Experiment Toward a Mutual Adaptive Speech Interface That Adopts the Cognitive Features Humans Use for Communication and
Induces and Exploits Users' Adaptation

Takanori Komatsu　Atushi Ustunomiya　Kentaro Suzuki　Kazuhiro Ueda
Kazuo Hiraki　Natsuki Oka

1. INTRODUCTION

- さまざまなペットロボット (例；アイボ) は老若男女の問わず人気
- 多くの人たちがペットロボットとインタラクティブを取り楽しむ
- 本物のペットとは違い結局は飽きる
- 本物の犬には " 待て " や " お座り " と言った言葉を教える
- 飼い犬は飼い主の声に適応し、飼い主は飼い主の声を飼い犬の行動に適応させる (Katcher&Beck,1983)。このようなことは、子供と親の関係においても観察される。
- ペットロボットには観察されない
- ペットロボットを本物のペットのようにするには２つの能力が必要であると考えられる
  - 人間の表現の情報であるパラ言語情報 (文字で書けば同じ文であっても、話し方によって印象が全く違ってしまいます。このような違いを伝える情報をパラ言語情報) を認識する能力
  - ２つの異なる報酬を使いこなす人間について学習システム

- 本研究では、音律情報ではなく韻律情報に焦点を当てる
  (音韻情報；言語の内容)
  (韻律情報；イントネーション)

- 発話・音声情報を媒介としたインターフェー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ースの多くは発話の音韻情報に注目して処理を行ってきた
- 音韻情報を基としてインターフェースは、音声認識率が低い
- 韻律情報を基にインタラクティブな処理を目指した音声インターフェースが研究されている

2. COMMUNICATION EXPERIMENT

2.1 目的と背景

- 発話中のパラ言語情報と機能との結びつきを学習することで発話の意味を理解していくような意味獲得プロセスを観察する実験を行った
- その結果から発話による意味獲得モデルの構築に対する知見を得た
- この条件を満たすために、相手が何か話したことは分かるがその意味は分からないような環境を設定し、その制限された環境下における被験者同士のコミュニケーション獲得を分析した
- 実験には、二人一組の被験者が参加
  - 一人が教示者、もう一人が操作者
- 教示者はＡの部屋。操作者はＢの部屋。 (Figure1)

- 被験者の目的は「ポン」というテレビゲームで協力してできるだけ高得点を狙うこと
- １０試行
  - 毎試行ごとに落下してくるボールを打ち返すことで得点を獲得する
- スコアがゲームウインドウの右端に表示される。
  - このシステムによって現在のアクションが成功したか失敗したか操作者が理解できる。
- 教示者の役割
  - 操作者に指示を出すことと
- 操作者の役割
  - ボールを打ち返すためにラケットを動かし、できる限り高いスコアを得ること
- 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ない
- 教示者の指示の意味や指示を理解する必要
- 教示者の指示は言語学的には理解不可能

2.2 被験者

- グループ１：日本人 22 人 11 組 (20~28 歳 ; 男性… 18 人、女性… 4 人)
  - 全ての被験者に十分にパソコンを使ってもらう経験をしてもらった。
  - 11 人が院生、５人が学部生、 3 人が大学の卒業生、３人が短大の卒業生
  - 全員が日本語を使う
  - ローパスフィルターを通した教示音声が与えられる。
- ローパスフィルターは音声のある周波数より高い周波数成分を除去する機能
  - 主に発話中の摩擦音が除去されるため、発話から音韻情報を獲得することが困難になる
  - ただし、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には影響がない
    - ローパスフィルターのカットオフ周波数は、教示者が男性の場合は役150Hz、女性の場合250Hz
  - 操作者は韻律情報から指示を推測しなければならない

☆注意

教示者・操作者は異なる画面設定を使用

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」という教示をしようし操作者は容易に推測できる。このような状況では実際にローパスフィルターを通された教示を効かされても、操作者は音声のモーラの違いから教示の意味を推定できる。そこで Figure1 のような画面にした。教示者は「上」「下」と発話するようになり、操作者はモーラでは教示を推定できなくなる。

モーラ：音韻論上の単位。１子音音素と１短母音音素とを合わせたものと等しい長さの音素結合。

- グループ２： 12 人 6 組 (23~26 歳；男性… 10 人、女性… 2 人)
  - 全ての被験者に十分にパソコンを使ってもらう経験をしてもらった。
  - 8 人が院生、 4 人が大学の卒業生
  - 共通の母語を持たず、操作者は教示者の母語の学習暦が無い条件
    - 被験者の組…インドネシア人 (教示者) アメリカ人 (操作者)
    - スペイン人 (教示者) フィリピン人 (操作者)
    - 韓国人 (教示者) 中国人 (操作者)
    - 3 組…中国人 (教示者) 日本人 (操作者)

2.3 結果

- 手順
  - 実験者はそれぞれのペアにこの実験の目的が協力してできる限り高得点を得ることであることを説明した。
  - グループ１では、実験者はペアにローパスフィルターありとローパスフィルターなしの場合でローパスフィルターの影響をデモンストレーションした
  - お互いの画面が異なっていることには言及しなかった
  - 説明後実験開始
  - 10 分間のゲームプレイを２回行い、その間に３分間の休憩を挟んだ
  - 実験中はいずれの被験者ペアにおいてもパートナーとの接触・会話の機会は与えられず、実験中に教示者と操作者の役割は固定したままである
  - ペアのパフォーマンスを評価するために、操作者のラケットの動きとボールのヒットそれぞれに関して得点を割り当てられる。
    - 方向正答値(CDV)：各試行に教示者の指示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
    - ヒット値 (HV) ：各試行に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

- 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指示する方向へ動く確率は 0.5 と考えられ、平均方向正答値が 0.8 以上になる確率は P<.0547→平均方向正答値が 0.8 以上の場合には教示者の指示する方向を操作者は理解していると考えられる。
  - 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、これから平均ヒット率が 0.7 となる確率は p<.023 →平均ヒット率が 0.7 以上の場合には、教示から移動方向だけでなく、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。

- 終了直前 10 試行の平均方向正答と平均ヒット値で３つのカテゴリーに別けた
  - カテゴリー 1 ：平均方向正答値が 0.8 以下の場合 (教示者の指示を理解していない)
  - カテゴリー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 (教示者の指示を理解したが、ボールを当てることができなかった場合)
  - カテゴリー 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 (教示者の指示を理解したが、ボールを当てた場合)

- 項目　グループ 1 の被験者の行動
  - 11 組の被験者のうち 9 組が方向教示を獲得し、残りの２組が獲得できなかった
  - 成功した 9 組は、殆ど同じようなプロセスを経て方向教示を獲得した
    - 実験開始直後は、操作者に対して教示者はさまざまな種類の指示した (ディスプレイの真ん中まで動かせ、ちょっとだけ上に動かせ、動くな)(図３)
    - 成功しているペアは、教示の種類を減らした。「うえ」「した」のような指示しかしなくなった
    - 失敗したペアは教示の種類を減らしていなかった。(figure4)
    - 使用される教示の実際の音声、ピッチ曲線もともに弁別しやすいものへと変化した。 (figure3 右)

  - 方向教示獲得の方法として、次のような二つの方法が考えられる
    - 操作者が試行錯誤を通し偶然にボールを当てたことで、指示を理解した場合
    - 教示者が教示中に高いピッチの発話をすることで操作者に注意を与える方法
      - 「ボールに当たる 」、「高いピッチ発話」という教示獲得方法を用いることで、操作者は方向教示を獲得していた
  - 指示を理解することは教示者の努力だけでなく、操作者の努力も必要あった

- 項目　グループ２の被験者の行動
  - 6 ペアの 4 のペアが、教示者の指示を理解した
  - グループ 1 、２の成功したペアの間にはほとんど行動の差は無かった。
  - 教示者は操作者にとって未知の言語で教示を行っていたが、この場合に与えられる教示音声は、その音韻情報の意味を理解できなくても「聞こえ方」によって、はじめから分別可能であった

- 成功したペアは以下のような行動が観察された
- 教示者
  - １：実験中、操作者の学習モデルに適するように指示の種類を減らした
  - ２：警告韻律によって操作者を行為に集中させた
- 操作者
  - １：指示が与えられたときはいつでも、ラケットをある特定の支持の理解を示すために動かした
  - ２：異なったタイプの指示に従ってラケットを別様に動かした
  - ３：警告韻律に応じてラケットの動かし方を訂正した

- 実験の結果から以下のことが明らかになった。
  - ほとんどの操作者は教示者の適応を引き起こし、意味獲得プロセスにおいてのこれらの適応を使うことによって教示者の言葉による意図と意味を理解できる
  - 警告音律の特徴は操作者の正確な行動への警告として操作者には解釈され、意味獲得プロセスにおいて負の報酬情報として重大な役割を果たしている。

3. PROPOSAL OF A MEANING-ACQUISTION MODEL

3.1 要件

- 前節のコミュニケーション実験の結果を基に、口語コマンドの意味を理解できる「意味獲得モデル」を構築する
- 本モデルは次の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つける能力
  - 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

- これらの能力を実現するために、次のような学習モデルを想定した
  - このモデルは、入ってくる指示音が正確なラケットの動きを示し、指示の意味を理解することは入ってくる指示と適切なラケットの動きの間で適切なマッピングすることと同等だと想定する。
  - 自分の行動に対して正の報酬を受けたとき(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、自分の行動を指示していると認識する。
  - 逆に、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とき)、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
  - それぞれの指示音は８次元ベクトル(8種類の韻律的特徴(ピッチ、ゼロクロス数など))によって表す(図6、4)。加えてそれぞれの行動はスカラー値によって表す(絶対値は速度、プラスマイナスは右左)。

- 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想定している
- このモデルは、どのデータがどの正規分布から生成されたのかが分かれば各混合分布のパラメータ(平均μ、標準偏差σ)を求めることができる

3.2 方法

- 適切な変数を得るための基本的な方法として、EMアルゴリズムを使う
  - このアルゴリズムは、直接観察できない変数を支配する確率分布の一般的な形態が知られていれば、値が直接観察されない変数に対して使います
  - 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
    - １．E(評価)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 z_{ij} の期待値を推定

$$E[z_{ij}] = \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_j)^2\right) \Big/ \sum_{n=1}^{m} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_n)^2\right)$$

✧ ２．M (見積もり) ステップ

- Ｅステップで求めた隠れ値 Z_{ij} の期待値 $\sum |Z_{ij}|$ から、各分布のパラメータを推定

$$\mu_j \leftarrow \sum_{i=1}^{m} \left(E\left[Z_{ij} \right] \times X_i \right) \Big/ \sum_{i=1}^{m} E\left[Z_{ij} \right]$$

- ＥＭアルゴリズムは、負の報酬を受けたときは使えない
- ＥＭアルゴリズムを次のように拡張することで負の報酬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 (警告韻律)

    $Z_{ij} \leftarrow 0$
  - それ以外

    $Z_{ik} \leftarrow 1 / N - 1$

3.3 検証実験

- 意味獲得モデルの能力を実際の人間とのインタラクションを通し評価した
- 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、このモデルが適しているのかどうかを検討する。
- このモデルのパフォーマンスを得るために、教示者は違った指示をモデルに与えた
- 次の５種類の指示である。
  1. 「あ～」と発音しながら、高いトーンで上を意味し、低いトーンで下を意味する
  2. 「あ～」と発音しながら、長い音で上を意味し、短い音で下を意味する。
  3. 「うえ」で上を意味し、「した」トーンで下を意味する
  4. 「 up 」で上を意味し、「 down 」で下を意味する
  5. 1 の逆バージョン
- この検証実験の教示者は十分にトレーニングをし、このモデル構造について理解している
- 最も新しい１０個のデータを使い、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意味を理解したとした。
- 詳細なデータ (table3)
- ２～５分で、モデルは指示を理解した
- このモデルは先行知識がなくても、人間とのインタラクティブを通して与えられた指示の意味を理解できる能力がある

4. INTRAATION BETWEEN THE PROPOSED MODEL AND USERS

4.1 目的と背景

- このモデルが近い将来、実際に仕事でつかえるようなインターフェースシステムとして適応されるためには、ユーザが特別な知識がなくても指示を理解させなくてはならない
- 先行研究
  - 被験者は、モデルの知識がなくても口語コマンドでモデルに教示できるようになった(Komatsu2003)
  - 人間は、コンピュータのようなエージェントに対し、自然な会話することをためらう (Harada2002)

4.2 被験者

- 被験者を２つのグループ (グループA，グループ B) に分けた
- グループA： 10 人(9人日本人、一人フィリピン人、21~29歳、 6 人男、４人女)
  - 全て、コンピュータ科学、認知科学の学生
  - コンピュータの利用経験は十分にあり、プログラムもできる
  - 被験者に次の教示をした。「ゲームのラケットは学習コンピュータによって作動し、 あなたの仕事はことばによる指示の使用によってこのコンピュータにラケットの動きかたを教えることです」

- グループB： 10 人 (１０人日本人22~39歳、８人男、２人女)
  - 全て、コンピュータ科学、認知科学の学生
  - コンピュータの利用経験は十分にあり、プログラムもできる。
  - 被験者に次の教示をした。「人と話すように教えてください」

4.3 結果

- モデルが指示を理解したかどうかについての水準は、前の実験と同じ水準
- 警告韻律の割合は、全ての指示数における警告韻律で計算
- 先行実験では、成功したペアで警告韻律は約 5 %。
  - 警告韻律が 5 %であることは、人間と話しているかのように話すことを意味している
- 約 30 分ゲームをやってもらい、最新の 10 個のデータでＣＤＶ値が 0.8 で終了
- その後、この実験の印象とパートナーについてどう感じたかについてアンケートをとった。

- グループAの被験者の行動
  - 10 人中 6 人成功
  - 全ての被験者に共通の行動が見られた。教示の数や種類を変化させない
  - 一貫して２種類の教示 (上、下)
  - モデルが被験者の適応 (被験者の教示の種類を減らすこと) を引き起こす機会はなかった
  - モデルが意味獲得プロセスでの適応を使う機会はなかった
  - 警告韻律をほとんど使わなかった
  - 感情的でない教示をした

- 実験後のインタビューで、被験者がストレスを感じていた
- 被験者はコンピュータには感情表現を理解できないと考えていた
- ラケットが教示と逆に動いても、感情的な表現やクレームを言わなかった
- 成功者と失敗者間で、振る舞いの違いは殆どない
- モデルは被験者の指示に対し正確に反応するけれど、モデルと被験者との関係と人間同士関係に違いが出てくる

・ グループBの被験者の行動
  - 10 人中５人成功
  - 成功した人たちは、ほぼ共通の行動をした
  - 教示の数を減らし、警告韻律を使った
    - ✧ 警告韻律の割合が約３%
  - だれもストレスを感じる人はいなかった
  - グループBで成功した人たちは、グループAの人よりも自然な指示を使っていた
  - 被験者のM・N（ともに成功）R（失敗）はモデルに指示を教えようとしているわけでなく、モデルの指示の受け取り方を見つけていた
  - 失敗した被験者の行動は、先行実験で失敗していた人たちと同じ
    - ✧ さまざまな教示を使い、教示数を減らさなかった。警告韻律での指示を使わないまたは、使いすぎる傾向にある

・ 実験の結果から、意味獲得モデルは十分にユーザの指示の意図や意味を理解する能力を持っていることが示される
・ モデルは、ユーザの意味獲得プロセスを利用することによって口語の指示を理解できる
・ 実験者の「人と話すように教えてください」という教示は、実際被験者に人と話すよに指示をした

5. DISCUSION

5.1 この意味獲得モデルの適用可能性

・ 適切な指示を与えると、意味獲得モデルは口語の意味を学びユーザが意図したようにふるまう
・ モデルと被験者は部分的な相互適応プロセスを形づくる
・ この結果を元に、パソコンやカーナビにおいて、ハンドフリースピーチへ応用
・ 関連技術
  - イントネーションを上げることによって画面が上げるインターフェース (Goto,Ito,Akiba,&Hayamizu,2001;Tsukahara&Ward,2001)
  - “上げろあ～ (Move up ahhh) ”のような口語コマンドをユーザが使用できるようにした
  - “上げろ (Move up) ”というコマンドによって画面が上がり、“あ～ (ahhh) ”の長さに語尾変化よって上がる程度が変わる。 (Igarashi and Hughes(2001))

・ この意味獲得モデルは、Igarashi ,Hughesの研究よりもパーソナライゼーション化・オートカスタマイズゼーション化という点でよい研究を提案している

- 感情表現 (警告韻律“ 負の報酬 ”) による指示を通し相互作用によって学ぶ
- ユーザが言語コマンドやスクロールの意味を個別にカスタマイズ化できる柔軟性を提供している
- ユーザがコンピュータの予期しない反応に対し、起こった声で不平をいうとモデルは負の報酬として学習する

5.2 モデルの不自然反応

- このモデルは少なくとも２つとも未解決の問題がある
  １：モデルの不自然な対応
  ２：多機能インターフェースとして実際利用のための拡張性

- この章では１つ目について言及する
- グループ A,B とも教示に成功している被験者にもかかわらず、モデルの不自然で、予測できない動きだと感じる
- それらの状態を２つのケースに別ける
- ケース 1 ：突然の策略の変化に対し対応ができない
  - 実験の最初、典型的な教示者は上の方へ動いて欲しい時は「うえ」、 下に動いてほしいときは「した」と言ったような教示を使う
  - 操作者のパフォーマンスが 2~3 分で指示に従わない時は、教示者は「うえ～」「した～」と言った教示を使う
  - 人間が教示者なら直ぐに教示を理解できるが、モデルはできない。

- ケース２：一つの方向にしか動かない
  - 被験者はさまざまな種類の教示をするのに一方方向にしか動かない
  - EM アルゴリズムがこの２つのケースを引き起こしているのではないかと思われる

- ケース１では、実例の数の問題である。
  - ユーザが教示戦略を変化させモデルに予期しない指示を与えると、モデルはどのように処理をして良いか分からない。
- ケース２では、局所的最小値
- 問題 (local minimum problem) である。
  - 統計上の学習特徴は、モデルに不自然な振る舞いを与える影響だと言われている
  - この問題を解決するために、補足的な仕組み、つまり混成の学習システムを導入しないといけない
  - この統計学の学習の仕組みだけでなく、他の種類の学習の仕組みである

5.3 実際利用のための拡張性

- このモデルはポンゲームと言った単純な環境での検証だけで、日常の多機能インターフェースでの適応に関しては確かではない
- 多機能インターフェースでは、ユーザはさまざまな種類の指示を使う
- 次のような障害が予見される。
- １：モデルが副詞的な指示や評価指示を理解できないこと

- 3.1 章で言及したように、このモデルは“ある指示がある確かな行動を示している”と言う想定を基に造られている
- “上げろ”“下げろ”といったような指示の意味だけは理解できるが、“ちょっと(a bit)”“いいよ (good) ”のような指示は利用できない。

・ ２：モデルが理解できる指示が限られていること
- モデルが理解できる指示の数は、EM アルゴリズムでの正規分布の混合によってできた分布の数と同じ
- 教示者はモデルが理解している指示よりも多くの指示を使う。

・ この意味獲得モデルは、近い将来多目的のインターフェースとして応用できる十分な能力を持ち合わせていない
・ 利用拡張のためには、この学習モデルが日常の多様なアプリケーションに適応される必要があるだろう
・ そのためには、少なくとも二つの補完的な機能が必要である。
- 分布数の問題を克服する機能である
  - この機能で克服するためには、正規分布の混成での分布数を自動的にカスタマイズ化する能力は、ユーザの指示パターンによって構築される必要がある
- 機能は副詞的な指示や評価指示を理解できるようにすることである
  - Suzuki などによる方法はエージェントに副詞的な指示や評価指示を理解されるが可能がある
  - もしこの方法を学習モデルに組み入れることができるなら、モデルは副詞的な指示や評価指示を理解できるようになるだろう。
  - それによって、モデルに新しい報酬情報を獲得させることができる。
    - たとえば、“よい(good)”という口語指示は正の報酬となるだろう

6. CONCLUSION

・ ペットロボットや適応的スピーチインターフェースといったインタラクティブ性のあるエージェントがユーザとの相互適応的処理が必要であるときには、エージェントは次の２つの能力を備え付けなければならない。
- パラ言語情報において表現された報酬情報を理解する能力
- 報酬情報の使用による人間について学習するシステム

・ この研究の目的は報酬の特別な内容と２人がどのように犬と主人との間のようなスムーズなコミュニケーションをするかを観察することによって学習システムの仕組みを明らかにすること
・ コミュニケーション実験では、人間同士のコミュニケーションの構築を評価する処理を観察した
・ それを基に、意味獲得モデルを構築
・ そのモデルがユーザの口語コマンドの意図や意味を理解できることを確認
・ そのモデルを使い、特別な知識のないユーザが指示をどうモデルに理解させたかの実験
・ 結果から被験者の適応行動は警告韻律を使うことと指示数を減らすこと
・ データを処理したがこの学習仕組みから問題を明らかに出来なかった。

- **モデルの不自然な振る舞い**
- **将来に向けての拡張性**

・ **この意味獲得モデルを改良し、機械とユーザとの間に本当の相互適応を形作り、スピーチインターフェースの自動カスタマイズゼーション化・パーソナライゼーション化の基礎的な技術となり、本当の犬と主人のような関係を作れるようなペットロボットのためのインターフェースの発展になる**