2A3-1 ヒューマノイドロボットを用いた発話と提示に基づく文法と挙動のオンライン学習

Developmental Word Acquisition And Grammer Learning by Humanoid Robots through SOINN

○非 小倉和貴(東工大) 非 賀 小淵(東工大)
非 佐藤彰洋(東工大) 非 長谷川修(東工大)

Tomotaka OGURA, Tokyo Institute of Technology, Nagatuta-tyo 4259, Midori-ku, Yokohama-shi, Kanagawa
Xiaoyuan HE, Tokyo Institute of Technology
Akihiro SATOU, Tokyo Institute of Technology
Osamu HASEGAWA, Tokyo Institute of Technology

Key Words : word grounding, SOINN, Mental Model, Top-down learning, Bottom-up learning

1. はじめに

近年，人間との共存を目指した知能ロボットの研究において，従来法に代わる新たな設計法の研究が盛んである。従来法では設計者がロボットの動作環境と行うべき動作を想定し(タスク依存)，あらゆる状況に対応できる処理を組み込むことで，知能ロボットを実現してきた．しかし，実世界の環境は常に変化するため，事前に組み込まれた処理で全ての状況に対応することはできず，そのようなロボットの行動できる環境は非常に限定されてしまう．このため，環境の変化に適応し，新規の問題に柔軟に対処できる知能ロボットが必要とされている．

これを実現するアプローチとして，「認知発達ロボティクス」がある．認知発達ロボティクスでは、ロボット自身が環境との相互作用を通じて学習・発達することで，実環境で行動可能なタスクに依存しない汎用の知能ロボットの構成が目指されている[(1)]．汎用の知能ロボットは以下の点で従来のロボットと異なる．(1) ロボットの行動環境や扱うタスクが限定されていないため，あらゆる状況を想定して処理を組み込むという従来法が通用しない．(2) 発達することが求められる．1990 年代後半から Autonomous Mental Development(AMD)[(2)]の重要性が指摘されているように，タスクに依存しない知能ロボットは様々な環境中で自律的に学習し，適応していくことを求められる．(3) 人間とのコミュニケーションが重要である．これを実現するには，人間との注意の共有や概念に対して人間と共通の理解を持つことが必要である．そのためには，環境中の事物に対して自律的に概念を形成すると共に，それを人間の持つ概念と適合させる機能が必要である．

1-1 従来研究 人間の概念形成における問題はシンボルグラウンディング問題[(3)]として今なお議論されている．人間は事物のシンボル的な表現を自身の内部に形成することで様々な単語を識別している．例えば，りんごについて想像すると，頭の中に仮想のりんごが現れ，りんごに触れたときの触感や食したときの味が浮かぶだろう．これは，我々が言葉と感覚器から得られる情報を対応付けて記憶していることを示唆する．つまり，ロボットが人間と概念を共有するには単語の音声パターンを認識し，それを感覚情報と対応付ける機能が必要であると考えられる．Siskind[(4)]は単語の意味獲得に関する問題について数学的に議論し，以下の 5 つの主要な問題を提起した．すなわち，単語の区切りの検出，概念の分類問題，事前知識なしでの学習，同音異義語，ノイズ処理である．

我々はロボット自身に単語と感覚情報の対応付けを行わせ，実環境で行動させることを試みており，上記の 5 つの問題のうち 4 つに挑んでいる．すなわち，ロボットは学習する言語に対する事前知識を持たず(事前知識なしでの学習)，日本語だけでなく英語や中国語などでも学習できる．また，感覚情報としての音声，画像の入力はオンラインで行い，事前にラベリングは行わない．このため，単語の区切りが不明瞭であり (単語の区切りの検出)，ノイズを含む場合や不完全な場合がある (ノイズ処理)．更に，単語がどの概念 (名前，色，形など) を表すかはロボット自身が獲得する (概念の分類問題)．しかし，これらを解決しても言語を理解するロボットは実現できない．これは，言語を理解するには，言語の構造を考慮した上で単語の意味を理解する必要があるためであり，本研究ではそれを実現する．

Roy，Pentland ら[(5)]は，音と映像のクロスモーダル情報の最大化によって単語グラウンディングを行う手法を提案した。しかし彼らが用いた映像は静止画像であった．岩橋ら[(7)(8)]は，物体が移動する 2 次元動画像とその動画を表現する文章の組みから動きの概念と文法の獲得を行う手法を提案した．この研究では，動画像から抽出された動きの軌道を隠れマルコフモデル (HMM) で学習し，それを概念としている．また，動きの概念の確率モデルから文法の語順生成の確率モデルを学習している．しかし，岩橋らの研究では学習データの動画像と文書が事前に対応付けされており，バッチ処理によって学習を行う．このため，実環境で人間とコミュニケーションしながらオンラインで追加的に学習するには適さない．Yu，Ballard ら[(6)]が提案したマルチモーダル学習システムは，物理的な世界の物体に発話された名前をグラウンディングでき，

発話中の言語学的なラベルに基づいて物体を分類できる．しかし，このシステムは一つの物体に対して一つの単語しか割り当てることができず，新たな物体の概念形成が既存の概念の影響によって阻害されるという問題があり，追加学習の能力に問題がある．

1·2 本研究のアプローチ 我々の研究目的は，実環境でロボットが追加的・自律的に概念形成を行うことであり，以下の4点が従来研究と異なる．(1) ロボットは事前にラベル付けされた学習データを必要とせず，実環境から逐次的に得られる音声・画像入力を用いて学習を行う．実世界のあらゆる事柄にラベルを付けることは不可能であるため，この機能は必須である．(2) 既存知識の影響で新しい知識の学習が阻害される，あるいは新しい知識の学習によって既存知識が失われるという追加学習の問題 (Stability-Plasticity Dilemma) を回避している．これは学習に筆者ら独自の自己増殖型ニューラルネットワーク (Self Organizing Incremental Neural Network：以下 SOINN) を用いることで実現している．(3) ロボットは実環境にある物体の特徴と音声を対応付けることで単語の意味を学習できる (シンボルグラウンディングの実現)．本研究で扱う単語の意味は静的な意味 (色や形、物体の属性) と動的な意味 (物体の動き) の2種類である．更に，ロボットは単語の種類 (色，形，動きなど) について事前知識なしで分類を行う．(4) ロボットは逐次的に与えられる音声の文法的な構造を分析し，言語における単語の意味 (主語や目的語といった役割) を学習でき，自らも簡単な言語を発したり，言語で指示された命令を理解して行動できる．この能力はボトムアップ学習とトップダウン学習の融合によって実現される．ボトムアップ学習では単語の種類 (単語クラス) に基づく統計的 Bigram モデルにより，ある単語クラスから別の単語クラスへの遷移確率を学習し，文章が特定の語順になる確率を計算する．トップダウン学習では「メンタルモデル」を用いて概念の観点から文法構造を学習する．これらにより，従来のバッチ的な学習と異なり，少数の例から文法構造の学習を可能としている．また，文書の構造をそのまま学習するアルゴリズムであるため，任意の言語での学習が可能である．

2. 提 案 手 法

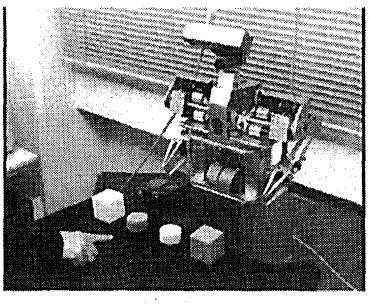

Fig. 1 Robot can survey its surroundings using its pair of CCD cameras

Fig 1 に実験環境を示す．テーブルの上には様々な色，形の物体が置かれている．ロボットは2自由度を持つ頭部に PointGrey 社製のステレオカメラを備え，連続的に映る物体の位置情報，色情報などを取得できる．また，5自由度を持つ腕と手を左右に備えており，手を使って目の前の物体を掴んで移動させることができる．ロボットは実験環境中に存在する物体の事前知識を持たず，学習によって獲得する．ここで知識とは「物体の色，形，名前」という静的な知識、及び「物体の動き」という動的な知識である．これらの知識を音声 (単語) 及び視覚的特徴の2種類から獲得する．音声はマイクから入力される音声特徴量からなる時系列データであり，視覚的特徴はカメラ画像から抽出される色や形の特徴量である．また，動的な知識の獲得では物体の軌跡 (座標の時系列データ) を視覚的特徴とした．ロボットは目の前にある物体の位置や数を記憶する「メンタルモデル」を持ち，テーブルの上にある物体に対して注意を向ける．物体の色や形に関する特徴量は画像処理モジュールによって抽出され，色，形，名前に対応した3つの SOINN で学習される．実験者がテーブル上の物体を指差しながら発話すると，ロボットは発話された音声とその音声が表現している視覚的特徴を対応付けようとする．ただし，ロボットは「"黄色" という音声は色を表す」といった事前知識を持たないため，音声と視覚的特徴を確率的に対応付け，学習によって修正していく．学習初期では誤った対応付けをすることがあるが，複数の物体に対して指差しと発話を繰り返すことで，ロボットは複数の物体間に共通する音声と視覚的特徴を発見し，正しい対応付けを獲得する．また，実験者がテーブル上の物体を動かしながら，例えば「メロン，ミカン，近づける」と発話すると，ロボットはボトムアップ学習とトップダウン学習のアルゴリズムから未知単語 "近づける" と物体の軌跡を対応付けて動きの概念を学習する．そして，実験者が動きの概念を獲得したロボットに同様の文法的構造を持つ文書を与えると，ロボットは自分の手を使ってその言語が意味する動きを生成できる．もちろん，学習時に用いた物体でなくても動きを生成できる．また，ロボットに物体の動きを見せると，DP マッチングにより獲得した動きと比較し，既知の動きであればその動きを表現する言語を発話できる．

2·1 視覚情報からの特徴量抽出 実験ではステレオカメラを用いる．原画像からカメラ付属の SDK(Software Development Kit) を用いて距離情報を取得し，テーブル上の黒色でない物体のみを切り出し，ラベル付けする．ただし，サイズが 100 ピクセル以下の物体はノイズとして削除する．切り出された画像は画像処理モジュールに引き渡され，物体の座標と特徴量の抽出が行われる．座標はロボットの胸部を原点とする座標系 (Fig 2) を利用する．また，物体の特徴量として色ベクトル，形ベクトル，物体ベクトルの3種類を抽出する．色ベクトルは画像の RGB 値をそれぞれ 256 で割った値を成分とする3次元ベクトルである．形ベクトルは物体の中心点とそこから最も離れた点との距離を 12 で割って，12 個の同心円をつくり，2つの同心円に挟まれる領域の面積のうち，物体が占める面積の割合を特徴量とする8次元ベクトルである．この特徴量は物体の大きさ及び回転に不変である．これら色

ベクトルと形ベクトルを併せて11次元ベクトルを生成し，物体自身を表す物体ベクトルとする．実験では実験者が物体を指差しながら，その物体に関する音声を発話する．画像処理モジュールは肌色の物体を人間の指と判断する．そして，指と平行な直線が他の物体と交差している場合に，その物体が指差されていると判断する．これは基本的な共同注視の役割を果たしており，実験者の指示でロボットが学習するために不可欠の機能である．各物体の座標と特徴ベクトルはメンタルモデルに送られ，SOINNを用いてロボットの記憶として保持される．そして，物体の属性と音声との対応付けや，座標の変化から物体の移動を検出するといった処理が行われる．

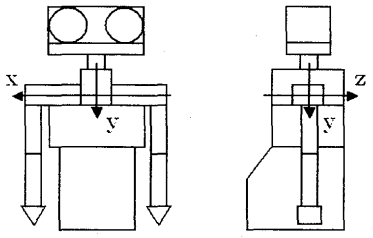

Fig. 2　The robot has own coordinate system

2·2　音声の特徴量抽出　本研究ではSpeech Signal Processing Toolkit(SPTK)[10]を基に音声処理モジュールを開発した．音声処理モジュールは，マイクを通じて入力された音声を単語単位に分割し，メルケプストラム係数を抽出する．これらはその後の音声合成に使用される．

音声はメルケプストラム係数の組合せで表現されるが，このパターン数は単語の長さに応じて指数関数的に増大するため，メルケプストラム係数をそのまま用いることは認識精度の低下を招く．これを防ぐために，VQ(Vector Quantization) 法[11]を用いてそれらを低次元に射影する．VQ法を適用すると，約100次元のベクトルが得られるが，これを単語のパターンとして扱う．

音声認識の際はDPマッチングを用いる．この手法は異なる時系列パターン間の距離を計算する一般的な手法であり，本研究ではVQ法によって得られた音声パターンに対して適用する．既知の音声のVQコードと，入力された音声のVQコードをDPマッチングで比較し，最も距離の小さいVQコードで表現される音声を認識結果とする．

2·3　SOINN　SOINN[12]はShenとHasegawaが開発した自己増殖型ニューラルネットワークと呼ばれる教師なし追加学習手法である．SOINNは通常のニューラルネットワークと異なり，ニューロンが自己増殖しながら入力ベクトルを連続的に近似することで，逐次的に与えられる入力の分布を表現するネットワークを追加的に生成する．SOINNは2層ネットワーク構造をしており，教師ラベルの付与されていないデータ群の位相構造及び適切な数のクラスタを逐次的に抽出できる．クラスタは連結したノードの集合として表現され，各ノードは入力に応じて生成・消滅を行う．各クラスタは1つのクラスを表し，クラス分類を行う上で有効なプロトタイプを持っているためコードブックとして有効な役割を果たす．このような性質を持つSOINNは，実環境で与えられるデータを追加的に学習させる本研究において極めて重要な役割を果たしている．

SOINNは次の3つの利点を持つ．(1)SOINNのクラスタは既存のクラスタに阻害されることなく成長していく．これは追加学習を行う際に不可欠な能力である．(2) 入力データに含まれるノイズを効果的かつ動的に除去する．(3) 逐次的に与えられる教師無しデータの位相構造を表現するネットワークを自律的に生成する．

本研究では上述の特徴を持つSOINNを用いて，次の4つの特徴量をクラスタリングする．(1) 色 (色ベクトル)，(2) 形 (形ベクトル)．(3) 色+形 (物体ベクトル)．(4) 軌跡 (物体の座標の時系列)．これら4つのベクトルは，入力画像から画像処理モジュールによって抽出され，各ベクトルに対応したSOINNへ入力され，分布の近いデータは集約されクラスタを形成する．このクラスタに音声をグラウンディングさせることで，概念が学習される．ただし，軌跡の学習ではオブジェクトの位置の時系列をクラスタリングするため，オーバーラップが発生し，通常のSOINNではうまくクラスタリングができない．そのため，オーバーラップに強く，教師あり学習ができるように拡張されたsupervised SOINN[13]を用いる．Fig3にsupervised SOINNの概略図を示す．このSOINNは教師なし学習，教師あり学習のどちらにも対応しているが，本研究では教師あり学習のみ行わせる．そのため，オブジェクトの軌跡と動きに関する音声情報を入力データとする．supervised SOINNは通常のSOINNと共通の入力層と1つ以上の中間層からなるニューラルネットワークで，入力された軌跡を中間層で適切な数のプロトタイプの集合としてクラスタリングし，それぞれのプロトタイプ集合と音声情報を対応づけることで動きの概念の獲得を行う．動作の生成を行う場合は音声情報を入力し，対応した軌跡を出力する．

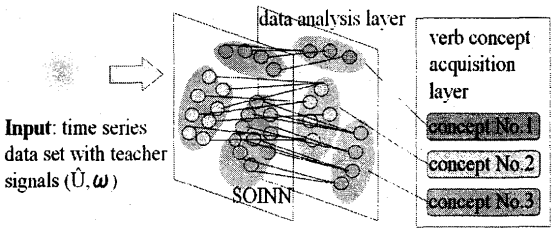

Fig. 3　The overview architecture of the supervised SOINN

2·4　メンタルモデル　メンタルモデルはロボットの内部世界として定義され，目の前に置かれている物体の数や位置情報，各物体の色や形などの情報を記憶し，状況の変化に応じて更新される．また，物体の位置変化から動きを検知し，その動きを軌跡として認識できる．メンタルモデルの主な機能は以下の3つである．(1) 物体の色と形の属性 (色ベクトル，形ベクトル) をSOINNを通して音声ラベルと対応づける．(2) 時刻Tにおける状況 (物体の数や位置など) を記憶する．(3) 時刻Tと直前の時

刻 T-1 における状況を比較することで物体の移動を検知する．

2·5 静的な概念の形成　筆者らのロボットは，指差された物体の属性に対応した SOINN 中のクラスタに，音声 (単語) を表す VQ コードのラベルを貼り付け，音声が物体の特徴を表現した言葉 (概念) だと解釈することで音声と物理的特徴とを対応づけ，概念の学習を行う．目の前に物体が置かれると，ロボットの画像認識モジュールは物体の属性に関する 3 つの特徴 (色ベクトル，形ベクトル，物体ベクトル) を抽出する．これら 3 つのベクトルは別々の SOINN に入力され，クラスタリングが行われる (Fig 4)．ただし，ニューロン数が 4 以下の小規模なクラスタは概念とはせず，ノイズとして処理する．本研究では，物体の 3 つの属性 (色，形，色+形) に関する単語は互いに異なっており，1 つの単語が複数の概念を表現すること (同音異義語) は無いと仮定している．また，物体の "色+形" という属性を表現する単語は "物体の名前" であるとしている．視覚的特徴量のクラスタリングが完了した後に実験者が物体を指差しながら発音すると，ロボットの画像認識モジュールは指差されている物体を検出し，音声認識モジュールは発音された音声を認識して音声と物体の特徴との対応付けを行う．ロボットは指差されている物体に対応する SOINN 中のクラスタ (物体の属性) のいずれかが発音された音声と対応づけられると判断し，各クラスタと音声との結合度を計算する．そして最大の結合度を示すクラスタに音声が対応していると判断する．学習初期ではすべてのクラスタが同程度の結合度であるが，例えば 2 つの異なる物体に対して指差しをしながら「黄色，四角，トパーズ」「緑，四角，エメラルド」と発音すると，「四角」という共通の単語が含まれているため，「四角」という音声と形を表す SOINN のクラスタとの結合度がより多く上昇し，正しい分類が可能となる．ロボットは上記の方法で音声と物体の特徴 (色，形，色+形) との対応付けを追加的に学習していく[9].

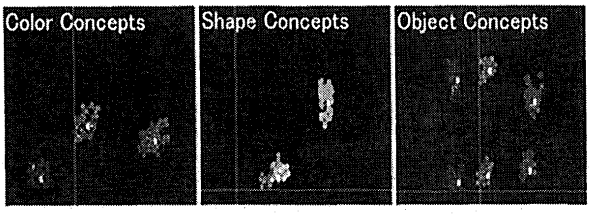

Fig. 4　Formation of static concept

2·6 動的な概念の形成　物体の静的な概念の形成が完了すると，ロボットはそれらの物体を用いて動的な概念 (動き) を学習できる．本研究では動きの概念が「動きの種類」「動作の主体」「動作の参照点」の 3 要素から構成されると仮定し，これらを学習する．例えば，「みかん，メロン，近づける」と発話しながら「みかん」を「メロン」に近づける動作を見せると，未知語である「近づける」を動きの概念を表す単語と解釈して，メンタルモデルから抽出された軌跡 (動きの種類) と対応付けて学習する．また，「みかん」が動き，「メロン」が静止していることを認識し，それぞれが動作の主体及び参照点であると解釈し，文章が「動きの主体＋動きの参照点＋動きの種類」という構造をしていることをボトムアップ及びトップダウンにより学習する．また，「みかん，上げる」のような文章の場合は動きの主体と参照点が同じであると解釈して学習する．動きの種類は軌跡を用いて学習するが，同じ動きであっても物体の位置によって軌跡は変化する．このため，軌跡の特徴量を一意に抽出するには以下で述べる座標変換が必要となる．変換後の特徴量は教師あり学習が可能なように改良された supervised SOINN によって学習される．

2·6·1 座標変換　物体 n の時刻 t におけるロボット座標系での位置を $O_{n,t}=(x_{n,t},y_{n,t},z_{n,t})$ とするとき，以下の手順で座標変換を行う．

まず，動作の主体となる物体 i の移動開始位置が原点となるように，座標系を平行移動する．

$$u,t = O_{i,t} - O_{i,0} \tag{1}$$

動作の主体と参照点が同じ物体の場合、これを変換後の座標とする。動作の主体と参照点が異なる物体の場合、これに加えて動作の主体となる物体 i と参照点となる物体 j を結ぶ直線が x 軸と平行になるように座標系を回転し、それらの距離が 1 となるように正規化する。

$$\theta = \arctan\left(\frac{z_{j,0}-z_{i,0}}{x_{j,0}-x_{i,0}}\right) \tag{2}$$

$$\tilde{u}_t = \frac{u_t}{|O_{i,0}-O_{j,0}|}\begin{pmatrix}\cos\theta & 0 & \sin\theta\\ 0 & 1 & 0\\ -\sin\theta & 0 & \cos\theta\end{pmatrix} \tag{3}$$

得られた変換後の座標を $\tilde{u}_t$ とするとき、次の式で表される軌跡 $\hat{U}$ と発話中の未知単語の音声情報 ω をペアにして、supervised SOINN に入力する学習データとする。

$$\hat{u} = (\tilde{u}_t, t) \tag{4}$$

$$\hat{U} = (\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{t-1}, \hat{u}_t) \tag{5}$$

$$Input = (\omega, \hat{U}) \tag{6}$$

2·6·2 ボトムアップ学習による語順の獲得　ボトムアップ学習では音声に含まれる単語の語順を、単語クラス間の遷移確率として学習する。本研究では全ての単語を 5 種類 (色、形、名前、動詞、未知語) の単語クラスのいずれかに分類する。発話内容を表す単語列 Q をセンテンスと呼ぶ。

$$Q = q_1, q_2, \cdots q_n \tag{7}$$

単語 q_i が属す単語クラスを $W(q_i)$ とすると、単語の語順を単語クラス間の遷移確率として扱うことができる。

$$P(W(q_{n-1})|W(q_n)) = C(W(q_{n-1})W(q_n))/C(W(q_n)) \tag{8}$$

ここで、$C(x)$ は現在の音声で発生した回数である。また、$P(W(q_n)|S)$ をセンテンスのはじめに単語クラス $W(q_n)$ が発生する確率、$P(W(E|q_n))$ をセンテンスの最後に単語クラス $W(q_n)$ が発生する確率とする。

$$P(W(q_{n-1})|S) = C(W(q_{n-1})W(q_n))/C(W(q_n)) \tag{9}$$

$$P(E|W(q_n)) = C(W(q_{n-1})W(q_n))/C(W(q_n)) \quad (10)$$

これらの遷移確率は音声が入力されるたびに更新される。ロボット自身が長さ n の文章を発話する場合、ボトムアップ学習では遷移確率を用いて語順の確率を計算する。

$$\gamma(Q) = \log P(W(q_1)|S) + \log P(E|W(q_n)) + \sum_{t=2}^{n} \log P(W(q_t)|W(q_{t-1})) \quad (11)$$

そして、確率が最も高い語順のセンテンスを出力する。

$$\hat{Q} = arg\max \gamma(Q) \quad (12)$$

2·6·3 トップダウン学習による語順の獲得 トップダウン学習では主語，目的語，動詞といった文章における単語の役割を学習する．例えば,「りんご，みかん，近づける」という文章は「名詞→名詞→動詞」としてボトムアップ学習されるが,「りんご」と「みかん」のどちらが主語 (動きの主体) でどちらが目的語 (対象物体) なのかは区別できない．トップダウン学習では，例えば,「りんご，みかん，近づける」という文章を発話しながら「みかん」が「りんご」に近づく動きを見せられた場合，ロボットはメンタルモデルを通して「みかん」が動いていることを認識し,「みかん」が主語にあたると解釈する．また,「りんご」は既知の単語であるので目的語にあたると解釈し，未知の単語である「近づける」が動詞であると解釈する．そして,「主語+目的語+動詞」という構造を獲得する．このようにしてトップダウンで文章の構造を保持し，ボトムアップが出力するセンテンスと組み合わせることで，文章の主語・目的語にあたる物体がどれであるか判断できる．これにより，人間が発話した通りの語順で学習でき，言語に依らず，英語のように動詞が目的語の前に来る文章にも，日本語のように動詞が文章の末尾に来る文章にも対応できる．

また，トップダウンで保持している構造を用いて，ボトムアップの出力を訂正することができる．例えばボトムアップが「動詞→名詞→名詞」というセンテンスを出力し，トップダウンは「主語+目的語+動詞」という構造を保持している場合，動詞は文章の主語にならないという評価基準によってボトムアップからの出力を誤りと判断し，次に高い確率のセンテンスを選択することができる．このような，トップダウン学習とボトムアップ学習の融合により，一般的なバッチ学習と異なり，逐次的に与えられる少数の例から語順を獲得することができる．

2·7 動きの認識と生成 ロボットは動きの概念を形成した後に同じような動きを見せられると，見せられた動きを表現する文書を発話することができる．また，音声によって命令された動きを実行することができる．

2·7·1 動きに関するシーンの説明 動きに関する発話を生成する際の手順を以下に示す．(1) メンタルモデルが動きを検出する．(2) その動きの軌跡を学習済みの動きの軌跡と DP マッチングを用いて比較する．(3)DP マッチングの結果が閾値以下であれば既知の動きであると判定する．(4) 動きの主語及び目的語となる物体の色，形，名前の概念とそれぞれに対応する音声を取得する．(5) ボトムアップの出力により語順を決定し，センテンスを生成する．(6) トップダウンで保持されている文章の構造を参照し，発話内容を決定する．(7) 発話により，シーンの説明を行う．

2·7·2 語順の認識と動作の生成 動きの生成は動きの学習結果と，語順の学習の結果から以下の手順で生成される．(1) 人間の発話した音声を単語に分解し，音声の VQ コードを得る．(2)VQ コードに対応した単語ラベルが概念としてグラウンディングされているか調べる．(3) グラウンディングされていたら，ボトムアップ学習により語順を決定し，センテンスを生成する．(5) トップダウンで保持されている文章の構造を参照し，ロボットが解釈可能な文書であるか判定する．(6) 解釈可能であれば，主語及び目的語に対応する語句を特定する．(7) それらの語句に対応する物体の位置と動きの軌跡を，メンタルモデル及び動きの概念から取得する．(8) 物体の位置を基に座標変換を行い，実際に生成する軌跡を計算する．(9) ロボットに動作コマンドを送信し，動作を実行する．

3. 実 験

Table 1 Event Triggers

	has audio input	**no audio input**
object pointed	static concept grounding & grammar learning	object description
has motion	dynamic concept grounding & grammar learning	motion description
no pointed no motion	point object & manipulate objects	no events

本システムの有効性を確かめるため，実験を行った．学習，認識，生成のフェーズは Table 1 のように定義する．

Step 1. 提案手法は，最大で 72 種類のオブジェクトを学習することができる[9]が，本実験ではそのうちの 9 種類 (色 3 種類，形 3 種類) を用いた．

Step 2. 実験者はオブジェクトを指差しながら色，形，名前に関する単語を発音し，ロボットに概念の学習を行わせる（例えば,"丸","赤","リンゴ" など）.9 種類のオブジェクトに対して学習させ終えたら，正しく覚えられているかを次の手順で判定し，正解率を算出した．(1) 実験者が 9 種類のオブジェクトを指差し，ロボットがオブジェクトの名前を正しく発音できるか確認する．(2) 実験者が色や形，名前に関する単語を発音し，ロボットに該当するオブジェクトを指差させて，正しく指差しをするか確認する．(3) 間違って覚えているオブジェクトがあった場合，そのオブジェクトだけ再び教える．

判定の結果，ロボットは 1 個を除いたすべてのオブジェクトに関して 1 回の学習で単語をグラウンディングした (Table 2)．間違って覚えていたオブジェクトも，再び教え

ることで正しく判定できた. ここで, 表中のPは実験者が指差しをしたという意味で,Cは色,Sは形,Nは名前に関する単語を実験者が発話したという意味である.

Table 2 recognition rates towards various combinations of object's attributes

C	S	N	P	C+N	S+N	C+S	C+S+N
100%	100%	89%	100%	100%	100%	100%	100%

Step 3. ロボットがすべてのオブジェクトを覚え終えたら,9種類のオブジェクトから3種類(動作主体, 参照点, 動作に関わらない物体)を選んで実験テーブル上に配置し, 6種類の動きをそれぞれ1回ずつ学習させた. 学習させた動きはFig 5の矢印で示すような軌跡で,6種類(近づける, 遠ざける, またぐ, まわす, 上げる, 下げる)である. 追加的に学習できる事を示すため,2種類の動きを教えるごとにStep 4の手順で学習した動きを正しく覚えているかの判定を行った.

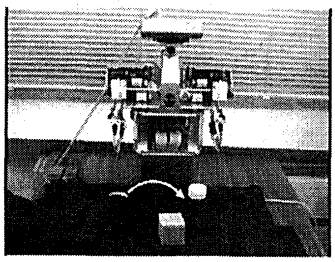

Fig. 5 Appearance of "Meron Mikan chikazukeru"

Step 4. 次の手順でロボットに動きの説明と生成を行わせた. (1) 実験者が学習させた動詞を含む文章(例:みかん メロン 近づける)を発話し, ロボットに動きの再現を行わせ, 正しく再現したかを確認する. (2) 実験者が学習させた動きを再現してロボットに見せ, 動きの説明を行わせて正しい説明か確認する.
ただし, 学習時に使用したオブジェクトは用いない. ひとつの動きに対して説明と生成をそれぞれ10回ずつ行わせ, 試行回数のうち正しく行った割合を算出した後, 平均を取って正解率とした(Fig 6).

6種類の動きを学習した場合, 動作の生成は100%, 動作の認識は92%の正解率となり, 既存の知識に学習が阻害されずに安定した学習が行えるということが示された. 以上の実験結果から, 提案手法は実環境で得られる逐次的なデータから学習が正しく行えること, 追加学習を行っても安定した認識率を示すことが示された.

4. まとめ

本研究では, 乳幼児の発達のように, 視覚, 聴覚から得られた情報を基に物体に関する言葉を学習し, 物体の動き, 発話の語順などを前提知識のない状態から発達的に獲得していく仕組みを提案した. 従来手法では物体の学習, 動きの学習, 文法(語順)の学習のための学習データをバッチ的に与えているため学習時間がかかり, オンラインでの学習や追加学習が困難であったが, 提案手法ではSOINNの特性を生かすことによって, 人間とコミュニケーションを行いながら少数のデータで学習を行うシステムを実現し, 優位性が示された.

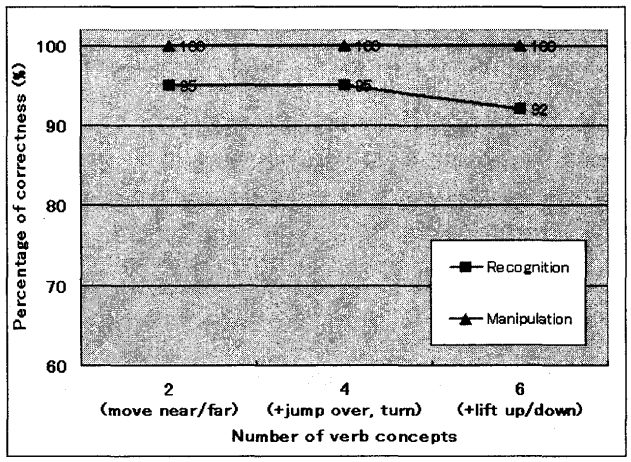

Fig. 6 Recognition of motion and manipulation

文献

(1) 浅田稔, "認知発達ロボティクスによる赤ちゃん学の試み", ベビーサイエンス, 4,(2004),pp2–27

(2) J.Weng, J.McClelland, A.Pentland, O.Sporns, I.Stockman, M.Sur, and E.Thelen, "Autonomous mental development by robots and animals", *Science*, **5504**-291, (2000)

(3) Steven Harnad, "The symbol grounding problem", *Physica*, D42, (1990), pp335–346

(4) J.M.Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings", *Cognition*, 61, (1996), pp1–2

(5) Deb Roy , Alex Pentland, "Learning words from sights and sounds: A computational model", *Cognitive Science*, 26(1), (2002), pp113–146

(6) Chen Yu and Dana Ballad, "On the integration of grounding language and learning objects", *In Nineteenth National Conference on Artificial Intelligence(AAAI '04)*, (2002)

(7) Naoto Iwahashi,"Active and unsupervised learning of spoken words through a multimodal interface", *Information Sciences*, , (2003), pp.109–121

(8) Naoto Iwahashi, "Language acquisition through a human-robot interface by combining speech, visual , and behavioral information", *13th IEEE Workshop Robot and Human Interactive Communication*, , (2004), pp.437–442

(9) Xiaoyuan He, Ryo Kojima, Osamu Hasegawa, "Developmental Word Grounding through A Growing Neural Network with A Humanoid Robot" , *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, (2006)

(10) S. Imai , T.Kobayashi , K. Tokuda, T. Masuko, "Speech signal processing toolkit: Sptk version 3.0", (2002)

(11) A. Gersho , R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Boston, (1992)

(12) Shen Furao and Osamu Hasegawa, "An Incremental Network for On-line Unsupervised Classification and Topology Learning", *Neural Networks*, (2006)

(13) Y.Kamiya,F.Shen and O.Hasegawa, "Non-stop Learning : A New Schem for Continuous Learning and Recognition", *Joint 2nd International Conference on Soft Computing and Intelligent Systems and 5th International Symposium on Advanced Intelligent Systems*, (2004)