論　文

Technical Papers

視聴覚情報の統合化に基づく概念と文法の獲得システム

An Acquisition System of Concept and Grammar Based on Combining with Visual and Auditory Information

中川　聖一*　　升方　幹雄*,†
Seiichi Nakagawa　Mikio Masukata

*　豊橋技術科学大学情報工学系
Dept. of Information and Computer Science, Toyohashi University of Technology, Toyohashi 441, Japan.

1994年6月7日　受理

Keywords: concept acquisition, grammar acquisition, visual and auditory information, image processing, speech processing.

Summary

Human beings accept external stimuli through the senses, and the information is transmitted to the brain. It seems that human beings acquire various concepts from these information. Particularly, auditory and vision stimuli play an important role in the concept acquisition.

From this point of view, we developed a system which acquired concepts and grammar from visual and acoustic information without a priori knowledge by comparing input information with acquired concepts. And, the system generates sentences (or speech) for input images and image concepts for input sentence. A concept consists of a set of relation between a figure feature and a speech event. A figure consists of four categories i. e. shape, size, color and position, and speech events consist of speech waves which are extracted from input speech. If the system perceives the acquired concepts in the input speech, the learning process begins from mapping between the input information and the perceived concept. The part of input which was not spotted by acquired concepts is registered as the new concept in the concept dictionary.

And, our system perceives the sequence of concept inputs by using acquired concepts, and models these sequences by using HMM, in other words, it acquires the grammar. Furthermore, our system generates an utterance which explains input images by using acquired concepts and grammar.

Our system acquired 11 out of 12 types of concepts from 100 pairs of utterance and images. Using the left-to-right HMM for grammar acquisition, the probability our system will generate correct sentences for input images is about 50%. We have realized the first stage of human's concept acquisition process on a computer system.

1. はじめに

人工知能の研究は，人間にのみ可能であった知識の保有・処理などを情報処理技術の立場から計算機処理の対象とし計算機に演繹・推論能力をも持たせようとするもので，将来的には人間の知的機能を模倣し，あるいはその代行をすることをも目指している．従来の人工知能の研究は，対象に関する概念・意味体系をあらかじめ分析し知識として記憶しておき，そのうえで推論操作などを行ういわば静的システムの研究にとどまっている．つまり，与えられた知識によって外的刺激に反応するが，外的刺激に対してシステムの知識を変更することはなかった．このため，環境に適応して能力・効率を高めていくことができないばかりでなく，まったく未知の外的刺激に対してなんら合理的な

† 現在，日本電気マイコンテクノロジー(株)

反応を行うことができない．しかも，外的刺激にはさまざまな種類があり相互に関連しているにもかかわらず，ただ1種類の外的刺激を扱うだけのものがほとんどであった．これらの困難を克服するには，新しい概念や意味体系を外的刺激から獲得する能力，および未知の複数の外的刺激に対する意味づけを行うメカニズムの解明が必要である．

本研究では，子供が概念と文法を獲得しているという事実から，それらの知識獲得をシミュレートしたシステムを作成した．具体的には，視覚と聴覚という二つの外的刺激を結びつけて未知の入力に対する概念を獲得し，その獲得された概念を用いて文法の獲得を行い，さらに，それらの獲得した知識を用いて外的刺激に応答できるシステムの作成を行った．

子供の概念の獲得過程は認知心理学の分野などで活発に研究が行われている．幼児の言葉の獲得では，音声情報や視覚情報だけでなく，対象物への運動機能が大きな役割を果たしているという報告があるが[正高 93]，我々が作成したシステムは，視覚情報と聴覚情報からの入力に対象を絞ったものである．複数の外的刺激からの音声言語の獲得は，A. L. Gorin らが文とアクションのペアを用いて試みている[Gorin 94]．Gorin らのシステムは，ニューラルネットワークを用いて単語と意味のあるアクションにマッピングすることで知識を獲得している．入力として音声とアクションのペアを用いた我々以外の唯一の研究例である Gorin らの方法は，入力文を単語ごとに区切って入力し，入力音声と単語辞書を音声認識手法を用いて照合し，照合の結果が悪かったら新しい単語とみなし辞書に登録し辞書とアクションの関係を獲得していく方法である．アクションが3種類(デパートの売場名)の場合は，テスト入力文の約半数に対して正しいアクションができるようになったと報告している．

Chomsky は言語生得説を唱え，人間は生まれながらにして言語によらない言語の獲得・習得能力を持っている(普遍的文法の存在)と提唱している．一方，Piaget は言語の認知・理解能力は記号処理能力による生後の学習能力であるという構成主義を唱えている[橋田 88]．我々の立場は両者のいずれでもありいずれでもない．つまり，Chomsky のいうほど普遍文法の存在を仮定しなく，だからといって Piaget のいうほど生後の記号処理による学習だけにも頼らない．我々は，生後獲得した言語に共通したメカニズムが遺伝的に組み込まれるようになり，生得的に獲得されるようになったと考える．つまり，先天的に得られるものも，原理的には後天的に得られ得るものと考え，これを工学的に実現する立場をとる．概念とは，あるものを観察することで得られる事象の情報と，周りの環境でそれを表現するのに使われ，知識情報が含まれる言語情報を関係づけたものといえる．言語情報の媒体となるものに，文字と音声があるが，幼児が多く接するのは音声であるので，本研究では言語情報の媒体として音声を用いる．また，概念獲得のメカニズムとして，文字を介する記号処理というよりも，概念そのものもパターン表現と考え，音声と画像のパターン処理に重点を置くという立場をとる．もちろん，皆無の知識から概念獲得するのは不可能なので，さまざまな予備的知識を利用している．それらは，画像のプリミティブな特徴を抽出するメカニズム，パターンの類似を判定するメカニズム，核心の概念獲得アルゴリズム，文法は確率正規文法で表現できるという仮定，などである．

本研究は，音声パターンと画像特徴の対応関係を概念とみなし，パターン照合法によってその関係を抽出していくもので，本質的には音声パターンや画像パターンを用いなくても記号列パターン上での概念獲得と同じと考えられる．これについての研究はすでに報告済みだが[中川 88a, 中川 89]，本研究は，一歩進めアナログパターンを用いた曖昧な照合結果，不確かな中間結果を用いて概念を獲得するという，より人間に近い問題を扱った点に特徴がある．Gorin らは単語ごとに区切って発声した音声を用いており，音声の概念単位(単語という概念)は獲得されているという仮定から出発している．我々は，連続して発声された音声を用いているので，さらに難しく，人間の概念獲得に即した問題設定となっている．

我々が，[中川 93]で報告した概念獲得システムは，概念の獲得だけを行ったものであり，入力も 20 組と少なかったが，本論文では，獲得した概念をもとに文法の獲得をも試み，入力も 100 組と規模を大きくした結果を報告する．

2. 本システムの概要

本研究では，具体的に人間の幼児がどのような情報によって概念形成を行っているのか考察し，これをヒントに工学的に概念形成メカニズムを計算機上で実現することを最終目標にしている．今回開発したシステム全体の構成を図1に示す．

人間の場合，いくつかの感覚器を単独であるいは組み合わせて使用し，外部からの刺激を感じてそれらの情報が脳に伝えられさまざまな概念を獲得していると考えられる．そのなかでも特に，事物の名前などを学

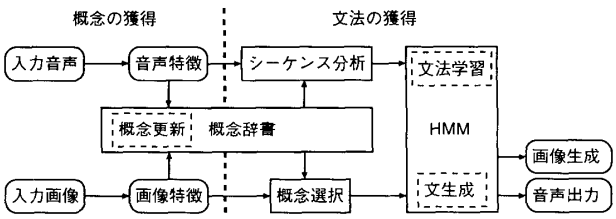

図1　概念獲得システムの構成図

習するためには視覚と聴覚が最も重要な役割を果たしているであろうことは容易に想像がつく．そこで，この視覚と聴覚によって得られる情報，つまり音声と画像の情報を用いて計算機に物の名前や位置などの概念を学習させるシステムを作成した．つまり，ある物を表現する画像があったとすると，その画像を説明する文を音声によって与えることにより，逐次，画像上の形状，色，大きさ，位置といった概念に対応する音声言語を獲得すること，逆にいえば，ある[音]に対応する形状の概念を獲得することが本研究の目標である．ただし，物の名前や位置などの概念を単語として与えるのではなく，簡単な文の音声データとそれに対応する画像データとを用いて，形状，大きさ，位置，色などの概念を形成することとした．ただし，大きさよりも形状のほうが早く獲得されるという子供の概念形成過程のモデル化は意図せず，これらの概念は互いに同質と考えて,同一のメカニズムで獲得するものとする．このことから，画像どうしの類似性の自動判定，音声どうしの類似性の自動判定，画像と音声の対応づけなどの機能が基本操作となる．

また，概念を獲得したシステムは，獲得した概念が入力にどのようなシーケンスで現れるかを検知することができ，それらをモデル化することで入力の文法を獲得する．さらに，概念と文法を獲得したシステムは，画像だけが入力されると，それらの獲得した知識を利用して，その画像に対応する文を生成しそれを発声することができる．これとは逆に，システムに音声だけを入力すると，システムはその文に対応する画像概念を出力することができる．

3．画像からの図形データの抽出

カメラで取り込んだ図形画像から，直接概念を形成することは難しい．そこで，概念形成に必要とされるパラメータを図形画像から抽出しなければならない．今回の実験で形成する概念は，図形の形状，図形の大きさ，図形の色，図形の位置の四つの概念グループに分類される．以下ではこれらの概念およびその概念を形成するために求めるパラメータについて述べる．このプリミティブな機構は先天的に人間に備わっている(あるいは，すでに概念獲得されている)と仮定している．図形は以下に述べるように(形状，大きさ，色，位置)と量子化される．この組を以後，単に画像特徴ということがある．詳細は[古部 91，中川 93]を参照されたい．

(a)　図形の形状

図形の形を表す概念グループであり，「丸」，「三角形」，「四角形」* が含まれる．これら3種類の図形の識別用のパラメータには，図形の輪郭線の屈曲点の数を用いた．

(b)　図形の大きさ

このグループには，図形が大きい，普通，小さいという概念が含まれる．そこで，パラメータとして，図形の絶対的な大きさ，つまり，2値化画像での図形の占める画素数(面積)を用いる．図形の大きさは，この面積をしきい値で区切って3段階(「小」，「中」，「大」に対応)に量子化する．

(c)　図形の色(グレーレベル)

図形が黒い，灰色，白いという概念のグループである．このグループに対しては，図形のグレーレベルの平均値をパラメータとする．図形の色はグレーレベルの平均値をしきい値で区切って3段階(「黒」，「灰色」，「白」に対応)に量子化する．

(d)　図形の位置

この概念グループには,図形の絶対的な位置の概念,すなわち，画像中の「上」，「右」，「真中」，「左」，「下」という概念が含まれる．この概念とマッチングをとるために，2値化した図形画像より，図形の重心座標をパラメータとして求める．これを適当なしきい値で横方向に3段階,縦方向に2段階の計5段階(「上」,「右」,「中」,「左」,「下」に対応)に量子化する．

4．音声データからの音声情報の抽出

入力音声中に登録辞書と共通の音声区間があるかどうかを検出するために，DPマッチングによるワードスポッティングアルゴリズム[中川 84]を用いた．これは，辞書の音声パターンと入力音声文をマッチングした際に，辞書のパターンに対し入力文の各フレームで終端するDPパスの最適な照合位置と照合距離を求

---

*　・画像の概念については「　」を使用．注：実際は画像の特徴パラメータ
・音声の概念については[　]を使用．注：実際は音声波形
・画像とか音声とかを特に意識しないで用いるときの概念については“　”を使用

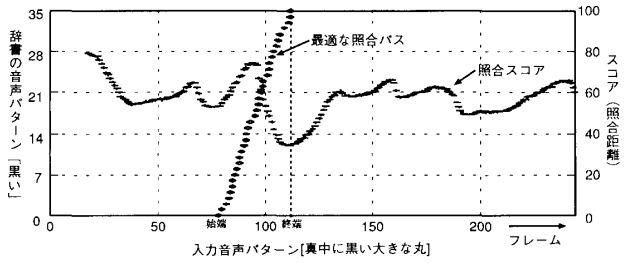

図2 スポッティングの例

め，照合距離を正規化したもの(スコア)があるしきい値以下であれば，入力文中で辞書と同じパターンが存在すると判断する方法である．詳細は[中川 88b, 中川 93]を参照されたい．図2は，この方法によって照合された例を示す．図中のスコアとは，入力文の各フレームで終端するDPパス内で最も良いスコアを示したものを示している．図中に示したDPパスとは，それらのスコア内で最もスコアの良い(値の小さい)結果の照合パスを示している．入力音声のうち，この部分に該当する部分が，辞書の音声パターンと同じ音声と判断され，スポッティングされる．

5. 概念と文法の獲得アルゴリズム

5・1 概念の獲得

本システムでは，入力された音声のなかに辞書(ここで，辞書というのは，一つの概念に対応すると仮定された音声パターンと画像パターンの対のことである)に登録された概念の音声があると検知されると，その概念に対応する画像特徴と入力の画像特徴が比較される．このとき，入力の画像特徴はその画像概念の履歴に加えられ，入力の画像特徴は，比較された概念の画像概念と等しければクリアされ，異なればそのまま残る．画像概念の履歴とは，それぞれの概念がそれまでに照らし合わされた特徴量とその学習回数を記録したものである．以前のシステム[中川 93]では，学習の更新回数(頻度情報)を用いないで，1回ごとに決定的な更新をしていた．人間はだんだんと確信しながら学習していくものと考えられるので，今回はこのメカニズムを導入した．各概念の画像概念には，学習回数が5回に満たなければ，最も多く出現した特徴量を選択する．学習回数が5回以上の場合は，それぞれの画像特徴において最も多く学習した特徴量の割合が80%以上であればその特徴量を選択する．それに該当する特徴量がなければどの特徴量も画像概念として選択しない．つまり，音声に対応する概念がその画像特徴には無関係であると判断される．例えば，表1の履歴を持つ概念は，表中に示したように画像概念が選択される．

表1 ある音声パターンに対応する画像概念の履歴(特徴量は実際に無意味な記号で表現されている)

画像特徴	特徴量の出現回数	選択画像概念
形	「三角形」(35)、「四角形」(2)	「三角形」
大きさ	「大きな」(15)、「小さな」(8)、「普通」(14)	-
色	「黒い」(25)、「白い」(12)	-
位置	「上に」(4)、「左に」(10)、「真中に」(9)、「右に」(8)、「下に」(6)	-

獲得されているすべての概念との比較の結果，まったくスポッティングされなかった入力音声の部分は入力の画像特徴と組になって，新しい概念として登録されることになる．こうして，システムは未獲得と判断される概念を次々と辞書に登録していく．

また，システムは概念獲得の過程において，獲得された概念どうしの画像概念と音声を比較し，同じ概念と判断される概念どうしを一つの概念に統合する．学習回数が5回以上の概念のうち，同じ画像概念を表しかつ同じ音声であると判断される概念どうしでは，一方の画像履歴は他方の画像履歴に足し合わされ，音声も一方の概念に移動することになる．こうして，概念は一つ減少し，一つの概念に複数の音声が登録されることになる．これにより，同じ概念に対する音声パターンの変動に対処することができる．

5・2 冗長な概念の削除

概念獲得の過程で，二つ以上の単位概念(例えば，“三角形”と“黒”)からなるまだ分化していないある一つの概念(“黒い三角形”)と，それらの単位概念が概念辞書中に同時に登録されるということが生じる．それらの単位概念が確定的なものであれば，これらの単位概念からなる未分化概念は冗長であるといえるので，この概念を記憶する必要はない．例えば，“白い丸”という概念の存在は，“白い”と“丸”の確定的な単位概念の存在により不要と判断される．そこで，ある概念が他の複数の確定的な単位概念の組合せで表せると判断できるものであれば，システムはその概念を冗長な概念と判断し，辞書から削除する．

冗長さの判断は，画像概念どうしの比較と音声どうしの比較の結果によって行われる．このときの音声の比較では，$O(n)$DP法[中川 88b]を用いて対象となる未分化概念の音声パターンと確定的な単位概念の音声パターンを任意の順序で連続して得られる音声パターン(上の例では，[白い]と[丸]をこの順で連結したり逆順に連結して得られるパターン)を照合し，最適な照合結果があるしきい値以上になるとこれらの確定的な単位概念に分解され得ると判定される．本システムにおいて，確定的な概念とは，以下の三つの条件を満たすものと定義する．ただし，しきい値は予備実験の検討

により決定した．

- 学習回数が8回以上．
- 四つの画像概念カテゴリ(画像特徴の属性)のうちで一つのカテゴリにだけ概念が現れている(例えば，形状にだけ何らかの概念が現れていて，その他の色，大きさ，位置のカテゴリには概念が現れていない)．
- ただ一つ現れている概念がそのカテゴリにおいて，0.7以上の割合で出現している(例えば，形状のカテゴリにのみ概念が現れているある概念において，「三角形」がそのカテゴリにおいて0.7以上の割合で学習されている)．

また，本研究では，概念とはある音声に対して何らかの画像特徴が対応づくものと定義しているから，学習回数が多いにもかかわらず何ら画像特徴と対応づけることのできない概念は信頼性が低く記憶するに値しないと思われる．本システムでは，学習回数が9回以上でかつ画像特徴の最大出現率が0.7未満の概念は記憶するに値しない概念と判断しシステムは辞書からそれを削除している．

5・3 文法の獲得

HMMの構造は確率正規文法と等価である[中川 88b]．したがって，システムはそれまでに獲得している概念を利用して，入力される音声中に現れる概念のシーケンスを同定し，それをHMMでモデル化すれば，誤りのある概念の例の集合からでも確率正規文法という枠組みではもっともらしい文法を獲得できると思われる．エルゴディックHMMを用いて言語構造を獲得する研究は[村上 92，上田 90]で見られる．

システムは入力文の音声に辞書中の概念がどのような順番で現れるか決定しなければならない．そこで，入力音声を獲得された概念の音声パターンの任意の連結パターンとを照合し，最適な連結パターンを$O(n)$ DP法により検知し，その概念シーケンスをHMMでモデル化することで文法を獲得する．

HMMの諸条件は表2に示すとおりである．エルゴディックHMMの初期モデルにおいて，状態遷移確率をすべての遷移で均等にし，シンボル出力確率の分布をすべての遷移で同じにすると，入力概念のシーケンスをうまくモデル化することができない．現在のところ，エルゴディックHMMの最適な初期モデルの決定方法は知られていない．よって，初期モデルのそれらの確率にはモデル化が十分に行われるだけの乱数を付加しなければならない．そこで，初期モデルの状態遷移確率は，まず各遷移で均等にし，その値に小さな乱数値δをそれぞれに付加し，シンボル出力確率は観測シーケンス中の出現率に小さな乱数値δを付加した．ここで，システムに入力される音声はすべて図3に示す文法に従っている．また，より先見的な文法の構造を取り入れたLeft-to-Right型HMMを用いても文法の獲得を行った．

表2 HMMの諸元

(a) エルゴディックHMM

HMMの構造	エルゴディックHMM
HMMの出力シンボル	単語(獲得された概念)
開始・終了状態	任意
状態数	4,5状態
学習終了条件	生成確率上昇率0.01%未満

(b) Left-to-Right HMM

HMMの構造	Left-to-Right HMM
HMMの出力シンボル	単語(獲得された概念)
状態数	4状態
学習終了条件	生成確率上昇率0.01%未満

S → A B C D
A → λ | 上に | 左に | 真中に | 右に | 下に
B → λ | 黒い | 白い
C → λ | 大きな | 小さな
D → 丸 | 三角形 | 四角形

図3 入力文の文法

5・4 文の生成

十分な概念と文法を獲得したシステムは，それらの情報を利用することで，入力される画像に対する文(音声波形の連結)を生成することが可能である．辞書中の概念から入力画像の各画像特徴に相当するものを一つずつ選びだし，それらの概念の並びによってできるすべてのシーケンスの生成確率を文法がモデル化されているHMMを用いて計算する．このときに選ばれる概念は，画像特徴がただ一つ現れている概念だけを選択し，それが同じ画像特徴に対して複数存在すれば，学習回数の多いほうを選ぶ．例えば，表3(a)で，形の特徴(屈曲点の数)が四であれば，[？]という音声が登録されている概念番号15(頻度7)よりも[四角形]という音声が登録されている概念番号12(頻度23)を選ぶ．現在の本システムは四つの画像特徴に対応しているため，最大で四つの概念による4!=24通りのシーケンスについて確率を計算する．そのうちで最も確率の高い概念のシーケンスが入力画像を表す文として生成され，システムは，そのシーケンスに従って選択された概念に付随している音声を連結し出力する．

5・5 画像概念の出力

5・4節とは逆に，入力される音声に対応する画像概念を出力することも可能である．これは，文法の学習のときと同様に，入力に現れる概念のシーケンスを検知し，それらの概念の持つ画像概念を組み合わせて，入力音声に対応する画像特徴量を出力する．このとき，各画像概念のカテゴリによっては，選ばれた概念の画像概念同士が矛盾する結果を示すときもあるが，そのような場合はその画像概念カテゴリに対しては画像概念を出力しない．現在のところ，画像概念(画像特徴量)から画像の生成は行っていないが，例からの学習法を用いれば可能だと考えている．

6. 実　　験

6・1 概念の獲得

システムに入力する音声の発話文は以下に示すような100文とそれに対応する図形である．

- 三角形
- 黒い三角形
- 黒い四角形
- 上に三角形
- 白い丸
- 左に小さな丸
- 黒い大きな丸
- 小さな三角形
- 右に黒い大きな丸
- 小さな丸

入力文はすべて図3に与えられた文法に従うものであり，対応する画像とは矛盾しないものとなっている．文中には，観察される図形の形状を表す言語情報が含まれる必要があるが，他の特徴については含まれる必要はない．幼児段階で提示される言語情報は，名詞が主なのと，本研究では動作概念の獲得は扱っていないため，例文はすべて体言止めの文を用いた．実験に用いた100文に含まれる総単語数は249単語である．入力文の初めの部分は，初期の学習がある程度能率良く進むように文の順番に注意を払ってあるが，それ以降はランダムになっている．もし，初めから複雑な文とその画像が入力されると，最初は暗記学習が中心となり，学習の進み具合が遅くなる．子供の言語獲得も最初は一語学習に，次に二語学習と段階を経て学習されるのと対応している．

概念をまったく獲得していないシステムが100組の画像と音声の入力によって得た概念の辞書を表3に示す．音声入力に対しては，男性話者2人が同じ100文の組を発声したものを用い，画像に対しては，共通の100組のデータを用いた．表3では，画像概念の部分をわかりやすくするため，「三」，「黒」などと表したが，システムは実際には無意味な記号として認識している．また，音声の欄でも便宜上，[三角形]，[黒い]などと表してあるが，実際は音声波形が対応しており，D-A変換して聴取するとそれぞれ[三角形]，[黒い]に聞こえたことを意味している．厳密にいえば，セグメンテーションの誤りで，一部の語頭や語尾部分が欠落したり，付加されているものもある．表3(a)では，“下に”以外はすべて完全な概念として獲得されている．表3(a)の概念番号2で，[黒い]という音声が二つ登録されているのは，概念獲得の途中で，同じ概念

表3 100組の入力で得られた概念辞書

(a) 話者1

	概念番号	画像特徴 形	大	色	位置	頻度	音声
◎	1	三	-	-	-	36	[三角形]
◎	2	-	-	黒	-	24	[黒い]、[黒い]
◎	3	-	-	-	上	13	[上に]
	4	丸	中	白	左	2	[白い丸]
	5	丸	小	白	左	1	[左に小さな丸]
◎	6	-	小	-	-	16	[小さな]
◎	7	-	-	-	右	16	[右に]
◎	8	丸	-	-	-	31	[丸]
◎	9	-	-	-	中	6	[真中に]
	10	-	小	黒	下	4	[下に]
△	11	-	大	-	-	14	[黒い大きな]
◎	12	四	-	-	-	23	[四角形]
◎	13	-	-	-	左	13	[左に]
◎	14	-	-	白	-	10	[白い]
×	15	四	-	-	-	7	[?]
×	16	-	-	-	-	5	[?]
	17	-	大	白	下	1	[大きな]
◎	18	-	大	-	-	14	[大きな]
◎	19	-	-	黒	-	3	[黒い]
	20	-	中	白	下	3	[下に]
	21	三	大	白	-	1	[白い三角形]
	22	丸	大	-	-	2	[丸]

(b) 話者2

	概念番号	画像特徴 形	大	色	位置	頻度	音声
◎	1	三	-	-	-	37	[三角形]
	2	三	中	黒	中	4	[黒い]
△	3	四	-	-	-	7	[黒い四角形]
◎	4	-	-	-	上	13	[上に]、[上に]
	5	丸	中	白	左	2	[白い丸]
	6	丸	小	白	左	4	[左に小さな丸]
△	7	-	大	黒	-	8	[黒い大きな丸]
◎	8	-	小	-	-	16	[小さな]
◎	9	-	-	-	右	15	[右に]
◎	10	-	-	黒	-	31	[黒い]、[黒い]、[黒い]
◎	11	-	-	-	中	5	[真中に]
×	12	四	-	-	-	28	[?しかっ]
◎	13	丸	-	-	-	26	[丸]
◎	14	-	大	-	-	22	[大きな]
◎	15	-	-	-	左	18	[左に]
	16	-	-	黒	上	6	[上に]
◎	17	-	-	白	-	8	[白い]
	18	-	-	-	-	8	[右に]
◎	19	-	-	-	下	6	[下に]
	20	-	大	白	-	6	[白い]
×	21	-	-	-	-	6	[?]
×	22	-	-	白	-	6	[?けい]

◎ … 画像概念と音声が一致おり，且つ，現れている画像概念が1つだけの概念
△ … 音声に対応する画像概念が一部現れていないもの
× … 無意味な音声(ノイズなど)が登録されたもの
無印 … 上記以外　? … 無意味な音声

を表すとシステムが判断したために統合されたからである．

6・2 文法の獲得

エルゴディック HMM の初期モデルの作成には乱数を使用しているため，初期値の異なる 10 個のエルゴディック HMM について学習を行った．図 4 は 100 組の入力データのうちで 2 単語以上を含む 87 組の入力データを誤りなく学習して得られたと仮定した場合に得られるエルゴディック HMM の一つの学習結果である．図中で，Pos. は位置のカテゴリに属する概念を表し，Col. は色，Size は大きさ，Shp. は形状をそれぞれ表す．図 5 中の遷移の側に記した数字はそのカテゴリに属する概念の出力確率であり，状態中の%の数字はその状態が初期状態となる確率である．同じカテゴリに属する単語は，固まって出力されており，入力の文法がこの HMM で獲得されていることが観察できる．

図 5 は実際の学習で得られた表 3 (a) の概念辞書をもとに，100 組の入力データのうちで 2 単語以上を含む 87 組の入力をシステムに入力することで得られた概念のシーケンスを学習したエルゴディック HMM の一つの学習結果である．この場合でも，先ほどのシミュレーション実験のときと同様に，エルゴディック HMM に入力の文法が獲得されている．

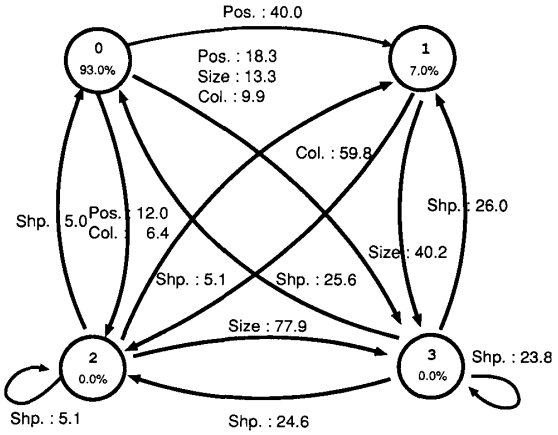

図 4 エルゴディック HMM による文法獲得の結果(シミュレーション：話者 1)

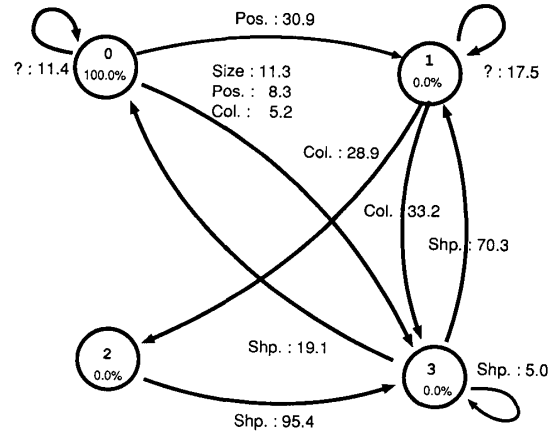

図 5 エルゴディック HMM による文法獲得の結果(実データ：話者 1)

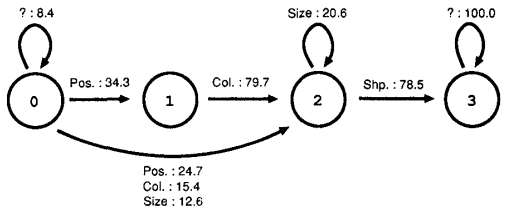

図 6 Left-to-Right 型 HMM による文法獲得の結果(実データ：話者 1)

図 6 は同様に概念のシーケンスを学習した Left-to-Right 型 HMM の学習結果である．Left-to-Right 型 HMM は，入力の文法を表現するのに適したものと思われる．図 6 を見ると，形状を表す単語の出力による遷移で終了状態に遷移し，その形状を修飾する単語はその前に現れ，色に関する概念は形に関する概念より先に現れるといった入力の文法がはっきり獲得されていることがわかる．

6・3 文の生成

異なる画像特徴の組合せ 5×2×2×3=60 個の入力画像に対して文生成を試みた．まず，誤りのない入力シーケンスを学習した HMM による文生成の結果を表 4 に示す．4 状態の Left-to-Right 型 HMM ではすべての入力画像に対して正しく文を生成した．エルゴディック HMM では，一部の初期状態から学習したものが 100%にならず平均すると 4 状態，5 状態とも正解率は 100.0%に届かなかった．

実際にシステムにデータを入力して得られた概念辞書と文法を用いた場合の文生成の結果を表 5 に示す．表 3 の概念辞書では“下に”の概念が未獲得である．よって，この概念辞書を用いて「下に」という画像特徴を含む入力画像に対応する文を生成しても位置の概

表 4 文生成のシミュレーション実験(正しい概念系列で学習し，正しい概念の組を入力)

HMM の種類	正解率
4 状態の Left-to-Right HMM	100.0%
4 状態の Ergodic HMM	平均 92.0%
5 状態の Ergodic HMM	平均 93.0%

表 5 文生成の実データでの実験

HMM の種類	話者 1	話者 2
4 状態の Left-to-Right HMM	50.0%	20.0 %
4 状態の Ergodic HMM	27.8%	31.0%

念を抜かした文を生成することになる．表中の正解率は，そうした場合に，[下に]という音声を含まなくても正解としてある．また，“大きな”に対応する概念として，概念番号11と概念番号18が存在する．概念番号11の音声は[黒い大きな]となっており，画像概念と一致していないが，概念番号18のほうの音声は[大きな]となっており，画像概念と一致している．システムはこのように同じ概念に複数の候補があった場合，頻度の多いほうを選択することにしていたが，両方の概念とも頻度は同じであった．このような場合，システムはより早い段階で登録されたほうを用いるように設定したため，[黒い大きな]が音声として登録された概念番号11が選択された．このため，この概念辞書を用いると，「大きな」という概念が入力画像に含まれていると正しい文生成がほとんどできなかった．このような点もあって，文生成の結果はまだ不十分なものであった．また，エルゴディックHMMの初期モデルの違いによって文生成の正解率にはばらつきが見られた．図6のLeft-to-Right型HMMでの文正解率は50.0%であった．これは，上記の理由により，「大きな」という概念を含む画像(全体の半分)に対しては，[黒い大きな]という文を生成してしまったためであり，それ以外の画像に対しては，すべて正しい文を生成できた．

話者2では“四角形”に関する概念が未獲得であった．よって，四角形の概念を含む入力画像に対してはまったく正しい文を生成することはできなかった．話者1では，4状態のLeft-to-Right HMMのほうが良い結果を示したが，話者2では4状態のErgodic HMMのほうが良い結果を示した．これは，上述のことに起因していると思われる．

6・4 画像概念の出力

文法の獲得の実験と同じ87文について画像概念の出力の実験を行った．得られた画像概念が入力音声と完全に一致した割合は，話者1では69.0%，話者2では60.9%の正解率であった．このように，画像データからの文生成よりも生の音声入力からの画像概念の生成のほうが正しくできた．これは，入力音声に現れる概念のシーケンスは，文頭や文末に無音声に検知されるものの，概ね正しい順序で検知されるので，検知された概念が正しい概念を表していれば，画像概念の生成の正解率は高くなると考えられるためである．一方，文生成では，辞書から各概念を代表して一つ選ばれる概念に誤った音声が登録されていれば，その概念を含む画像に対してはまったく正しい文を生成できなくなるため，正解率は低くなると考えられる．

7. ま と め

画像情報と音声情報から概念と文法を獲得するシステムを作成し実験を行った．音声パターンや画像パターンの照合は，シンボルレベルと比べて曖昧な照合結果しか得られないが，このような不確かな結果を用いても単純な概念の獲得では目的の概念を概ね獲得することができた．Left-to-Right型HMMでは，正しく獲得された概念を含む画像に対してはすべて正しい文(音声波形の連結)を生成できた．しかし，エルゴディックHMMによる文の生成ではまだ芳しい結果が得られなかった．当然ではあるが，正しく獲得されていない概念を含む画像に対しては，文生成は正しく行えない．本システムでは，音声のスポッティングが動機となって学習が行われるため，音声のスポッティングの性能が概念および文法の獲得に大きく影響する．

現在，文生成の際は，選択された概念で生じるすべてのシーケンスについて確率を計算しているが，画像特徴が多くなると計算量が膨大になる．画像特徴が多くなった場合でも対応できる文生成のアルゴリズムを考える必要があろう．現在，本システムをもとに，さらに図形の相対位置の概念の獲得法を検討している[中西 94]．

謝 辞

有益な議論をしていただいた協同研究者の中西宏文愛知教育大学助教授に感謝致します．

◇参 考 文 献◇

[古部 91] 古部好計, 中西宏文, 辰巳昭治, 中川聖一: “音声と画像の対応付けに基づく概念の獲得”, 人工知能学会全大, 2-10 (1991).

[Gorin 94] Gorin, A. L., Levinson, S. E. and Sankar, A.: An Experiment in Spoken Language Acquisition, *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 1, PART II, pp. 224-239 (1994).

[橋田 88] 橋田浩一: 言語生得説—チョムスキーをめぐって, 人工知能学会誌, Vol. 3, No. 2, pp. 139-148 (1988).

[正高 93] 正高信男: 身ぶりの行動発達学, 科学, Vol. 63, No. 8, pp. 499-507 (1993).

[村上 92] 村上仁一, 山本寛樹, 嵯峨山茂樹: Ergodic HMMによる確率つきネットワーク文法の獲得の可能性について, 人工知能学会研資, SIG-SLUD-9204-3(1992).

[中川 84] 中川聖一：拡張連続DP法による連続音声認識アルゴリズム，信学論，Vol. 67-D, No. 10, pp. 1242-1249 (1984).
[中川 88a] 中川聖一，若原一彰：自然言語の構文・意味解析規則の主観的確率を用いた帰納的学習システム，人工知能学会誌，Vol. 3, No. 6, pp. 773-782 (1988).
[中川 88b] 中川聖一：確率モデルによる音声認識，電子情報通信学会 (1988).
[中川 89] 中川聖一，山本幹雄，若原一彰：自然言語の文法と意味解析規則の帰納的学習システム，情処学論，Vol. 30, No. 1, pp. 72-80 (1989).
[中川 93] 中川聖一，中西宏文，古部好計，板橋光義：視聴覚情報の統合化に基づく概念の獲得，人工知能学会誌，Vol. 8, No. 4, pp. 499-508 (1993).
[中西 94] 中西宏文，中川聖一：音声と画像情報からの相対概念の獲得，人工知能学会全大，8-3 (1994).
[上田 90] 上田佳央，中川聖一：HMMを用いたテキスト中の音韻，音節，品詞の予測，日本音響学会講論集，3-8-9 (1990-9).

〔担当編集委員・査読者：白井克彦〕

著者紹介

中川 聖一（正会員）

1976年京都大学大学院博士課程修了．工学博士．同年，京都大学情報学科助手．1980年豊橋技術科学大学情報工学系講師，1983年助教授，1990年教授．1985～86年カーネギーメロン大学客員研究員．音声情報処理，自然言語処理，人工知能の研究に従事．1977年電子通信学会論文賞．1988年度IETE最優秀論文賞．著書に「確認モデルによる音声認識」電子情報通信学会 (1988)，「情報理論の基礎と応用」近代科学社 (1992) など．電子情報通信学会，日本音響学会，情報処理学会，計量国語学会，神経回路学会，言語処理学会，IEEE，ESCA，ACL各会員．

升方 幹雄

1993年豊橋技術科学大学情報工学課程卒業．1995年同大学院修士課程情報工学専攻修了．同年，日本電気マイコンテクノロジー(株)入社．在学中は，音声言語の概念獲得の研究に従事．