モデル選択による言語獲得手法とその評価

Proposal and Evaluation of a Language Acquisition Method Based on Model Selection

田口 亮[*1 *2] Ryo Taguchi　岩橋 直人[*1*3] Naoto Iwahashi　能勢 隆[*1*4] Takashi Nose　船越孝太郎[*5] Kotaro Funakoshi　中野幹生[*5] Mikio Nakano

[*1] (株)国際電気通信基礎技術研究所 Advanced Telecommunications Research Institute International

[*2] 名古屋工業大学 Nagoya Institute of Technology

[*3] (独)情報通信研究機構 National Institute of Information and Communications Technology

[*4] 東京工業大学 Tokyo Institute of Technology

[*5] (株)ホンダ・リサーチ・インスティチュート・ジャパン Honda Research Institute Japan Co., Ltd.

This paper proposes a method for the unsupervised learning of lexicons from pairs of a spoken utterance and an object as its meaning without any a priori linguistic knowledge other than a phoneme acoustic model. In order to obtain a lexicon, a statistical model of the joint probability of a spoken utterance and an object is learned based on the minimum description length principle. This model consists of a list of word phoneme sequences and three statistical models: the phoneme acoustic model, a word-bigram model, and a word meaning model. Experimental results show that the method can acquire acoustically, grammatically and semantically appropriate words with about 85% phoneme accuracy.

1. はじめに

現在，多くの対話システムでは開発者が言語知識を用意しているが，その全てを網羅することは不可能であり，システムがユーザとのインタラクションを通して自ら知識を学習していくことが望まれる．特に，家庭用のロボットは，未知の人や物，場所等に遭遇する機会が多く，それらの名前を学習・発話するためには，ユーザの発話から，未知の単語（未登録語）の正しい分節とその音素系列，およびその意味（単語が指示する対象）を学習する必要がある．

発話から未登録語の音素系列を推定することは非常に困難であり，「私の名前は○○です」等のように，未登録語が受理可能な定型文を予め規定する必要があった [Holzapfel 08, 杉浦 08]．定型文を規定せず，人の自然な発話から意味的に有用な音素系列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[Gorin 99, Roy 02]．また，音声認識の分野では，発話から未登録語を検出するために，未登録語のクラス（人名や地名など）が持つ音響的，文法的なモデルを学習・利用する手法[Asadi 91, Schaaf 01, Bazzi 02, 山本 04]が提案されているが，単一の発話を対象としており，複数発話を用いた音素系列の精度向上は行われていない．

そこで本稿では，単語の知識を持たないロボットが，人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（意味）を学習するための手法を提案する．

2. 問題設定

ユーザがある対象をロボットに提示し，音声でその対象の名前を教示する課題を扱う．教示音声はユーザが日常的に用いている表現で発話され，「これはボールペンです」等のように，対象の名前以外の語を含んでいるとする．本稿では，対象の名前を**キーワード**，発話に含まれるキーワード以外の表現を**言い回し**と呼ぶ．両者は互いに独立であり，同じ言い回しで複数のキーワードが教示されること，一つのキーワードが複数の言い回しで教示されることを仮定する．ロボットは，音声を音素列として認識するための音響モデルだけを持ち，単語に関する知識は与えられない．人の発話のどの部分がキーワードであるかロボットには未知である．与えられた複数の音声-対象ペアから，キーワードの音素系列とその意味を学習し，各対象を表すキーワードの正しい音素系列を出力することが目的である．

連絡先：田口亮 Email: taguchi.ryo@nitech.ac.jp
TEL&FAX:052-735-5552

3. アプローチ

本タスクでは，発話からの単語の分節化と，各単語の正しい音素系列の推定が最大の問題となる．認識された音素系列は誤りを含むため，複数の音素系列間で完全一致する区間を分節するのではなく，なんらかの類似性の尺度を用いる必要がある．[中川 95]は音響的な類似性に基づいて，音声を直接分節するが，複数単語による教示だけでは正確に分節することが困難であることが報告されている．[Roy 02]では，音響的な類似性だけでなく，意味的な情報も利用して類似区間を検出しているが，獲得された単語の約7割で付加・脱落が発生している．また，[Roy 02]では獲得する単語の取捨選択について，古い知識を削除することしかできないため，複数の候補の中からより良い音素系列の単語候補を選定するといったことはできない．

そこで本稿では従来使用されていた音響，意味情報に加えて，文法的な情報も利用することを考える．先行研究では，キーワード以外の単語は無視されていたが，言い回しの持つ文法的な規則性を学習・利用することで，キーワードの前後に制約を与え，始端終端に生じる付加・脱落の抑制が期待できる．音響，文法，意味の各モデルは，発話と対象の共起確率モデルとして統合される． そして，この共起確率モデルを最小記述長原理（MDL）に基づいて最適化することで，キーワードのより正確な音素系列が獲得される．

4. 発話と対象の共起確率モデル

音声 A と対象 O の共起確率 $P(A|O)$を，図 1 のグラフィカルモデルで表現する．図中の各ノードは確率変数，エッジは確率の依存関係を表わしている．W_j は単語である．このモデルは音声 A が単語列 $S=(W_0,W_1,\ldots,W_{L+1})$から出力されること，各単語は対象 O を指示することを表しており，指示発話とその意味をモデル化している．図 1 のモデルを次式のように定式化する．

$$
\begin{aligned}
&\log p(A=a,O=o;\theta)\\
&=\log\sum_s p(A=a,O=o,S=s;\theta)\\
&=\log\sum_s \{p(A=a\mid S=s;\theta)p(S=s;\theta)p(O=o\mid S=s;\theta)\} \qquad \ldots(1)\\
&\approx \max_s \Big\{ \omega_1 \log p(A=a\mid S=s;\theta)\\
&\quad + \omega_2 \log\prod_{l=0}^{L^s} p(W_{l+1}=w_{l+1}^s \mid W_l=w_l^s;\theta)\\
&\quad + \omega_3 \log\sum_{l=1}^{L^s} \gamma(w_l^s,s,\theta)\,p(O=o\mid W_l=w_l^s;\theta) \Big\}
\end{aligned}
$$

但し，θ はモデルのパラメータ，L^s は単語列 s の単語数，w^s_l は単語列 s の l 番目の単語，w^s_0 は始端，w^s_{Ls+1} は終端を表わす．

この式の第一項は音響スコアであり，音素 HMM（音響モデル）を連結して作られた単語 HMM が出力する単語列 s の尤度として計算される．第二項は文法スコアであり，単語 bigram（文法モデル）を利用して計算される．但し，後述の式(3)でキーワードと判定された単語はキーワードクラスとして扱う．すなわち，全キーワードを一つの単語とみなし，各単語 bigram を統合する．第三項は意味スコアであり，各単語が示す対象の確率（語意モデル）を重み $\gamma(w^s_l, s, \theta)$によって加重平均している．重み $\gamma(w^s_l, s, \theta)$は次の式で計算する．

$$
\gamma(w_l^s,s,\theta)=\frac{\text{単語}w_l^s\text{の音素数}}{\text{発話}s\text{に含まれるキーワード の総音素数}} \qquad \ldots(2)
$$

但し，単語 w^s_lがキーワードでない場合，$\gamma(w^s_l, s, \theta)$を 0 とする．これは，発話の意味を，発話に含まれるキーワードのみから推定することを意図している．

なお，各確率モデルはモデリングの精度が異なるので，各スコアを ω_1～ ω_3 で重み付けて調整した．

5. 言語獲得手法の詳細

図 2 を用いて，処理の流れを説明する．提案手法は，（STEP1）発話を音素認識し，その結果から初期の単語リストを作成する．（STEP2）単語リストを用いて発話を単語列として認識し直し，その結果から語意モデルと文法モデルを学習する．（STEP3）統計的なモデル選択の基準に基づき単語リストを最適化する．STEP2 のモデル学習と STEP3 の単語リストの最適化を交互に繰り返すことで，各キーワードのより正しい音素系列が獲得される．各 STEP の詳細は次節以降で述べる．

5.1 STEP1：初期単語リストの生成

音素認識結果の音素列をモーラ列に変換し，その統計量に基づいて初期単語リストを生成する．具体的には，教示された全モーラ列に含まれる部分列の頻度から，各部分列の前後に

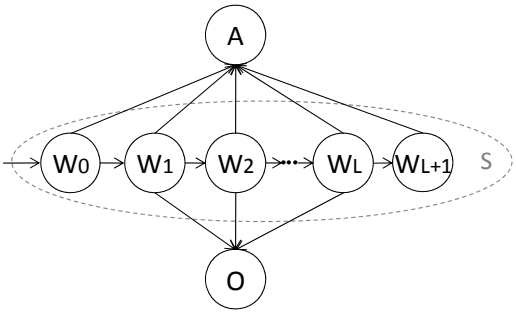

図 1：発話と対象の共起確率モデル

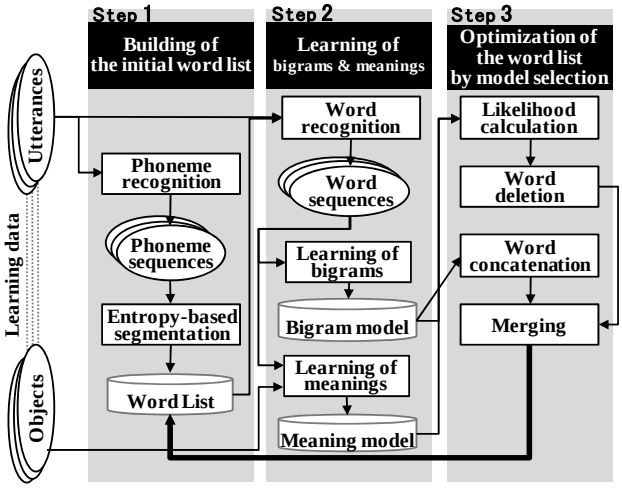

図 2：言語獲得手法の概要

接続されるモーラのエントロピーを算出する．エントロピーが非ゼロ，かつ出現頻度が 2 回以上である場合に，そのモーラ列を単語候補として単語リストに登録する．

5.2 STEP2：文法モデルと語意モデルの学習

生成された単語リストを用い，学習データの全音声を単語認識し，結果を N 個の候補（NBest）として得る．NBest の全単語列を用いて文法モデルと語意モデルを学習する．文法モデルは，単語 bigram とし，単語の並びの頻度から計算する．語意モデルは，単語 W と対象 O の条件付き確率分布 $P(O|W)$とし，単語と対象の共起頻度から算出する．そして，対象 O のエントロピー$H(O)$と条件付きエントロピー$H(O|W=w)$の差 $I(O|W=w)$が閾値よりも大きければ，単語 w をキーワードと判定する．

$$
\begin{aligned}
&I(O\mid W=w)\\
&=H(O)-H(O\mid W=w)\\
&=-\sum_o p(O=o)\log p(O=o)\\
&\quad+\sum_o p(O=o\mid W=w)\log p(O=o\mid W=w)
\end{aligned}
\qquad \ldots(3)
$$

5.3 STEP3:単語リストの最適化

統計的モデル選択の基準の一つである最小記述長原理（MDL）[下平 04]を用いて，単語の数および各単語の音素系列，すなわち単語リストの構成を最適化する．本稿では，モデルパラメータ θ（単語リスト，および各確率モデルのパラメータが含まれる）と観測データの記述長 DL を次のように定義する．

$$\mathrm{DL}(\theta) = -\mathrm{L}(\theta, \mathbf{D}) + \frac{f(\theta)}{2}\log M \quad \text{...(4)}$$

$$\mathrm{L}(\theta, \mathbf{D}) = \sum_{i=1}^{M} \log p(A = a_i, O = o_i; \theta) \quad \text{...(5)}$$

$$f(\theta) = K^2 + CK \quad . \quad \text{...(6)}$$

但し，L(θ , **D**)は θ の対数尤度，**D** は学習データセット（$\mathbf{D} = \{d_i \mid 1 \le i \le M\}, d_i = (a_i, o_i)$ ）である．$f(\theta)$は θ の自由度であり，第一項が文法モデル，第二項が語意モデルのパラメータ数である（音響モデルは学習していないため除外する）．K は単語数，C は対象数である．

単語セットを上記基準で最適化するためには，その組み合わせ全てに対して尤度を計算する必要があるが，現実的ではない．そこで本手法では，STEP2 で得た NBest を用いて，単語の有無による DL の差分を近似的に求める．音声を単語認識した結果の NBest には，ある単語 w が含まれる候補と，含まれない候補が存在する．それらの候補間の尤度差から，単語 w を削除した場合の DL が近似的に求まる．それが元のモデルの DL よりも小さければ，実際に単語 w を削除する．このように一つずつ単語を削除する場合，その順番によって結果が変わる．本手法では，削除の影響の少ない単語，すなわち尤度差が最小となる単語から削除していくこととした．

また，STEP2 で学習した bigram 確率が閾値以上（実験では 0.5 とした）となる単語のペアがある場合，それらを連結し，新たな単語を生成する．これにより，STEP1 で誤って分節された単語を復元することができる．

両者の結果をマージして新たな単語リストを生成する．その後 STEP2 に戻り，学習をやり直す．

記述長 DL と単語数との関係を図 3 に示す．図は実験の一事例である（50 語以上は省略）．単語削除の 1 回目（図中"1st"）では，31 単語の DL が 32 単語の DL を上回ったため，32 単語で削除をストップした．得られた単語セットに，単語連結によって作られた 14 単語を統合することで，46 単語の単語リストが生成される．得られた単語リストを用いてモデル学習（STEP2）が実行され，その後，単語削除の 2 回目（図中"2nd"）が実行される．図から，モデル学習（STEP2）と単語リスト最適化(STEP3)を繰り返すことで，単語連結で生成される単語が徐々に少なくなり，単語数が収束していくことがわかる．

5.4 キーワードの出力

ロボットがある対象 o の名前を正しく発話するためには，複数獲得されている単語の中から，対象 o を最も良く表すキーワード $\tilde{w}$ を一つ選び出力する必要がある．そのための関数を次にしめす．

$$\tilde{w} = \arg\max_{w \in \Omega} \left\{ \omega_2 \log p(W = w; \theta) + \omega_3 \log p(O = o \mid W = w; \theta) \right\} \quad \text{…(7)}$$

但し，Ω はキーワード集合である．この式は，複数単語を対象とする式(1)を，一単語に限定化したものである．なお，本稿では議論の混乱を避けるため音声合成の問題は扱わず，音素系列はシンボル列としての精度で評価する．そのため，式(7)では音響スコアの項が省略されている．

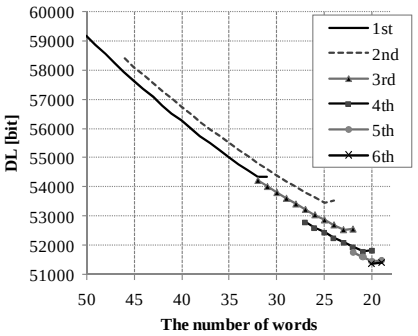

図 3：単語削除における記述長 DL の推移

表 1：対象の番号と対応するキーワード

対象	キーワード X	対象	キーワード X
1	会議室の前	6	竹内さんのブースの南
2	辻野さんのブース	7	工作室
3	フロアの真ん中	8	アシモの部屋
4	学生部屋の前	9	スマートルーム
5	お茶飲み場	10	スマートルームの入り口

表 2：言い回しのパターン

X の所に行って	今から X へ行って
X へお願い	この場所は X
ここは X です	ここの名前は X

6. 実験と考察

6.1 実験 1：重みパラメータ検討

まず，式(1)に用いる重み ω_1，ω_2，ω_3，を実験的に求めるため，1 名の発話データを用いて重みの影響を調査した．対象の数は 10，言い回しのパターン数は 6 とし，その全ての組み合わせとなる 60 発話を収集した．対象番号と対応するキーワードを表 1 に，言い回しのパターンを表 2 に示す．今回の実験では，文法重み ω_2=語意重み ω_3 と探索範囲を限定し，調整を行った．ω_2，ω_3を変化させた場合の実験は今後の課題とする．

モデル学習と単語リスト最適化を 10 回繰り返した結果を図 4，5 に示す．図 4 は単語リストに登録された単語数である．重みを設定しない場合 （ω_1= ω_2= ω_3=1），100 語以上の単語が残ったのに対し，重みを設定することで， 単語の取捨選択が行われることがわかる．図 5 は 10 個の対象のそれぞれに対し，式(6)でキーワードを出力し，その音素正解精度を算出した結果である．重みを使用しない場合，70%未満の精度しか得られないが，ω_1=0.0001，ω_2= ω_3=10 とした場合に最も高い 90%の精度が得られた．

6.2 実験 2：単語リスト最適化の効果の検証

各重みを実験 1 でキーワードの正解精度が最も高くなった値（ω_1=0.0001，ω_2= ω_3=10）に設定し，モデル選択による単語リストの最適化の効果を検証した．実験 1 と同じ発話セットを話者 17 名分収録し，話者毎に学習と評価を行った．17 名分の結果を平均したものを図 6 に示す．図の横軸は単語リスト最適化の回数になっている．図中のヒストグラムは獲得単語数と，そこに

含まれるキーワード数を表わしており，最適化を繰り返すことで単語数が減少していくことがわかる．最終的にはキーワードとして平均 13 語が得られた．これは真のキーワード数(10 語)とほぼ同数まで絞り込むことができたことを示している．

初期の言語知識を用いて 60 発話を音素認識した際の音素正解精度は 82%であった(図中の"phoneme accuracy for utterances")．一方，出力キーワードの音素正解精度(図中"phoneme accuracy for keywords")は，単語リスト最適化を行わない場合(図中 0 回目)に 50%以下であるのに対し，最適化を繰り返すことで 85%まで上昇した．最適化を行わない場合に，キーワードの音素正解精度が，ベースラインとなる発話の音素正解精度を大きく下回るという結果となったのは，出力されたキーワードの両端に付加・脱落誤りが生じているためである．表 3 に単語リスト最適化なしでの出力キーワードの例と，単語リスト最適化を 10 回繰り返した後の例を示す．表から，正しく得られていなかった単語境界が，単語リスト最適化によって修正された事がわかる．このように提案手法は，モデル学習と単語リスト最適化を繰り返すことで，キーワードの始端・終端を正しく推定できるようになる．

7. まとめ

本稿では，多様な言い回しでの教示から発話と対象の関係や単語の音素列を学習できる言語獲得手法を提案した．実験の結果から，単語の知識を与えることなく，平均 85%の精度でキーワードの音素列を獲得できること示した．研究で用いた基本原理は，音声と対応する非言語的意味情報のみから音声言語の形態素を抽出するという，言語獲得問題の解決への糸口になると考えられる．今後は，非言語的意味情報をより複雑にし，非限定的な発話を対象とした実験を行う予定である．

参考文献

[Holzapfel 08] Holzapfel, H., Neubig, D. and Waibel, A., "A Dialogue Approach to Learning Object Descriptions and Semantic Categories", Robotics and Autonomous Systems, Vol. 56, Issue 11: 1004-1013, 2008.

[杉浦 08] 杉浦孔明, 水谷了, 中村友昭, 長井隆行, 岩橋直人, 岡田浩之, 大森隆司, "音声からの未登録語切り出しと画像からの物体抽出の統合による新規物体の学習", 第 26 回日本ロボット学会学術講演会, 1N1-05, 2008.

[Gorin 99] Gorin, A. L., Petrovska-Delacretaz D., Wright, J. H. and Riccardi, G., "Learning spoken language without transcription", Proc. ASRU Workshop, 1999.

[Roy 02] Roy, D., and Pentland, A., "Learning words from sights and sounds: A computational model", Cognitive Science, 26, 113-146, 2002.

[Asadi 01] Asadi, A., Schwartz, R. and Makhoul, J., "Automatic Modeling for Adding New Words to a Large Vocabulary Continuous Speech Recognition System", Proc. ICASSP91:305-308, 1991.

[Schaaf 01] Schaaf, T., "Detection Of OOV Words Using Generalized Word Models And A Semantic Class Language Model", Proc. Eurospeech 2001.

[Bazzi 02] Bazzi, I., and Glass, J., "A multi-class approach for modelling out-of-vocabulary words", Proc. ICSLP02: 1613-1616, 2002.

[山本 04] 山本博史, 小窪浩明, 菊井玄一郎, 小川良彦, 匂坂芳典, "複数のマルコフモデルを用いた階層化言語モデルによる未登録語認識",電子情報通信学会論文誌 D-2, Vol.J87-D-2, No.12, pp.2104-2111, 2004.

[下平 04] 下平英寿, 久保川達也, 竹内啓, 伊藤秀一, "モデル選択 予測・検定・推定の交差点", 岩波書店, 2004.

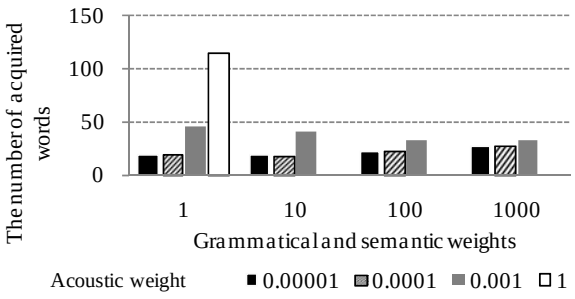

図 4: 重みと獲得単語数の関係

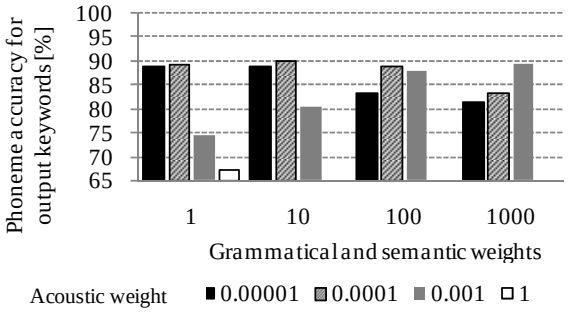

図 5: 重みとキーワードの音素正解精度の関係

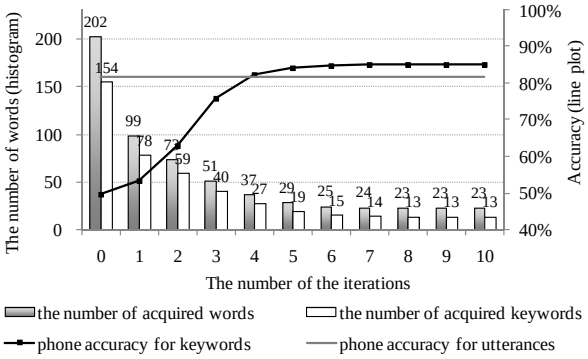

図 6: 単語リスト最適化が与える影響

表 3：出力されたキーワードの例

O	単語リスト最適化なし	単語リスト最適化 10 回
1	かいすのまえ	かいすのまえ
2	つじのさ	つじのさうのぶす
3	なか	ふろあどまんなか
4	がくせえべや	がくせえべやのまえ
5	おちょ	おちゃのみま
6	み	たきょいつさんのぶすのみなみ
7	こおさくしつ	こおさくひつ
8	あしものへや	あしものへや
9	む	すもあとるむ
10	ち	すまとるむのいいぐち