システム制御情報学会論文誌, Vol. 16, No. 6, pp. 260–269, 2003

論　文

発話理解学習を利用した適応的インターフェイス —人間同士のコミュニケーション成立過程からの知見*

小松　孝徳† ・鈴木健太郎† ・植田　一博† ・開　一夫†‡ ・岡　夏樹§

An Adaptive Interface System Based on a Speech Meaning Acquisition Model —From a Perspective of Human-Human Communication Establishment Process*

Takanori KOMATSU†, Kentaro SUZUKI†, Kazuhiro UEDA†, Kazuo HIRAKI†‡ and Natsuki OKA§

The purpose of this study is to propose a speech meaning acquisition model, which can be applied for an adaptive interface system, from a perspective of human-human communication establishment process. The model was designed to discriminate the types of instructions based on salient prosodic features and to recognize the given instructions using a positive reward (given for its successful action) and a negative reward (from the utterance of a instructor which draws listener's attention). As a result of a test, this model could eventually learn to recognize the given instruction from an actual human instructor. It is expected that the constructed meaning acquisition model can be applied for an adaptive interface system, which provides a natural interaction environment for its user.

1. 序論

複雑な機能をもつ機械であってもユーザが効率よく操作できる環境を提供するためのインターフェイスが，さまざまな分野で活発に研究されている．その中でも，人間が最も自然に用いることができる「発話・音声情報」を媒介としたインターフェイス技術が近年注目を集めている．

従来の音声インターフェイスの多くは，発話のうち文字として表現される音韻情報に注目して処理を行ってきた（たとえば[1]を参照のこと）．このようなインターフェイスでは，音声認識により発話を文字情報に変換し，その変換された文字情報から適切な機能を選択する．しかしこのような手法では，音声認識などのプロセスに多くの計算を必要とし，また音声認識率の低さのため，発話情報の入力から適切な機能を選択するまでに時間がかかってしまい，その応答の遅さが問題とされていた．この問題に対して，発話情報のうち文字に転写することで失われる韻律情報（イントネーション，音量，話速など）に注目することで計算時間を短縮し，インタラクティブな処理を目指した音声インターフェイスが研究されている[2–4]．たとえば，Igarashi and Hughes[4]は，地図画面を表示したナビゲーションシステムに対してユーザが「move up, ahhhh」と発話し，「ahhh」音声が継続している間は「move up」という発話の意味する「画面上方向にスクロール」という行動をとり続け，そのピッチの値を高くすることでスクロールの動きを加速できるという音声インターフェイスを開発した．ここで，「move up 20cm」のような発話を扱う従来型の音声インターフェイスの手法とIgarashiらの提案する手法を比較してみると，前者では「move up 20cm」という一連の発話の意味を解析した後に行動が開始されるのに対し，後者の手法では「move up」までを解析しただけで行動を開始することができ，どの程度動けばよいのかというユーザの意図（前者の場合は「20cm」という文字情報に含まれている）は「ahhh」というユーザの発話からインタラクティブに獲得することができる．よって後者では発話の継続時間とピッチ情報を用いることで，インタラクティブな応答を実現することに成功している．

しかしこのインターフェイスでも，「move up＝上スク

* 原稿受付　2002年6月3日

† 東京大学　大学院総合文化研究科　Department of General System Studies, The University of Tokyo; Komaba 3-8-1, Meguro-ku, Tokyo 153-8902, JAPAN

‡ 科学技術振興事業団　JST/Presto

§ 松下電器産業(株)　先端技術研究所　Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd.; 3-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-2579, JAPAN

Key Words: speech meaning acquisition, prosodic information, EM algorithm, mutual adaptation, adaptive interface.

ロール」といった発話と行動とのマッピングを設計者があらかじめ定義しておく必要がある．したがって，ユーザは自分の行いたい教示方法を自由に行えるのではなく，定義された教示方法に習熟していく必要がある．また，もしユーザが発話と機能のマッピングを自由に定義できたとしても，違和感のないインタラクション環境を実現するために，ユーザ自身が試行錯誤的にそのマッピングを何度も何度も修正していく必要が生じる場合があると考えられる．よって，ユーザにとって違和感のない自然なインターフェイスを実現するためは，インタラクションを通じてユーザの用いる音声教示の意味をインターフェイス自身がリアルタイムに学習していくことが重要だと考えられる．この状況を「発話理解学習」という状況に置き換えると，ユーザが発話者，インターフェイスが学習者という関係になる．

本研究の目的は，ユーザとのインタラクションを通じて発話と機能との結び付きをインターフェイス自身が学習することを可能にするための基礎技術を提案することである．具体的には，教示の種類を発話の韻律的特徴から弁別し，その弁別された情報と機能との結び付きを学習することで発話の意味を理解する学習モデルを構築した．そのために，まず，相手が何かを話していることはわかるがその意味はわからないような状況を設定し，そこで話し手の発話意味をどのようにして聞き手が理解していくのかを観察するための人間同士のコミュニケーション実験を行った．そして，そこで観察された人間の発話理解プロセスをもとに，発話中の韻律情報と機能とを結び付けることで発話の意味理解を可能とするモデルを提案・構築し，実際に人間からインタラクティブに教示を受けた場合に，モデルが教示の意味を理解できるかどうかを調べることでその学習能力を検証した．この際，この学習モデルが実際の人間の教示を理解するようになれば，ユーザに対して自然なインタラクション環境を提供する適応的インターフェイスを実現するための基礎技術として，このモデルは活用できると考えられる．

本論文は以下のような構成からなる．**2.** ではコミュニケーション実験について，**3.** では実験の結果をもとに提案した発話意味学習モデルについて述べる．**4.** で，提案されたモデルの学習能力検証実験を行い，**5.** で，適応的インターフェイスの実現へむけて本モデルが貢献できると期待される点を議論し，最後に **6.** で本研究をまとめる．

2. コミュニケーション観察実験

2.1 実験概要

実験には二人一組の被験者が参加し，そのうちの一人（操作者）がTVゲームを操作し，もう一人（教示者）が音声による教示を与えた．本実験で用いたTVゲームは，画面上のラケットを左右に動かし落下してくるボールを打ち返すことで得点を獲得し，失敗すると減点されるという，スカッシュのようなゲームである．教示者と操作者はそれぞれ別々の部屋に配置され，操作者は教示者から与えられる音声教示に基づいてラケットを操作し，教示者は操作者がラケットでボールを打ち返せるように音声で指示を出した．この実験環境をFig. 1に示す．

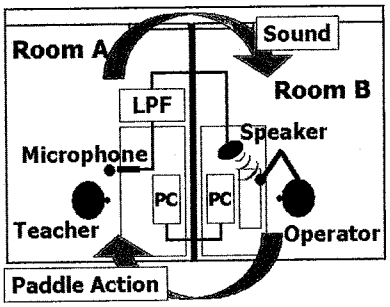

Fig. 1 Experimental environment

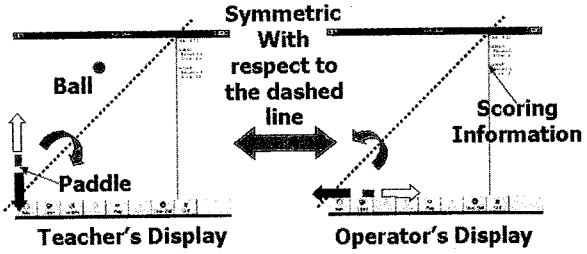

Fig. 2 Display setting

この際，教示者の教示の音韻的な意味を操作者が理解できないようにした．具体的には，被験者ペアが共通の母語を持つ場合には，教示音声にローパスフィルタを通した音声[1]が操作者に与えられ，一方，共通の母語を持たない場合には，操作者にとって未知の言語である教示者の母語によって教示が与えられた．これにより操作者は，教示者が何かを話していることはわかるがその意味はわからないような状況となる．なお，教示者は使用する教示の種類，手法には制限は加えられておらず，自由に教示を行うことができる．

実験中，操作者と教示者は，ゲームの状況をそれぞれのディスプレイで見ることができるが，操作者の画面にはラケットで打ち返すべき目標のボールは表示されていない（Fig. 2）．この状況を「すいか割り」ゲームにたとえると，操作者が目隠しをしてスイカを叩く役で，教示者が周りで指示を出す役ということになる．このため，操作者がラケットでボールを打ち返すためには，何を言っているのかわからない教示音声の意味

[1] ローパスフィルタは，音声中のある周波数より高い周波数成分を除去する機能がある．そのため，主に発話中の摩擦音が除去され，発話から音韻情報を獲得することが困難になる．一方，低周波成分は保持されるために，発話の基本周波数成分やイントネーションなどの韻律情報はある程度保持される．なお，本実験におけるローパスフィルタのカットオフ周波数は，教示者が男性の場合約150Hz，女性の場合約250Hzとした．

Table 1 Correct direction value of and hit value of subject pairs in Experiment 1

グループ	(平均方向正答値, 平均ヒット値)
グループ1・教示無理解(2組)	(0.5, 0.5), (0.3, 0.2)
グループ2・方向教示獲得(5組)	(0.9, 0.3), (1.0, 0.2), (1.0, 0.5), (0.8, 0.6), (1.0, 0.6)
グループ3・距離教示獲得(4組)	(1.0, 0.9), (1.0, 0.7), (1.0, 0.7), (0.9, 0.8)

Table 2 Correct direction value of and hit value of subject pairs in Experiment 2

グループ	(平均方向正答値, 平均ヒット値, 教示者-操作者の母語)
グループ1・教示無理解(2組)	(0.5, 0.5, 中国語-日本語), (0.4, 0.4, 中国語-日本語)
グループ2・方向教示獲得(2組)	(1.0, 0.6, インドネシア語-英語), (0.8, 0.6, 中国語-日本語)
グループ3・距離教示獲得(2組)	(0.9, 0.7, スペイン語-タガログ語)
	(0.9, 0.9, ハングル語-中国語)

を理解する必要がある[2]. したがって, この環境下で操作者がラケットでボールを打ち返せるようになれば, 未知の音声の意味を獲得したとみなされ, この二者はある種のコミュニケーションを成立させたと考えられる. その成立プロセスを観察することが本実験の目的である(この実験についての詳しい説明は, 参考文献[5, 6]を参照されたい).

2.2 被験者

被験者ペアは共通の母語を持つか持たないかで二群に分けられた. 実験1には, 共通の母語を持つ被験者ペア22人11組(20~28歳, すべて日本人)が参加した. また, 実験2には, 共通の母語を持たず, かつお互いに相手の母語を理解できない被験者ペア12人6組(20~32歳)が参加した.

2.3 実験結果

被験者ペアのパフォーマンスを評価するために, 方向正答値, ヒット値という二種類の指標を導入した. **方向正答値**とは, 各試行で教示者の意図した方向に操作者がラケットを動かした場合に1点, そうでない場合に0点を与えたものである. また, **ヒット値**とは, 各試行でラケットにボールが当たった場合に1点, 当たらなかった場合に0点を与えたものである. 各被験者ペアのゲーム終了直前10試行の平均方向正答値と平均ヒット値を比較することで, 被験者ペアを大きく三つのグループに分けることができた.

グループ1 平均方向正答値が0.8以下の場合.

グループ2 平均方向正答値が0.8以上で, 平均ヒット値が0.7以下の場合.

グループ3 平均方向正答値が0.8以上で, 平均ヒット値が0.7以上の場合.

被験者ペアの分類は, 二項分布による仮説検定によって行われた. 方向教示を理解していない場合に教示の意図する方向へ動く確率は0.5と考えられ, このとき最終平均方向正答値が0.8以上になる確率は$p<.0547$となる. よって, 平均方向正答値が0.8以上の場合には教示者の意図する方向を操作者は理解していると考えられる. また, 方向教示の意味を獲得した場合にラケットとボールが偶然に当たる確率は0.34となり(ラケットと画面の幅, ラケットとボールの速度から計算される), ここから平均ヒット値が0.7以上となる確率は$p<.023$となる. よって, 平均ヒット値が0.7以上の場合には, 教示から移動方向のみならず移動距離も理解してボールをラケットで打ち返していると考えられる.

各実験における被験者ペアの, 平均方向正答値と平均ヒット値をそれぞれTables 1, 2に示す. これらの表から以下のことが理解できる. 実験1の11組の被験者ペアのうち, 9組が「どちらの方向へ動けば良いのか」という教示者の意図を理解し(方向教示理解), そのうちの4組は,「どの地点に動いたら良いのか」というより深い意味を理解できていた(距離教示理解). また, 実験2の6組の被験者ペアの場合は, 4組が方向教示理解を達成し, そのうちの2組が距離教示理解を達成していた.

このように, 両実験の多くの被験者ペアが, 未知の音声の意味を獲得することで, ゲームで効率よく得点を獲得していた. そして, それらの意味理解プロセスにおい

[2] それぞれの被験者の画面はFig. 2にように点線に対して対称になっている, 異なる画面設定が使用されている. その理由は以下の通りである. ラケットが左右に動く画面を両者ともが見ているような状況だと, 教示者が「左」「右」という教示を使用するであろうと操作者は容易に推測できる. このような状況で実際にローパスフィルタを通された教示を聞かされても, 操作者はその音声のモーラ数の違いから教示の意味を容易に推定することが可能となる. これに対して, Fig. 2のような画面設定にすると, 教示者は「上」「下」と発話するようになり, 操作者はモーラ数では教示者の発話の意味を推測できなくなるため(操作者は「左」「右」という教示を期待しているため), 発話の音韻情報からの推測の影響を除くことができる. 被験者ペアは, お互いの画面がこのように異なっていることは実験終了まで知らされなかった.

て，以下の二点が共通に観察された．

(1) 韻律情報による注意喚起

操作者は未知の教示であっても，その音の「聞こえ方」から教示の種類を区別していた．また，教示音声中における「ピッチの上昇」や「声のうねり」として観測される声を荒げるような韻律パターンが，操作者に対して注意を喚起していたことが観察された．Fig. 3の上図は，実際に声を荒げている音声のピッチ情報とパワー情報を縦軸に，時間を横軸にプロットしたもので，下図はその音声に対応した画面上のラケットの位置を縦軸にプロットしたものである．このFig. 3によって，教示者の音声に対応した操作者のラケット動作を観察することができる．そしてここから，教示音声のピッチが急激に上昇すると，それに対応して操作者のラケットの移動方向が反転していたことが理解できる．つまり，教示音声のこのような韻律的特徴が操作者の行動に対して注意を与えていたことが観察された．このような効果を持つ韻律情報を**警告韻律**（attention prosody）とよぶ．このような警告韻律に対する操作者の反応は，実験開始当初からすべての操作者において観察されていたため，その役割は普遍的なものだと考えられる．

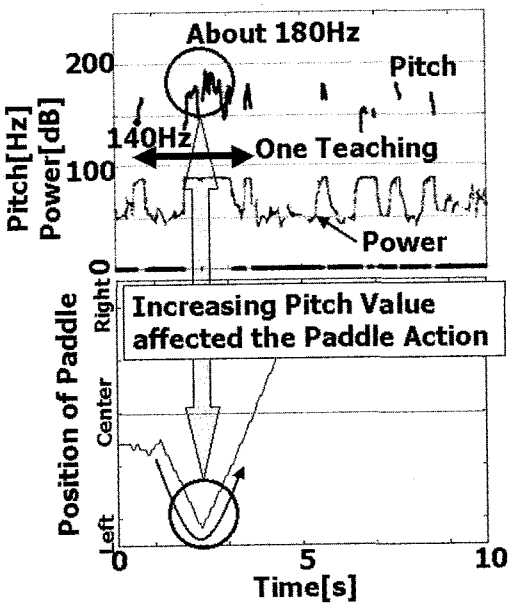

Fig. 3 Attention prosody and paddle action

(2) 複合報酬による強化学習的な意味獲得プロセス

被験者は音声教示とラケットの行動を対応させて教示の意味を獲得していた．その際，行動と教示の結び付きを評価する情報が必要となるが，本実験においては二種類の情報（報酬）が用いられた．一つは，「ボールをラケットに当てる」という目標を達成したことにより得られる正の報酬であり，もう一つは教示音声の警告韻律から与えられる負の報酬である．したがって，本実験で観察された教示の意味獲得プロセスは，これら複合報酬に基づく強化学習的なプロセスだと考えられる．

以上より，本実験の操作者は，

- 韻律的特徴の差異から教示の種類を区別する
- ボールを当てることで得た正の報酬と，警告韻律による負の報酬，という複合的な報酬を利用して教示の意味理解を行う

という二つのプロセスによって，未知の発話意味を理解していたと考えられる．音声学などの分野では発話を用いたコミュニケーションにおいて韻律情報がその成立に重要な役割を果たしているという知見が得られているが[7,8]，本実験の結果においても，韻律情報が発話理解というコミュニケーションの成立に大きな影響を与えていたことが観察された．よって，この実験で得られた知見は，発話の中の韻律情報を用いた適応的インターフェイスの開発に寄与できると期待される．次節では，この知見に基づいた教示の意味学習モデルを提案する．

3. 意味学習モデルの提案

3.1 モデルの概要

前節の実験結果から，韻律情報がコミュニケーション成立過程に与える影響についての知見が得られた．これらの知見をもとに，実際に人間の音声から意味学習できる操作者モデルを構築し実装する．まず，本モデルは次のような能力を持つことが求められる．

1) 自分の行動と音声教示とを結び付ける能力．
2) 教示の区別の決め手となる音響的特徴を見つけ出す能力．
3) 発せられる教示の意味を前もって限定せず，学習によって獲得する能力．
4) 警告韻律を抽出し活用できる能力．

そして，これらの能力を具体的に実現するために，次のような学習モデルを想定した．

- 自分の行動に対して正の報酬を受けた時（本ゲーム環境ではラケットにボールが当たった時），自分の行動の直前に発せられた教示音声の意味は，自分のとった行動（行動速度）を指示していると認識する．
- 自分の行動に対して負の報酬を受けたとき（本ゲーム環境では警告韻律を与えられた時），行動の直前に発せられた教示音声の意味は，自分の行動を指示していないと認識する．
- 教示音声はある程度の誤差をもって発せられると仮定する．報酬を受けたときの教示音声・行動のセットは蓄積され（音声–行動データ），そのデータはいくつかのクラスタに分類される．一つのクラスタが一つの教示の意味に相当し，各クラスタはパラメータとしてそれぞれ平均値と分散を持つ．
- 音声–行動データがある程度蓄積され，クラスタリングされたら，そのクラスタ情報をもとに新たに与

えられた音声の意味を推測する．

3.2 音声–行動データの形式

コミュニケーション実験における実際の操作者は，自分の行動に対して報酬を与えられた際，その直前に発せられた教示音声を記憶し，その行動と共にその音声データを蓄積すると考えられる．そして，その音声–行動データに共通する特徴から，教示の意味を判断していたと考えられる．本モデルでも，報酬が与えられたときの直前の教示音声データおよび直前の行動データをクラスタリングして，そのクラスタ内のデータの平均値を記憶する．

教示音声は，無音区間が1秒以上あった場合にその前と後の音声は別の教示であるとみなした．4秒以上の長さにわたる教示音声の場合は，直近の4秒の音声–行動データのみを利用した．最終的に一つの教示音声は，八種類の韻律的特徴（ピッチ，ピッチ一次微分，ピッチ二次微分，ピッチの急激な変化回数，ピッチ一次微分の急激な変化回数，ピッチ二次微分の急激な変化回数，ゼロクロス数，有声率[1]）の教示区間における時間平均として表された[2]．

また，行動は時間あたりの左・右の移動（キープッシュ）頻度で示されるが，報酬を受ける直前の行動が最も重視されるように係数をかけた加重速度を用いた．時刻 $t=T$ で報酬を受け取った場合の（平均）加重速度を次式で定義する．ここで T_0 は，教示の長さか4秒のいずれか小さい方の値を取る．

$$W\_speed = \frac{\sum_{t=T-T_0}^{T} action_t \cdot (t-(T-T_0))}{T_0} \qquad (1)$$

ただし，

$$action_t = \begin{cases} 1 & (\text{時刻}\,t\,\text{で右へ移動}) \\ 0 & (\text{移動なし}) \\ -1 & (\text{時刻}\,t\,\text{で左へ移動}) \end{cases}$$

である．

3.3 クラスタリングの方法

本モデルでは正規混合分布から音声–行動データが生成されたと仮定した．この場合，どのデータがどの正規分布から生成されたのかがわかれば各混合分布のパラメータ（平均値・分散）を求めるのは非常に容易である．しかし，「どのデータがどの分布から生成されたのか」ということは，実際には観測不能な値（隠れ値）である．この隠れ値とは，i 番目のデータが j 番目の正規分布から生成された場合は $Z_{ij}=1$，生成されない場合には $Z_{ij}=0$ となる値である．

そこで本モデルでは，EMアルゴリズム [9] を用いることで，このような不完全データから混合分布のパラメータ（平均値・分散）を推測し，同時に隠れ値の推定を行った．EMアルゴリズムでは，適当な初期値を与えたパラメータから，Eステップ（Expectation Step）とMステップ（Maximization Step）とよばれる二つの手続きを繰り返すことにより，パラメータの値を逐次更新する．簡単に表すと，

Eステップ: 現在のパラメータから，入力された音声データに対する隠れ値の期待値を推定

Mステップ: Eステップで求めた隠れ値の期待値から，各分布のパラメータを推定する

という操作の繰り返しとなり，パラメータと隠れ値が同時に推定できる．

しかし，このような従来のEMアルゴリズムでは，負の報酬を受けた時の音声–行動データを扱うことはできない．なぜなら，負の報酬時の音声–行動データは，「モデルが推定した隠れ値を使用してとった行動が，教示者の意図とは違う行動である」ことを示しているからである．そこで本モデルでは，従来型のEMアルゴリズムのEステップを以下のように拡張することで失敗例を扱えるようにした．音声データ i が与えられた時，分布 j に属する行動をとることで失敗例となった場合（教示者から警告韻律を与えられた場合），$Z_{ij}=0$ として，残りの分布の隠れ値を $Z=1/(\text{分布数}-1)$ と修正する．成功例の場合には，従来の方法と同様に現在のパラメータから隠れ値を推定する．

本来，EMアルゴリズムはすべてのデータを獲得してからバッチ処理する計算であるが，本モデルでは実際にインタラクションしている発話者から与えられる発話の意味をインタラクティブに学習したいので，オンラインで音声–行動データを処理する必要がある．オンラインEMアルゴリズムについてはいくつか提案されているが [10]，ここでは簡単のために (2) 式のようなパラメータ更新式を用いる．

$$\Theta(t+1) = \alpha\Theta(t) + (1-\alpha)\Theta'(t+1) \qquad (2)$$

ただし $\Theta'(t)$ は時刻 t に計算されたパラメータ Θ の値，$\Theta(t)$ は時刻 t における (2) 式で擬似推測されたパラメータ Θ の値である．本モデルでは $\alpha=0.85$ とした．本モデルのアルゴリズムの全体像は Fig. 4 のようになる．

3.4 モデルを利用した教示の意味判定

発せられた音声の意味は，その音声が最も高い確率で生み出される分布の平均加重速度であるとし，その速度で行動する．この実験環境ではゲームソフトの仕様上，(1) 式の W_speed が負の場合にはラケットを左方向に，W_speed が正の場合には右方向に一定の速度で移動さ

[1] 有声率とは，ある教示音声の中で実際にピッチを持つ音声が発生されている割合のこと．

[2] 本モデルには，これらの韻律的特徴量のほかにパワー情報も加える予定であったが，Fig. 3 からわかるように，今回の実験系においてその値は音声区間の有無しか示していなかった．よって，本論文ではピッチを中心とした韻律的特徴量を入力情報とした．

1. 音声，行動，報酬データ読み込み
2. 突然0になる，あるいは突然大きな値になっているピッチデータを補正
3. 補正データをさらに0.05[s]間隔幅の移動平均で円滑化
   →このデータに対して以下の処理を実行
4. 差分，2次差分，極端な変化の頻度のデータ，有声率を計算
5. 教示の開始，停止のポイントをチェック（1.0[s]の無音区間があれば別教示とみなす）
6. 報酬があった場合
   - 6-1 教示区間の行動をサーチして，現在に近い行動に重み付けをして左右行動値を計算
   - 6-2 それをもとに加重速度を計算
   - 6-3 教示区間の音声データの平均を計算（6-2，6-3がクラスタリングの元データとなる）
   - 6-4 混合ガウス分布を仮定してEMアルゴリズム開始
   - 6-5 計算が一定回数を越える，または，平均・分散データの変化が一定以下になるまで計算をする
   - 6-6 各パラメータΘを(2)式で更新
7. 1. に戻る

Fig. 4 Learning procedure of constructed model for speech meaning acquisition

せた．

3.5 警告韻律の抽出

本モデルでは，教示者から与えられる音声から警告韻律を検出し，教示の意味学習に負の報酬として利用する．**2.**のコミュニケーション実験にて録音された警告韻律の音響的な性質を調べたところ，すべての警告韻律に以下のような特徴のいずれかが観察された．

ピッチ値の相対的増加 同一教示内において発話の開始時と比べてピッチの値が約20%増加していた（Fig. 3に示した警告韻律では，教示開始時のピッチ値が約140Hzで，実際に警告韻律と認識されていた部分では約180Hzであった）．

ピッチ二次微分値の絶対値の相対的増加 ピッチ二次微分値はピッチの「うねり」を表しているため，その絶対値が大きければピッチの上下動が激しい音声となる．

有声率の相対的増加 有声率の高い音声は圧力を感じる音声となる．

しかし逆に，これらいずれかの特徴量が観察されたとしても，必ずしも操作者は警告韻律と認識していない場合があることが別の実験から明らかになった．現段階においては，これらの，あるいはその他の特徴量が具体的にどのような値をとれば警告韻律として認識されるかという問題は完全には解決していない．今後，この問題に対しては，聴覚心理，音響心理的な側面からの検討が必要となるであろう．そこで本モデルにおいては簡単のため，これら三つの特徴のうち最も検出が容易な一つ目の特徴量である「ピッチ値の相対的増加」を検出した際に，それを警告韻律と認識することとした．

4. 人間とのインタラクションによる意味学習モデルの評価

4.1 実験目的

前節で提案した学習アルゴリズムをもとに，実際に人間の音声から意味学習を行うモデルを構築し，その学習能力を検証した．この検証実験によって，教示者とインタラクションしながら発話の意味を学習していく操作者のモデルとして，本モデルが適しているのかどうかを検討できる．具体的には，コミュニケーション実験で人間の操作者が操作していたラケットに提案された意味学習モデルを実装した．

4.2 実験概要

本モデルに実装された学習アルゴリズムでは，混合正規分布中の正規分布数を6個と設定して学習を開始した．クラスタリング計算は報酬を得たときの音声－行動データを10個獲得してから開始し，その際に選択された分布から生成される行動を行う．音声－行動データを10個獲得するまでは，教示を与えられた際にランダムに混合分布中の分布を選択し，その分布から生成される行動を行う．クラスタリング計算に使用される音声－行動データは最近の10データとし，新しいデータを受け取るたびに，最も古いデータを削除した．

実際にこのモデルに教示を行ったのは，事前に教示の練習を十分に行った教示者1名であり，以下のような五種類の教示をモデルに与えた．

(1) 日本語の「右」「左」を使用する．
(2) 英語の「right」「left」を使用する．
(3) 「あ～」と発音しながら，高いトーンで右を意味し，低いトーンで左を意味する．
(4) 「あ～」と発音しながら，高いトーンで左を意味し，低いトーンで右を意味する（(3)の逆）．
(5) 「あ～」と発音しながら，長い音で右を意味し，ぶ

つぶつ途切れる音で左を意味する．

2. のコミュニケーション実験で，まず初めに観察された現象は，操作者の教示理解を助けるために，教示者が使用する教示の種類を減少させるという「教示者の操作者に対する適応学習」である．一方，本節の評価実験における教示者は，実験開始時から固定した教示方略を用いているため，上記のような適応学習をすでに済ませた教示者と同等であるといえる．よって，この評価実験における教示学習は，コミュニケーション実験で一番初めに観察された現象以降の，操作者（意味学習モデル）の教示学習に相当する．

このようなさまざまな教示方略による発話の意味を本モデルが理解できるようになれば，本モデルは，インタラクションを通じて言語的な情報ではなく韻律情報を利用してその意味を獲得できるといえる．具体的な手順として，これら五種類の教示者と本モデルとで10分間のゲームを行い，その際の方向正答値とヒット値を用いてそのパフォーマンスを評価した．なお，方向正答値とヒット値がそれぞれ0.8，0.7を超えた場合は，モデルが教示の意味を学習したとみなし，その時点でゲームを終了した．この基準は人間同士のコミュニケーション実験において距離教示を理解したと判定した基準値と同じものである．

4.3 実験結果・考察

実験結果をTable 3に示す．ここからいずれのタイプの教示方略に対しても，本モデルは与えられる教示の意味を理解していたと考えられる．すなわち，各実験結果を人間同士のコミュニケーション実験のそれと比較すると，いずれの場合においても距離教示を理解したレベルに達していたといえる．モデルの操作するラケットが一定の速度でしか行動できないにもかかわらず距離教示を理解した被験者と同等のパフォーマンスを達成できた理由は，モデルに対して教示を与えるタイミングを教示者が学習していたからだと考えられる．

またTable 3から，それぞれの教示タイプに対して，教示理解の達成速度に差が生じていることが理解できる．Fig. 5に，達成速度が速い場合と遅い場合の方向正答率の経時的な推移を示す．白丸でプロットされたグラフはタイプ(2)の教示者の場合，黒角でプロットされたものはタイプ(3)の場合である．この図より，達成速度が速いタイプ(2)の教示者の場合，クラスタリングが始まるとそのパフォーマンスが一気に上昇しているのに対して，達成速度が遅いタイプ(3)の場合には，クラスタリングが始まってもパフォーマンスがすぐには上昇せずに，ある期間を経てから上昇するという経過をたどっている．このパフォーマンスの差の要因を探るために，モデル内部の混合分布の状態を観察した．タイプ(2)のモデル内部の混合分布の遷移をFig. 6，タイプ(3)のそれをFig. 7を示す．これらの図は，教示の弁別に最も関与

Table 3 Correct direction value and hit value by instruction types

教示タイプ	(平均方向正答値, 平均ヒット値，教示回数)
(1) 日本語「右」「左」	(0.9, 0.7, 72)
(2) 「right」「left」	(1.0, 0.8, 56)
(3) 音程「高」「低」	(1.0, 0.7, 82)
(4) 音程「低」「高」	(1.0, 0.7, 60)
(5) 有声率「密」「疎」	(0.9, 0.7, 73)

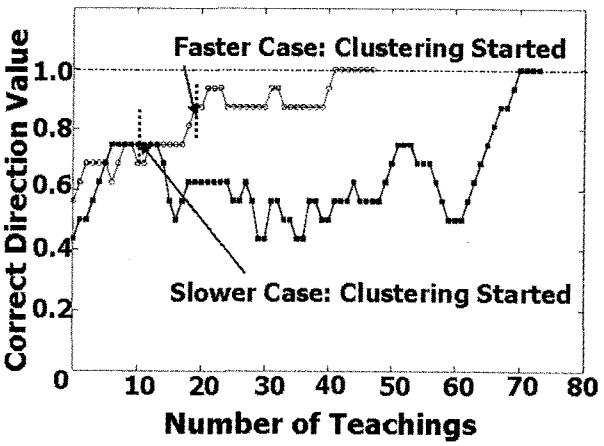

Fig. 5 Comparison of slower and faster performance

していたと考えられる混合分布中の代表的なパラメータ値を横軸，行動速度を縦軸にとり，この二軸からなる平面上に，学習ステップごとの各分布のパラメータと行動速度値との組をプロットしたものである．つまりこの図は，教示学習過程においてモデル内部の六つの分布がどのような遷移をしたかを示したものである．また代表的なパラメータとして，タイプ(2)の場合は有声率，タイプ(3)の場合はピッチ値を選択した[1]．

たとえば，タイプ(2)のような教示を理解するには，「高い有声率＝正の行動速度」「低い有声率＝負の行動速度」といった組合せを学習モデルが獲得する必要がある．Fig. 6中のDistribution#1と示した分布は，行動速度が7前後で有声率が0.2前後の地点に実験の初期状態としてランダムに配置されていた．しかし学習が進むにつれてこのモデルは，分布Distribution#1が高い有声率を扱えるようにパラメータの更新を行い，最終的にこの分布は有声率が0.9前後の地点に移動していた．この位置は，「高い有声率＝正の行動速度」という組合せを表しており，結果としてこのモデルは，タイプ(2)の教示方略

[1] タイプ(2)の教示者が使用した「left」という教示音声は，その「ft」部分が摩擦音であるのでピッチ値は検出されない．よって，摩擦音の存在しない「right」教示と比べ「left」教示の有声率は低くなるため，有声率に注目することで教示の弁別が可能となる．一方，タイプ(3)の場合は，設定された教示方法である音の高低に注目することで教示の弁別が可能である．

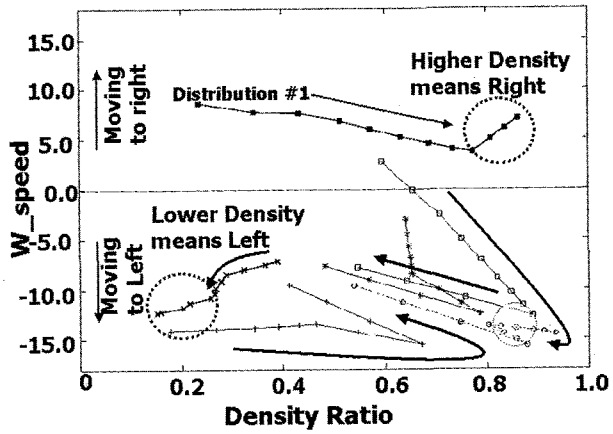

Fig. 6 Transition of model's parameters (Fast learned case)

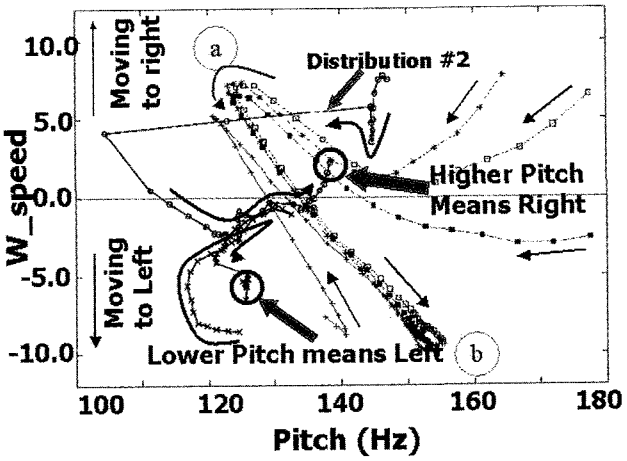

Fig. 7 Transition of model's parameters (Slow learned case)

である「長い音で右を意味する」という意味の一つを表現することに成功しているのである．

達成速度の速いタイプ(2)の場合，Fig. 6から教示者の使用する二種類の教示パターン（「right」,「left」）に対応する二種類のクラスタが学習開始時から安定して学習されていたことが観察された．また，これら以外のクラスタは，負の報酬を受けたときに分配される微小な隠れ値の影響を受けることで，徐々に灰色の円で示された負の報酬を得た際のデータ（負例）に近づいていく．そして，モデルの教示理解が進んで負の報酬が使用されなくなると，教示とは無関係な地点に急激に移動していた．よって，この場合のモデルの学習は，局所解に陥ることなくスムーズに最適解に達したと考えられる．

それに対して，達成速度の遅いタイプ(3)のモデル内部の混合分布のパラメータの遷移は非常に複雑なものであった．特に注目すべき点は，負例に対する各分布のパラメータの変化である．学習開始直後，低いトーンの教示音声に対してモデルが右移動した際に，教示者はモデルに対して警告韻律を与えていた（Fig. 7のa地点付近に該当するデータ）．そのため，実際に負の報酬を受けた分布以外の分布のパラメータはその負例に近づいていく．その後，高いトーンの教示音声に対してモデルが左移動した際に教示者は警告韻律を与え始めたため（b地点に該当するデータ），a地点に向かっていた分布のパラメータがb地点に向かって急激に向きを変えているのが観察できる．このような，負例による急激なパラメータの変化が及ぼす影響は，図中にDistribution#2と示された分布のパラメータ値の変遷から理解できる．この分布は学習開始直後から多くの負の報酬を直接受けたため，そのパラメータ値は急激に降下していた．しかしその後，他の分布の行動に対して負の報酬が与えられ始めると，その方向（b地点の方向）に向けてパラメータが急激に移動していた．結果として，このパラメータ値の急激な移動によって，このパラメータは再び学習に使用されるような値をとることができ，最終的にこの分布は教示意味を示すクラスタを形成していた．このように，モデル外部の教示者から与えられる負の報酬により，モデル内部のパラメータ値は劇的に変化することができ，その結果として，一度陥った局所解から抜け出し最適解に到達したと考えられる．

結果としてタイプ(3)の学習は，最適解にたどり着くまでに局所解に陥ったため，スムーズに最適解を得ることができたタイプ(2)の場合と比べて学習に時間がかかったと考えられる．よって，Table 3に見られるパフォーマンスの差異は，教示者の教示方略に依存しているのではなく，局所解にどれだけ陥るかに依存していると考えられる．EMアルゴリズムの学習能力は，初期値に大きく依存することが既に知られているため，その学習過程において常に局所解に陥らないとは限らない．よって，本研究で拡張されたEMアルゴリズムは，時間がかかりながらも教示者の教示によって局所解を脱することができるという特性を持つことから，教示者とのインタラクションによる意味学習モデルの基本アルゴリズムとして有用だと考えられる．

4.4 評価実験のまとめ

モデルの学習能力評価実験より，本研究にて提案された発話理解モデルでは以下の点が実現できた．

(1) 八種類の韻律的特徴からその差異を見いだすことによる教示種類の弁別．
(2) 成功例と失敗例とを報酬として活用することで，行動を通じた未知の教示意味の獲得．
(3) 警告韻律でモデルに負の報酬を与えることで，モデルのパラメータを局所解（たとえば，すべてのデータが一つのクラスタで説明されてしまうような状態）から脱出させることができ，その結果，最適解に達するまでの継続的な学習を実現．

よって，本研究にて提案された意味学習モデルは，コミュニケーション実験の結果から推測された意味理解プロセスを反映したものであるといえる．ここから本モデ

ルは，教示者とのインタラクションを通じて発話の意味を学習していくモデルとして有用だといえよう．

また，実際に教示を与える教示者は，教示を与えるタイミングを学習することで，一定の速度でしか動くことのできないラケットに対して効率よく正の報酬を獲得させていた．よって，モデルが教示者に適応するだけではなく，教示者もモデルに適応すること，つまりお互いが相手のことを相互に学習していくことが，最終的に二者のスムーズな関係を成立させていたと考えることができる．このような現象は人間同士のコミュニケーション実験でも実際に観察されていたので[5,6]，二者間での相互的な適応現象は，Human-Agent Interaction (HAI) 技術を考える上で重要な要素になると考えられる．

また，現段階の本モデルに対しては以下の改良点が考えられる．

(1) 失敗例の認識をより自然に行うための，より詳細な警告韻律抽出方法の検討．
(2) ピッチ以外の韻律情報（例えばパワー情報）の利用．
(3) ラケットが一律の速度で動作するのではなく，速度を変化させる機能の追加．
(4) コミュニケーション実験で観察された，教示種類を減少させるという「教示者の適応学習」プロセスを含めた学習モデルの評価．

5. 議論

一般的にインターフェイスに発話理解などの学習機能を実装しようとした場合，その学習時間が長くかかると考えられるため，実用化には不向きだとされている．よって，実用化されているインターフェイスは，前述のIgarashi and Hughes[4]のように，発話と機能とのマッピングをあらかじめ与えておくことで「ユーザが機械に適応する」手法をとったものがほとんどである．しかし，本研究で提案された発話理解モデルでは，長くても5分弱（教示回数にして80回前後），短ければ2分程度（30回前後）で教示の意味を理解することができた．このインターフェイスの学習時間を長いと感じるユーザも存在するかもしれないが，ユーザが機械に適応する労力を考えた場合，前者の方がユーザに対する負担が少ないと考えられる．つまり，従来のインターフェイスの方式だと，ユーザはあらかじめ設定された「発話と機能とのマッピング」を学習しなければならないのに対して，本研究の方式では，ユーザは「教示のコツ」のようなものを学習するだけでよいのである．よって，試行錯誤的に自分でインターフェイスを繰り返しカスタマイズしていく手法よりも，本研究で提案したようなユーザの発話をインターフェイスが学習していく手法の方が，自然なインタラクション環境を効率よく実現できると考えられる．このような発話理解学習の機能を音声インターフェイスに利用し，5分程度の学習時間でユーザにとって自然なインタラクション環境を提供できるのであれば，本研究で提案された「発話理解学習を利用して適応的インターフェイスを実現する手法」は実用性のあるものだと考えられる．

本研究では韻律情報のみに注目した発話理解技術を紹介したが，応用されるシステムの機能が複雑になった場合にはその対応には限界があるとも想像できる．しかし，本研究で提案した方式と画像処理・音声認識技術などを併用し，それに合わせて学習モデルの拡張・改良を行っていくことで，実環境での使用に耐えられるような，より適応的なインターフェイスの構築が期待される．

6. 結論

本研究では，適応的インターフェイスを実現するために，人間同士のコミュニケーション成立過程から得られた知見を利用して，ユーザとのインタラクションを通じて発話の意味を学習することを可能とする基礎技術の構築を行った．この学習モデルは，教示の種類を韻律的特徴の差異から弁別し，自分のとった行動と対応づけることでその意味を学習できる．その学習は，自分の行動が成功した際に得られる「正の報酬」と，教示者から警告的な意図を持つ教示を受けることによる「負の報酬」という複合報酬をもとに行われ，その結果，本モデルは，与えられる未知の教示の意味を教示者とのインタラクションを通じて理解することができた．音声を媒介としたインターフェイスにこのような技術を応用することで，ユーザの使用する教示の意味を理解しながらユーザにとって自然なインタラクションを提供できる適応的インターフェイスを実現することが期待される．

参考文献

[1] 中野，堂坂：音声対話システムの言語・対話処理；人工知能学会誌，Vol. 17, No. 3, pp. 271–278 (2002)
[2] M. Goto, K. Itoh, T. Akiba and S. Hayamizu: Speech completion: New speech interface with on-demand completion assistance; *Proceedings of HCI International* (2001)
[3] W. Tsukahara and N. Ward: Responding to subtle, fleeting changes in user's internal state; *Proceedings of CHI 2001*, pp. 77–84 (2001)
[4] T. Igarashi and J. F. Hughes: Voice as sound: Using non-verbal voice input for interactive control; *Proceedings of 14th Annual Symposium on User Interface Software and Technology (ACM UIST'01)*, pp. 155–156 (2001)
[5] T. Komatsu, K. Suzuki, K. Ueda, K. Hiraki and N. Oka: Mutual adaptive meaning acquisition by paralanguage information: Experimental analysis of communication establishing process; *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, pp. 548–553 (2002)
[6] 小松，鈴木，植田，開，岡：パラ言語情報を利用した相互適応的な意味獲得プロセスの実験的分析；認知科学，

Vol. 10, No. 1, pp. 121–138 (2002)
[7] 小磯，伝：円滑な話者交替はいかにして成立するか —会話コーパスの分析にもとづく考察；認知科学，Vol. 7, No. 1, pp. 93–106 (2000)
[8] J. B. Pirrehumbert and J. Hirschberg: The meaning of intonational contours in the interpretation of discourse; *Intentions in Communication*, (Editors, P. R. Cohen and M. E. Pollack), MIT Press, pp. 271–311 (1990)
[9] A. Dempster, N. Laird and D. Rubin: Maximum likelihood from incomplete data via the EM algorithm; *Journal of Royal Statistical Society B*, Vol. 39, pp. 1–38 (1977)
[10] 赤穂，速水，長谷川，吉村，麻生：EM 法を用いた複数情報源からの概念獲得；電子情報通信学会論文誌 A, Vol. J80-A, No. 9, pp. 1546–1553 (1997)

著者略歴

小松 孝徳 (学生会員)

1974年8月16日生．1997年 芝浦工業大学工学部機械第二学科卒業．1997～1999年 オーストリア工科大学客員兼Profactor Research GmbH EU 圏外研究員．現在，東京大学大学院総合文化研究科広域科学専攻 博士課程在籍．2001年より理化学研究所脳科学総合研究センター・ジュニアリサーチアソシエイト．人間との円滑なコミュニケーションを可能とするようなインターフェイス，エージェントの研究開発に従事．IEEE，情報処理学会，日本ロボット学会，日本認知科学会などの学生会員．

鈴木健太郎

1977年5月27日生．2002年 東京大学大学院総合文化研究科広域科学専攻修士課程修了．現在，同研究科博士課程在籍．脳神経科学の知見に基づき，構成論的手法を用いて学習現象へのアプローチを行っている．人工知能学会会員．

植田 一博

1963年6月14日生．1993年 東京大学大学院総合文化研究科広域科学専攻 博士課程修了．現在，東京大学大学院情報学環・学際情報学府助教授．科学的発見（特に類推と協同），図的推論，認知的インタフェース，人工社会・経済，マルチエージェント・システム，人間＝人工物間コミュニケーションなどの研究に従事．博士（学術）．著書に『科学を考える：人工知能からカルチュラル・スタディーズまで14の視点』（共著，北大路書房），『協同の知を探る：創造的コラボレーションの認知科学』（共編著，共立出版），"Evolutionary Computation in Economics and Finance" (co-authored, Springer Verlag) など．人工知能学会，情報処理学会，Cognitive Science Society，AAAI 各会員．

開 一夫

1963年8月11日生．1993年 慶應義塾大学大学院理工学研究科計算機科学専攻博士課程修了．電子技術総合研究所（現，産業技術総合研究所）情報科学部主任研究官を経て，現在東京大学総合文化研究科助教授．科学技術振興事業団さきがけ21「協調と制御」領域研究員およびATR メディア情報科学研究所非常勤研究員兼務．乳幼児の認知発達過程（特に自己認知の発達）に興味があり，発達認知神経科学的研究を行っている．博士（工学）．日本ロボット学会，日本赤ちゃん学会，SRCD，OHBM，AAAI などの会員．

岡 夏樹

1956年5月27日生．1979年 東京大学工学部計数工学科卒業．(株) 島津製作所，東京大学に勤務後，松下電器産業 (株) に入社．(財) 新世代コンピュータ技術開発機構研究員などを経て，現在松下電器産業 (株) 先端技術研究所主席研究員．学習，発達，インタラクションを中心とした認知モデル構築とその工学的実現に関心があり，自然な日常的インタラクションを通して言語獲得するシステムの構築を目指している．工学博士．情報処理学会の会員．