B-07

相互模倣を通じた相手との声域の対応付けと母音獲得

Matching Voice Region and Acquiring Vowel Categories through Mutual Imitaion

石原 尚 † 吉川 雄一郎 ‡ 三浦 勝司 ‡† 浅田 稔 ‡†

Hisashi Ishihara Yuichiro Yoshikawa Katsushi Miura Minoru Asada

† 大阪大学大学院 (Graduate School of Osaka Univ.), ‡JST ERATO

1 はじめに

ロボットが人とコミュニケーションをするためには，人に違和感を持たせないコミュニケーション行動を獲得している必要があるが，身体構造の違いにより人のコミュニケーション行動をロボットに再現させることは難しい．一方で人の乳児も身体構造が親と異なり，自身の行動が親のどの行動に対応するのかをあらかじめ知らないため，親の行動を完全には再現できない．しかし，それにも関わらず乳児は親との関わりあいの中で親がするようなコミュニケーション行動を獲得していく．このような，親とのインタラクションを含めた乳児の発達過程をモデル化し，人とのインタラクションを通じて発達するロボットを実現することは，ロボットに人にとって違和感のないコミュニケーション行動を持たせる手法としてだけでなく，人の乳児のコミュニケーション能力獲得に至る認知発達過程の理解に対する構成的アプローチ [1] としても期待される．

乳児の観察研究から，乳児の発声が音韻様であると知覚した場合母親は高頻度で乳児の音声を模倣すること [2]，また母親の模倣が乳児の音韻様の発声の頻度を高めること [3] が知られている．このように人の母子間には相互に模倣しあう様子が見られ，養育者側の模倣は，乳児の行動に対応する養育者の行動，すなわち身体構造の違いを吸収する対応づけを正しく学習するためのリファレンスを乳児に呈示する効果を持つと考えられる．さらに養育者の模倣音声は知覚のマグネット効果と呼ばれる現象により実際の乳児の発話音声よりも，養育者自身が普段使用する発話カテゴリである母音に近づけて生成されると考えられるため，乳児の発話カテゴリを母音に誘導する働きも持つと考えられる．

この考えに基づき Miura et al.[4] は，人との相互模倣を通じてロボットが明瞭な母音を獲得することが可能であることを示した．しかし，身体構造の対応付け，すなわち人とロボットの声域をおおまかに対応付ける写像が与えられることが仮定されていた．また，人の模倣が誘導的な働きを持つことが仮定されていたが，それがどの程度強いものであるかも議論されていなかった．

そこで本研究では 2 節でまず，人の模倣特性を考慮した模倣モデルを導入し，これに基づき人と相互に模倣しあう過程を通じ人との声域の違いを吸収する対応付けと母音獲得を同時に行う学習モデルを提案する．次に 3 節では計算機による様々な合成音声を被験者に模倣させる実験を行い，提案する模倣モデルの妥当性を検証する．最後に，提案する模倣モデルによる相互模倣シミュレーションによりロボットの同時学習が可能であることを示し，この時マグネット効果が学習の精度を高める要素となっていることを確認する．

2 声域の対応付けと母音獲得の同時学習モデル

ロボットが人の母音を獲得する際に問題となるのが人との発声可能な音声 (以下声域と呼ぶ) の違いである．人は相手の発声した母音が自身の発声不可能な音声であっても，自身の発声可能な母音として認識することができる (例えば男性が女性の「あ」を聞いて「あ」だと認識できる)．これは相手の声域と自身の声域の対応がついているからであると考えられ，人と声域の異なるロボットに母音を獲得させる場合，この対応付けをどのように与えるかが大きな課題である．

本研究ではこの問題に対して，人との相互模倣を通じて自身の声域と人の声域の対応付けを推測させながら同時に人の母音獲得もさせる同時学習モデルを提案する．この学習モデルでは，人の模倣が持つ二つの働きがこの同時学習を可能にすると考えられる．一つは，ロボットの発声に対応する人の発声を呈示することで声域の対応関係を教える働きであり，もう一つは相手の発声を再現する時に完全には再現できず，無意識のうちに自身の普段よくする発話，つまり母音に置き換えた模倣をすることでロボットの発声を母音に導く働きである．

本節ではまず，このような働きをもつ模倣メカニズムのモデルを構築し，これに基づく相互模倣を通じた同時学習モデルを提案する．

2.1 マグネット効果を表現する模倣モデル

人が聞いた音声を模倣する際，聞いた音声が実際の音声よりも自身の母音に近い音声として認識されるマグネット効果と呼ばれる現象の影響を受け，模倣音声は実際に聞いた音声よりも母音に近い音になると考えられる．以下ではこのマグネット効果を表現する模倣モデルについて説明する．

ロボットと人の音声の N 次元音響特徴ベクトルをそれぞれ $\boldsymbol{x}$，$\boldsymbol{y}$ とし，ロボットの模倣メカニズムを以

下のようにモデル化する．すなわち，人の発声 $\boldsymbol{y}$ を聞いたとき，ロボットの出力すべき音声 $\boldsymbol{x}$ を，

$$\begin{aligned}\boldsymbol{x} &= {}^r\boldsymbol{f}(\boldsymbol{y};{}^r\boldsymbol{x}_i,{}^r\alpha) \\ &= \sum_i^l \frac{\exp(-|{}^r\boldsymbol{g}(\boldsymbol{y};{}^r\alpha)-{}^r\boldsymbol{x}_i|^2/{}^r\sigma^2)}{\sum_i^l \exp(-|{}^r\boldsymbol{g}(\boldsymbol{y};{}^r\alpha)-{}^r\boldsymbol{x}_i|^2/{}^r\sigma^2)}{}^r\boldsymbol{x}_i\end{aligned} \tag{1}$$

と定める．ここで，${}^r\boldsymbol{x}_i(i=1,2,\cdots,l)$ はロボットの i 番目の発話カテゴリの代表ベクトルである．また ${}^r\boldsymbol{g}$ は声域の違いを大まかに対応づける写像であり，${}^r\alpha$ はそのパラメータである．本研究では，${}^r\boldsymbol{g}$ として，

$${}^r\boldsymbol{g}(\boldsymbol{y};{}^r\alpha) = {}^rA\begin{pmatrix}\boldsymbol{y}^{\mathrm{T}} & 1\end{pmatrix}^{\mathrm{T}}$$

のような線形写像を用いる．この模倣モデルでは発話カテゴリの線形和で模倣音声を決定しており，それぞれの発話カテゴリの重みとして写像後の音声との近さに応じた値が設定される．ここで式中の ${}^r\sigma$ は発話カテゴリの重ね合わせの重みのばらつきを決めるパラメータであり，この値によって模倣におけるマグネット効果の程度の強さを変えることができる．例えば ${}^r\sigma$ が小さいとき，近くの発話カテゴリの寄与がより大きくなるので，発話カテゴリにより近い音声が模倣音声となる，すなわちマグネット効果の程度が強い模倣のモデルとなる．

2.2 ロボットの行う同時学習

相互模倣の流れを Fig.1 に示す．学習ステップが t の時，人はロボットの発話 $\boldsymbol{x}(t)$ を，模倣し，音声 $\boldsymbol{y}(t)$ を発声する．ロボットはこの人の発声を聞き，学習中の声域の写像を用いて $\boldsymbol{y}(t)$ を ${}^r\boldsymbol{g}(\boldsymbol{y}(t);{}^r\alpha(t))$ に変換した後，式 1 の模倣メカニズムによりマグネット効果の影響を受けた音声 $\boldsymbol{x}(t+1)$ を次のステップの発声とし，それを人が再び模倣するということを繰り返す．ただし，ロボットは二回に一回ランダムな発声をするものとする．

この相互模倣過程でロボットは，人の発話 $\boldsymbol{y}(t)$ と先ほどの自身の発話 $\boldsymbol{x}(t)$ の組の履歴を用いて声域の写像のパラメータである ${}^rA(t)$ を忘却係数 0.95 の忘却係数付き逐次最小自乗アルゴリズム [5] で毎ステップ更新する．またロボットは発話カテゴリを，写像後の人の発話 ${}^r\boldsymbol{g}(\boldsymbol{y}(t);{}^r\alpha(t))$ を入力とした 2 次元 SOM[6] のコードベクトルとして毎ステップ更新を行う．

3 合成音声模倣実験

模倣モデルで仮定されているマグネット効果の存在を確認するため，5 名の日本人男子大学生および大学院生を被験者として様々な合成音声に対して人がどのように模倣を行うかを調査した．

3.1 実験概要

ヘッドフォンを通じて被験者に合成音声を一秒間聞かた後，その音声をマイクに向かって模倣させる．これを 1 回として，聞かせる音声を毎回変えながら計 400 回繰り返す．ここで合成音声は第一，第二フォルマント空間* において日本人男性の各母音の平均値を中心とする半径 180[mel]† の円内でランダムに決定した (Fig.2(a))．

本実験では音声合成，音声分析ライブラリとして Snack Tool Kit‡ を利用した．また，音声合成の際に固定のパラメータとして，F1 のバンド幅を 50[Hz]，F2 のバンド幅を 75[Hz] とし，また F3 と F4 もそれぞれ 2500[Hz](バンド幅 100[Hz])，3500[Hz](バンド幅 150[Hz]) とした．音声分析の際の録音，処理はサンプリング周波数を 10[kHz]，量子化ビット数を 16[bit] とした．またフォルマントの推定には 15 次の線形予測分析を用いた．

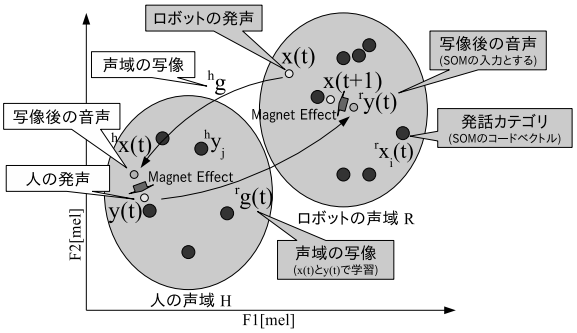
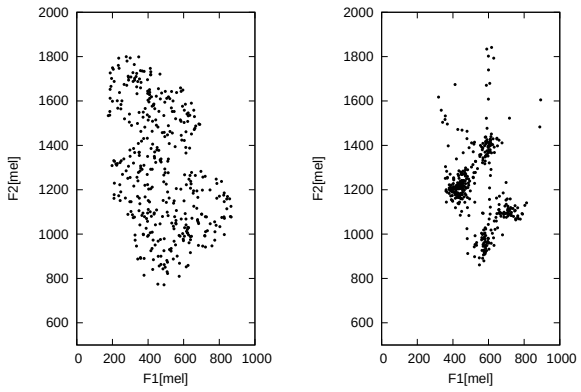


Fig.1 相互模倣の流れ



(a) 聞かせた音声　(b) 模倣音声

Fig.2 被験者に聞かせた音声とそれに対する模倣音声の 1 例

3.2 実験結果

被験者の模倣音声は例えば Fig.2(b) のようにいくつかの母音周辺に偏って分布しており，他の被験者でも程度の違いはあるが同様の傾向が見られた．聞かせる音声を選択した各円内において，聞かせた音声の散らばりの程度と模倣音声の散らばりの程度を比較した (Fig.3)．ここで，散らばりの程度は第一主成分の分散とした．Fig.3 では横軸が母音の種類を示し，各母音

* 人の母音はフォルマントと呼ばれる共鳴周波数のピークの値によって分類可能であることが知られている

† 人間の音の高さの知覚特性に合わせた尺度であり，周波数 f[Hz] に対するメル周波数 m[mel] は，
m = (1000/log2)log(f/1000 + 1) で与えられる．

‡ http://www.speech.kth.se/snack/

についてその母音を中心とする円内で選んで聞かせた音声の分散を左側に，またその円内に入る模倣音声の分散を右側に示しており，全ての母音について聞かせた音声よりも模倣音声の散らばりが小さくなっていることが分かる．ここで，分散は被験者 5 名の平均値である．

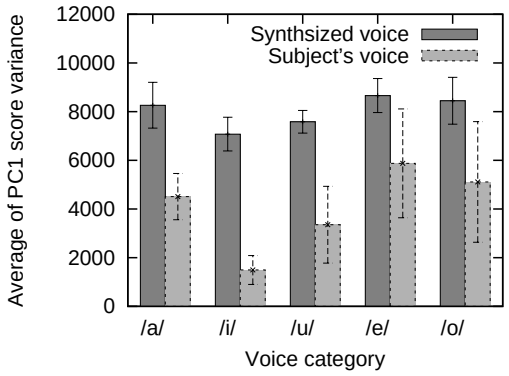


Fig.3 聞かせた音声と模倣音声の分散比較

この結果により，人は聞いた音声の模倣をする際に聞いたそのままの音声を発声せず，母音周辺で偏った音声を発声することが確認できた．

4　相互模倣シミュレーション

人もロボットと同様の模倣モデルにより模倣をするとして計算機による相互模倣シミュレーションを行い，ロボットの同時学習が可能であるかを検討した．

4.1　シミュレーション設定

今回のシミュレーションでは音響特徴量の次元を 2 とし，周波数の単位はメル周波数 [mel] とした．ロボットと人の発話カテゴリの数はそれぞれ $l = 20$，$m = 5$ とし，ロボットの発話カテゴリを 4×5 の 2 次元 SOM のコードベクトルとした．

今回のシミュレーションでは Fig.4 のように人の声域 H を日本人男性の声域に近い範囲とし，ロボットの声域 R は H を F1 に関して +200 ，F2 に関して +600 平行移動させた範囲として互いの声域が異なるように設定した．また，人の発話カテゴリは日本人男性の母音の平均値とし学習を通して不変であるとする．ロボットの発話カテゴリの初期値は R の中でランダムに決定する．

また，人は聞いた音声を線形変換により自身の発声可能な音声に対応付けていると仮定して，人が行う写像の係数行列を

$$ {}^{h}\hat{A} = \begin{pmatrix} 1 & 0 & -200 \\ 0 & 1 & -600 \end{pmatrix} $$

とし，学習を通して不変であるとする．よって，ロボットの獲得すべき写像の係数行列は人の写像の逆写像であるので

$$ {}^{r}A_{target} = \begin{pmatrix} 1 & 0 & 200 \\ 0 & 1 & 600 \end{pmatrix} $$

であるが，初期値は

$$ {}^{r}A(0) = \begin{pmatrix} N(0, 0.5^2) & N(0, 0.5^2) & N(0, 500^2) \\ N(0, 0.5^2) & N(0, 0.5^2) & N(0, 500^2) \end{pmatrix} $$

のように全要素を 0 を中心とする正規分布乱数によりランダムに与える．

このとき獲得目標母音はロボットの声域の音声の中で人が最も母音らしいと知覚する音声，すなわち

$$ {}^{r}\boldsymbol{x}_{i,target} = {}^{r}A_{target} \begin{pmatrix} {}^{h}\boldsymbol{y}_i^{\mathrm{T}} & 1 \end{pmatrix}^{\mathrm{T}} $$

である．

また，ロボットが二回に一回行うランダムな発声は学習中の発話カテゴリのいずれかを中心としてガウス分布となるように選ぶとする．そしてこのガウス分布の σ は，総学習回数を T として

$$ \sigma = 400.0 \times (\frac{T - t}{T}) $$

のように減少させ，学習が進むにつれ発話カテゴリから大きく離れた音声は選ばなくなるようにする．学習の総回数は 400 回とした．

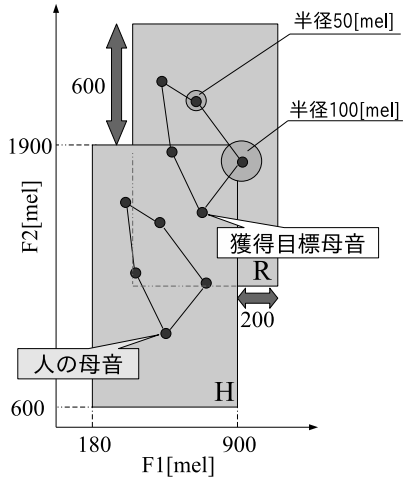


Fig.4 人とロボットの声域及び人の母音の設定位置関係

4.2　結果

人の模倣の仕方を決める ${}^{h}\sigma$ とロボットの模倣の仕方を決める ${}^{r}\sigma$ を様々に変え，それぞれの組合せについて 100 セットずつの結果の平均を比較する．ここで ${}^{h}\sigma$ は 10 から 210 まで 40 ずつ変化させた．${}^{h}\sigma = 10$ の時，模倣モデルのマグネット効果はほぼ最大となる．この場合，入力される音声が複数の母音が混ざったような曖昧な音であってもその中で一番近い発話カテゴリをそのまま出力とするような模倣の仕方を示す．それに対し ${}^{h}\sigma = 210$ の時，模倣モデルのマグネット効果は弱く，曖昧な音声ばかり出力されることになる．また ${}^{r}\sigma$ は 10 から 130 まで 20 ずつの 7 通りとした．

学習終了時に獲得目標母音を中心とする 100 [mel] の範囲に発話カテゴリが入っていれば母音が獲得でき

たとし，獲得母音の数を比較した．Fig.5 では x 軸 ${}^h\sigma$ ，y 軸 ${}^r\sigma$ として，学習終了時に獲得された母音数の平均が色の濃さとして示されている．この図より ${}^r\sigma$ が小さい程母音が多く獲得でき，また ${}^h\sigma$ は 50 から 90 程度の値で母音の獲得率がよくなっていることが分かる．

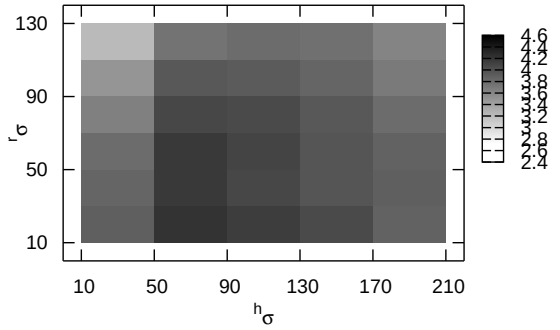


Fig.5 人とロボットのマグネット効果の強さを変えた場合の母音獲得数の比較

4.3 考察

シミュレーションにおいて人の模倣の仕方を決める ${}^h\sigma$ による学習結果の違いが見られた．人の模倣におけるマグネット効果の強さを決定する ${}^h\sigma$ は 50 から 90 程度の値で多くの目標母音を獲得でき，それより大きくても小さくても学習結果は悪くなっている．ここで写像がどの程度正しく推測されたかを評価するため，${}^r\sigma = 10$ に関して学習終了時の写像で人の母音を変換した推測母音と，真の写像で人の母音を変換した獲得目標母音の距離が ${}^h\sigma$ によりどう変わるかを比較した (Fig.6)．縦軸は推測母音と獲得目標母音の距離の母音毎の平均を示しており，${}^h\sigma$ が 130 程度の値で他と比べて小さい値となる，すなわち精度の高い写像の推測ができているといえる．

マグネット効果が強すぎる時に写像の推測精度がよくならないのは，人が聞いた音を大きく歪めてロボットに呈示するためであると考えられ，また反対にマグネット効果が弱すぎても精度がよくならないのは，発話カテゴリが母音に近付かないため学習後半でも人が曖昧な音声ばかりの精度の悪い模倣ばかりをしているためと考えられる．一方，${}^h\sigma = 50, 90$ の場合，写像の推測精度は ${}^h\sigma = 130$ の場合より劣っているにも関わらず，母音の獲得数が多くなっている．このことはマグネット効果が母音獲得の精度を高める要因となっていることを示している．

5 まとめ

本研究では人との相互模倣を通じてロボットが自身の声域と人の声域の対応付けを推測しながら人の母音を獲得する同時学習モデルを提案した．そして計算機シミュレーションにより，人が聞いた音を模倣する際に知覚のマグネット効果の影響を受けて，聞いた音声

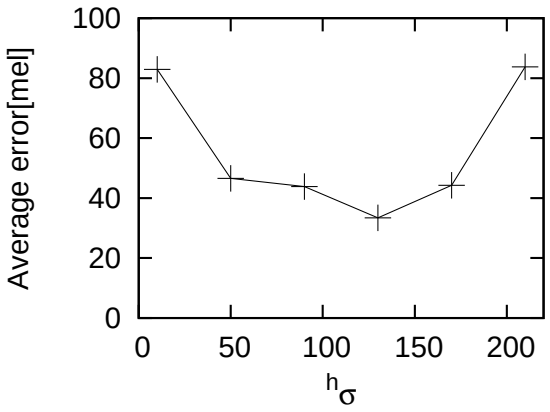


Fig.6 声域の変換の推測誤差．人の逆写像が正しく獲得されているほど誤差は小さい値となる

を適度に母音に近付けてロボットに呈示することがロボットの同時学習を助け，母音獲得を可能にすることが確認できた．

今回，人が自身の声域外の音声を聞いたとき，線形変換により自身の声域の音声に対応付けて知覚すると仮定し，ロボットに線形変換で声域の対応付けを学習させている．これは日本人の発声する 5 母音の位置関係が話者によらず維持され，線形変換で近似的に対応付けることが可能であるためにおいた仮定であるが，人が実際にロボットとの相互模倣過程において線形変換で 5 母音を対応付けるかは明らかでない．従って，今後の課題として実際の人との相互模倣実験を行い同時学習が可能であるかを検討する必要がある．

参考文献

1) Minoru Asada, Karl F. MacDorman, Hiroshi Ishiguro, and Yasuo Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous System*, Vol. 37, pp. 185–193, 2001.
2) N. Masataka and K. Bloom. Accoustic properties that determine adult's preference for 3-month-old infant vocalization. *Infant Behavior and Development*, Vol. 17, pp. 461–464, 1994.
3) M. Peláez-Nogueras, J. L. Gewirtz, and M. M. Markham. Infant vocalizations are confitioned both by maternal imitation and motherese speech. *Infant behavior and development*, Vol. 19, p. 670, 1996.
4) Katsushi Miura, Minoru Asada, Koh Hosoda, and Yuichiro Yoshikawa. Vowel acquisition based on visual and auditory mutual imitation in mother-infant interaction. In *The 5th International Conference on Development and Learning (ICDL'06)*, 2006.
5) 飯國洋二．適応信号処理アルゴリズム．培風館, 2000.
6) T.Kohonen. *Self-Organizing Maps(2nd Ed.)*. Springer, 1997.