インタラクションの方策をオンライン学習する センサ情報を用いた単語のインタラクティブ学習法

長友 謙治† 岩橋 直人‡ 長井 隆行† 樽松 明†
†電気通信大学 ‡(株)ソニーコンピュータサイエンス研究所

1 はじめに

環境から必要な情報を抽出し状況に応じて適切な処理を行うコンピュータやロボットは，近い将来に日常生活の場において身近な存在になるであろう．それらのインターフェースとして対話は，従来のコンピュータの場合よりもいっそう重要になる．ダイナミックに変化する環境の中で，ユーザとコンピュータが共有する経験が適切に反映される対話を実現するための技術が必要となる．

従来の言語処理技術は，あらかじめ意味を与えられた記号体系に拘束されるため，ダイナミックに変化する環境に柔軟に適応することができない．経験を反映する対話を実現するためには，まず，実世界の事物を指示する単語とその意味をコンピュータが適応的に学習する必要がある．この学習は，センサを通して観測した特徴量を用いて，事物と音声の間の対応を求めるものであり，従来法は大まかに，バッチ学習(赤穂 97，Roy 99，金 01)とインタラクティブな逐次学習(Gorin 94，稲邑 00，Steels 01)[*1] に分けられる．それぞれ異なる学習の局面で有効に機能すると考えられるが，特にインタラクティブな逐次学習には，(1)コンピュータが必要な情報をユーザに要求できる，(2)ユーザがコンピュータの学習状態に応じて教示の仕方を変えることで効率的に学習が行える，という良い特徴がある．本稿では，コンピュータがユーザとのインタラクションの方策を学習することで，語彙を効率的に逐次学習できる方法について述べる．

2 提案法

単語学習のためのユーザとコンピュータのインタラクションは次の通りである．ユーザは，コンピュータにオブジェクトを見せながら単語を発話する．コンピュータは，発話された単語がコンピュータにとって未知の単語であると判断した時，その単語を語彙に登録する．コンピュータは，発話された単語が未知であるか既知であるか高い精度で判断できないと判断した時，既知の特定の単語であるかどうかをユーザに質問する．ユーザは，これに対して Yes/No の返答をする．以上を繰り返す．

このような学習を実現する際の問題は，(1)音声認識のあいまいな結果を用いて，発話された単語が未知か既知かの判断をいかにして正確に行うか，(2)この判断の精度をコンピュータ自身がいかに推測するか，(3)これらの判断と推測を，システムの使用環境，ユーザが提示するオブジェクト，単語の意味などに応じていかに適応させるか，の三つである．提案法は，これらの問題に対処するものである．まず，第一の問題は，発話の未/既知の判断を音声認識によって得られた音韻列だけに頼って行うのではなく，音声とともに提示されたオブジェクトの画像情報も利用して判断することで対処する．例えば，コンピュータがユーザの発話を [mi:tan]と認識した時，ユーザがみかんを見せていたならば，コンピュータはそれが既知の単語 [mikan] であったと判断し，ユーザがぬいぐるみを見せていたならば，その単語がぬいぐるみを指示する未知の単語 [mi:tan]であると判断できる．第二の問題には，未/既知の判断の精度を過去の質問の応答の結果に基づいて統計的に推測することで対処する．そして，第三の問題には，前述した判断と推測を行うためのシステムパラメータをオンラインで適応的に学習することで対処する．

提案法のアルゴリズムは次のとおりである．語彙中 i 番目の単語を W_i とする．各単語 W_i は音声の確率分布（音韻 HMM 列）q_i とオブジェクト画像のメンバーシップ関数（ガウシアン）r_i の組で表される．ユーザから音声 s とオブジェクト画像 v の与えられた時，音声 s が既知の単語である確率を次の決定関数 F で表す：

$$F(o_s, o_v) = \frac{1}{1+\exp(-\alpha_1(M(o_s)-\beta_1))} \cdot \frac{1}{1+\exp(-\alpha_2(L(o_v)-\beta_2))}$$

ここで，o_s，o_v はそれぞれ s と v の特徴量である．$M(o_s)$ は，o_s に対して最も尤度が高い音韻 HMM 列と，既知の単語の中で最も尤度が高い音韻 HMM 列 q_k の対数尤度比（音声マージンと呼ぶ）である．$L(o_v)$は，o_v に対する単語 W_k のオブジェクト画像メンバーシップ関数 r_k の尤度（最近傍単語画像尤度）を表す．α_1, β_1, α_2, β_2 は F のパラメータである．システムは，F の値が閾値 T_h 以上となった場合は s が既知単語 W_k である，閾値 T_l 以下となった場合は s が未知単語である，と判断する．そして，その他の場合はユーザに質問をする．質問として「今発声したのはこの単語ですか？」という意味で単語 W_k の音声を出力する．ユーザは，その音声が s と同じ単語であったらキーボードで Yes ，そうでなければ No と応答する．システムは，応答 Yes と No に対してそれぞれ s が既知または未知の単語であると判断する．

各単語 W_i の q_i と r_i は，W_i の発話であると判断された音声サンプルとそれと同時に提示されたオブジェクト画像サンプルを用いて統計的に学習する．また，α_1, β_1, α_2, β_2 の値は，F の出力が，質問に対するユーザの応答が Yes の時は 1，No の時は 0，に近くなるよう LMS アルゴリズムにより逐次的に学習する．

3 実験

オブジェクトとして図1のぬいぐるみ等を用い，システムへの提示は図2のように行った．学習実験では，あらかじめ収録しておいた，音声サンプルとオブジェクト画像サンプルの組からなるデータセット $\mathcal{O} = \{(s_1,v_1), (s_2,v_2), \ldots, (s_{280},v_{280})\}$を用いた．音声データは，各ぬいぐるみの全体的な特徴を指示する 32 単語と，色，形，

*1 ただし，(稲邑 00, Steels 01) はあらかじめ音声モデルが与えられた単語の意味を学習するものである．

†Kenji Nagatomo, Takayuki Nagai, Akira Kurematsu, The University of Electro-Communications.
‡ Naoto Iwahashi, Sony Computer Science Labs Inc.

大きさのそれぞれを指示する 8 単語を発話した音声 7 サンプルずつからなる．画像サンプルは，オブジェクトを提示する位置や照明を変えながら収録した．音声は MFCC，オブジェクトは色（L*a*b），大きさ(1 次元)，輪郭（複素フーリエ級数 8 次元）を特徴量とした．

まず，音声マージン M(s)と最近傍単語画像尤度 L(v)を用いてどの程度，未/既知単語の判断ができるか調べた．事前に正しい単語に分類したサンプルで単語を学習したのち，未知および既知の単語と判断すべきそれぞれの場合の音声マージンの分布（図３）および最近傍単語画像尤度の分布（図４）を求めてみた．分布の重なりの様子から，それぞれのパラメータだけでは未/既知の判断の誤りがかなり生じてしまうことがわかる．

次に，提案法による単語学習実験の結果を示す．データセット *O* から，それぞれの単語の音声・画像ペアを７ペアずつ連続して提示していった．T_h = 0.8，T_l=0.1 とし，F のパラメータの初期値は学習の初期段階でシステムが頻繁に質問するように経験的に決めた．図５は，決定関数Ｆを構成する音声マージン M(s)と最近傍単語画像尤度 L(v)のそれぞれを入力とするシグモイド関数の形状の変化を示す．図６は，質問が行われたサンプルの割合を，教示者が Yes/No の応答をした場合に分けて 10 サンプルごとに示したものである．決定関数Ｆの形状が変化するに伴い，質問の頻度が減っていった．質問に対する教示者の応答は，はじめは Yes ばかりであったが，しだいに Yes と No の割合が等しくなってきて一回の質問でシステムが得られる情報量が増えている．最終的にすべてサンプルを正しく単語に分類できた．

さらに，システムの振舞いに応じて教示者が教示の仕方を変える単語学習実験を行った．この実験では，同じ単語のサンプルの提示に対して３回連続してシステムから質問を受けなかった場合，ユーザはシステムがすでにその単語を正しく学習したものと判断するとみなし，以後この単語のサンプルを提示しないようにした．その結果，学習は 178 音声・画像サンプルのみ使用して終了した（図７）．少ないサンプルでも画像メンバーシップ関数が適切に学習できたかどうかを確認するために，この実験で学習された各単語の画像メンバーシップ関数を用いて，図４の場合と同様に最近傍単語画像尤度の分布を求めた（図８）．図４の場合と同じような形状になっていることから，比較的良好な学習が行えたことがわかる．この結果は，提案法において，ユーザがシステムの振る舞いに応じて適切に教示の仕方を変えて学習の効率を高められることを意味している．

4　さいごに

環境の変化に柔軟に適応できる単語の学習を実現するための計算機構を示した．閾値 T_l，T_h の適応，オブジェクト画像のメンバーシップの表現などが今後検討すべき課題である．

参考文献

赤穂，速水，長谷川，吉村，麻生 (1997) EM 法を用いた複数情源からの概念獲得，電子情報通信学会誌 J80-A(9)，p.1546-1553.

稲邑，稲葉，井上.(2000)個人に適応した語彙を獲得するロボットとの自然言語対話処理．ロボティクスシンポジア予稿集，pp.146--151.

Gorin, A. ,Levinson, S.and Sanker, A. (1994) An Experiment in SpokenLanguage Acquisition, IEEE Trans. Speech and Audio Processing, Vol.2, No1,pp.224-240.

金，岩橋．(2001) 知覚情報の統合に基づく階層構造を有する音声単位の獲得，日本音響学会春季講演論文集，pp.98-99.

Roy, D. (1999) Learning Words from Sights and Sounds: A computational model, Ph.D. Thesis, MIT.

Steels,L. and Kaplan,F. (2001) AIBO's first words: The social learning of language and meaning, *Evolution of Communication*, 4(1).

図１使用したオブジェクト例

図２ 提示の様子

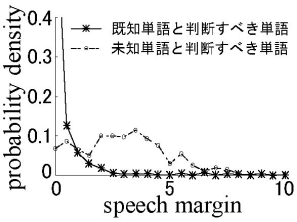

図３ 音声マージンの分布

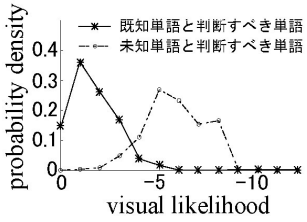

図４ 最近傍単語画像尤度の分布

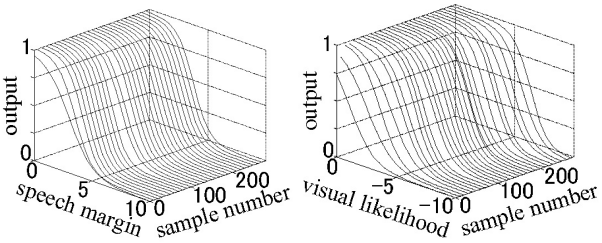

図５　決定関数Ｆの変化

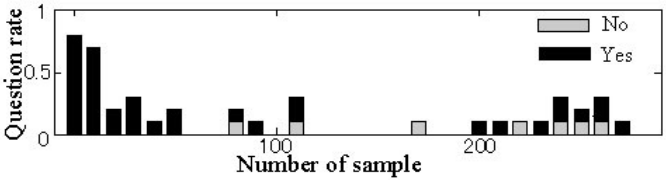

図６ 質問が行われた割合の変化

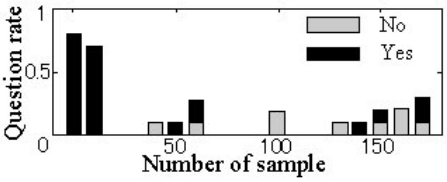

図７ ユーザが教示の仕方をシステムに適応した場合

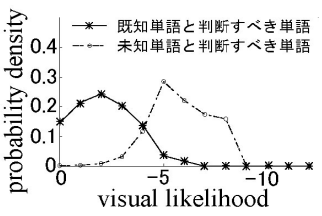

図８ 最近傍単語画像尤度