1ZB-3

連続音声から場所の名前を学習する自律移動ロボット

山田雄治[*1]　服部公央亮[*1]　田口亮[*1]　梅崎太造[*1]　保黒政大[*2]　岩橋直人[*3]　船越孝太郎[*4]　中野幹生[*4]
[*1]名古屋工業大学　[*2]中部大学　[*3](独)情報通信研究機構　[*4](株)ホンダ・リサーチ・インスティチュート・ジャパン

1. はじめに

家庭やオフィス等で人と対話するロボットは，その環境に固有の言語知識（人や物，場所の名前など）をユーザとのインタラクションを通して学習できなければならない．我々は，単語の知識を持たないロボットが，ユーザの多様な言い回しの発話から，単語の正しい分節とその音素系列，および，単語と対象の間の直接的な対応関係（本稿ではこれを意味と呼ぶ）を学習するための手法を提案している[1]．シミュレーション実験の結果，83%の音素正解精度で単語が学習できることが示された．この実験では指示対象が ID（離散値）として正しく認識出来ると仮定していた．しかし，実際のロボットが取得できる情報は，画像特徴量や自己位置の座標といった連続ベクトルであり，そのカテゴリ化は語彙学習と同時に行われるべきである．そこで本稿では，連続ベクトルを指示対象として扱えるように先の手法を拡張すると共に，実ロボットでその有効性を評価する．

2. 語彙学習タスクの概要

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．本稿では，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．

Learning Names of Places from Spoken Utterances for Autonomous Mobile Robot
Yuji Yamada[*1], Koosuke Hattori[*1], Ryo Taguchi[*1], Taizo Umezaki[*1], Masahiro Hoguro[*2], Naoto Iwahashi[*3], Kotaro Funakoshi[*4] and Mikio Nakano[*4]
[*1] Nagoya Institute of Technology, [*2] Chubu University, [*3] National Institute of Information and Communications Technology, [*4] Honda Research Institute Japan Co., Ltd.

3. 提案手法

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[2]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．

3.1 発話と指示対象の共起確率モデル

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．[1]では，$P(\mathbf{o}|w)$を離散確率分布としていたが，本稿では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．

3.2 語彙学習法

本学習は 4 つのステップに分けられる．Step 1 は，学習データの全音声を音素列として認識し，その統計量から初期の単語リストを生成する．Step 2 は，単語リストを用いて発話を単語列として認識し直し，その結果から意味モデルと文法モデルのパラメータを学習する．Step 3 では学習結果を利用し単語リストの再構築を行う．ここでは，MDL 原理に基づいた単語削除と，bigram 確率に基づいた単語連結を行う．これらの処理により，不要な単語の削除や，分割された単語

図 1：ASKA

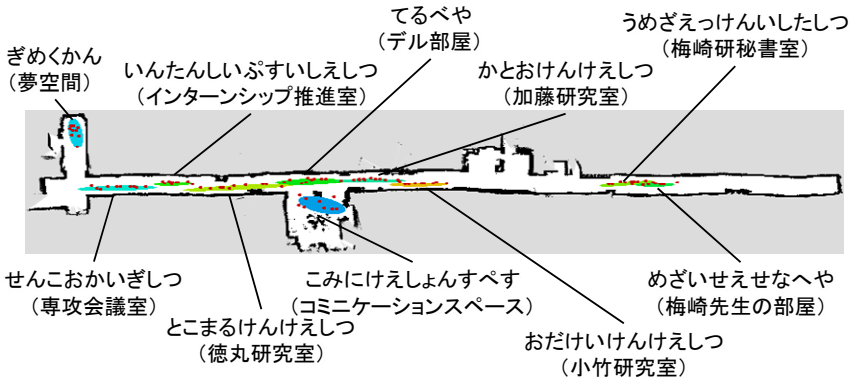

図 2：教示したキーワード（括弧書き）と出力されたキーワード（かな表記）

の復元ができる．Step 4 では，再構築された単語リストを用いてモデルパラメータの再学習を行う．Step 3，Step 4 を交互に繰り返すことで，より正しい音素系列が獲得される．手法の詳細は[1]の通りである．ただし，意味モデルの変更に伴い，モデル θ のパラメータ数 $f(\theta)$は次式のように修正した．

$$f(\theta) = K + (K^2 + 2K) + (K(m + m(m+1)/2)) \quad \ldots(4)$$

K は単語数，m は $\mathbf{o}$ の次元数である．

3.3 キーワードの出力

対象 $\mathbf{o}$ が入力された場合，対応するキーワード $w_{\mathbf{o}}$ を次式により出力する．

$$w_{\mathbf{o}} = \arg\max_{w \in \Omega} \left\{ \log P(w) + \log P(\mathbf{o} \mid w) \right\} \quad \ldots(5)$$

Ω はキーワード集合である．キーワードの判定にはエントロピーの減少量 $I(w)$を用いる．

$$I(w) = -\int P(\mathbf{o}) \log P(\mathbf{o})\, d\mathbf{o} + \int P(\mathbf{o} \mid w) \log P(\mathbf{o} \mid w)\, d\mathbf{o} \quad \ldots(6)$$

$I(w)$が閾値以上の単語をキーワードと判定する．

4. 実験と考察

対象 $\mathbf{o}$ を 2 次元平面上の位置座標(x,y)とし，車輪移動型ロボット ASKA（図 1）を用いて場所名の学習を行う．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[3].

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．作成した地図と教示キーワードを図 2 に示す．位置を変えながら各場所で 10 箇所（学習用 9，評価用 1），計 100 箇所の位置情報を取得した（図 2 の丸印）．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，各場所の名前を 9 種類の言い回し（表 1）で発話した．

3.2 の Step 3, Step 4 を 10 回繰り返し，語彙を学習させた．その後，学習に用いていない位置情報を入力し，対応するキーワードを出力させた．出力されたキーワードを図 2 に示す．他のキーワードとの混同や，大きな脱落・湧き出し誤りは無く，ほぼ正しくキーワードが学習できた．出力キーワードの平均音素正解精度は 81%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 74%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．ただし，全ての「研究室」が「けんけえしつ」となっているように，話者の発声に癖がある場合には，そのまま学習結果に反映される．また，図 2 に出力キーワードの確率値が閾値以上の部分を楕円で示す．この結果から，場所のカテゴリ化が行われていることがわかる．

表 1：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	

5. まとめ

本稿では，多様な言い回しでの教示から，指示対象のカテゴリとそれを表す音素系列を同時に学習する手法を提案した．今後は，複数の単語の組み合わせにより成り立つキーワードの学習を行う予定である．

参考文献

[1] 田口 他: 統計的モデル選択に基づいた連続音声からの語彙学習，人工知能学会論文誌，Vol.25, No.4, pp.5491-5501, (2010-6).

[2] Rissanen, J.: A universal prior for integers and estimation by minimum description length, The Annals of Stat., Vol. 11, No. 2, pp. 416–431, (1983)

[3] Hahnel et al.,: An Efficient FastSLAM Algorithm for Generating Maps of Large-Scale Cyclic Environments from Raw Laser Range Measurements, In Proc. of the IEEE/RSJ Int. Conf. on IROS, (2003).