様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年　５月　１日現在

機関番号：１３９０３
研究種目：若手研究（Ｂ）
研究期間：2011～2012
課題番号：２３７００２００
研究課題名（和文）　自由な連続音声からの単語音素系列と指示対象カテゴリの学習

研究課題名（英文）　Learning of Phoneme Sequence and Indicated Category from Spoken Utterances

研究代表者
田口　亮（TAGUCHI　RYO）
名古屋工業大学・工学研究科・助教
研究者番号：　70508415

研究成果の概要（和文）：
本研究では，単語知識を持たないロボットが，音声と指示対象のペアから，各単語の音素系列とその意味（対象のカテゴリ）を自動的に学習するための手法を開発した．家庭用ロボットなどのように多様な環境で用いられるロボットの場合，ユーザとのインタラクションを通した単語学習能力が必要不可欠である．従来は単語を教示する際に単語単位で区切って発話するか，事前に決められた言い回しで発話する必要があったが，提案手法は，自由な言い回しから単語学習が可能である．

研究成果の概要（英文）：
This report proposes a method for unsupervised learning of phoneme sequences of words and the categories indicated by the words from pairs of spoken utterances and feature vectors, which are gotten through human-robot interaction in the real-world, without any priori linguistic knowledge other than a phoneme acoustic model. Domestic robots must be able to learn phoneme sequences of unknown words and them meanings through human-robot interaction. In previous works, when users teach novel words to robots, they have to use isolated words or fixed phrases. However, in our method, robots can learn novel words from user’s free utterances.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,500,000	1,050,000	4,550,000

研究分野：総合領域
科研費の分科・細目：情報学，知覚情報処理・知能ロボティクス
キーワード：音声情報処理，言語獲得，シンボルグラウンディング

１．研究開始当初の背景

ロボットが人と対話するためには，言葉と実世界の事物・事象の対応関係をロボットが理解できなければならない．家庭やオフィスなどでは，未知の人や物，場所等に対応する必要があるため，それらを表す単語知識，すなわち語彙をユーザとのインタラクションを通して学習できることが望まれる．

ロボットによる語彙学習に関する先行研究では，人がロボットに物や動作を見せながら対応する単語を発話することで，「箱」や「青い」といった物を表す単語や，「乗せて」や「近づけて」といった動作を表す単語を学習させた．申請者はこれまで，幼児のバイアスを適用した効率的な単語学習アルゴリズムや，単語学習を効率化するための対話戦略の獲得手法，「なに？」や「どれ？」といったコミュニケーションに意味付られる単語の学習手法を提案してきた．これらの研究や，他の多くの先行研究では，単語単位に区切られた発話や，決められた文法に沿った発話が学習に用いられてきた．しかし，実運用を考慮すると，ユーザの自然な発話から学習できることが望ましい．単語区切りのない連続音

声から単語を切り出す実験も行われているが，高精度な音素系列の獲得までは至っていない．また，未登録語（辞書に登録されていない単語）のクラス（人名や地名など）が持つ音響的，文法的なモデルを学習・利用することで，発話に含まれる未登録語を抽出する手法が提案されているが，未知の言い回しが入力された場合には，未登録語の境界を判定することはできない．

こうした背景から申請者はユーザの自由な発話から語彙を学習する手法を開発している．先行研究では，発話と指示対象（物や場所など）の対応関係を，音響，文法，意味を統合した確率モデルで表現し，それを統計的モデル選択に基づいて最適化することで，単語の音素系列と意味（単語と指示対象の直接的な対応関係）を学習する手法を提案した．この研究には次に示す3つの課題が残っている．（1）指示対象が事前にカテゴリ化されていることを仮定している．実際にロボットが取得するセンサ情報は，画像特徴量や自己位置の座標など連続量であり，それらのカテゴリ化も語彙学習と同時に行わなければならない．（2）色や形といった物体の属性を表す単語は学習できず，物体そのものを表す単語のみ学習可能である．（3）実験はシミュレーションで行っており，実環境での評価は行っていない．

2．研究の目的

本研究では上記の課題を解決するために，これまで提案してきた統計的モデル選択に基づいた語彙学習手法に，（1）連続量として与えられるセンサ情報を適切にカテゴリ化する機能と，（2）単語の対象となる属性を判定する機能（例えば，色か形かを判定する）を開発・付加する．そして（3）実ロボットによる実験を通し提案手法の有効性を評価する．また，色や大きさ，位置などの概念は，他の物体との比較や，典型的な概念との比較によって表されるが，従来研究では単語と共起するセンサ情報を直接的に学習していたため，それら相対的な概念は学習することができなかった．そこで，（4）相対的な概念を学習できる機構を検討する．

本研究で扱う語彙学習のタスクを説明する．ユーザがある対象をロボットに提示し，音声でその名前を教示する．「これはボールペンです」等のように，教示には対象の名前以外の語を含む．本研究では対象の名前を【 キーワード 】，キーワード以外の表現を【 言い回し 】と呼ぶ．言い回しとキーワードは独立であると仮定し，同じ言い回しで複数のキーワードが発話され，一つのキーワードが複数の言い回しで発話されるものとする．ロボットの初期知識は，各音素の音響モデルと，音素間の遷移モデル（有限状態オートマトン）の二つだけであり，単語の知識は持っていない．教示された複数の音声-対象ペアから，単語の音素系列とその意味を学習する．未知の対象が入力された時に，正しいキーワードを出力することを目標とする．

3．研究の方法

（1）センサ情報のカテゴリ化

センサ情報のカテゴリ化に関しては，従来法で用いた3つの確率モデルのうち，意味のモデルを連続確率分布に拡張することによりカテゴリ化の機能を実現する．そしてシミュレーション実験を通して提案手法の有効性を評価する．

（2）属性判定

これまでは発話が指示する対象は1つに限定されていた．本研究では，意味のモデルを拡張し，「これは赤いボールです」のように，一つの発話で教示対象が持つ複数の特徴（物の色や形など）を教示できるようにする．そして，各単語が指示する特徴を選択する手法を提案する．

（3）実ロボットによる実験

車輪移動型の警備用ロボットASKA（図1）を用いて場所名の学習を行う．ASKAは夜間の巡回警備用として開発され，頭部に全方位カメラ，胸部に前方カメラを搭載している．また，超音波センサとレーザ距離センサにより障害物との距離を計測することで，衝突を回避する．また，ASKAはレーザ距離センサを用いて地図作成と自己位置推定を行う．語彙学習実験ではレーザ距離センサを用いて推定された自己位置を用いて，場所の名前と場所の範囲（カテゴリ）を学習する．

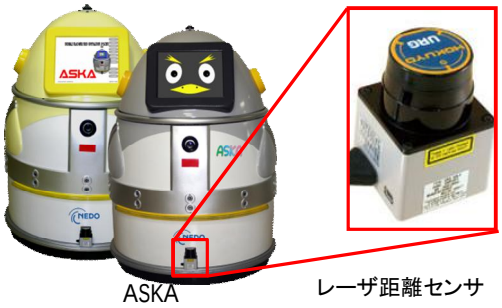

図 1 自律移動型ロボット ASKA

（4）相対的な概念の学習

相対的な概念を学習するためには，学習時に参照点（比較対象となる物体や概念）が必要となる．しかし，人間同士の対話においては，「右の箱を取って」などのように，どこから見て右なのか，すなわち参照点（比較対象となる物体や概念）が明示されない場合がある．そこで，EMアルゴリズムを用い，場面毎に，参照点の推定と語意の学習とを同時に行う手法を開発する．

４．研究成果

（１）センサ情報のカテゴリ化

①提案モデル

与えられた発話と指示対象の対応関係を，隠れ変数である単語列（単語ラベルの列）を介した共起確率モデルとして表現する．単語ラベルとその音素系列のペアは単語リストに記述する．単語リストを MDL 原理に基づいて最適化することにより，発話と対象の対応関係を少ない単語数でうまくモデル化できるような単語集合を得ることができる．

発話 **a**（1 発話分の音声の特徴ベクトル）と対象を表す m 次元の連続ベクトル $\mathbf{o}=(o_1,o_2,...,o_m)^{\mathrm{T}}$ の共起確率モデルを次式に示す．

$$\begin{aligned}&\log P(\mathbf{a},\mathbf{o})\\&=\log\sum_s\{P(\mathbf{a}|s)P(s)P(\mathbf{o}|s)\}\\&\approx\max_s\{\alpha\log P(\mathbf{a}|s)+\log P(s)+\log P(\mathbf{o}|s)\}\end{aligned}\quad\cdots(1)$$

s は単語列である．$P(\mathbf{a}|s)$は音響モデルであり音素 HMM の連結として表現される．$P(s)$は文法モデルであり，単語 bigram として表現される．$P(\mathbf{o}|s)$は意味モデルであり次式で表す．

$$P(\mathbf{o}|s)=\sum_{i=1}^{n}\gamma(s,i)P(\mathbf{o}|w_i)\quad\cdots(2)$$

w_i は s に含まれる i 番目の単語，$P(\mathbf{o}|w_i)$は単語 w_i の意味，$\gamma(s,i)$は各単語の重み(単語の音素数より決定する)である．対象を連続ベクトルとして与えるため，$P(\mathbf{o}|w)$を次式のように多次元正規分布で表す．

$$P(\mathbf{o}|w)=\frac{1}{(\sqrt{2\pi})^m\sqrt{|\mathbf{S}|}}\exp\left(-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu})^T\mathbf{S}^{-1}(\mathbf{o}-\boldsymbol{\mu})\right)\quad\cdots(3)$$

$\boldsymbol{\mu}$ は平均ベクトル，$\mathbf{S}$ は共分散行列である．

②実験条件

シミュレーション実験には，国立情報学研究所で開発された社会的知能発生学シミュレータの SIGVerse を用いた．実験の様子を図 2 に示す．シミュレータ上の仮想空間内でロボットを移動させ，任意の点で場所名を発話する．発話した音声はロボットの位置座標(x, y)と共に保存される．実験では「本棚」「観葉植物」「テレビの前」「ソファーの所」の 4 つの場所名を各 4 回，位置を変えながら教示した．対象のカテゴリ化に焦点を絞るため，発話は言い回しを含めないものとした．

③実験結果と考察

実験の結果を図 3 に示す．図中のひらがなは獲得されたキーワード，括弧書きは教示したキーワード，小さな丸は教示の際の位置座標を表し，楕円は各地点を入力とした時に式(4)によって出力されるキーワード（閾値以上の確率を持つもの）を表している．

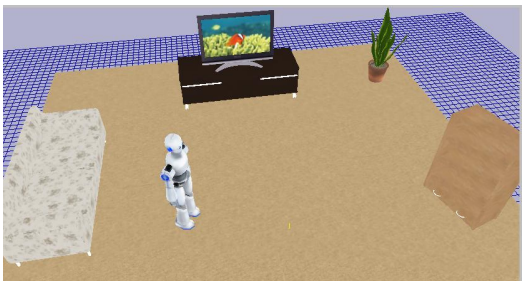

図 2：シミュレーションの実行画面

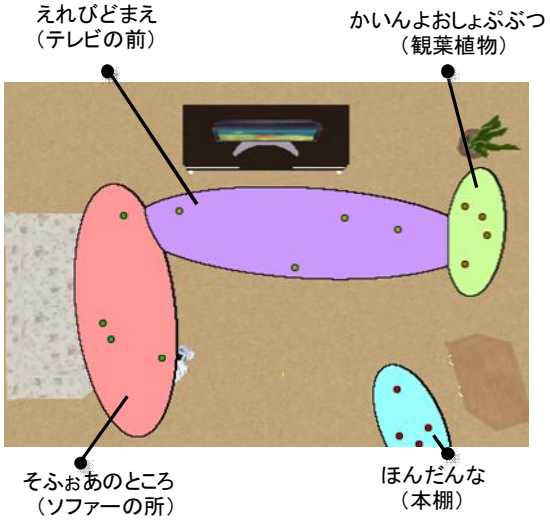

図 3：シミュレーションでの学習結果

$$w_{\mathbf{o}}=\arg\max_{w\in\Omega}\left\{\log P(w)+\log P(\mathbf{o}|w)\right\}\quad\cdots(4)$$

Ω は獲得したキーワード集合である．ただし，キーワードの判定には情報量に基いて判定する．次に示すエントロピーの減少量 $I(w)$を用いる．

$$\begin{aligned}I(w)=&-\int P(\mathbf{o})\log P(\mathbf{o})\,d\mathbf{o}\\&+\int P(\mathbf{o}|w)\log P(\mathbf{o}|w)\,d\mathbf{o}\end{aligned}\quad\cdots(5)$$

$I(w)$が閾値以上の単語をキーワードと判定する．

ロボットは教示に対応する 4 単語を獲得し，各単語が表す場所の範囲も同時に学習できたことがわかる．

（２）属性判定

①提案モデル

先の実験では発話が指示する対象は 1 つに限定されていた．ここでは，「これは赤いボールです」のように，色や形といった対象の特徴を表すキーワードが 1 つ以上含まれるものとする．各キーワードが対象のどの特

徴を表しているのかという情報は明示的に与えられないため，ロボットは学習を通し，各単語が表している特徴を選択しなければならない．

提案モデルでは，(2)式を拡張し，複数の単語が異なる特徴を示すことを可能にする．具体的には下記の２手法を提案し，実験によりその性能を比較した．

方法 1: 各単語の意味を重み和で統合

$$P(\mathbf{o} \mid s) = \sum_{i=1}^{n} \gamma(s,i) \left\{ \prod_{j=1}^{m} P(o_j \mid w_i)^{a_j} \right\} \quad \ldots(6)$$

方法 2: 各特徴に対応する単語のみで意味を統合

$$P(\mathbf{o} \mid s) = \prod_{j=1}^{m} P(o_j \mid \tilde{w}_j) \quad \ldots(7)$$

$$\tilde{w}_j = \arg\max_{w} I(o_j, w) \quad \ldots(8)$$

w_i は s に含まれる i 番目の単語，$P(o|w_i)$は単語 w_i の意味である．方式 1 の $\gamma(s,i)$は各単語の重み(単語の音素数より決定する)である．ただし，特徴 o_j に関するエントロピーの減少量 $I(o_j,w)$が閾値以上の単語を j 番目の特徴を表すキーワードと判定し $a_i = 1$，閾値未満の場合 $a_i = 0$ とする．また，全ての特徴において閾値未満となる場合, $\gamma(s,i)=0$ とし計算から除外する．

方式 1 では固定的に特徴選択を行うが，方式 2 では発話に含まれる単語のうち各特徴を表すのに最も適切な単語を動的に選択する．また，方式 1 では発話に含まれる複数のキーワードが同じ特徴を説明することを許すが，方式 2 では各特徴につき 1 単語に限定される．

②実験条件

提案手法の有効性を確かめるため実験を行う．実験では画像の 1 点を指示し，「これは明るい緑色」や「暗い赤です」，「これは青です」などのように，音声でその色を教示する．対象を表す特徴は色相，明度，彩度の 3 つとした．使用した単語は，「黄色」，「緑色」，「青」，「赤」，「明るい」，「暗い」，「これは」，「です」の 8 単語で，発話には 1 語以上の色名と言い回しを含むものとし，男性話者 1 名 36 発話を収録した．

③実験結果と考察

学習の結果を表 1 に示す．表中の「選択特徴」は $I(o_j, w)$が閾値以上となった特徴を表す．表から方法 1 では緑色に対応する単語が脱落しており，また「です」が色を表す単語として獲得されているのに対し，方法 2 では全ての単語が獲得された．これは方法 2 の方が，キーワードが脱落した場合の $P(\mathbf{o}|s)$の低下が大きいためであると考えられる．ただし，教示した色に偏りがあったため，選択される特徴が複数になる傾向がみられた．また，色を表す単語は，純粋に色相だけではなく，明度や彩度の影響を受けることが確認された．一方，明るさを表す単語は，明度だけでなく彩度や色相も影響を与えることや，同じ明度であっても色相によって呼称される単語が異なることが確認された．これは明るさを表す単語が，色情報の絶対値に基いて使用されるのではなく，他の色や，各色の平均的な明度との比較によって使用されることを示唆しており，このような単語を正しく学習するためには，相対的な概念の学習方法が必要不可欠である．

表 1：学習結果

正解	方法 1		方法 2	
	獲得単語	選択特徴	獲得単語	選択特徴
黄色	りいろ	色	しりいろ	色, 明
緑色			みどりいろ	色, 明
青	あお	色	あお	色
赤	あか	色, 彩	あか	色, 彩
明るい	あかるい	明, 彩, 色	あかるい	明, 彩, 色
暗い	くらい	彩	ふらい	彩
です	です	色	です	なし
これは	これわ	なし	これわ	なし

（３）実ロボットによる実験

①実験条件

警備用ロボットの ASKA を用いて，（１）で実施した場所名の学習実験を実環境で行う．

ASKA はレーザレンジファインダを用いて地図作成と自己位置推定を行う．地図作成には格子ベース FastSLAM を用いた．実験ではまず ASKA をリモコンで操作しながら建物内の地図を作成する．その後，ASKA を所定の場所に移動させ場所名を教示する．教示する場所は 10 箇所とした．図 4 に作成した地図と教示キーワード（括弧内で表記）を示す．位置を変えながら各場所で 9 箇所，計 90 箇所の位置情報を取得した．本実験では音声の収録と位置情報の取得は別々に行なった．収録した音声は男性話者 1 名であり，各場所の名前を 9 種類の言い回し（表 2）で発話した．

表 2：言い回しの種類（X はキーワードを表す）

ここが X です	ここが X
ここの名前は X だよ	ここの名前は X
この場所は X っていうんだ	この場所は X
X です	X っていうんだ
X だよ	

②実験結果と考察

図 4 に学習結果を示す．ひらがな表記の単語が獲得されたキーワードの音素系列，図中の小さな丸が教示した際の自己位置推定結果，楕円が楕円は各地点を入力とした時に式(4)によって出力されるキーワード（閾値以上の確率を持つもの）を表している．

学習の結果，20 単語が獲得された．そのうち，キーワードと判定された単語は 12 単語であった．本実験条件では，真のキーワードの数が 10，真の単語数（キーワード数＋言い回しに使われる単語数）が 16 であるため，正解に近い数の単語が得られたことがわかる．また，キーワードと判定された 12 単語の平均音素正解精度は 80%であった．学習なしで収録音声を音素認識した際の音素正解精度は 76%であり正解精度の向上が見られた．これは単語リストの再構築時に音響的に有用な単語が取捨選択された結果である．一部の単語が重複して獲得されたが，先の実験と同様に各単語が表す場所の範囲も同時に学習できることが示された．

（4）相対的な概念の学習

①提案モデル

ユーザがある対象をロボットに提示し，対象が持つ特徴のラベルを教示する．その際，ユーザはある参照点を基準として，ラベル（例えば「明るい」なのか「暗い」なのか）を決定する．各場面における指示対象と参照点の候補は既知とするが，候補のいずれかが真の参照点なのかは未知である．そのため，本タスクでは，学習サンプル毎に真の参照点を推定し，学習していくことが望まれる．

提案手法は，学習サンプル n において参照点候補 k を基準に得られる特徴量を x_{nk}，参照点候補 k が真の参照点である確率を π_{nk} と置くそして式(9)に示すモデル尤度を最大化するパラメータ $\theta = (\mu, \sigma^2, \pi_{nk})$ を EM アルゴリズムにより求める．

$$\theta = \arg max \sum_{n=1}^{N} ln \sum_{k=1}^{M_n} \pi_{nk} \mathrm{N}(x_{nk} | \mu, \sigma^2) \quad \ldots(9)$$

EM アルゴリズムは隠れ変数を含む確率モデルのパラメータを，最尤法に基づいて推定する手法であり，混合正規分布の学習にも用いられる．本手法ではサンプル毎に参照点の確率 π_{nk} を定義しているが，混合正規分布では各サンプルに対して独立な混合重み π_k を用いる．また，混合正規分布の場合には複数の正規分布を用いるが，本手法では単一の正規分布となっている．

②実験結果と考察

「上，下，左，右」の 4 単語の学習実験を実施した． 1 単語につき 10 枚の画像データを用意した．画像データの例を図 5 に示す．1 つの画像には 1～4 つの物体を含む．学習に用いる特徴量は指示対象の位置座標（x，y）とした．参照点候補は画面中心および他の物体の重心位置である．本画像は Microsoft Kinect を用いて撮影しており，RGB 画像の他に，深度情報が得られる．この深度情報を用いて物体の切り出し，各物体の重心位置を取得する．本実験では音声は使用せず，ラベルを表す文字列をロボットに直接与える．

「左，右」の学習結果を図 6,7 に示す．図の横軸は X 軸の相対位置（原点は参照点の重心位置），縦軸はラベルが与えられた際の確率である．すなわち，「左」または「右」というラベルが与えられた際に，指示対象が参照点から見てどの位置にある確率が高いのかを表している．図 6 に示すように，提案手法を用いずに全ての特徴量候補を用いると，適切に学習できないことがわかる．一方，図 7 に示すように，提案手法では X 軸において「左，右」の分布が分離されており，相対的な意味の学習ができたと言える．また，「上，下」についても同様に学習できることが確認された．

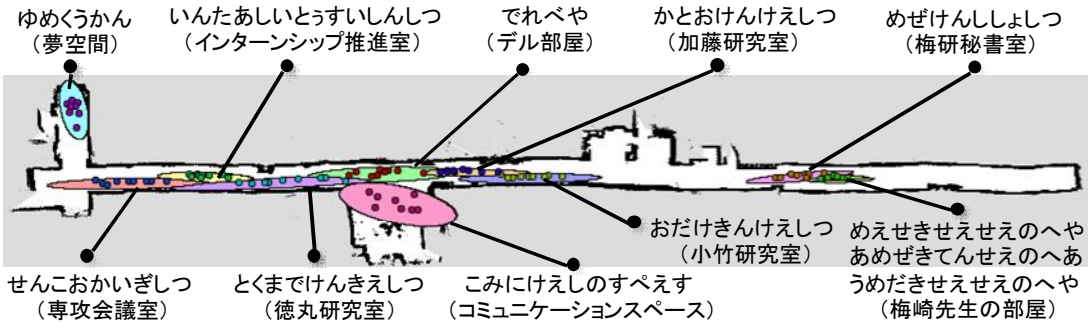

図 4：実ロボットを用いた実験の結果

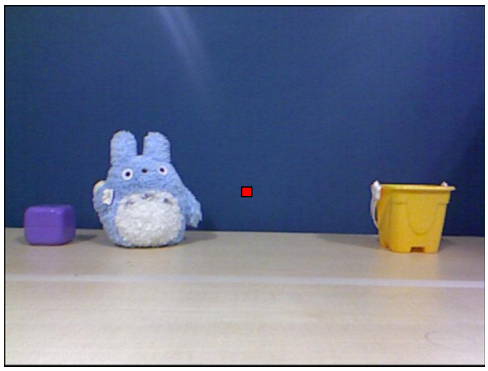

図 5：相対的な概念の学習に用いた画像データ

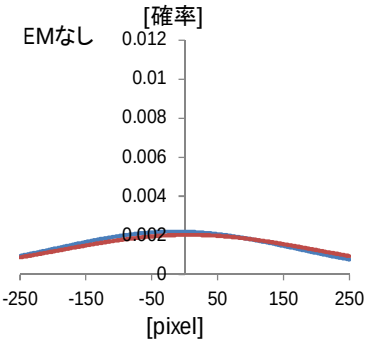

図 6：「左」「右」の学習結果（EM なし）

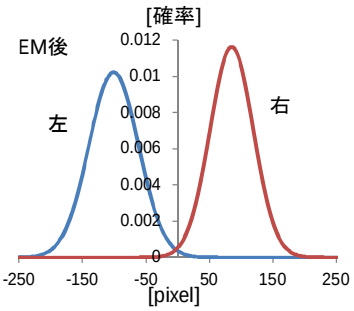

図 7：「左」「右」の学習結果（EM 後）

（5）成果のまとめと今後の展望

本研究では，（1）連続量として与えられるセンサ情報を適切にカテゴリ化する機能を実現した．実ロボットによる実験を通し，その有効性を確認した．また，（2）単語の対象となる属性を判定する機能を開発した．実験の結果，一つの発話で複数の特徴同時に教示した場合でも，単語の切り出しが可能なことを示した．しかし，明るさなど相対的な意味を持つ単語については，語意が正確に学習できないことが確認された．そこで（3）相対的な概念の学習手法を開発した．提案手法は EM アルゴリズムを用いることで，教示で参照点が明示されない場合でも，学習が可能である．実験の結果，「上下左右」の単語の意味が正しく学習できた．

これら（1）～（3）の成果は，語彙学習手法の拡張となっているが，それらを統合した実験は行なっていない．今後は，３つの手法を統合し，より多くのデータを用いて，提案手法の有効性を評価する必要がある．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

[1]田口 亮：連続音声からの語彙学習（特集記号創発ロボティクス），人工知能学会誌 27(6), 594-599, 2012-11

〔学会発表〕（計７件）

[2]田口亮，保黒政大，梅崎太造：連続音声からの語彙学習と特徴選択，第 12 回 計測自動制御学会 システムインテグレーション部門講演会講演論文集，2L2-4，2011-12.
[3]東拓実, 加藤嗣, 服部公央亮, 田口亮, 保黒政大, 梅崎太造，自律走行型ロボット ASKA による自動巡回システムの開発，情報処理学会第 74 回全国大会,1ZF-7，2012.3.6
[4]東拓実, 服部公央亮, 田口亮, 保黒政大, 梅崎太造：音声対話による場所名の学習と巡回経路指定，第 18 回 画像センシングシンポジウム（SSII2012),IS2-14，2012.6.6 - 2012.6.8
[5]田口亮，連続音声からの語彙学習と特徴選択，第 18 回 創発システムシンポジウム，P15，2012.9.1.（優秀ポスター賞）
[6]田口亮，東拓実，梅崎太造，保黒政大：連続音声からの語彙学習と自動巡回ロボットへの応用，日本ロボット学会 第 30 回記念学術講演会，3N2-2，2012.9.19
[7]東拓実, 服部公央亮, 田口亮, 保黒政大, 梅崎太造：音声対話による場所名の学習と巡回経路指定, 平成 24 年度電気関係学会, 東海支部連合大会,D2-2，2012.9.24-2012.9.25
[8]田口亮，WANG DING，YU QIYUE，保黒政大，梅崎太造：EM アルゴリズムを用いた参照点に依存した語意の学習，情報処理学会 第 75 回全国大会，4D-2，2013.3.7

〔その他〕

ホームページ

http://taguchi-lab.com/

６．研究組織

(1)研究代表者

田口　亮（TAGUCHI　RYO）
名古屋工業大学・工学研究科・助教
研究者番号：　70508415