論　文

人間との相互作用に基づくヒューマノイドロボット上の語順と挙動のオンライン学習

佐藤　彰洋[†]　賀　小淵[†]　小倉　和貴[†]　長谷川　修[††a)]

Developmental Word Acquisition and Grammar Learning by Humanoid Robots through SOINN

Akihiro SATOU[†], Xiaoyuan HE[†], Tomotaka OGURA[†], and Osamu HASEGAWA[††a)]

あらまし　実環境においてロボットと人間が円滑なコミュニケーションを行うには，新しい言語獲得手法を開発する必要がある．本研究では，人間とのコミュニケーションを通じて動きの概念と文章の構造を獲得する手法を提案する．提案手法は自己増殖型ニューラルネットワーク（SOINN）とトップダウン処理，ボトムアップ処理を組み合わせることで，実環境から逐次的に得られる少数のデータを用いて単語の意味や文法構造，動きの概念を追加的に学習できる．更に動きを自身で生成したり，動きを言語によって説明することができる．実験により，提案手法がオンラインで与えられる少数の学習データから正しく言語の学習が行えることを示す．

キーワード　自己増殖型ニューラルネットワーク（SOINN），メンタルモデル，インタラクティブラーニング

1. まえがき

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），その環境内で起こる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．例えば，これまで，歩く，走る，人間と共同作業を行うなどの多彩な行動が可能な人間型ロボットが開発されているが，そうしたロボットのほとんどはあらかじめ人間が組み込んだ動作や行動を再現している．実世界の環境は常に変化し，極めて複雑なため，事前にすべての状況に対する処理を組み込むことは事実上できない．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汎用の（タスクに依存しない）知能ロボットの構成が目指されている [1]．人間と共存する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）　行動環境や扱うタスクが限定されておらず，あらゆる状況を想定して処理を組み込むという従来法では設計できない．

（2）「発達」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，適応していくことが求められる．

（3）　人間とのコミュニケーション能力が必要不可欠である．

これらを実現するには，概念に対して人間と共通の理解を発達とコミュニケーションを通じて獲得することが必要と考えられる．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内部

† 東京工業大学大学院総合理工学研究科知能システム科学専攻，横浜市
Department of Computational Intelligence and System Science, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama-shi, 226–8503 Japan
†† 東京工業大学像情報工学研究施設，横浜市
Imaging Science and Engineering Laboratory, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama-shi, 226–8503 Japan
a) E-mail: hasegawa@isl.titech.ac.jp

に形成することで様々な単語を識別している．そのため，ロボットに事物のシンボルと単語との対応付けを獲得させることで単語の意味を獲得し，人間と概念を共有するという手法が考えられる．

また，Siskind [4] は単語の意味獲得に関する問題について議論し，五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．

単語の意味獲得に関する研究は広く行われている．Roy，Pentland ら [5] は，音声情報と視覚情報の相互情報量を定義し，これを最大化するというアルゴリズムを提案した．しかし彼らが用いた視覚情報は静止画像であり，自然なインタラクションから得られた音声情報を用いているものの，それらを用いたオンライン学習及び追加学習は行われていない．Yu，Ballard ら [9] が提案したマルチモーダル学習システムは，実世界の物体に発話された単語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ないほか，新たな物体の学習が既存の学習結果に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．

ここまで述べてきた従来の研究はマルチモーダルな知覚を通じて単語単体の意味を学習することに主眼を置いており，文脈の中で表現される単語の文法的な意味を扱っていない．文章における単語の意味を理解するには単語単体での意味を理解するだけでなく，文章の構造（文脈）を考慮した上で単語の意味を理解する必要があり，本研究ではこれを実現する．

文章における単語の意味獲得に関する従来研究として，岩橋らは，物体が移動する二次元動画像とそれを表現する文章の組から動きの概念と文法の獲得を行う手法 [6] 及び，能動学習と教師なし学習を用い，視覚概念と音声概念のペアを追加学習可能な手法 [7] を提案した. [7] では，音声データが Hidden Markov Model（HMM）によって，また視覚データがガウス関数によってモデル化され，オンラインで追加的に単語の意味を獲得するシステムが提案されている. [6] では同様に，HMM ベースで物体の静止画像と物体に関する単語の発話から単語の意味を獲得し，その後物体の操作と操作を説明する文章から文法を獲得するシステムが提案されている．

しかし, [8] で指摘されているように，HMM によるオンライン追加学習は原理的には可能なものの，あらかじめ十分な訓練データを与えなければ初期認識精度が低く，更に追加学習による認識精度の向上も望めない．

我々は Siskind が提起した五つの問題のうち三つに挑んでいる．すなわち，ロボットは学習する言語に対する事前知識をもたず（事前知識なしでの学習），日本語のように「主語—目的語—述語」という語順の言語であっても，英語のように「主語—述語—目的語」という語順の言語であっても学習できる．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，事前の対応付けは行わない．実環境から得られる情報はノイズを含む場合や不完全な場合があるため，ノイズを軽減する機能（ノイズ処理）及び再学習により誤った学習を修正する機能をもつ．更に，単語がどの概念（名前，色，形，動きなど）を表すかはロボット自身が獲得する（概念の分類問題）．

1.1 本研究のアプローチ

本研究では [6] 同様，物体の静止画像と物体に関する単語の発話から単語単体の意味を獲得した後に，物体の操作と操作を説明する文章から単語の文章における意味を獲得するシステムを提案する．本研究の特徴は以下の 3 点である．

（1） 提案手法は事前知識なしの（訓練データを用いない）状態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．これは物体及び動きの特徴ベクトルの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 [11]）を用いることで実現する．実世界のあらゆる事柄について，事前に訓練データを用意しておくことは事実上不可能であり，実環境で動作するシステムにこの機能は必須である．

（2） SOINN を利用することで，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 [10]）を回避できる．また，SOINN は学習した知識を十分に表現できるネットワークまで自己組織的に成長するため，HMM を用いる [6], [7] の手法のように，学習したいデータに応じてモデルのトポロジーを事前に決定しておく必要がない．

（3） ロボットは物体の静的な特徴（色や形などの属性）を表す単語と動的な特徴（物体の動き）を表す単語の両方を学習することができる．更に，多くの従来手法と異なり逐次的に与えられる音声の文法的な構

造を分析し，文章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，自らも文章を発したり，文章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処理とトップダウン処理の融合によって実現される．ボトムアップ処理では単語の種類（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，語順の対数ゆう度（以後，ゆう度）を計算する．トップダウン処理では「メンタルモデル」を用いて単語の文章における役割の観点から文法構造を学習する．従来のバッチ的な学習と異なり，少数の例から（一つの例からでも）文法構造の学習を可能としている．また，文章の構造をそのまま学習するアルゴリズムであるため，任意の語順（文法）での学習が可能である．

以下，**2.** を通じてシステムの動作する環境とシステムを構成するモジュールについて述べ，**3.** を通じてシステムの学習・認識時の動作について述べる．

2. 提案手法

2.1 概　　要

本研究では，様々な色，形の物体が置かれた黒いテーブルが実験環境である（図 1）．上半身と腕からなるヒューマノイドロボット IKR1 は 2 自由度をもつ頭部に市販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，腕とグリッパで 5 自由度をもち，グリッパを使って物体をつかんで移動させることができる．提案手法は以下に示すいくつかのモジュールからなる．

（1） 画像処理モジュール，音声処理モジュール．

（2） SOINN：物体の画像から得た特徴量をクラスタリングでき，静的な概念の学習に用いられる．

（3） メンタルモデル：外界に対応したロボット内部の表現である．

（4） supervised SOINN：教師あり学習が可能なように拡張した SOINN であり，動きの概念の学習に用いられる．

これらの関係を図 2 に示す．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」をもち，テーブルの上にある物体に対して注意を向ける．次に，目の前に置かれた物体の色と形に関する特徴量を画像処理モジュールで取得し，未知の物体ならば物体の色，形，色＋形という特徴に対応した三つの SOINN でクラスタリングを行う．物体の特徴のクラスタリングが完了した後，ロボットは **3.2** で述べる手順で物体の視覚的特徴と教示された音声を対応づけ，「“黄色” という音声は色を表す」といったように物体の特徴を表す単語を学習する．

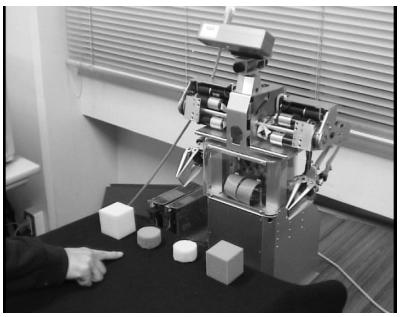

図 1　実験に使用するロボット
Fig. 1　The robot used for the experiment.

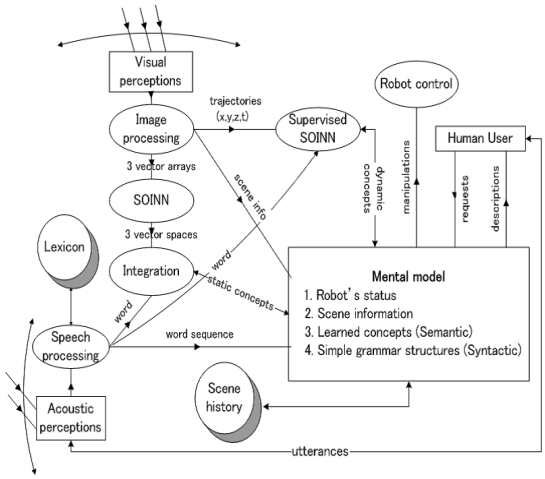

図 2　システム概要図
Fig. 2　The architecture of the proposed system.

提案手法が扱うタスクは，物体の視覚的特徴を表す単語の学習が完了したという状況において，単語の文章における意味を獲得するというものである．ロボットは物体の特徴を表す単語の学習が完了した後，**3.4** で述べる手順を通じて単語の文章における意味を獲得する．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，みかん，近づける」のような文章を発話すると，ロボットは **3.4.2** で述べるボトムアップ処理と **3.4.3** で述べるトップダウン処理を組み合わせ，文章のどの部分が主語の役割を果たし，どの部分が述語の役割を果たすというような単語の文章中における意味を学習する．

学習後，同様の文法的構造をもつ文章を与えられる

と，ロボットは自分のグリッパを使ってその文章が意味する動きを生成することができる．もちろん，学習時に用いた物体でなくとも，単語の単体での意味と文章における意味を正しく解釈し，動きを生成することができる．また，既知の動きを見せられると，その動きを表現する文章を構成して発話することができる．なお, [6] では文章における単語の意味の学習後，提案手法が用いる文章と同様の文章を与えて動きを生成する実験については記述があるが，既知の動きを見せられた場合に文章として発話するという実験については記述がない．

2.2 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．カメラが取得した画像に対し，付属の SDK（Software Development Kit）を用いてしきい値以上の距離にある領域を処理対象から除外する．次に，黒色でなく，かつ 100 ピクセルを超える領域を物体とみなす．

そして，物体の RGB 値を色ベクトル（三次元）とし，小島らと同様の手法 [12] を用いて形ベクトルを抽出する（八次元）．この色ベクトルと形ベクトルを併せて物体ベクトル（十一次元）とする．

これら 3 種類のベクトルは **2.5** で述べるメンタルモデルに送られ，SOINN によりクラスタリングされる．更に，各物体の座標はメンタルモデルにおいて 1 時刻前の座標と比較して物体の移動を検出するといった処理が行われる．物体の特徴量である 3 種類のベクトル及び座標の時系列は **3.2** 及び **3.4** で述べる処理を通じて音声との対応付けが行われる．

2.3 音声情報からの特徴量抽出

本研究では Speech Signal Processing Toolkit（SPTK）[13] をもと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入力された音声にしきい値以上の間隔があればそこで単語に分割してメルケプストラム係数を抽出する．次に，VQ（Vector Quantization）法 [14] を用いてメルケプストラム係数の量子化を行いコード化する．以後，このコードを単語の音声ラベルとする．

音声認識には DP マッチングを用いる．テンプレートとして登録されている既知の音声の 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，算出した最小の距離がしきい値を下回った場合にその VQ コードで表現される音声を認識結果とする．しきい値を下回るテンプレートがなかった場合は新たな音声と認識し，音声の VQ コードをテンプレートに追加する．

2.4 SOINN

SOINN [11] 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 は結合荷重をもつノードとノードを結ぶエッジからなるネットワークで，事前知識なしで学習を行う．ネットワークは空の状態から学習を開始し，ノードとエッジが入力に応じて増殖・消滅を繰り返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ために最適なネットワークを生成する．SOINN は教師ラベルの付与されていない入力ベクトル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．クラスタはエッジにより連結されたノードの集合として表現される．各クラスタは一つのクラスを表し，クラス分類を行う上で有効なプロトタイプをもっているためコードブックとして有効な役割を果たす．

このような性質をもつ 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重要な役割を果たしている．

SOINN は次の三つの利点をもつ．

（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なしデータのトポロジーを過不足なく表現するネットワークを自己組織的に生成する．

本研究では上述の特徴をもつ SOINN を用いて，次の四つの特徴量をクラスタリングする．（1）色ベクトル，（2）形ベクトル，（3）物体ベクトル（色ベクトル+形ベクトル），（4）軌跡（物体の座標の時系列）．四つのベクトルは，画像処理モジュール及びメンタルモデ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．このクラスタに単語の音声ラベルをグラウンディングさせる．

ただし，軌跡の学習では物体の位置の時系列をクラスタリングするため，複数の軌跡の間でオーバラップが発生し，SOINN ではうまくクラスタリングができない．そのため，オーバラップに強く，教師あり学習ができるように拡張した supervised SOINN を用いる．図 3 に supervised SOINN の概略図を示す．super-

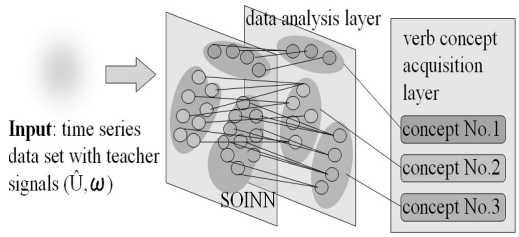

図 3　supervised SOINN の概要
Fig. 3　The overview architecture of the supervised SOINN.

vised SOINN は SOINN 同様，ネットワークが空の状態から学習を開始し，ノードとエッジが学習に応じて生成・消滅を繰り返しながら入力データを近似するために最適なネットワークまで成長する．supervised SOINN は大別してデータ分析層（図中 data analysys layer）と概念獲得層（図中 verb concept acquisition layer）からなる．

データ分析層は SOINN を二層重ねた構造をしている．第二層には第一層がクラスタリングしたノードのもつベクトルを一定時間ごとに入力する．入力データにノイズが乗っている場合，第一層だけではノイズを除去しきれず，ノイズをノードとして記憶してしまう可能性がある．そのため，第一層に形成されたノードのベクトルを第二層に入力し，クラスタを代表するのによりふさわしい少数のノードを SOINN のアルゴリズムにより残し，ノイズの影響を軽減する．実際, [11] では二層構造のネットワークを用いる事で，ノイズ耐性が向上したと報告されており，supervised SOINN はこれを継承して二層構造としている．

概念獲得層はデータ分析層の第二層が生成したクラスタを概念とみなし，教師信号との対応付けを行う．本研究では，物体の動いた軌跡 $\hat{U}$ を入力データ，単語の音声ラベル ω を教師信号として軌跡をクラスタリングし，単語の音声ラベルを教師信号としてクラスタに対応づける．動作の生成を行う場合は動きに関する単語の音声ラベルを入力し，軌跡を出力する．

2.5　メンタルモデル

我々はロボットの内部世界 [2], [15] を表現するためにメンタルモデルを開発した．ロボットはコンピュータビジョンの技術を用いて，目の前に置かれている物体の数や位置，各物体の色や形などの情報を逐次的に取得する．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する．

メンタルモデルの主な機能は以下の三つである．

（1） 目の前に置かれた物体の色と形の特徴ベクトル（色ベクトル，形ベクトル，物体ベクトル）を事前に用意された三つの SOINN にクラスタリングさせる．生成されたクラスタは **3.2** で述べる手順で単語との対応付けが行われる．

（2） 時刻 T と T − 1 における状況（物体の数や位置など）を記憶する．これらの情報は **3.4.1** で述べる単語の文章における意味の獲得時や，**3.4.4** で述べる動きに関する単語の意味獲得時に利用される．また，時刻 T と T − 1 の情報を比較し，同一物体の移動を検知した場合は静止するまで追跡し，移動した軌跡を記憶する．

（3） **3.4.2** で述べるボトムアップ処理，**3.4.3** 及び **3.6.1** で述べるトップダウン処理において，物体の特徴ベクトル及び動きの軌跡に対応づけられた音声ラベルを提供する．あるいは音声に対応づけられた特徴ベクトル及び軌跡の情報を提供する．

3.　システムの動作

3.1　概　　要

提案手法は音声入力の有無，動作教示の有無と指差しの有無によって表 1 に示すように学習フェーズ，認識フェーズを決定し動作する．

3.2　フェーズ 1：静的な特徴のシンボルグラウンディング

メンタルモデルによって物体の視覚的特徴が SOINN にクラスタリングされた後，教示者が物体を指差し（表 1 中 object pointed），物体の静的な特徴（色，形，名前）に関する文章を発話した場合（表 1 中 has audio input），システムは以下で述べる手順によって静的な特徴のシンボルグラウンディングを行い，**3.4.2** で述べるボトムアップ処理によって語順のゆう度を更新する．

ロボットは，指差された物体の特徴ベクトル（色ベクトル，形ベクトル，物体ベクトル）に対応した SOINN のクラスタに，単語の音声ラベルを対応づけ，音声が物体の特徴を表現する単語だと解釈することで単語単体の意味を学習する．詳細は [12] を参照されたい．

本研究では，すべての単語は 5 種類（色，形，名前，動詞，未知語）の単語クラスのいずれかに属すると仮定している．そして，単語の属するクラスは学習により獲得される．静的な特徴のシンボルグラウンディングにおいては，単語と物体の属性（特徴ベクトル）と

表 1　学習・認識フェーズの選択
Table 1　Selection of learning phase or recognition phase.

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

を学習により対応づけ，色ベクトルに対応づけられた単語は “色” という単語クラスに属するというように解釈することで単語を三つ（色，形，名前）の単語クラスに分類する．ここで，物体の三つの属性に関する単語は互いに異なっており，一つの単語が複数の概念を表現すること（同音異義語）はないと仮定している．また，物体の “物体ベクトル” という属性に対応づけられる単語は “名前” という単語クラスに属すると解釈する．

ただし，システムは発音された単語が物体の三つの属性のうち，どの属性に対応する単語なのかという情報は与えられず，対応づけを学習により獲得する．システムは物体の特徴ベクトルに対応して用意された三つの SOINN 中に存在するクラスタのいずれかが発音された単語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確信度という値で表現し，計算する．そして最大の確信度を示すクラスタに単語が対応していると判断する．システムは上記の方法で音声と物体の属性（色ベクトル，形ベクトル，物体ベクトル）との対応づけを追加的に学習していく．

3.3　フェーズ 2：物体の説明

物体の指差し（表 1 中 object pointed）のみで発話がない（表 1 中 no audio input）場合，システムは指差された物体の静的な特徴に関する文章を発話する．物体の説明を行う場合の語順はボトムアップ処理を通じて得られた語順であり，例えば「赤」や「赤　丸」，「赤　丸　りんご」といった語順の文章が考えられる．指差された物体に対応づけられた単語が存在しない場合，システムは発話を行わない．

3.4　フェーズ 3：動的な特徴のシンボルグラウンディング

物体の静的な特徴のシンボルグラウンディングが完了した後，教示者が物体を操作し（表 1 中 has motion），動きを表現する文章を発話した場合（表 1 中 has audio input），システムは物体の静的な特徴に関する知識を用いて動きに関する知識を学習する．

本研究では二つの物体がかかわる「りんごをみかんに近づける」というような動きと，一つの物体がかかわる「りんごを上げる」というような動きを対象とする．

3.4.1　単語の文章における意味の獲得

例えば，既知の物体を用いて,「りんご」を「みかん」に近づける動作を見せながら「りんご，みかん，近づける」という文章を人間が発話すると，システムは見せられた動作と文章が対応していると解釈し，**3.4.2** で述べるボトムアップ処理と **3.4.3** で述べるトップダウン処理を通じて文章を解析し，動作の主体となる単語はどれか，動詞にあたる単語はどれかといったように文章の意味を認識しながら学習を行うことで少数の例から単語の文章における意味を獲得する．

3.4.2　ボトムアップ処理

ボトムアップ処理では，すべての単語を 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．ここで，3 種類（色，形，名前）の単語クラスへの分類は静的な特徴のシンボルグラウンディングを通じて獲得された分類を用いる．また，単語クラス “動詞” への分類は，動的な特徴のシンボルグラウンディングの過程で行う．ここで，物体の動いた軌跡に対応づけられる単語クラスを “動詞” とする．

ロボットは一つの単語は一つの単語クラスに属する（一つの属性に対応する）という判断基準をもち，この基準に基づいて動詞を判定する．例えば，既知の物体名「みかん」と「メロン」を用い,「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,「みかん」及び「メロン」は物体ベクトルに対応づけられた既知の単語であると判定し，物体の動いた軌跡と「近づける」という未知単語が対応づけられると判断し,「近づける」を “動詞” という単語クラスに分類する．

そして文章における単語の語順を，統計的 Bigram モデルを用いて「単語クラス」間の遷移確率として学習する．

$$Q = q_1, q_2, \cdots q_n \tag{1}$$

ここで，発話内容を表す単語列 Q をセンテンスと呼ぶ．単語 q_i が属す単語クラスを $\mathrm{W}(q_i)$ とすると，単語クラス間の遷移確率を使って語順のゆう度を計算で

きる．

$$P(W(q_n)|W(q_{n-1})) = C(W(q_{n-1}) \rightarrow W(q_n))/C(W(q_{n-1})) \quad (2)$$

ここで，$C(x)$ はこれまでに x が登場した合計回数である．例えば，$C(W(q_{n-1}) \rightarrow W(q_n))$ はこれまでに発話された全文章の中で単語クラス $\mathrm{W}(q_{n-1})$ の単語の後に単語クラス $\mathrm{W}(q_n)$ の単語が登場した回数である．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．これらの遷移確率は文章が入力されるたびに更新される．

フェーズ 2（物体の説明）やフェーズ 4（動作の説明）においてロボット自身が長さ n の文章を発話する場合，ボトムアップ処理では遷移確率を用いて語順のゆう度を計算する．

$$\gamma(Q) = \log P(W(q_1)|S) + \log P(E|W(q_n)) + \sum_{t=2}^{n} \log P(W(q_t)|W(q_{t-1})) \quad (3)$$

そして，ゆう度が最も高い語順のセンテンスを出力する．

$$\hat{Q} = argmax\gamma(Q) \quad (4)$$

しかし，単語の組合せの多様性のため，長さ “n” を一意に決めることができないという問題がある．例えば，“みかん” について発話する場合,「みかん」と一つの単語で表現する場合や「丸い，みかん」と二つの単語で表現する場合などが考えられる．また，“近づける” という動きを含む文章を発話する場合でも,「みかん，メロン，近づける」という発話や「丸い，みかん，緑，メロン，近づける」という発話のように，様々な長さの文章が考えられる．本研究では，ロボットが発話を行う際，次の条件をすべて満たす語順を適切な語順とする．

（1） 学習時に使用された語順
（2） 長さが最大の語順
（3） 登場回数が最多な語順

すなわち，ロボットは人間が学習時に用いた語順を正しい語順として扱い，これを逸脱する語順を棄却する．ここで，長さが最大という条件を用いているのは，単語を多く使用している文章の方が情報量が多いためである．そして，適切な語順を以下のアルゴリズムで決定する．ただし，ロボットが文章を受理する場合（人間が文章の発話を行う際）は学習時に使用された語順であれば，どのような長さの語順であっても適切な語順となる．

ボトムアップ処理

input：

（1） *target*：発話の目的
（2） *History of input*：学習の際に入力された音声の語順
（3） *set of verbs*：記憶している「動きの種類」,「動作の主体」,「動作の参照点」からなる文章構造
（4） *trajectry*：現在見えている物体の軌跡（動きの説明を行う場合のみ）

if（*target* が物体の説明ならば）
  then $n \leftarrow$ 対象となる物体に関する既知の静的な概念の数
  loop do
  $Q_n \leftarrow$ Bigram モデルを用いて生成した，長さ n の語順のうち最もゆう度が高い語順
  if（Q_n が *History of input* に含まれていれば）
    then return Q_n
    else $n \leftarrow n-1$
if（*target* が動きの説明）
  then $V_i \leftarrow$ *trajectry* を DP マッチングで比較し，最も似ている動きのインデックス
    if（*set of verbs* が存在すれば）
      then $L_{max} \leftarrow$ *History of input* 中の最も長い文章の長さ
      $Q_{L_{max}} \leftarrow$ Bigram モデルを用いて生成した，長さ L_{max} の語順のうち最もゆう度が高い語順
      loop do
      if（$Q_{L_{max}}$ が *History of input* に含まれており，かつ $Q_{L_{max}}$ 中に動詞が含まれていれば）
        then return $Q_{L_{max}}$
        else $Q_{L_{max}} \leftarrow$ Bigram モデルを用いて生成した，長さ L_{max} の語順のうち次にゆう度が高い語順

3.4.3 トップダウン処理の更新

トップダウン処理では文章が「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，文章における単語の意味をこの 3 要素で表現

する．そして，トップダウン処理はこれら 3 要素で表現された文章の構造を記憶する．

例えば,「りんご，みかん，近づける」という文章はボトムアップ処理を通じて「名前 → 名前 → 動詞」という単語クラスの遷移として認識されるが,「りんご」と「みかん」のどちらが動作の主体（主語）でどちらが動作の参照点（目的語）なのかは区別できない．トップダウン処理では「りんご，みかん，近づける」という文章を発話しながら「りんご」が「みかん」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利用することで動作の主体と動作の参照点を推定する．

まず，メンタルモデルに対して目の前にある物体に対応する単語を問い合わせる．メンタルモデルは視界に存在する物体の特徴ベクトルを学習済みの三つの SOINN に対して入力し，クラスタとして存在するかを問い合わせる．クラスタが存在すれば，対応づけられた音声ラベルを取得し，発話された文章中の音声ラベルと比較し，同一の単語が存在するかを確認する．

以上の処理を通して文章中に登場する「りんご」及び「みかん」という単語は “名前” という単語クラスの単語であることを認識する．(「りんご」「みかん」以外の物体が視界内にあったとしても，この処理を通じて文章に登場する物体だけをトップダウン処理の対象とする）また,「りんご」は視界内で動いている物体であることを認識し，文章を構成する三つの単語のうち，1 番目に位置する単語は「動作の主体」という意味をもつと解釈する．更に,「みかん」という名前の物体が視界内に存在し，動いていないことを認識する．そして文章を構成する三つの単語のうち，2 番目に位置する単語は「動作の参照点」にあたると解釈する．

そして，残った単語「近づける」について既知か否かの判定を行い，この単語を割り当てられたクラスタは存在せず，かつボトムアップ処理を通じて「近づける」という単語が「動詞」という単語クラスに属すと判定された場合，トップダウン処理はこの単語を「動きの種類」であると解釈する．

総じて，トップダウン処理は「動作の主体+動作の参照点+動きの種類」という文章の構造を獲得し記憶する.「りんご，上げる」のように「動作の参照点」が存在しない文章の場合,「動作の主体」と「動きの種類」からなる構造を獲得し記憶する．ただし,「りんご，みかん，近づける」と発話しながら「みかん」が「りんご」に近づく動きを見せられた場合，上記の手順と同様の手順により，トップダウン処理は「動作の参照点+動作の主体+動きの種類」という文章の構造を獲得し記憶する．

また,「赤　丸　りんご　黄色　丸　みかん　近づける」という文章を発話され,「りんご」が「みかん」に近づく動きを見せられた場合，ボトムアップ処理は「色 → 形 → 名前 → 色 → 形 → 名前 → 動詞」という単語クラス間の遷移と認識する．トップダウン処理はメンタルモデルへの問合せを通じて,「赤」「丸」「りんご」及び「黄色」「丸」「みかん」という単語がそれぞれの物体に対応づけられることを認識し,「動作の主体 → 動作の主体 → 動作の主体 → 動作の参照点 → 動作の参照点 → 動作の参照点 → 動作の種類」という構造を記憶する．

3. 4. 4　単語単体での意味の獲得

単語の文章における意味の獲得が完了した後，以下に述べる手順で物体の動いた軌跡を supervised SOINN によりクラスタリングし，動詞にあたると判定された単語の音声ラベルをクラスタに対応づけることで動きに関する単語単体の意味を学習する．

まず，同じ動きであっても物体の位置関係によって軌跡の見え方は変化するため，以下で述べる座標変換により軌跡の特徴量を一意に抽出する．

物体 n の時刻 t における位置を $O_{n,t} = (x_{n,t}, y_{n,t}, z_{n,t})$ とするとき，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．

$$u_t = O_{i,t} - O_{i,0} \tag{5}$$

一つの物体がかかわる動きの場合，これを変換後の座標とする．二つの物体がかかわる動きの場合，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，直線の距離が 1 となるように正規化する．ここで,「動作の主体」及び「動作の参照点」となる物体はボトムアップ処理，トップダウン処理を通じてシステムが推定した物体である．

$$\theta = \arctan\left(\frac{z_{j,0} - z_{i,0}}{x_{j,0} - x_{i,0}}\right) \tag{6}$$

$$\tilde{u}_t = \frac{u_t}{|O_{i,0} - O_{j,0}|} \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \tag{7}$$

得られた変換後の座標を $\tilde{u}_t$ とするとき，軌跡 $\hat{U}$ を

次の式で表す.

$$\hat{u} = (\tilde{u}_t, t) \tag{8}$$

$$\hat{U} = (\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{t-1}, \hat{u}_t) \tag{9}$$

得られた $\hat{U}$ に対して乱数による微小な摂動を加え，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を 200 個用意する．ここで，乱数の大きさは実験的に定めた値を用いる．

$$Input_i = (\omega, \hat{U}_i) \quad (i = 1, 2, \ldots, 200) \tag{10}$$

supervised SOINN は学習データを入力として少数のノードからなるクラスタを形成し，音声ラベル ω とペアにして記憶する．

3.5 フェーズ 4：動作の説明

フェーズ 3（動的な特徴のシンボルグラウンディング）を通じて動きに関する単語の単体での意味と文章中における意味を学習した後に，学習した動きと同じような動きを見せられ（表 1 中 has motion)），かつ発話がない場合（表 1 中 no audio input），以下の手順で見せられた動きを表現する文章を発話する．

（1） メンタルモデルが物体の動きを検出する．

（2） その際，音声入力がなければ動きの説明を求められていると判断する．

（3） **3.4.2** のボトムアップ処理により単語クラスの語順を決定し，センテンス Q_k を生成する．

（4） **3.6.1**（トップダウン処理を用いた文章構造の決定）の手順により，発話内容 U_k を決定する．

（5） 発話により，シーンの説明を行う．

3.6 フェーズ 5：動作の生成

フェーズ 3（動的な特徴のシンボルグラウンディング）を通じて動きに関する単語の単体での意味と文章中における意味を学習した後に，学習時に用いた文章と同様の構造をもつ文章が発話され（表 1 中 has audio input），かつ動作も指差しもない場合（表 1 中 no motion, no pointed），以下の手順で発話された文章が表現する動作を生成する．

（1） 人間の発話した文章 U を単語に分解し，各単語の VQ コードを得る．文章を構成する単語の個数を k とする．

（2） **3.4.2** のボトムアップ処理により単語クラスの語順を決定し，長さ k のセンテンス Q_k を生成する．

（3） Q_k 中に単語クラスが「動詞」の単語が含まれていれば，動きの生成を求められていると判断する．

（4） **3.6.1**（トップダウン処理を用いた文章構造の決定）の手順により，文章 U の「動作の主体」,「動作の参照点」及び「動きの種類」を表す単語を特定する．

（5） 「動作の主体」,「動作の参照点」を表す単語に対応する物体の位置を，メンタルモデルから取得する．

（6） 「動きの種類」を表す単語に対応する軌跡を supervised SOINN から取得する．

（7） supervised SOINN から取得した軌跡は，物体の位置関係に依存しない座標系での軌跡になっているため，後述する手順に従って座標変換を行い，軌跡を求める．

（8） ロボットに動作コマンドを送信し，動作を実行する．

supervised SOINN から取得した軌跡を，物体の位置関係に応じた軌跡へ変換する方法を示す．メンタルモデルから取得した「動作の主体」,「動作の参照点」を表す単語に対応する物体の座標をそれぞれ次のように定義する．

$$O_s = (x_s, y_s, z_s) \tag{11}$$

$$O_o = (x_o, y_o, z_o) \tag{12}$$

また「動きの種類」を表す単語に対応する軌跡を次のように定義する．

$$\hat{U} = (\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{n-1}, \hat{u}_n) \tag{13}$$

$$\hat{u}_n = (x_n, y_n, z_n, t_n) \tag{14}$$

ここで，n は系列の数である.「動作の参照点」が文章中に存在しない場合，物体の位置が原点になるよう座標系を水平移動させて生成する軌跡を決定する．そうでない場合，次の式を用いて生成する軌跡を決定する．この軌跡は物体の位置関係に応じた軌跡である．

$$\omega = \arctan\left(\frac{z_{j,0} - z_{i,0}}{x_{j,0} - x_{i,0}}\right) \tag{15}$$

$$scale = |O_s - O_0| \tag{16}$$

$$\tilde{u}_n = \hat{u}_n \begin{pmatrix} \cos(-\omega) & 0 & \sin(-\omega) \\ 0 & 1 & 0 \\ -\sin(-\omega) & 0 & \cos(-\omega) \\ 1 & 1 & 1 \end{pmatrix} scale \tag{17}$$

$$u_n = \tilde{u}_n + (O_s, 0) \tag{18}$$

$$U = (u_0, u_1, \cdots, u_n) \tag{19}$$

ここで，$(O_s, 0)$ は動作の主体となる物体の現在位置 O_s と時間 $t(=0)$ の組である．変換された軌跡をサー

バからロボットにソケットとして送ることにより，ロボットは動きの生成を行う．

3. 6. 1 トップダウン処理を用いた文章構造の決定

トップダウン処理はフェーズ 3（動的な特徴のシンボルグラウンディング）で獲得した文章の構造を用い，フェーズ 4（動作の説明）及びフェーズ 5（動作の生成）において以下のアルゴリズムで文章構造を決定する．

トップダウン処理を用いた文章構造の決定

input：

（1） *target*：発話の目的

（2） *mm*：メンタルモデル

（3） *set of verbs*：記憶している「動きの種類」,「動作の主体」,「動作の参照点」からなる文章構造

（4） *sentence*：人間が発話した文章（動きの生成時のみ）

（5） *trajectry*：mm から取得する物体の軌跡

output：発話する文章 U_k

loop do

$Q_k \leftarrow$ *target*, *set of verbs*, *trajectry*（動きの生成時のみ）を入力としてボトムアップ処理から得られる語順

if *(Check(Q_k , mm , set of verbs) is ok)*

then $U_k \leftarrow$ *GenerateUtterance*(Q_k, *mm* *set of verbs* , *sentence*)

return U_k

else $Q_k \leftarrow$ ボトムアップ処理から得られる次にゆう度の高い語順

ここで，*Check(...)* は Q_k と *mm*，*set of verbs* を用いて不正なセンテンスを棄却する処理である．この処理では，学習時の語順と文章構造との対応関係を逸脱するセンテンスを不正と判定する．例えば,「名前 → 名前 → 動詞」という語順で「動作の主体+動作の参照点+動きの種類」という文章構造を学習していた場合，ボトムアップ処理から「動詞 → 名前 → 名前」というセンテンスが出力されても，それを不正として棄却する．また，メンタルモデルにより視界内に物体が置かれていないと認識しており,「名前」にあたる物体が存在しない場合もセンテンスを不正と判定する．

また，*GenerateUtterance(...)* は以下の二つの状況に対応する処理である．

（1） フェーズ 4（動作の説明）における動作

set of verbs から学習済みの文章構造（「動作の主体＋動作の参照点＋動きの種類」など）を取得し,「動作の主体」と「動作の参照点」に対応する物体及び「動きの種類」に対応する軌跡を *mm* から取得し，物体及び軌跡にグラウンディングされた単語の音声ラベルを割り当てて U_k を決定する．

例えば，ボトムアップ処理が「名前 → 名前 → 動詞」というセンテンスを出力し，*set of verbs* に「動作の主体＋動作の参照点＋動きの種類」という文章構造が記憶されていた場合，まず *mm* より動いている物体の物体ベクトルを取得する．そして学習済みの SOINN に同様の物体ベクトルをもつクラスタに対応づけられた単語を取得して「動作の主体」の位置に代入する．

次に，視界内の物体と動いた物体の間で式 (5)〜(7) による座標変換を行い，変換後の軌跡と supervised SOINN が獲得した軌跡とを DP マッチングで比較する．そして誤差が最小の軌跡において選ばれた物体の名前を「動作の参照点」の位置に，軌跡に対応する単語を「動きの種類」の位置に代入し，U_k を決定する．

（2） フェーズ 5（動作の生成）における動作

各単語に対応づけられた物体若しくは軌跡を SOINN 及び supervised SOINN より取得し，*mm* を通じてそのような物体が視界内に存在するかを検出する．そして *set of verbs* から学習済みの文章構造を取得し，文章構造と発話された文章とを対応づけて発話された文章の意味を認識する．

例えば,「動作の参照点＋動作の主体＋動きの種類」という文章構造を学習し，人間が「りんご みかん 近づける」と発話した場合,「みかん」という名前の物体を操作し,「りんご」に対して動詞「近づける」が表す軌跡を再現する命令と解釈する．

ここまで述べたように，トップダウン処理はメンタルモデルを通じて得られる目の前の状況と発話された文章とを対応づけ，システムが「動作の主体」及び「動作の参照点」を推定しながら学習することで少数の学習例から文章構造を獲得する．また，トップダウン処理で文章の構造を学習し，ボトムアップ処理が出力する語順と組み合わせることで，動作の生成命令として発話された文章の主語（動作の主体）・目的語（動作の参照点）にあたる物体を正しく選択する．

総じて，人間が発話したとおりの文章構造を学習でき，英語のように動詞が目的語の前にくる文章にも，日本語のように動詞が文章末尾にくる文章にも対応できる．

4. 実　　験

本システムの有効性を確かめるため，実験を行った．図 4 に示す 9 種類（色 3 種類，形 3 種類）の物体を実験に用いる．

提案手法は物体の静的な特徴のシンボルグラウンディングが完了したという状況を前提とし，更に単語の文章中における意味を学習させるというものである．そのため，まず物体を実験環境に一つずつ配置し，実験者が物体を指差しながら色，形，名前に関する単語を 1 回ずつ発話することで物体の静的な特徴のシンボルグラウンディングを行わせ，単語単体の意味を学習させる．提案手法と同様の手法で単語単体の意味を学習した小島らの手法では，70 種類程度の物体に関して単語単体の意味を学習する能力を有することが実験で示されている [12].

9 種類の物体に対して音声の教示を終えた後，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物体を指差させて，学習させたとおりに指差しをするか判定した．

判定の結果，ロボットはすべての物体に関して 1 回の学習で単語のグラウンディングを成功させた．これは，小島らの実験結果である，12 種類の物体に対して音声の教示を終えた時点ですべての物体に関して単語のグラウンディングが成功したという結果と比較しても妥当な結果だと考えられる．

Step 1.　ロボットがすべての物体に単語をグラウンディングさせた後，動きの学習及び単語の文章における意味の学習を行わせた．このとき，学習させた動きは図 5〜図 7 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．また,「近づける」,「遠ざける」,「またぐ」の動きはそれぞれ「名前 → 名前 → 動詞」,「色 → 名前 → 形 → 名前 → 動詞」,「色 → 形 → 名前 → 色 → 形 → 名前 → 動詞」という語順で「動作の主体＋動作の参照点＋動きの種類」という構造の文章を発話し，「まわす」,「上げる」,「下げる」の動きはそれぞれ「名前 → 動詞」,「色 → 名前 → 動詞」,「色 → 形 → 名前 → 動詞」という語順で「動作の主体＋動きの種類」という構造の文章を発話して学習させた．ただし，どの動きも 1 回ずつ学習させた.「動作の主体＋動作の参照点＋動きの種類」という構造の文章で学習させる動きの場合，9 種類の物体から 3 種類（動作の主体，動作の参照点，動作にかかわらない物体）をランダムに選んで図 5 及び図 6 左で示される位置に配置し，動作の主体を人間が操作しながら文章の発話を行うことで動きの学習及び単語の文章における意味の学習を行わ

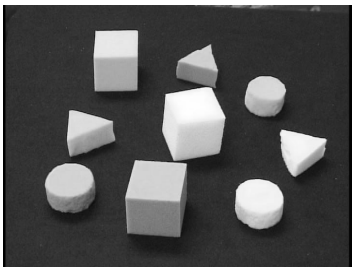

図 4　実験に使用したオブジェクト
Fig. 4　Nine objects we used in this experiment.

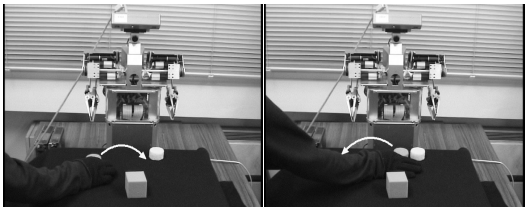

図 5　左「近づける」，右「遠ざける」
Fig. 5　Left “chikazukeru” (move near to), right “toozakeru” (move far from).

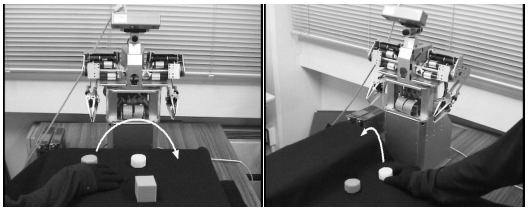

図 6　左「またぐ」，右「上げる」
Fig. 6　Left “matagu” (jump over), right “ageru” (lift up).

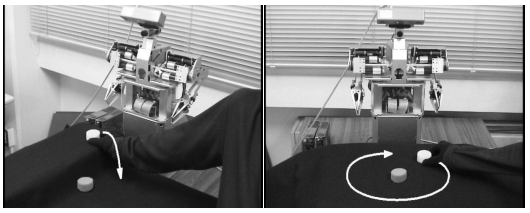

図 7　左「下げる」，右「まわす」
Fig. 7　Left “sageru” (lift down), right “mawasu” (turn around).

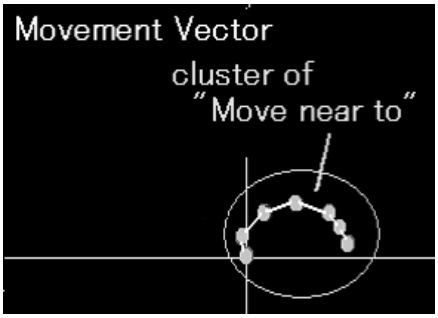

図 8 「近づける」を学習したときの supervised SOINN の様子
Fig. 8 The output of the supervised SOINN after "chikazukeru" was taught by the show-and-tell procedure.

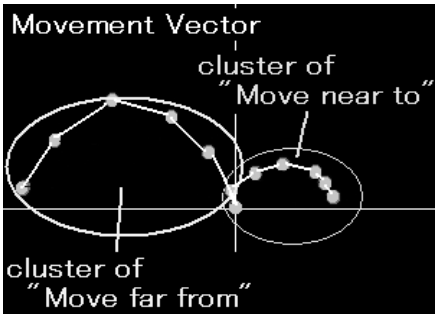

図 9 「遠ざける」を追加学習したときの supervised SOINN の様子
Fig. 9 The output of the supervised SOINN after both "chikazukeru" and "toozakeru" was taught by the show-and-tell procedure.

せた.

「動作の主体＋動きの種類」という構造の文章で学習させる動きの場合，動作の主体と動作の参照点が同じとみなし，9 種類の物体から 2 種類（動作の主体，動作にかかわらない物体）をランダムに選んで図 6 右及び図 7 の位置に配置し，動作の主体を人間が操作しながら文章の発話を行うことで動きの学習及び単語の文章における意味の学習を行わせた.

追加的に学習できることを示すため，2 種類の動きを教えるごとに Step 2 の手順によって学習した動き及び文章における単語の意味を正しく覚えているかの判定を行った.

例として,「近づける」を学習させたときの supervised SOINN の様子と，これに加えて「遠ざける」を学習させたときの supervised SOINN の様子を図 8 と図 9 に示す．図中の丸が supervised SOINN のノードに対応し，丸を結ぶ白線がエッジに対応する．ノードとエッジからなるクラスタが学習した軌跡に対応する．図 9 では新たな軌跡に対応するクラスタが形成されており,「遠ざける」の追加的な学習が確認できる．また，既存の軌跡はそのまま保持されており，新たな学習に影響を受けていないことも確認できる.

Step 2. 学習した動き及び単語の文章における意味を正しく学習しているかを確認するため，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み，かつ学習時に使用した語順を用いた文章（例えば,「みかん，メロン，近づける」,「赤，りんご，丸，みかん，またぐ」）を発話し，ロボットに動きの再現を行わせる．ただし，一つの動詞に関して学習時に用いた 3 種類の語順を用いてテストを行う．ロボットが単語単体の意味及び文章における単語の意味を正しく解釈して動きを再現すれば正解とみなす．上記の一つ目に例として挙げた文章を発話した場合,「動作の主体」にあたる "みかん" を操作し，"メロン" に対して "近づける" によって表される動きを生成すれば正解とする.

（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．次の点を満たす文章を発話すれば正しいと判定する.

（a） 動く主体となる物体の名前が文章構造の「動作の主体」に相当する位置にある.

（b） 動きの主体ではない物体の名前が文章構造の「動作の参照点」に相当する位置にある.

（c） 見せた動きを表現する動詞が文章構造の「動きの種類」に相当する位置にある.

ただし，学習時に用いた物体に依存せず生成，認識が行えることを示すため，動きの学習時に使用した物体と異なる物体を用いて動きの説明と生成を行わせる．また，物体の位置関係，見せる動きの角度は試行のたびに変化させ，物体の位置関係によらない認識及び生成ができることを示す.

例として,「ルビー，メロン，近づける」,「メロン，ルビー，遠ざける」と発話したときにロボットが生成した動作を図 10，図 11 に示す．ここで，三角の物体がルビーであり，ロボットから見て近い方の丸い物体がメロンである．ロボットから見て遠い方の丸い物体は発話中に登場せず，動作に関係ない物体である．ロボットは動作の主体となる物体と動作の参照点となる物体を学習した文章構造どおりに判定し，動作を生成していることが確認できる.

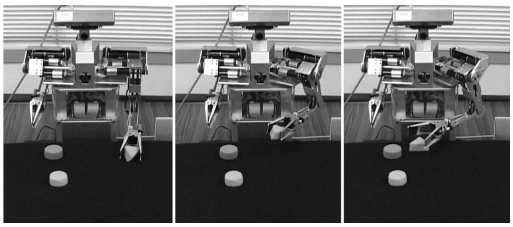

図 10 「ルビー，メロン，近づける」の生成
Fig. 10 The motion performed by the robot when it heard “Ruby meron chikazukeru.”

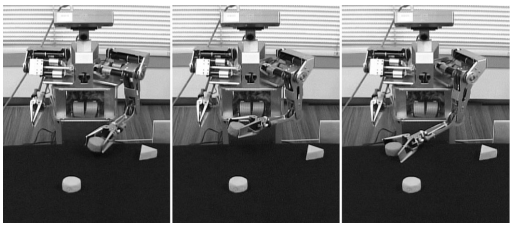

図 11 「メロン，ルビー，遠ざける」の生成
Fig. 11 The motion performed by the robot when it heard “Meron Ruby toozakeru.”

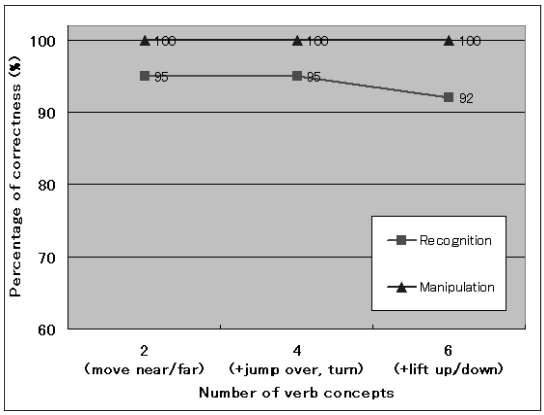

図 12 動作の説明と動作の生成の正解率
Fig. 12 Recognition rates towards motions and manipulations over object.

六つの動きに関して説明と生成をそれぞれ 10 回ずつ（生成に関しては一つの動詞に関して 3 種の語順をそれぞれ 10 回，計 30 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，平均をとって正解率とした（図 12）．

6 種類の動きを学習した場合，動作の認識は 92%，動作の生成は 100%の正解率となり，2 種類，4 種類の動きを学習した場合の正解率と比べてほとんど低下していないことが分かる．この結果により，既存の知識に学習が阻害されずに安定した追加学習が行えるということが示された．また，逐次的に与えられた少数の教示で文章構造を獲得し，正しく生成，認識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．

5. 考　　察

実験の結果より，提案手法は事前知識のない状態からの追加的な学習が可能であることを示した．以下では，大きく二つの節に分けて考察を行う．まず，提案システムの課題について議論する．次に，日常環境において言語獲得を行うために必要な機能について議論する．

5.1 提案システムの課題

本研究では，特定の人間とのコミュニケーションを通じてその人間の挙動，発する単語の意味を獲得するシステムを提案した．学習した単語の文章における意味は新規の単語に対しても適用することができるため，再学習の必要がない．例えば,「りんご，みかん，近づける」という文章が「りんご」を「みかん」に「近づける」という意味をもつと学習した場合，新規単語「メロン」「トマト」を学習すれば「メロン，トマト，近づける」という文章を扱うことが可能である．

しかし，日常環境で動作する場合を想定すると，多数の人間と接する可能性があり，発話者による音声の違いや挙動の違いが考えられる．提案システムの機能は，音声処理モジュールと視覚情報処理モジュールの機能に主に依存しているため，この二つの機能に問題が生じた場合（別の単語として誤認された場合，あるいは別の軌跡として誤認された場合）システムとして正しく動作しない．

本研究では比較的簡易な動作（物体の動いた軌跡）及び音声（連続音声ではなく分節化された単語を使用）を用いて実験を行った．このことにより起き得る問題及び必要な機能を，音声処理モジュールと視覚処理モジュールに分けて議論する．

5.1.1 音声処理モジュールにおける課題

本研究の目的はロボットと人間のコミュニケーションを通じて，視覚情報及び音声情報に関する事前知識をもたない状態から（訓練データを用いずに）追加的に言語獲得を行うことである．そのため，音声に関して音素モデルの構築や，訓練データを用いた学習器の

構築を必要としない手法として VQ コードによるテンプレートと DP マッチングを用いて音声認識を行った．

音声情報処理に用いられる代表的な手法として HMM があるが，HMM では一般に多数の訓練データを用いたバッチ学習や，入力音声に応じたトポロジー（状態数などのパラメータ）の事前決定が必要なため，本研究の目的には適さないと判断した．

しかし DP マッチングは，HMM と異なり特徴空間上の分布をモデル化できないため，発話者が増えるに従い音声の認識精度が低下する可能性がある．また，本研究では教示者が単語ごとに区切りを入れて文章を発話することとしているが，一般に文章は区切りのない連続音声として与えられる．

これらを踏まえ，今後事前知識のない状態からの連続音声の頑健な学習・認識が可能な手法の導入を検討する必要がある．

5.1.2 視覚処理モジュールにおける課題

本研究では，事前に決められた座標変換により変換された軌跡を supervised SOINN により学習し，DP マッチングを用いて動作認識を行った．

また，提案手法は二つの物体及び一つの物体がかかわる動きのみを対象とし，座標変換後の軌跡を用いて学習・認識を行う．そのため，軌跡の回転（右回り，左回りで近づける）や三つ以上の物体の相互関係を必要とする動き（並べる，など）に対応することができない．

更に，多数の行為者がいる場合，DP マッチングだけでは対応できない状況が起こり得る．例えば，同じ「近づける」という動作であっても，ある行為者は「持ち上げて近づける」軌跡を行い，別の行為者は「押して（持ち上げず）近づける」軌跡を行った場合，軌跡として比較すると異なると判定され得る．

よってこのような動きにも対処できる特徴量（例えば物体の位置関係）の導入が必要である．

また，本研究では supervised SOINN に学習させるテンプレートをロバストにするため，パラメータに依存する乱数を学習データに加えて入力する．このパラメータは実験的に定めた値であり，論文中のすべての実験は同一のパラメータで行っている．一方，軌跡のような時系列データの認識にも用いられる HMM は，多数の事前知識（訓練データ）を用いてバッチ学習を行えば，テンプレートの分散を学習により自動決定するため，行為者が増えても認識精度の低下を招きにくい．しかし，実環境を想定した場合，事前知識を準備することができず，HMM のような手法では高い認識精度が得られない [8]．そのため，事前知識のない状態から追加的に学習を行い，頑健なテンプレートを獲得できる時系列データ認識手法の導入を検討する必要がある．

5.2 言語獲得における課題

本研究ではメタ知識として物体の視覚的情報，聴覚情報の分類（クラス化）能力を与えている．情報の分類能力は，乳幼児の発達の極めて早い段階で獲得され，言語の獲得にとって前提になる能力と考えられるため [16], [17]，**5.1.1**，**5.1.2** で議論した課題を解決できるようシステムを拡張する必要がある．

また，乳幼児は語意の獲得を行う段階で言語獲得以前，若しくは生得的に備わっている認知機能をバイアスとして探索の範囲を狭め（認知的制約），言葉を獲得していくことが乳幼児の発達研究で指摘されている．例えば, [18] では「未知の事物に対してつけられたラベルを，その事物を含むカテゴリーのラベルと解釈する」という事物カテゴリーバイアスが提案されている．本研究では単純なルール（一つの単語は一つの属性のみに対応する）に基づくバイアスを用いたが，より複雑な語意の獲得を行うために適切なバイアスの設計が重要な課題になると思われる．

また，提案システムは事物の具体的概念（色や形）と単語の対応関係を獲得するという点に主眼を置いたが，子どもは目に見える物体の属性による単語の獲得から，物体が属するカテゴリーを用いた単語の獲得（例えば，異なる犬種の犬に同一の「犬」というカテゴリーラベルを付与できる）へと移行していくことが報告されている [19]．そのためカテゴリーの獲得はロボットの言語獲得にとっても重要な課題になると考えられ，カテゴリーの獲得（「赤」や「黄色」は「色」というカテゴリーに属する，等）を可能とするよう提案手法を拡張していく必要がある．

6. むすび

本研究では，乳幼児が発達していくように，視覚，聴覚から得られた情報をもとに物体についての言葉を学習し，物体の動き，発話の語順などを事前知識のない状態（バッチ的に学習する訓練データを用いない）から発達的に獲得していく仕組みを提案した．

本システムは SOINN の特性を生かし，事前知識のない状態から物体の色や形，動きに関する概念を逐次的に与えられるデータを用いて追加的に学習する．ま

た，メンタルモデルを用いて眼前の状況を常に認識することで，トップダウン処理を通じて文章構造を解析し，ボトムアップ処理と組み合わせることで一つの例からでも物体の動きを表現する文章構造を学習できる．そして，自らも文章を構成して発話したり，言語的な指示を解釈して行動することができる．

提案手法は SOINN とメンタルモデルの特性を生かすことによって，訓練データを用いず，人間とコミュニケーションを行いながら逐次的に与えられる文例からリアルタイムの学習が可能なシステムを実現し，優位性が示された．

今後，**5.** で述べた課題を解決し，日常生活環境において発達的な言語獲得を行うシステムの実現を目指す．

謝辞　本研究の実施にあたり NEDO 産業技術研究助成事業から支援を頂きました．記して感謝致します．

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，vol.4, pp.2–27, 2004.

[2] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, “Autonomous mental development by robots and animals,” Science, vol.291 no.5504, pp.599–600, 2000.

[3] S. Harnad, “The symbol grounding problem,” Physica, vol.D42, pp.335–346, 1990.

[4] J.M. Siskind, “A computational study of cross-situational techniques for learning word-to-meaning mappings,” Cognition, vol.61, pp.1–2, 1996.

[5] D. Roy and A. Pentland, “Learning words from sights and sounds: A computational model,” Cognitive Science, vol.26, no.1, pp.113–146, 2002.

[6] N. Iwahashi, “Language acquisition through a human-robot interface by combining speech, visual, and behavioral information,” Inf. Sci., vol.156, pp.109–121, 2003.

[7] N. Iwahashi, “Active and unsupervised learning of spoken words through a multimodal interface,” 13th IEEE Workshop Robot and Human Interactive Communication, pp.437–442, 2004.

[8] 篠田浩一，“確率モデルによる音声認識のための話者適応化技術，” 信学論（D-II），vol.J87-D-II, no.2, pp.371–386, Feb. 2004.

[9] Y. Chen and B. Dana, “On the integration of grounding language and learning objects,” Nineteenth National Conference on Artificial Intelligence (AAAI ’04), pp.25–29, 2004.

[10] G.A. Carpenter and S. Grossberg, “The ART of adaptive pattern recognition by a self organizing neural network,” Computer, vol.26, pp.77–88, 1988.

[11] S. Furao and O. Hasegawa, “An incremental network for on-line unsupervised classification and topology learning,” Neural Netw., vol.19, no.1, pp.90–106, 2006.

[12] 小島　量，長谷川修，“ヒューマノイドロボット上の自己増殖型ニューラルネットワークを用いた視聴覚情報からの能動的・追加的概念獲得，” 信学技報，PRMU2005-57, 2005.

[13] S. Imai, T. Kobayashi, K. Tokuda, and T. Masuko, “Speech signal processing toolkit: Sptk version 3.0,” 2003, http://kt-lab.ics.nitech.ac.jp/˜tokuda/SPTK/release/SPTKref-3.0.pdf

[14] A. Gersho and R.M. Gray, Vector Quantization and Signal Compression, Kluwer Boston, 1992.

[15] D. Roy, K. Hsiao, and N. Mavridis, “Mental imagery for a conversational robot,” IEEE Trans. Syst. Man Cybern. B, Cybern., vol.34, no.3, pp.1374–1383, 2004.

[16] P.D. Eimas and P.C. Quinn, “Studies on the formation of perceptually based basic-level categories in young infants,” Child Development, vol.65, pp.903–917, 1994.

[17] C.L. Stager and J.F. Werker, “Infants listen for more phonetic detail in speech perception than in word-learning tasks,” Nature, vol.388, pp.381–382, 1997.

[18] E.M. Markman and J.E. Hutchinson, “Children’s sensitivity to constraints on word meaning: Taxonomic versus thematic relations,” Cognitive Psychology, vol.16, pp.1–27, 1984.

[19] N.N. Soja, S. Carey, and E.S. Spelke, “Discussion:Perception, ontology, and word meaning,” Cognition, vol.45, pp.101–107, 1992.

（平成 19 年 9 月 5 日受付，20 年 1 月 31 日再受付）

佐藤　彰洋　（学生員）

2006 電通大・電気通信・情報工学卒．現在，東京工業大学大学院知能システム科学専攻修士課程在学中．ニューラルネットワークとヒューマノイドロボティクスに興味をもつ．

賀　小淵

中国浙江大・情報工学卒．2006 東京工業大学大学院知能システム科学専攻修士課程了．同年，（株）NEC に就職，現在に至る．在学中は，ヒューマノイドロボットを用いた自然言語獲得の研究に従事．

小倉　和貴

2005 明大・理工・情報科学卒．2007 東京工業大学大学院知能システム科学専攻修士課程了．同年，(株）ソニーに就職，現在に至る．在学中は，ニューラルネットワークと教師なし学習に関する研究に従事．

長谷川　修　（正員）

1993 東京大学大学院電子工学専攻博士課程了，博士（工学），同年電子技術総合研究所，1999 から 1 年間米国カーネギーメロン大学客員研究員，2001 産業技術総合研究所主任研究員，2002 東京工業大学像情報工学研究施設准教授，2003〜2006 JST さきがけ研究 21 研究者（兼任），情報処理学会，日本認知科学会，人工知能学会，IEEE-CS 等各会員．