社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

乳幼児の音声模倣能力の獲得過程における調音ジェスチャの役割

金野　武司†　　錦戸　信和††　　党　　建武††

† 北陸先端科学技術大学院大学　知識科学研究科　〒 923–1292 石川県能美市旭台 1–1
†† 北陸先端科学技術大学院大学　情報科学研究科　〒 923–1292 石川県能美市旭台 1–1
E-mail: †{t-konno,a-nishi,jdang}@jaist.ac.jp

あらまし　乳幼児が言語的な発声を獲得する過程には，そもそも「あ」を/a/であると認識する知識をどのように獲得するのかという疑問がある。本論では，乳幼児の発声と大人の発声が音響特性的にかなり異なることから，音響特性によって親の/a/と乳幼児 (自分) の/a/が同じであると判断できない場合であっても，乳幼児が音声模倣を獲得できるのかを問い掛ける。この問いに対して，本論では音響特性での同一性判断に代わって，親が返す真似行動を手掛かりにした音声模倣の獲得モデルを構築し，計算機シミュレーションと，ヒトとのインタラクション実験によってその動作を確認した。結果として，計算モデルにおける乳幼児エージェントは，親がいつでも真似行動を返さない場合であっても，真似しようとする目標状態を内部に維持することによって，音声模倣を獲得できる可能性があることを確認した。この結果に基づいて，乳幼児は，獲得した音声模倣時の調音ジェスチャの同一性を手掛かりに，複数の話者に対する同一性認識の形成を可能にするのではないかということを議論する。
キーワード　音声模倣，音声知覚の運動理論，調音ジェスチャ，乳幼児の発達過程，調音モデル

Role of Articulatory Gesture in Acquisition Process of Infants Vowel Imitation

Takeshi KONNO†, Akikazu NISHIKIDO††, and Jianwu DANG††

† School of Knowledge Science, Japan Advanced Institute of Science and Technology(JAIST)
†† School of Information Science, Japan Advanced Institute of Science and Technology(JAIST),
Asahidai 1–1, Nomi-Si, Ishikawa, 923–1292 Japan
E-mail: †{t-konno,a-nishi,jdang}@jaist.ac.jp

Abstract　How can infants acquire the ability to identify his/her articulation of /a/ as the same category of adult articulation of /a/ in spite of differences of acoustic characters during developmental processes of infants. In this paper, we question whether or not the infants can acquire the ability of vocal imitation even if the infants cannot judge the vowels as the same category. We construct an acquisition model of the vocal imitation based on adults' imitative actions instead of the judgments on acoustic characters. We confirmed that the model can acquire a possibility of the vocal imitation both by computer simulation and by interactive experiments with human. As a result, if an infant agent of the model has a purpose of imitation as an internal state, the agent can acquire the ability of the vocal imitation although the adults do not always response with imitative voices. Based on these results, we discuss a possibility of acquisition how the infants can judge vowels as the same category using a similarity of articulatory gestures at the vocal imitation.
Key words　Vowel imitation, Motor theory of speech perception, Articulatory gesture, Developmental process of infants, Three dimensional articulation model.

1. 背　　景

乳幼児の発声と大人の発声は音響特性的にかなり異なるにも関わらず，乳幼児は親から聴いた「あ」に対して，同じく「あ」と発声する調音能力を獲得する (本論では，言語的な音声模倣能力の獲得によって形成される，聴いた音と調音の対応関係を音声言語カテゴリと呼ぶ)。多くの非言語的な発声も可能な乳幼児が，言語的な発声を獲得する過程には，そもそも「あ」を/a/

であると認識する知識をどのように獲得するのかという疑問がある。

乳幼児が聴いた声を真似るとき，そこには音響的な特徴を抽出して，それを脳内に蓄えられたパターンと照合する仕組みを考えることができる [1]。このとき，もし特徴抽出において生得的に音声言語カテゴリの識別が可能であるとするならば，真似るための調音ジェスチャ(音素に固有な調音動作の特徴 [2]) を，その識別能力に従って調整すれば良いことになる。

ところが柏野 [1] は，これまでの研究では音素に一対一に対応するような音響的な特徴 (音響的不変量) が見つからないことを指摘している。そうだとすれば，音声を知覚する特徴抽出の機能が音声言語カテゴリを担うとは考えにくくなる。

この問題に対する 1 つの回答として，Liberman [3] は，調音ジェスチャを参照しながら音声を知覚するのではないかとする仮説 (音声知覚の運動理論) を提唱している。この仮説を支持する研究として，母音の音響的カテゴリが，発声に関わる筋電特性と相関関係をもつことを明らかにした党らの研究 [4] や，母音のホルマント周波数が，聴覚の周波数分解能から規定されるのではなく，調音ジェスチャの変化に対するホルマント周波数の変化量と相関することを見い出した廣谷らの研究 [5] がある。これらの研究は，音声言語カテゴリが調音ジェスチャを基に形成されているのではないかということを示唆する。しかし，この考えを支持するには，そもそも調音ジェスチャに基づく音声言語カテゴリが，特徴抽出の機能に依らずに形成できるのか，という疑問に答える必要がある。

本論では，この疑問に答える鍵として，認知発達心理学で指摘されている随伴関係の強化傾向に着目する。乳幼児はその発達過程において，自分の行動に対して時間的に接近して起こる事象を，自分の行動の結果として記憶する傾向があると言われている [6], [7] (注1)。この傾向に基づいた学習によって，子どもは，親の/a/に対して親が/a/と認められるような/a/を返すことができる調音ジェスチャを獲得するのではないかと考える。

この考えの検証には，2 つのアプローチが考えられる。1 つは，実際に乳幼児の内部で何が起こっているのかを調べる方法である。これは，問題とする現象 (音声模倣) を直接観察し，その神経的基盤を調べていく方法である。近年では fMRI のような脳活動のイメージング機器によって，よりヒトの内部に踏み込んだ調査ができるようになってきている。しかし，乳幼児の内部状態の調査はまだ多くの困難を抱えており，脳機能の局在性が明らかになりつつあっても，聴覚から調音までの経路を包括的に捉えることは難しい。他方，随伴関係の強化傾向という仮定の下で，音声模倣を学習によって獲得する計算モデルを構築し，その計算モデルが組み込まれたシステムとヒトとのインタラクションを観察することで，仮定として妥当であるような随伴関係の強化傾向を調べる方法が考えられる。これは，人工的なシステムを作って動かすことによって，対象とする現象を理解しようとする方法である。この方法は，問題とする現象を直接扱っていないことを根本的な問題として抱えながらも，抽象化された計算モデルを試行錯誤的にテストできることから，音声模倣に関わる本質的な能力を見つけ出すことに貢献できると考えられる。

そこで本論では，随伴関係の強化傾向を，自分が発する声に対して親から返される声を因果関係として強化する傾向として仮定し，特徴抽出での音響特性による判別を必要とせずに，音声言語カテゴリを学習によって獲得する計算モデルの構築を試みる。この構築を通じて，音声模倣を実現する計算モデルに，どのような能力の仮定が要求されるのかを明らかにすることを目的とする。

本論の構成は以下の通りである。まず，2. 節において，仮定する随伴関係の強化傾向が具体的にどのような能力として実現可能であるのかを探るための抽象的な計算モデルを構成し，コンピュータシミュレーションによって，その基本的な学習過程のメカニズムを確認する。3. 節において，その計算モデルをヒトとインタラクション可能なシステムに実装し，音声模倣の可否と，獲得される音声言語カテゴリを調査する。4. 節において，得られる音声言語カテゴリに関する議論を進め，5. 節において，本論から得られる結論を述べる。

2. 音声模倣の学習モデル

ここでは，音声模倣のために用意する相互発声環境 (2.1 節) と構築する学習モデル (2.2 節) を説明し，音声模倣の学習過程がどのように推移するのかを示す (2.3 節)。

2.1 相互発声環境の抽象化

乳幼児が，聴いた声を真似る環境として，乳幼児と親が向き合って声を出し合う状況を想定する。計算モデルにおいては，親と子はそれぞれをエージェントとして単純化する。親エージェントは子エージェントが発した声を聴き，同一カテゴリの声を返す。子エージェントは自分が発した声を聴き，次いで親の発した声を聴く (図 1) (注2)。また，本論で扱う音声言語カテゴリは，日本語の 5 母音 (/a/,/i/,/u/,/e/,/o/) とそれ以外 (/other/) とする。

2.2 子エージェントの学習モデル

子エージェントは**特徴抽出**，**記憶**，**構音**，**評価**の 4 つのモジュールで構成する (図 2)。

特徴抽出は，聴いた声 (/a/,/i/,/u/,/e/,/o/,/other/) を特定の番号 (={1,..,6}) に割り当てる。今，子エージェントは特徴抽出において，自分の発した/a/と，親の発した/a/が同じであることを判断できないと仮定するので，親の発する声には 7 から 12 までの番号を順に割り当てる。これにより，特徴抽出は，

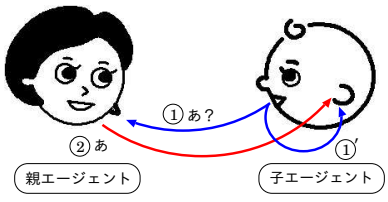


図 1　相互発声環境
Fig. 1　Talking environment.

(注1)：ここで興味深いのは，自分の行動と時間的に接近して起こる事象の間に物理的な結び付きがない場合にも，関係を強化してしまう傾向があることである。本論はその傾向に着目する。

(注2)：ここにはターンテイキングの問題が潜んでいるが，対象とする問題を明確にするため，本論では交互に声を発することを仮定して，ターンテイキングの問題には触れない。

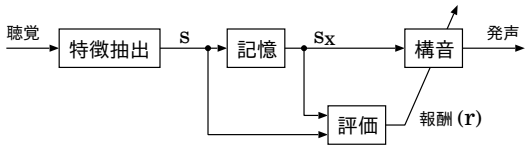


図 2　子エージェントのシステムブロック図
Fig. 2　System block diagram of infant agent.

1 から 12 までの番号を s に出力する。

記憶は，特徴抽出から受け取る番号 (s) を保存することによって，親の/a/,/i/,/u/,/e/,/o/(=$\{7,..,11\}$) をレパートリに持つことを仮定し，その中から 1 つをランダムに選び出力する (s_X)。

構音は，記憶の出力を受け取る。受け取る声番号には N_a 個の発声選択肢を割り当て，それぞれの選択肢には一様乱数によって/a/,/i/,/u/,/e/,/o/,/other/の音素 (=$\{1,..,6\}$) を割り当てる。子エージェントは，受け取った声番号に対して，発声選択肢の中から発声音素を選択する。選択する発声音素を試行錯誤的に学習する仕組みには，強化学習 (特に Q-Learning) [8] の考え方を用いる。強化学習は，内部状態に対する行動に価値を割り当てて，ある行動を起こしたときに得られる報酬によって，その行動の価値を更新する考え方である。内部状態を記憶からの出力である s_X とし，N_a 個から選択される発声番号を行動 a として，その行動の価値を実数値 $Q(s_X, a)$ で表現したとき，構音で選択される発声を，次式で確率的に決定する。

$$p(a|s_X) = \frac{Q(s_X, a)}{\sum_{a'=1}^{N_a} Q(s_X, a')}. \qquad (1)$$

評価は，記憶の出力状態 (s_X) と，逐次聴くことになる声番号 (s) との一致／不一致を次式で判断し(注3)，それを報酬 (r) として構音に伝える。

$$r = \begin{cases} R & \text{if} \quad s = s_X \\ -R & \text{otherwise} \end{cases} \qquad (2)$$

ここで，R は報酬定数として実験の中で任意に設定する。価値 (Q) は，評価から伝えられる報酬 (r) によって次式で更新する。

$$Q_{t+1}(s_X, a) = Q_t(s_X, a) + r, \qquad (3)$$

2.3　子エージェントの学習過程

図 3 に，音声模倣の成功比率の推移を示す。発声選択肢の数 (N_a) は 30 (注4)，報酬定数 (R) は 0.025 である。図は，親エージェントと 4000 回インタラクションしたときの，10 回毎の成功比率の推移を 1 セットとし，これを 100 セット実施したときの平均値である。価値分布は 1 セット毎に一様乱数 ($[0, 1]$) で初期化し，発声選択肢に割り当てる音声も一様乱数で再設定する(注5)。

(注3)：ここで仮定するのは，記憶した親の声と，もう一度聴いた親の声が同じかどうかを判断できることだけであり，親の声と自分の声が同じであることは判断できないと仮定する。
(注4)：発声選択肢を複数用意しても，1 つの音素 (例えば/a/) が割り振られたそれぞれの選択肢には違いがない。このため，選択肢を複数用意するのは，主にインタラクション回数への影響を調べることが目的になる。
(注5)：発声選択肢には，1 セット毎に一様乱数で/a/,/i/,/u/,/e/,/o/,/other/が割り当てられており，記憶が出力する状態 ($s_X = \{7, .., 11\}$) それぞれに価値分布 ($Q(s_X, a)$) を持つ。

一回のインタラクションでは，子どもは，親の声と自分の声を評価する。自分の声の評価結果は，親の声との一致を判断できないように設定しているので常に負の報酬となるが，図 3 は，それでも音声模倣が学習できることを示している。これが可能なのは，子が発しようとした声 (記憶の出力状態 (s_X)) と親から帰ってきた声が一致する場合だけ報酬 (r) が $R + (-R) = 0$ で，それ以外ではマイナスの報酬になり，結果として，声が一致する場合の価値が相対的に高くなるからである。親の間違い率 (子どもの声を聴いたとき，同一のカテゴリではない声を返す確率) を高くしたとき，親が 60%程度間違えて返しても，音声模倣は 90%程度の成功率にまで上昇する。これは，親が間違えても，子エージェントが自身の記憶が出力した状態 (s_X) に対して評価を下していることで，間違いの影響が小さくなるためである。例えば，親の間違い率が 60%の場合は，子エージェントが s_X に/a/= 7 を選択する確率は，記憶の持つレパートリが/a/,/i/,/u/,/e/,/o/の 5 つであることから 1/5 なので，成功確率は $1 - 0.2 \times 0.6 = 0.88$ となる。逆に，自分が発した声 (a) に対して，親から返ってきた声 (s) で無条件に価値付けを行なう ($Q_{t+1}(s, a) = Q_t(s, a) + R$ で更新する) と，親の間違い率は，直接音声模倣の成功率に反映するようになる。

親が 60%程度を間違って返す場合の，記憶が出力する親の発声/a/($s_X = 7$) に関する価値分布の変化は図 4 のようになる。図 4 は，発声選択肢にある/a/〜/o/それぞれの発声選択頻度の推移を示したものである。図中に同時に表示した成功比率の推移と比較すると，/a/の選択確率の上昇と共に，成功比率も上昇することが見て取れる。

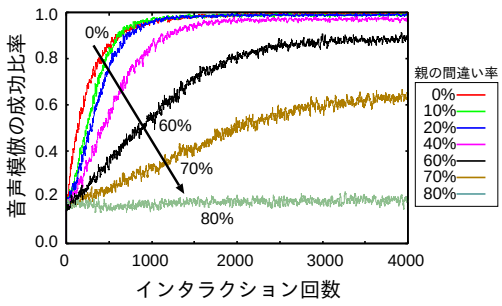


図 3　学習曲線 (音声模倣の成功比率の推移)
Fig. 3　Learning curves: success ratio of vowel imitation.

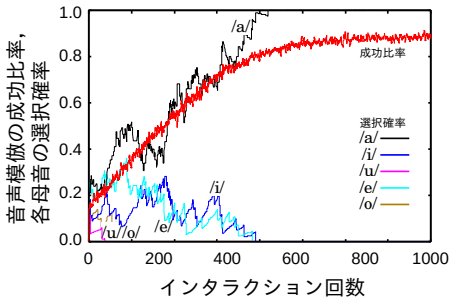


図 4　/a/に関する音声模倣の成功比率と，母音毎の選択確率の推移
Fig. 4　Learning curves for /a/: success ratio of vowel imitation and probability of vowel selection.

3. ヒトとのインタラクション

前節で構成した計算モデルでは，子エージェントが選択する音素 (例えば/a/) には 1 つの番号 (/a/=1) を割り当てた。そのため，発声選択肢の中に同一音素が複数あっても，それらの音素間には違いがない。しかし実際には，乳幼児は同じ音素カテゴリであっても，音響特性の異なる声を発することができる。ここでは，音響特性の異なる声が発声選択肢の中に複数存在する状況で，子エージェントが音声模倣を行なえる発声価値分布を学習できるかどうかを確認する。このために，図 2 の構音部を，3 次元声道モデル (以降，調音モデルとする) を用いた音声合成システム [9] に変更し，発声選択の価値分布がどのように形成されるのかを調査する。

3.1 節に，子エージェントの発声選択肢となる音素を調音モデルでどのように用意するのかを説明し，3.2 節で，その学習過程を示す。

3.1 調音モデルによる音声合成

図 5 の調音モデルを用いて，300[msec] の音声を約 13 万個合成し，ホルマント周波数で 5 母音にカテゴライズされる音素をピックアップする [10]。これらの音声の音響特性は，筋肉に収縮力を設定して変形させた声道形状に基づいて算出されており，調音ジェスチャとして表現される特徴パラメータが非常に多い。特徴パラメータを舌尖と舌背の水平および垂直位置に絞るため，以下の操作を行なう。

ピックアップされた音素の声道形状を目標到達位置に設定し，初期位置 (ニュートラルポジション) から声道形状を変形させたときの音声を合成する。設定する目標位置は，調音モデルの正中矢状断面 (x-y 断面) での下顎，舌尖，舌背の水平および垂直位置と，2 つの口唇開口面積である。下顎と口唇開口面積は，ピックアップした音声を合成する声道形状の平均値に設定する。舌尖と舌背は，それぞれの到達位置が分布する領域を 100 個の正方区画に分割すると，設定位置が入る区画の舌尖と舌背の組み合わせが 852 個得られる。この 852 個の区画それぞれに存在する位置データの平均位置を算出し，これを舌尖と舌背の到達目標位置とする (図 6)。到達位置を設定して合成した音声を著者が聴き，明らかに非言語音である 284 個の音声を除いた 568 個の音声を，子エージェントの発声選択可能な音素とする。参考として，表.1 に 568 個の音声の母音カテゴリの音素数を，図 7 にホルマント周波数と調音ジェスチャの分布を示す(注6)。

ホルマント周波数は，合成音声の線形予測分析 (Linear Predictive Coding:LPC) を行ない，得られる全極型フィルタの分母多項式の根から求めた。300[msec] 間で前後 10%を削った 240[msec] の音声データに対してホルマント周波数を算出し，その平均値を 1 つの音声のホルマント周波数とする。FFT および LPC の分析条件を表 2 に示す。

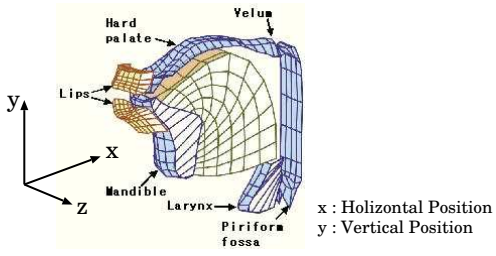


図 5　3 次元声道モデル

Fig. 5　Three Dimensional Articulation Model.

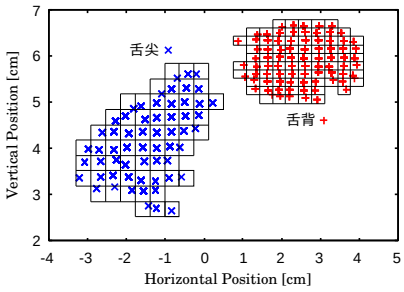


図 6　舌尖と舌背の到達目標設定位置

Fig. 6　Setting of target position of tongue tip and dorsum.

3.2 子エージェントの学習過程

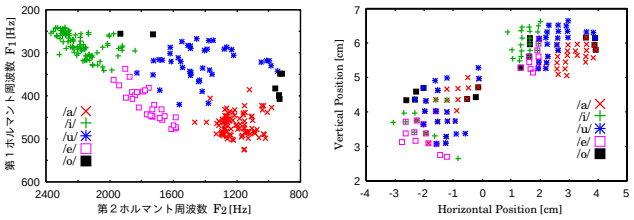
前節で用意した 568 個の音素を子エージェントの発声選択肢に設定し，図 8 に示すインターフェイスを用いて，インタラクション実験を行なう。被験者は子エージェントが選択した音素をイヤフォンから聴き，聴いた音が属すると考えるカテゴリのボタンを押す。Replay ボタンによって，再生回数は任意とするが，後戻りは不可とした。また，Replay ボタンによって同じ音素を何度再生しても，インタラクション回数は 1 回とする。

実験は 5 人の成人男性被験者に対して実施した。特徴的な結果として，学習がうまくいった被験者 A と，あまりうまくいかなかった被験者 B の，2 人の被験者に関する音声模倣の成功比率の推移を図 9 に，発声選択番号の推移を図 10 に示す。 被験者 A(図 9 左) とのインタラクションでは，子エージェントの音

表 1　5 母音の音素数

Table 1　Number of vowels.

母音	/a/	/i/	/u/	/e/	/o/	/other/
音素数	96	94	41	28	10	299



音響特性　　調音ジェスチャ

図 7　合成音素の音響特性と調音ジェスチャ

Fig. 7　Vowels formant frequencies and articulatory gestures of synthesis voices.

図 8　実験用インターフェイス

Fig. 8　Interaction interface for experiments.

(注6)：音素のカテゴリ判断は非言語音の判断と同時に著者が行なった。

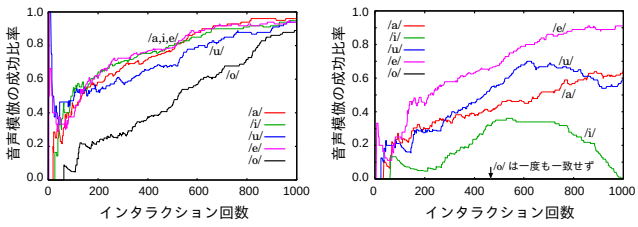
表 2　合成音声の分析条件

Table 2　Analysis condition of synthetic voice.

合成音声	:	サンプリング周波数 16[kHz]
FFT	:	窓長 20[msec]，シフト長 4[msec]
LPC	:	線形予測係数の次数 18

声模倣が実現しているが，被験者 B(図 9 右) とのインタラクションでは，特に/i/と/o/に関して，音声模倣が実現していない。これを反映するように，被験者 B に対して子エージェントが選択する音声番号は，/i/に関してはインタラクション回数の 300 から 800 ステップでは収束しつつも最後には拡散し (図 10 右 2 段目)，/o/に関してはインタラクション回数の全体に渡って広範囲に分布している (図 10 右下段)。

被験者 A のように，音声模倣が実現する場合がある一方で，被験者 B のように，特定の音素に関して，音声模倣が実現しない場合のあることが確認された。/o/に関しては，用意した音素の中で/o/であろう音素が極端に少ないことが影響していると考えられる (表.1)。/i/に関して，被験者 B がインタラクションの最後に/i/ではないと判断するようになったのは，本人への聞き取り調査によれば，単音を何度も聴いているうちに，そもそも聴いている声がどのカテゴリに入るのかが分からなくなってきたせいではないかということであった。対して，被験者 A は，単音での判断はあまり重視せず，前の音との比較によってリズム良く判断したことを報告しているのは興味深い。



被験者 A　　被験者 B

図 9　音声模倣の成功比率の推移

Fig. 9　Learning curves of success ratio of vowels imitation.

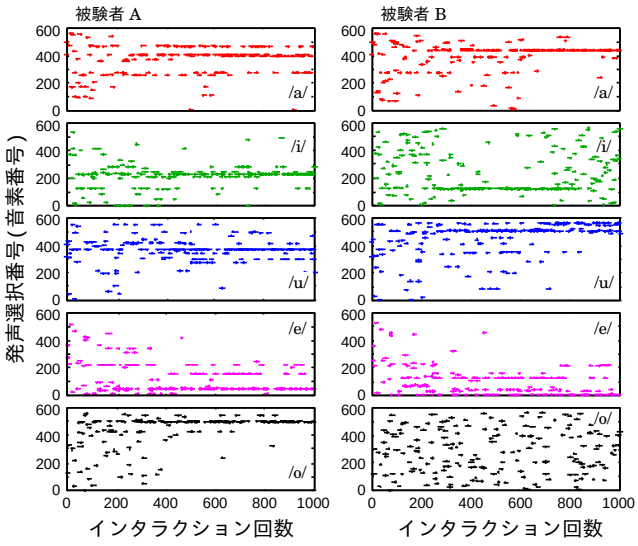


図 10　発声選択番号の推移

Fig. 10　Selective number of vowels to times of interaction.

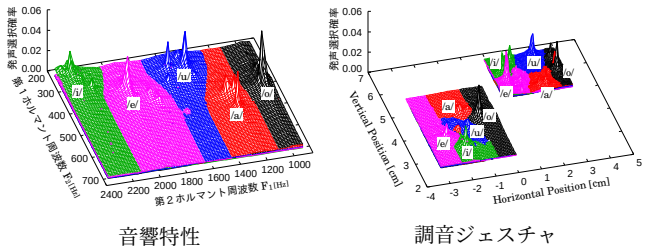
音声模倣が実現した被験者 A の発声価値分布は，音響特性，調音ジェスチャのそれぞれで図 11 のようになる。この図は，ホルマント空間と位置空間を微少区間に分割し，その微少区間内の音響特性あるいは調音ジェスチャを持った音素の選択確率を平均化して表示したものである。

音響特性に基づいて学習したのではないにも関わらず，標準的な 5 母音のホルマント周波数に価値のピークが現われている。これは，被験者の持っている 5 母音それぞれのカテゴリが，子エージェントの発声価値分布に取り込まれた結果である。このことから，逆に子エージェントは，この発声価値分布を使って，5 母音それぞれのカテゴリを判断することが可能になるのではないかと考えられる。

被験者 5 人の音声模倣の成功比率の推移を図 12 に示す。図 12 は，母音カテゴリ毎に 5 人の平均値を算出したものである。/o/の成功比率が少ないものの，それ以外の音素に関しては，子エージェントは音声模倣の成功比率を高められることが確認された。

4. 議　　論

乳幼児の発達過程において観察される音声模倣の基本的なフレームワークは，親の声を聴いて，それを真似しようとして子が声を発する，というものだろう。真似しようとして，という部分が，本論の学習モデルでは，図 2 で記憶が出力する番号 (s_X) に相当する。記憶の出力番号から，子エージェントは価値分布に基づいて声を発する。ヒトとのインタラクション実験においては，子エージェントが発した声を被験者が聴いたとき，被験者はその声が属すると判断するカテゴリのボタンを押す。子エージェントは，押されたボタンの番号を被験者の声と判断することはできるが，自分の声と被験者の声が同じであるとは判断できない (2. 2 節：学習モデルの仮定)。このように，音響



音響特性　　調音ジェスチャ

図 11　被験者 A の発声価値分布

Fig. 11　Value distibution, Q, of subject A.

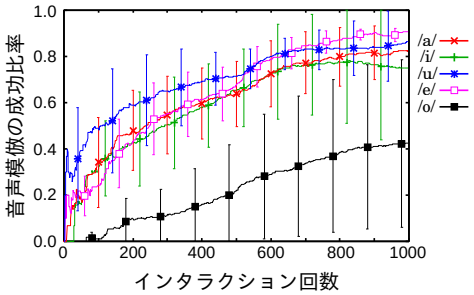


図 12　被験者 5 人の音声模倣の平均成功比率の推移

Fig. 12　Average success ratio of vowel imitation for five subjects to times of interaction.

特性において，被験者 (親) の声と自分の声の一致を判断できないと仮定したとき，被験者の持つ母音カテゴリを，子エージェントが学習によって獲得することができるのかを問うことが，本論の主題である。

結果として，音響特性による判別に代わって，自分の声に対して返される親の随伴的発声を評価の基準に置けば，子エージェントは，被験者が同じ母音カテゴリに属すると判断できるだろう発声選択の価値分布を学習することが確認できた。シミュレーションにおいては，自分の声と親の声の不一致に負の評価を与えたとしても，音声模倣の学習は可能であり，さらには，子エージェントの発声に，親エージェントが 60%程度の確率で異なるカテゴリの声を返しても，真似しようとした親エージェントの声 (s_X) に対して，子エージェントは，親エージェントが同じカテゴリであると認識できる声を 90%程度の確率で発声できる価値分布を形成することが確認できた。この結果から，親がいつでも真似を返すことがない状況においても，子が記憶の出力状態 (s_X) を維持していれば，音声模倣を学習によって獲得できる可能性が示唆される。

ただし，計算モデルでは，親の 5 母音を再現目標とすることが予め決まっているので，子エージェントが形成するカテゴリも 5 母音に固定されている。聴いた声によって，子エージェントの再現しようとする目標が決まる過程を考えると，一人の人間の/a/は，声を発する状態や環境条件によって異なる母音に識別される場合があるだろうし，被験者 A の/a/と被験者 B の/a/も，当然それぞれに異なる可能性があるだろう。聴いた声が子エージェントの内部で全て異なる状態に区別されるとき，子エージェントは，それぞれの状態に価値分布を割り当てて学習することになる。この状態で，子エージェントはそれぞれの/a/が同じであることをどのように知ることができるだろうか。

音響特性での同一化判断は行なわれないという仮定の下での 1 つの考え方として，異なる識別状態から選択される音声の調音ジェスチャが類似性を持つことを手掛かりに，それぞれの音声の同一性を判断できるという機構が有り得るのではないだろうか (図 13)。被験者 A の/a_1/と，被験者 B の/a_2/が同じ調音ジェスチャの音声を選択する価値分布を持っているとき，両者を同じカテゴリの声であると認識するのである。こういった方法での音声言語カテゴリの形成過程を支持するのが，調音ジェスチャを基にした音声知覚の運動理論 [3] であると考えられる。

5. 結　　論

本論では，乳幼児の発声と親の発声が音響特性的にかなり異なるにも関わらず，乳幼児が同じカテゴリの声として親との音声模倣を成立させていく過程に注目した。特に，同じカテゴリの声であることを音響特性によっては判断できないと仮定した場合に，親の声を真似る行動が獲得可能かどうかを問い掛けた。この問いに対して，本論は随伴関係の強化傾向を仮定した計算モデルを構成することで，その可能性を検討した。構築した計算モデルは，親とのインタラクションによる学習過程を持ち，随伴的に返される親の声に基づいた選択音声 (調音) の価値修正を行なう。

この計算モデルによるコンピュータシミュレーションと，ヒトとのインタラクション実験において，カテゴリが曖昧な子エージェントの発声に対して，親エージェント (ヒト) が，その声が属するカテゴリの声を返すことで，子エージェントは音声模倣を可能にする音素の選択価値分布を学習することが確認された。また，コンピュータシミュレーションにおいては，子が真似しようとする声に基づいて親の声を判断していれば，親がいつでも真似を返すことがない状況にあっても，音声模倣を獲得できる可能性が示唆された。

音響特性でのカテゴリ判断ができないという仮定の下での，複数の話者に対する共通認識を形成する問題に関しては，学習された価値分布における調音ジェスチャの同一性によって，カテゴリの同一性を形成していくことが可能なのではないかということを議論した。我々は，他人の「あ」と自分の「あ」は違うけれど，同時に同じ「あ」であることを知っている。違う状態として特徴抽出される音響特性に，調音ジェスチャの同一性に基づいて同じであるというラベルを貼ることで，受け取る感覚状態は異なるけれども，それらは同じであるという認識状態が実現されているのではないだろうか。

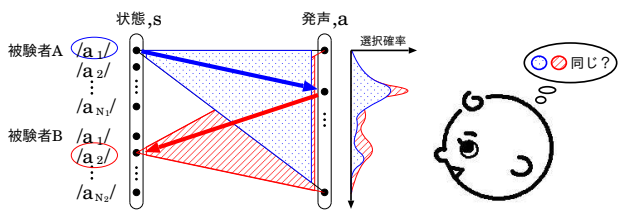


図 13　調音ジェスチャによる共通認識の形成

Fig. 13　Cognition of common category by articulatory gestures.

文　　献

[1] 柏野 牧夫, “音声知覚の運動理論をめぐって,” 日本音響学会誌,vol.62,5,pp.391-396,(2006).

[2] 誉田 雅彰, “人に迫る発話工学 —人の発話をいかに観測し真似るか—,” 日本音響学会誌,vol.55,11,pp.777-782,(1999).

[3] Liberman, A., and Mattingly, I., “The motor theory of speech perception revised,” Cognition,vol.21,pp.1-36,(1985).

[4] Dang, J., Akagi, M., and Honda, K., “Communication between speech production and perception within the brain - Obseravation and Simulation,” J.Comput. Sci. & Technol.,Vol.21,pp.95-105,(2006)

[5] Hiroya, S., Mochida, T., and Kashino, M., “Articulatory gestures, not audiotory frequency resolution, determine formant frequency discrimination thresholds in vowels.” The 29th MidWinter Meeting on Association for Research in Otolaryngology(ARO),Vol.29,p.249(2006)

[6] 正高 信男, “言語音声の獲得,” 脳科学大事典,pp.225-228,(2000).

[7] Premack, D., Premack, A., Original intelligence : Unlocking the mystery of who we are, McGraw-Hill, New York,(2003), (訳書:心の発生と進化, 長谷川 寿一 監修, 新曜社,(2005)).

[8] Sutton,R.S. and Barto,A.G., Reinforcement Learning, A Bradford Book, MIT Press,cambridge, MA,(1998), (訳書:強化学習, 三上 貞芳, 皆川 雅章共訳, 森北出版,(2000)).

[9] Dang,J., and Honda,K., “Construction and control of a physiological articulatory model,” Journal of Acoustical Society of America,vol.115,2,pp.853-870,(2004).

[10] 錦戸 信和, 党 建武, “日本語 5 母音の調音・音響的観測とモデルシミュレーションとの比較,” 電子情報通信学会技術研究報告 SP,Vol.106,No.177,pp.5-10,(2006).