知 能 と 複 雑 系　130−14
（２００２．１１．１２）

ユーザとのインタラクションを通じた発話意味獲得モデル

小 松 孝 徳[†]　鈴木 健太郎[†]　植 田 一 博[†]
開 一 夫[†,††]　岡 夏 樹[†††]

本研究の目的は，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ユーザの教示の意味を学習することを可能とする基礎技術を構築することである．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する．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」という複合報酬を基に行われ，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の実現が期待される．

Speech Meaning Acquisition Model by Interaction with its User

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. はじめに

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，様々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（例えば文献[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．さらにこれらの手法では，ある発話の意味するインターフェイスの機能を「発話と行動のマッピング」として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．

この問題に対して筆者らは，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た．この状況を「発話理解学

† 東京大学 大学院総合文化研究科
†† 科学技術振興事業団さきがけ研究 21
††† 松下電器産業（株）

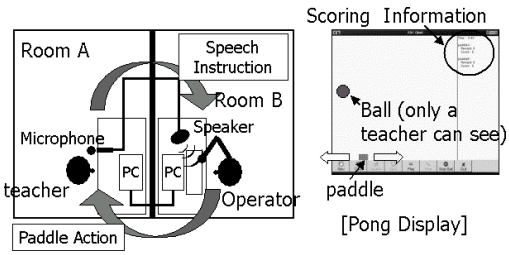

図 1 Experiment Setting

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ことは分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基に，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．

2. コミュニケーション実験

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．この実験環境を図 1 に示す．

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☆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分かるがその意味は分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．この状況を「すいか割り」ゲームに例えると，操作者が目隠しをしてスイカを叩く役で，教示者が回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を理解する必要がある．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，文献[2] を参照されたい)．実験に参加した被験者ペアは，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，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．

2.2 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，ヒット値，方向正答値という二種類の指標を導入した．**ヒット値**とは，各試行でラケットがボールに当たった場合に 1 点，当たらなかった場合に 0 点を与えたものである．また，**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．各被験者ペアのゲーム終了直前 10 試行の平均ヒット値と平均方向正答値を用いることで，平均方向正答値が 0.8 以上

☆ 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．

表 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)

表 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)

の場合「どちらの方向へ動けば良いのかという教示者の意図を操作者が理解しており（方向教示理解），平均ヒット値が 0.7 以上の場合は「どの地点に動けば良いのか」という意味を理解している（距離教示理解）とみなすことができる☆．この値を用いて被験者ペアを大きく 3 つのグループに分けることができた（表 1,2）．これらの表より，実験 1 の 11 組の被験者ペアのうち，9 組が方向教示理解，そのうちの 4 組が距離教示理解を達成しており，実験 2 の 6 組の被験者ペアのうち，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とが理解できる．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，声を荒げるような音声の含まれる韻律パターンが，操作者に対して注意を喚起していたことが観察された．このような効果を持つ韻律情報を**警告韻律 (Attention Prosody)** と呼ぶ．図 2 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．この図 2 によって，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．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．

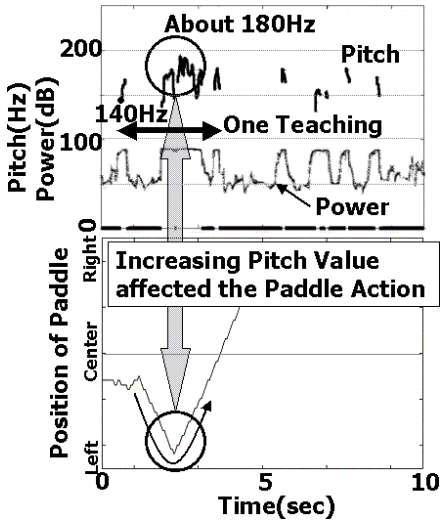

図 2 Attention Prosody and Paddle Action

(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．
- ボールに当てることで得た正の報酬と，警告韻律

☆ これらの値は，操作者が教示者の意図する方向へ移動する確率，ラケットがボールに当たる確率とを用いた二項分布による仮説検定によって設定された．

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．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. 発話意味獲得モデルの提案

3.1 モデル概要

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．具体的には，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（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し，負の報酬を受けたときには（警告韻律を与えられた時），教示の意味は，自分の行動を指示していないと認識する．
- 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．

一つの教示音声は，8 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☆）の教示区間における時間平均として表され，モデルの行動速度（右方向は正，左方向は負）は一つの実数で表した．

3.2 クラスタリング方法

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．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）☆☆であるため，本モデルでは，EM アルゴリズム[3] 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）を推測した．手順としては，適当な初期値を与えたパラメータから，入力音声に対する隠れ値の期待値を推定する E ステップ (Expectation Step) と，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 M ステップ (Maximization Step) と呼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ものである．

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取った行動が，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 $Z = 1/(N-1)$ と修正する（N: 混合分布中のクラスタの数）．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．

また，一般的な EM アルゴリズムはバッチ処理に用いられる計算手法であるが，本モデルでは実際にリアルタイムに発話者から与えられる発話の意味を学習する必要があるので，オンラインでデータを処理できるように EM アルゴリズムを拡張したものを使用した．最終的な本モデルのアルゴリズムの全体像は図 3 のようになる．

3.3 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．第 2 節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は，ピッチ値の相対的増加，ピッチ二次微分値の絶対値の相対的増加（ピッチの「うねり」に相当），有声率の相対的増加，といった特徴のいずれかが観察された．しかし，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．本稿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「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．

☆ 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．

☆☆ 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．

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 0.05sec 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（1sec 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づ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 を式 (2) で更新
7. 1. に戻る

図 3 Learning Procedure of Constructed Model for Speech Meaning Acquisition

3.4 発話意味獲得モデルの評価実験

提案した学習アルゴリズムを基に，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．

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，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，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，ぶつぶつ途切れる音で左を意味する．

このような様々な教示方略による発話の意味を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．具体的には，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から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の移動平均値がそれぞれ，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，その時点で実験を終了した．

表 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)

3.5 評価実験結果

実験結果を表 3 に示す．ここからいずれのタイプの教示方法に対しても，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．

実際に教示学習がどのように行われていたかを示すため，タイプ (2) の教示を受けた際のモデル内部の混合分布の状態を図 4 に示した．この図は，教示の弁別に最も関与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（この場合は有声率）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，学習ステップ毎の各分布のパラメータと行動速度値との組を

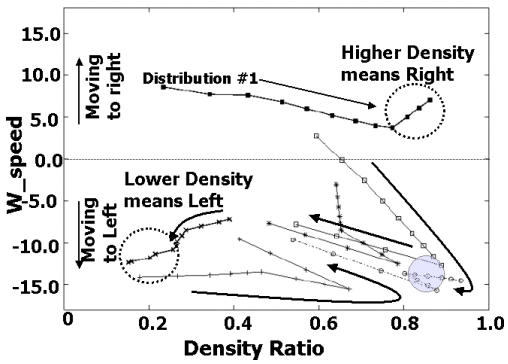

図 4 Transition of Model's Parameters

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６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．例えば，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」といった組み合わせを学習モデルが獲得する必要がある☆．図 4 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 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」という組み合わせを表しており，結果としてこのモデルは，タイプ (2) の教示方略である「長い音で右を意味する」という意味を表現することに成功しているのである．

よって，モデルの学習能力評価実験より，この発話意味獲得モデルは，1). 与えられた教示音声の韻律的特徴の差異を見出すことで教示の種類を弁別し，2). 成功例と失敗例とを報酬として活用した学習手法を用いて教示の意味を獲得している，ということが確認された．よってこのモデルの機能は，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言える．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言えよう．

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．

☆ ここで使用された「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は検出されない．よって，摩擦音の存在しない「right」教示より「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．

(2) さらに多くの種類の教示に対する意味獲得モデルの動作確認．

4. 議論・おわりに

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，その学習の信頼度，学習時間といった問題から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，Igarashi and Hughes[4] に代表されるように，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．このような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．しかし，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良いためユーザへの負担が少ない．よって，本研究で提案したようなユーザの用いる発話をインターフェイスが学習していく手法により，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．

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．

謝辞 日頃，本稿の内容に関連する研究の議論に参加頂いている IΔEA(Interaction DEsign for Adaptation) 研究会のメンバに感謝します．

参 考 文 献

1) 中野，堂坂．(2002)．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.
2) 小松，鈴木，植田，開，岡．(2002)．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』，(to appear).
3) A. Dempster, N. Laird and D. Rubin.(1977). Maximum Likelihood from Incomplete data via the EM Algorithm, *Journal of Royal Statistical Society B*, Vol. 39, pp.1–38.
4) T. Igarashi and J. F. Hughes.(2001). 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.