）OCR（Optical Character Reader）という装置が技術的に完成しており，文字を読むことにほとんど問題はない。最近では見出しや，図面及びそのキャプションなどを本文から区別するような機能もできている。この文字の読み取り部分についても後述するニューラルネットの適用が試みられている。

ここではこのような文字に対応する符合列が入力された後の読み上げ機能の実現について述べる。このような機能はテキストからの音声合成と呼ばれており，一応実用となるものが作られている。例えば，最近新聞記事は電子編集となっているので文字に対応する符合入力がそのまま合成器の入力となるので新聞社において記事の校正をするのに使われている。しかし，その音声品質や読みに関する性能は人間のそれと比較するとかなり大きな差があるので，高品質，高性能化のための研究が進められている。

さて図1に見られるごとく，このようなテキスト読み上げのシステムでは，入力文章はあらかじめ装置に記憶されている文章で規則と辞書を用いて解析される。通常日本語の文章では分かち書きの習慣がないので，解析によって分かち書きし，それぞれの単語に対して辞書に書かれているアクセントを与える。しかし意味理解のような高次レベルでの解析は行なわれないので複合語の適切な区切り方や，漢字単語に存在する複数の読みの正しい選択（例えば，人気：ひとけ，にんき）などの問題は充分には解決されない。この段階で入力された漢字かな混じりの文章は各単語ごとに切り分けられてかな表記され，それぞれの単語単独のアクセントが辞書によって与えられる。しかし実際に文章として読まれるときには，動詞に種々の助動詞が接続する場合など単独の単語としてのアクセントとは異なったアクセントの移動が生じる。また格助詞の「ハ」などは実際には「ワ」と読まれるし，東京方言では文末の「デス」などの「ス」は母音部が発声されないなど，一連のかな文字表記から音素への変換規則が必要となる。個々の音素の継続時間の設定もかな表記から音声への変換に重要である。句，文などの単位でのイントネーションや息継ぎも，指定する必要がある。このように文字表記された文章を自然な朗読音声に変えるためには実に多くの知識や規則が必要とされており，かなり長い間の研究によってより良い規則の作成が進む一方，半導体技術の進歩によってより大きな知識データを装置に持たせることが可能にはなってきているが，まだ性能の面で改善の余地を残している。

このようにして声の高さ，速さ，大きさ，音韻性の決定が行なわれると後は人間の調音器官（舌，顎，口唇，声帯）などの役割を担う音声合成器が駆動され朗読音声が出力される。この最後の段階では出力音声を高品質化する上での課題が残されているが，全体としては先にも述べたように新聞の校正作業に実際に使用されるなど了解性の良い合成音が得られている。文章を読み上げるという過程だけでも上記のように多くの知識や規則が関わって文章の読み上げが



ニューロンモデル

図2 ニューラルネットを構成するしきい素子

工学的に実現されている。

ニューラルネットについては上記に述べた文字表記から音素表記への変換に関わるステップの実現を例にとってそれらの機能がいかにして学習されるか，またその学習にはどのような特徴があるかを説明する。

III ニューラルネットとは

ニューラルネットとはしきい素子が網状に接続されているものをいっている。このしきい素子は通常 Macaloch と Pitz（1943）が神経細胞をモデル化した図2のようなものが用いられる。図2に示すしきい素子は複数の入力を受けて一種類の出力を出す。入力（xi）にそれぞれ重み付け（wi）をして加算した後，閾値（θi）を差し引いたものを関数（f）によって非線形変換したものが素子の出力となる。この関数（f）としては図中に示すような単調に増加し，飽和するような形を持つシグモイド（S字状）関数が通常多く用いられる。このような飽和的特性は，神経細胞が強い入力を受けてもその出力が一定値で飽和するような特性を表現したものとなっている。このようなしきい素子が複数結合して得られるニューラルネットには接続の仕方によって網状のものや層状のものなどさまざまな型式のものがあり，それぞれに応じて異なった機能を持たせることができる。ここで

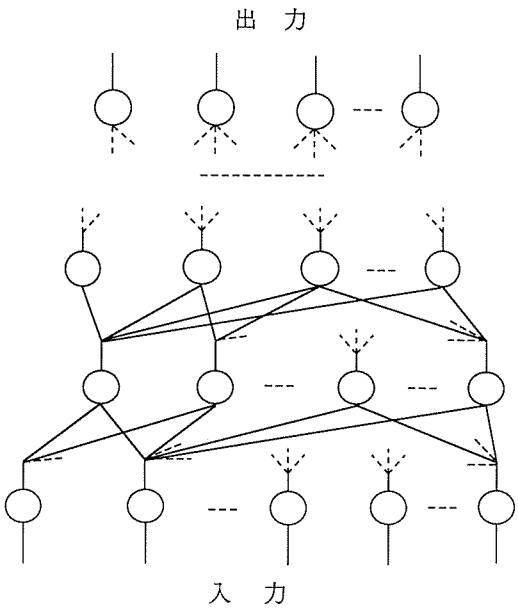


図 3　層状のニューラルネット

は，その中で最も代表的で，誤差逆伝播学習法（Error Back-Propagation Learning Method）が可能な層状神経回路について説明する（Rumelhart, McClelland, 他，1987）。

層状のニューラルネットとは図 3 に示すように同じ層内での素子同士の結合はなく，素子の出力は全て一つ上の層の素子の入力になっている。このようなニューラルネットは，ある入力をある出力へと写像する特性があり，分類の機能を持たせることができる。入力が与えられたとき，それに対して正しい（望ましい）出力を導くような規則や関数は分かっていないが，正しい出力は分かっているとき，正しい例を示すことによって学習を行なわせる方法を教師ありの学習と呼んでいる。この層状回路では上述した誤差逆伝播学習によりそれが可能となる。

簡単な例として 5 母音の認識の場合を図 4 に示し説明する（入野，河原，1989）。母音/e/の音声波形それ自身には相当の冗長性があるため，波形の自己相関関数を作り，相関関数の値

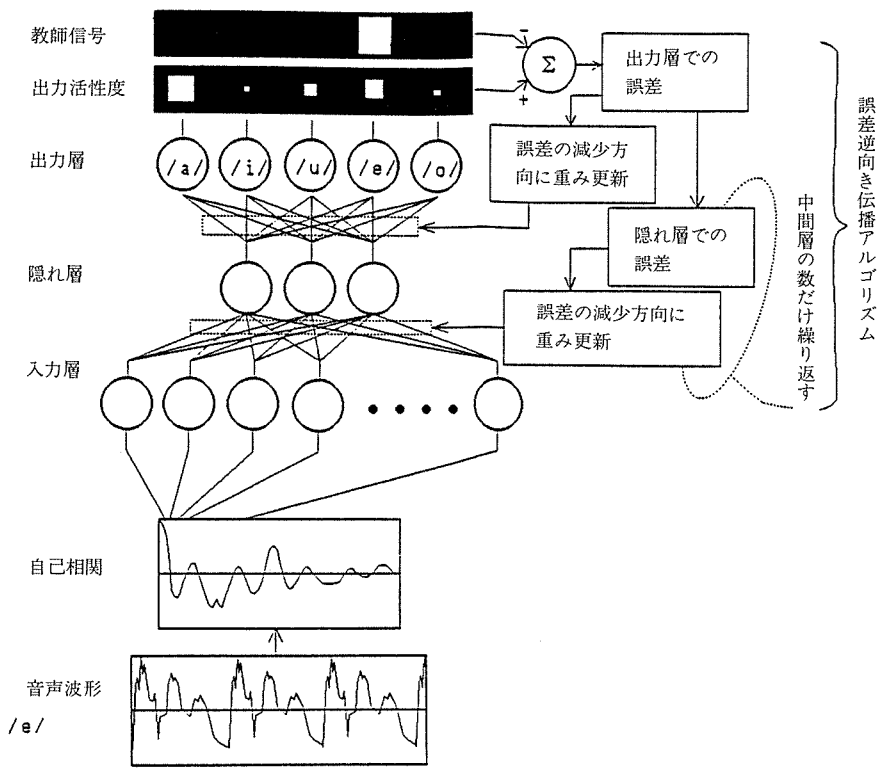


図 4　5 母音の識別を行なうニューラルネット

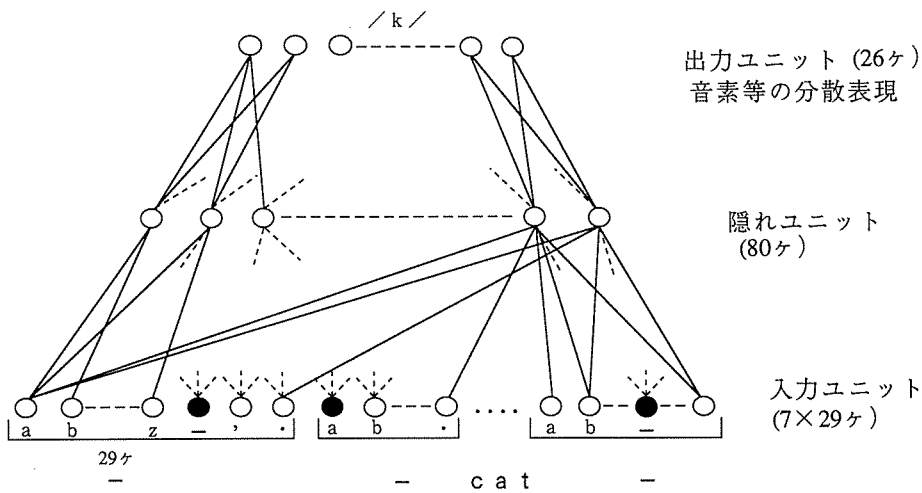


図5 英語の発音を学習する並列ネットワーク
（T. J. Sejnowski & C. R. Rosenberg）

を第一段目の入力とする。入力層の素子はその出力を二層目の隠れ層に送る。各隠れ層の素子は先に説明したごとく入力に重み付けをして加算し，非線形変換した後にその出力を出力層の素子に伝える。このようにして得られた出力層の素子の活性度（出力値）が四角形の大きさで図示されている。この各出力層の素子をそれぞれ各母音/a/, /i/, /u/, /e/, /o/に対応して活性化する素子だと考える。この図では，/e/を入力したにもかかわらず，/a/の素子の活性度が一番高く，その次が/e/となっている。最も望ましい反応は/e/の素子のみが活性化される状態であるから教師信号の欄に示すように，/e/の素子が最大に活性化され，他の素子は活性化されない状態である。そこで今得られた出力信号と教師信号の2乗誤差（R）を計算し，誤差が減少するように各入力に対する重みの値を更新する手続きをとればよい。更新すべき重みの程度は，誤差（R）を各重みで偏微分することにより求められる。この計算は出力層から入力層に向けて順次逆向きに行なうことができ，これが誤差を順次逆向きに伝播するような操作として記述されるため，誤差逆向き伝播法の名前が付けられている。ここではごく原理的な点のみを述べたがこのような学習を繰り返して行くことによって，ニューラルネットは与えられた入力に対し準最適な解を出すようになる。

IV ニューラルネットによる読みへのアプローチ

前述の層状ニューラルネットを読みの機能実現に対して適用した最初のものが，NETtalk（Sejinowski & Rosenberg, 1986）である。日本語の読みに関する工学的実現法のところでも述べたように，表音表記であるかな表記のような場合でも実際に発声される音素系列に直すためには変換規則が必要である。一般に英語ではアルファベットによる綴字と発声されるべき音素系列の対応には多くの例外を含み，日本語のかな表記の場合よりずっと不規則である。NETtalk はこのような複雑な読みの規則を学習させようとしたもので，そのニューラルネットの構成を図5に示す（Sejinowski & Rosenberg, 1987）。入力は26のアルファベットと単語間の区切りを表わす空白（スペース），カンマ（,）及びピリオド（.）をそれぞれ代表する入力素子を設け，それらを一組として全部で7組合計7×29ヶの素子で入力層を形成する。これによって，入力文の記号も含め7文字の系列が同時に入力される。この7文字の真ん中の文字に対する読みを定めるのに必要な前後の文

表1 調音点，調音様式，有声/無声，区切り記号等による音素の分散表現の例

／b／	Voiced(有声)，Labial(唇音)，Stop(閉鎖音)
／k／	Unvoiced(無声)，Velar(軟口蓋音)，Stop(閉鎖音)
／e／	Medium(中位音)，Tensed(緊張母音)，Front 2(前舌)

脈情報としてこの同時提示の文字数がほぼ充分ということと，ネットワークのシミュレーションに使った計算機の能力の限界から7という数が決められている。スペースを含む7文字「-a-cat-」の真ん中にある「c」の読みを求めようとするときの入力は図に示されるようになり，黒丸の素子が入力によって活性化される。隠れ層は80ヶの素子で構成され，出力層は26ヶの素子で構成されている。しかし出力の表現は一つの音素記号，例えばこの場合/k/に対応した素子が活性化するというような構成ではなく音素の分散的表現が用いられている。音素の分散表現とは表1に示すように調音点，調音様式，有声／無声のそれぞれの特徴要素の有無によって一つの音素を表現するやり方である。例えば表に見られるように無声，軟口蓋音，閉鎖音をそれぞれ示す出力層の素子が同時に活性化した場合に，それは音素/k/を出力したことになる。

このようにネットワークに対して，発話データの学習（文章の実際の発話を音素表現して教師データとして用いる）や，辞書データの学習が行なわれた。

1,024語の文を用いて学習させ，学習外の同一話者による439語の文とネットワークによる音素出力を比較したところ78%の正解率が得られた。これはネットワークが丸暗記を行なったのではなく，一応規則の学習が行なわれたとSejnowski らは考えている。

従来の工学的アプローチのところで述べたと同じような仕組みで DECtalk と呼ばれる英語のテキストからの音声合成装置が実用化されている。この装置の文字から音素への変換部分をこのニューラルネットで置き換えて合成音を出すデモンストレーションが行なわれた。これは学習が進むに従って合成音が聞き取れるようになることが実感でき，NETtalk の評判を高め，多くの人の関心をニューラルネットに引きつけた。

V ニューラルネットによる学習

ニューラルネットでは，前述のように読みの規則の学習が行なわれていると見られる。それではこのような学習がどのような特徴を持っているかについて NETtalk に関する観測に基づく結果を以下に示す。

①辞書の学習を行なわせたとき，ニューラルネットでは誤り数と提示単語数の関係はべき乗則に従う。これは人間の技能の学習と同様である。

②学習済みのネットワークの各素子の結合重みを乱数で変化させたような障害を加えると，性能劣化は緩やかであり障害に強い。

③学習済みのネットワークに障害を与えたものと初期学習からある程度の学習が進み，障害を与えたものと同じ性能に達したものとのその後の学習過程を比較すると，障害の程度がひどくない場合には再学習のネットワークの方が学習が完了するまでに必要な単語提示回数がずっと少なくて良い。

④ネットワークの初期状態が異なっても学習の結果は同じようなところに収束する。

⑤新しい単語を学習させるには新しい単語と既に知っている単語を交互に学習させるのが良い。これは Ebinghaus によって指摘され，その後検証されている人間の記憶に関する一般的な現象と良く対応している。

VI まとめ

テキストの読み上げに関してはその他にもニューラルネットでの韻律の学習を試みたものなどがある(鷲坂，河原，1988)。人間と機械の知能の橋渡しをするニューラルネットという観点から，種々の人間の知能をニューラルネットで実現することが試みられ，学習過程や人間の知能との差異などを検討する試みが引続き検討されている(McClelland, Cleeremans, Servan-Schreiber, 1990)。しかしまだそれらは人間と

の類似性の発見の段階をそれほどは出ていない。人間の脳はその素子（細胞）数は現在試みられているニューラルネットのそれより圧倒的に多いものの，やはり分散，並列処理が行なわれているわけで，どの程度の階層性構造等があらかじめ組み込まれており，学習によって形成されるものは何かといったことは難しい問題であるが，ニューラルネットのような構成的手法はそれを解明する一つのアプローチであり，今後の発展が期待される。

文　献

1）McCleeland, J. L., Cleeremans, A., and Servan-Schreiber, D.：Bridging the Gap between Human and Machine Intelligence. 人工知能学会誌，5(1)；2-14，1990.

2）Rumelhart, D. E., McClelland, J. L., and The PDP Research Group：Parrallel Distributed Processing. MIT Press, Cambridge Massachusetts, 1；318-362，1987.（(部分訳)甘利俊一監訳：PDP モデル. 認知科学とニューロン回路網の探索，pp. 321-366，1989.）

3）Sejinowski, T. J. and Rosenberg, C. R.：Parallel networks that learn to pronounce English text. Complex Systems, 1；145-168, 1987.（麻生英樹訳：英語の発音を学習する並列ネットワーク. bit 臨時増刊，21；1482-1495, 1989.）

4）入野俊夫，河原英紀：多層神経回路網の多変量解析による構成法と不特定話者母音認識への適用. 電子情報通信学会　音声研究会資料，sp.；88-123，1989.

5）佐藤大和，匂坂芳典，小暮潔，嵯峨山茂樹：日本語テキストからの音声合成. 通研実用化報告，32；2243-2252，1983.

6）鷲坂光一，河原英紀：ニューラルネットによる韻律の学習. 未発表資料，1988.

Engineering approach to realization of reading function
——An application of a neural network——

Kazuhiko Kakehi

NTT Basic Research Laboratories

The realization of reading function by machine is described. A machine which synthesizes speech from input text has been developed. This machine, called a text to speech synthesizer, uses many rules and knowledge such as lexicon, grammar, letter-to-sound rules, and accent and intonation rules. Those rules and knowledge are extracted from human spoken and written language behavior by long intensive study in linguistics and phonology. This indicates that many rules and knowledge are needed to realize the reading ability, even though the rules and knowledge for the synthesizer are still not perfect.

Recently neural network have been studied as a means to emulate cognitive functions. A brief introduction of the concept of a neural network is briefly introduced. The clustering function and the learning method are explained using a speech recognition neural network for the five Japanese vowels as an example. NET talk (Sejinowsky & Rosenberg, 1986), which is a neural network for letter-to-sound conversion, is described. In the learning procedure for NETtalk, the relationship between pronunciation error rate and the number of words to be learned is log linear. The acquired function shows graceful degradation for defection. Relearning is easier than the first learning. The acquired function reaches to almost the same state regardless of the initial state of the neural network. To teach a new word, it is better to alternatively present the new word to be learned and a word already learned. These characte-

ristics resemble human learning behavior.

At this stage, only the analogy of cognitive ability of human to that of neural network is asserted. Human cognition seems to be based on parallel distributed processing as occurs in a neural network. So neural network is one constructive approach to clarification of the human cognitive mechanism.