教示者の音韻情報と唇の形状を利用したロボットの音韻獲得

Vowel Acquisition based on the Phoneme and Lip Shape information from a Caregiver.

○学　三浦　勝司　(阪大院，科学技術振興機構 ERATO)
正　浅田　稔　(阪大院，科学技術振興機構 ERATO)
正　細田　耕　(阪大院，科学技術振興機構 ERATO)
正　吉川　雄一郎　(阪大院)

Katsushi Miura[1)2)], Minoru Asada[1)2)], Koh Hosoda[1)2)], and Yuichiro Yoshikawa[2)]
[1)]Asada Synergistic Intelligence Project, ERATO, JST (www.jeap.org)
[2)]Graduate School of Engineering, Osaka University
{miura,asada,hosoda,yoshikawa}@er.ams.eng.osaka-u.ac.jp

A pioneering constructivist approach to building a robot that reproduces a developmental process of infants' vowel acquisition has been conducted by Yoshikawa et al. [1)] inspired by the observation in infant study. They have constructed a mother-infant interaction model with robot learning capability and parrot-like teaching by caregiver. However, the robot has not listened his/her own voice, therefore nor actively explored more natural vowels similar to the caregiver. The study presented in this paper extends the previous work in the following manners seeking for more natural interaction. First, imitation of the caregiver's lip. Second, utilizing the pentagon which the caregiver's vowels construct in the formant space. Third, hypothesizing that mutual imitation between the robot and the caregiver is introduced to obtain more natural vowels. Through this process, the desired formants are gradually shifted, expecting to more natural ones. The experimental results are shown and the future issues are discussed.

Key Words: Vowel Imitation, Formant Space, Lip Shape, Maternal Imitation

1　緒言

言語による人とのコミュニケーションはロボットにとって最も困難な課題の一つである．また，人の乳児がどのようにして言語を獲得するかは，人の認知発達における謎の一つである．この発達過程に対し，発話ロボットを使用した構成論的アプローチは認知発達ロボティクスの観点から有望であると考えられる [2)]．

人の乳児と同様に，自身のセンサモータ出力と音韻との関係を与えられていない発話ロボットが音韻を獲得するためには，環境とのインタラクション，特に教示者とのインタラクションを通じて自身のセンサモータ出力と音韻との関係を学習しなくてはならない．従来研究で，声道と蝸牛を備えた複数のエージェントがシミュレーション上で互いにインタラクションすることで音韻を獲得するモデルが提案されている [3, 4)]．この imitation game[3)] や magnet effect [4)] では，音韻に関する知識をエージェントにあらかじめ与えていないが，エージェント同士で同じ発声が可能であると仮定しているため，共通の音韻を獲得することが出来た．しかし，構音器官の未発達な乳児が教示者と同じ発声をおこなうことは不可能であるため，我々は乳児の未成熟さを考慮した音韻獲得モデルを構築する必要がある．

そこで，Yoshikawa et al. [1)] は，乳児のクーイングが母親の模倣的応答を引き出す [5)]，母親の模倣が乳児の発声を促す [6)] という 2 つの乳児の発達に関する知見から，母子間のインタラクションをモデルとしたロボットの音韻獲得を提案した．このモデルでは，乳児の音韻様の発声に対し教示者が対応する音韻を返すことが音韻獲得において重要であると仮定している．そして，教示者のオウム返し教示によるロボットの音韻獲得を行った結果，ロボットは日本語 5 母音のうち 4 つを獲得している．しかし，ロボットが自身の発声を聞いていないため，獲得した音韻を教示者のような自然な発声になるように改善することは出来ない．

本論文では，Yoshikawa et al. [1)] のモデルを以下のような手法を用いて改良する．まず始めに，ロボットに唇を付け，教示者の口唇形状を模倣させることでフォルマント空間上での初期の探索範囲を狭め，短時間での学習を可能にする．この口唇形状の模倣により各音韻の初期位置がフォルマント空間上で決まる．次に，教示者の日本語 5 母音がフォルマント空間上で形成する五角形をロボットの音韻の初期位置の中心へシフトさせることでロボットの目標フォルマントとして利用する．最後に，教示者の模倣が無意識に自身の音韻に近い発声になるはずと仮定し，ロボットと教示者の相互模倣によってロボットに自然な音韻を獲得させる．これらの過程を通して，ロボットの目標の音韻がより自然なものへ徐々に変化すると予想される．実験結果および今後の課題を示す．

2　母子間インタラクションモデルに基づく音韻模倣

システムの概要を Fig. 1 に示す．発話ロボットは構音および聴覚機構を持っており，教示者と互いの発声を模倣しあう．本論文で使用される発話ロボットは Yoshikawa et al. [1)] で使用されたロボットを改良したものであり，口唇部に水平・鉛直方向の開閉で 2 自由度，声道部の変形で 4 自由度もつ．

教示者には 2 つの役割がある．1 つは，ロボットが口唇形状を模倣できるように音韻発声時の口唇形状を見せるこ

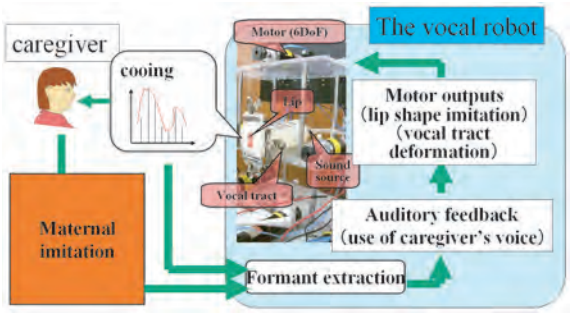


Fig.1 An overview of the whole system

とである．発声と口唇形状の関係については，Carre[7] が口唇部の開閉とフォルマントの変化について単一声道管モデルのシミュレーションで示している．また，Patterson and Werker [8] は乳児が口唇形状と音韻との対応関係を認識していることを示している．そのため，ロボットが教示者の音韻発声時の口唇形状を模倣することで，フォルマント空間上で音韻探索をするための適当な初期値をロボットに与えることが出来ると考えられる．

2 つ目は相互模倣，つまりロボットの発声を教示者が模倣することである．ロボットも教示者の発声を模倣するため，ロボットの発声を自然なものへと引き込むような相互模倣が起きると期待される．ただし，教示者が模倣する際に無意識に音韻よりの発声をしてしまうという仮定があり，結果としてロボットの発声がより自然なものへと徐々に変化していくと考えている．

次節で口唇形状の模倣による音韻探索範囲の限定，教示者の日本語 5 母音がフォルマント空間上で形成する五角形の利用，相互模倣による学習法について述べる．

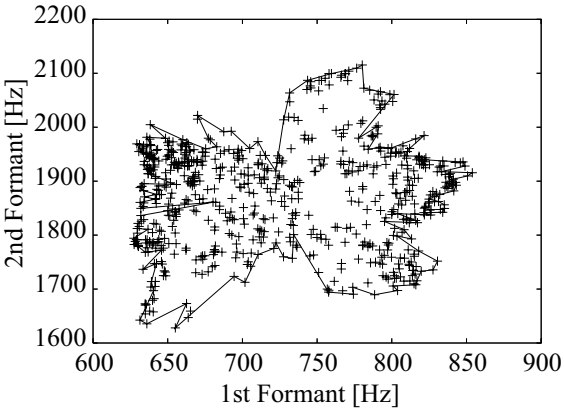


Fig.2 Formant distribution of robot's voice

3　口唇形状の模倣

ここでは，人とロボットの各音韻の相対的位置関係がフォルマント空間上で類似すること，および，口唇形状の模倣がフォルマント空間上での音韻探索範囲をどの程度制限するかを示す．

3.1　口唇形状とフォルマントとの関係

発話ロボットは 6 自由度を持ち，そのうち 2 自由度は唇の開閉に利用する．まず初めに，ロボットの発話能力について調べた．各モータ出力を 0(変形なし)，0,5(0 と 1.0 の間の変形)，1.0(最大の変形) の 3 段階に正規化し，729(3^6) 通りの発声を行った．この中からフォルマントの抽出結果が不安定となった発声を除く 711 通りの発声を選び出し，第 1・第 2 フォルマントの値を Fig.2 に示す．

Table 1 Relation between robot's lip shape and motor outputs

	/a/	/i/	/u/	/e/	/o/
モータ出力 (水平)	1.0	0.0	0.0	0.5	0.5
モータ出力 (鉛直)	1.0	1.0	0.0	1.0	0.0

次に，ロボットが持つ 6 自由度のうち口唇部の 2 自由度を人の口唇形状を模倣するように固定し，声道部のみを 81 通り (3^4) に変形させて発声し，口唇形状の効果を実験する．Table 1 にロボットが口唇形状を模倣したときの口唇部のモータ出力を，Fig.3 に口唇形状を模倣して発声したときの第 1 第 2 フォルマントを示す．上段はフォルマントの分布，下段は人とロボットの音韻発声時の口唇形状である．

口唇形状の模倣により，発声した回数は 711 回から 394 回 (フォルマントが不安定な 11 回の発声を除く) に減少しており，発声回数を約 55%に減少させることが可能である．

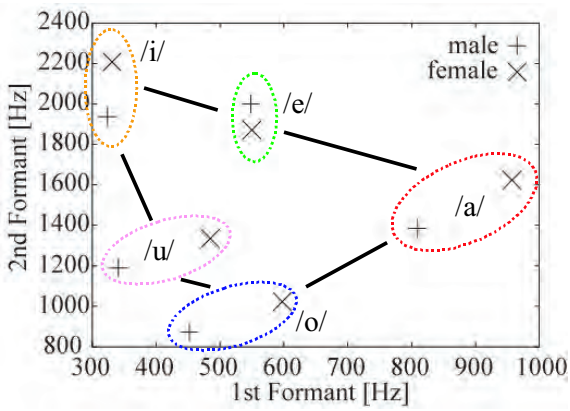


Fig.4 Formant distribution of human's vowel

3.2　人とロボットの音韻のフォルマント

人の音韻発声時のフォルマントを Fig.4 に示す．日本語 5 母音の相対的位置関係は五角形となっており，この相対的位置関係はロボットの 5 母音と似ている (Fig.3 参照)．例えば，/a/の第 1 フォルマントは高く，第 2 フォルマントは中央付近に集まっているなどである．

この結果から，人の各音韻のフォルマント空間上での位置関係をロボットが音韻を獲得するための目標値として利用できると考えられる．しかし，人とロボットとの間で構音可能な領域の広さや位置が異なるため，人の音韻 (C/a/-/o/) を Fig.5 に示すようにロボットの音韻 (L/a/-/o/) 方向へ，中心が一致するように移動させる．そして，移動後の人の音韻 (C/a/'-/o/') をロボットが自然な音韻を獲得するための目標値として利用する．

3.3　相互模倣による自然な音韻の獲得

口唇形状を固定したまま初期の音韻の位置 (L/j/, j=a, i, u, e, o) から目標の音韻の位置 (C/j/', j=a, i, u, e, o) への軌跡をたどる事で，ロボットは簡単に音韻を探索できる．しかし，目標の音韻が必ずしも自然に聞こえるという保証はない．そこで，相互模倣により目標音韻の位置を調整することで，ロボットにさらに自然な音韻を獲得させる．ただし，教示者がロボットの発声に対し無意

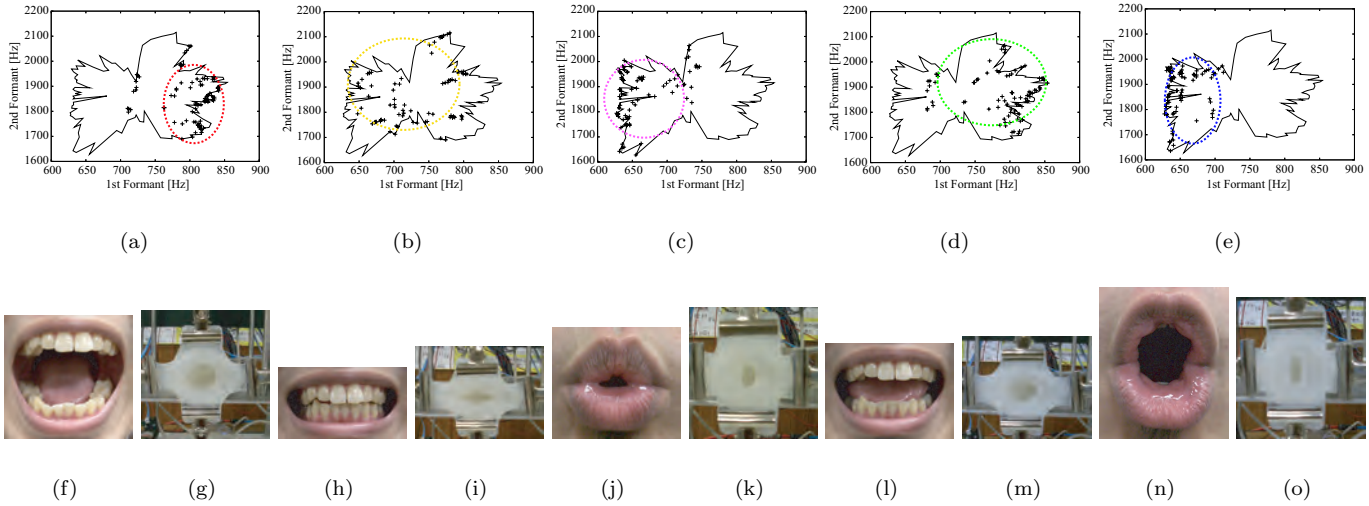


Fig.3 Formant distribution of the utterances (top: /a/, /i/, /u/, /e/, /o/) from lip shape imitation (bottom)

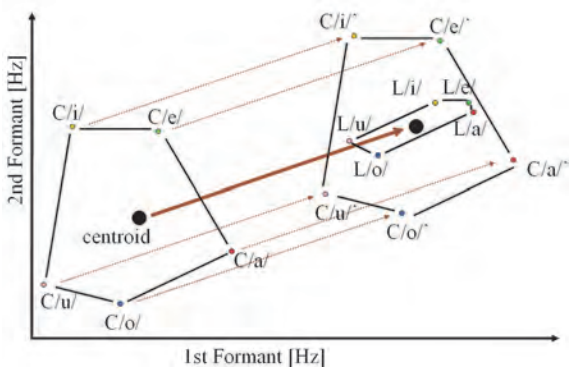


Fig.5 Transformation of human's vowels to the robot utterance area

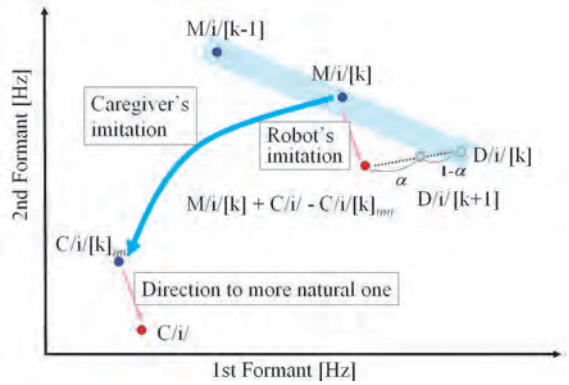


Fig.6 Mutual imitation process on the formant space

識に音韻よりの模倣をしてしまうという仮定が含まれている．

Fig.6 は相互模倣によってどの様にロボットの発声が変化するかを示している．D/j/[k], M/j/[k], C/j/[k]$_{imi}$ (j=a, i, u, e, o) はそれぞれ目標値，現在値，時刻 k の教示者の発声である．初めは D/j/[0] = C/j/'， M/j/[0]= L/j/， C/j/は教示者の音韻である．

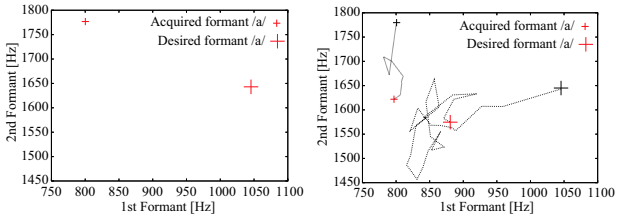
1. M/j/[k-1] と D/j/[k] との軌跡上にあるような M/j/[k] をロボットが探索し，教示者に向かって発声する．
2. 教示者はロボットの発声 M/j/[k] を模倣した C/j/[k]$_{imi}$ を発声する．
3. C/j/[k]$_{imi}$ から C/j/までのベクトルの差を求め，ロボットの発声 M/j/[k] と目標値 D/j/[k] との差として評価．
4. ロボットの目標値を M/j/[k] + C/j/ - C/j/[k]$_{imi}$ とする．
5. 急激な目標値の変化を防ぐために，新たな目標値 D/j/[k+1] を D/j/[k+1] = α D/j/[k] + (1-α) (M/j/[k] + C/j/ - C/j/[k]$_{imi}$) に設定する．
6. M/j/が変化しなくなる (構音能力の限界) まで上の過程を繰り返す．

4 実験結果

α = 0.7，発声回数は 500 回と設定し，実ロボットによる学習を行った．Fig.7 の (a) は母音/a/の目標値，(b) は相互模倣によりロボットが獲得していった母音の変化を示している．(a) は初期値であり，M/a/[0] は小さい+，D/a/[0] は大きい+で示している．また，(b) は学習結果であり，M/a/[500] は小さい+， D/a/[500] は大きい+で示している．

Fig.8 は日本語 5 母音についての学習結果である．細い点線は相互模倣による目標値や獲得した音韻の変化を示しており，D/j/の初期値が黒い五角形，学習後の D/j/が黄色の五角形，獲得した 5 母音 M/j/[500] が水色の五角形である．

ロボットがどの程度自然な発声を獲得したかを示すことは難しいため，相互模倣を用いない (目標値 D/j/が変化しない) 実験結果との聞き比べにより評価を行った．その結果，評価者 40 人中約 7 割が相互模倣ありのほうが自然であると評価した．



(a) Desired and acquired vowel /a/ by a robot at the learning step 0

(b) Desired and acquired vowel /a/ by a robot at the learning step 500

Fig.7 The learning process in the case of the vowel /a/

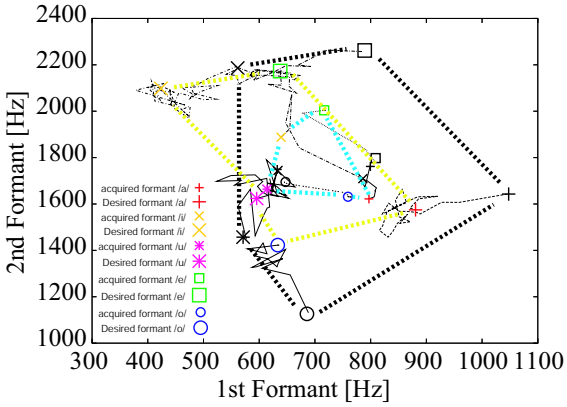


Fig.8 Desired and acquired vowels by a robot at the learning step 500

5 討論

現在，ロボットの口唇形状は教示者によってあらかじめ与えられている．実際には，ロボットはカメラ画像を用いて教示者の顔から口唇形状を抽出し模倣すべきである．そのためにはロボットは目，口，鼻などの顔の各部位の対応関係について知らなければならない．まだ議論はあるが，新生児模倣に関する知見 [9] は出生後間もない乳児がこのような対応関係を獲得している可能性を示している．しかし，この乳児の身体表象獲得メカニズムはいまだ解明されていない．ロボットの身体表象を設計者があらかじめ与える場合，ロボットのセンサやモータ数が増加するほど，ロボットに身体表象とモータコントロールとの関係を与える設計者の負担は大きくなる．そのため，このようなモータコントロールと身体表象の発達は認知発達ロボティクスにとって鍵となる問題の一つであり，近い将来解決しなくてはならない．

もう一つの問題は教示者の振る舞いが学習に与える影響についてである．本論文では教示者は 1 人だけであり，ロボットに対しどのような模倣をしたか記録してない．このようなインタラクションを解析することで，人の音韻獲得過程に対する理解や学習ロボットの設計方針の確立に役立つことが期待される．

本研究では，口唇形状の情報が乳児の音韻獲得に役立つ，教示者の模倣が自身の音韻に近い発声になりやすいとの仮定を用いた．また，これらの仮定に基づく人とロボットとのインタラクションモデルを構築し，実験結果を示した．しかし，これらの仮定の証明は出来ていないため，乳児研究や発達心理学との協力によってこれらの仮定の証明だけでなく，インタラクションモデルの改良をしていく必要がある．

参考文献

[1] Y. Yoshikawa, M. Asada, K. Hosoda, and J. Koga. A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science*, Vol. 15, No. 4, pp. 245–258, *December* 2003.

[2] Minoru Asada, Karl F. MacDorman, Hiroshi Ishigro, and Yasuo Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, pp. 185–193, 2001.

[3] B. de Boer. Self-organization in vowel systems. *Journal of Phonetics*, Vol. 28, pp. 441–465, 2000.

[4] P.-Y. Oudeyer. Phonemic coding might result from sensory-motor coupling dynamics. In *In Proceedings of the 7th international conference on simulation of adaptive behavior (SAB02), OPTcrossref = , OPTkey = , pages = 406-416, year = 2002, OPTeditor = , OPTvolume = , OPTnumber = , OPTseries = , OPTaddress = , OPTmonth = , OPTorganization = , OPTpublisher = , OPTnote = , OPTannote = .*

[5] N. Masataka and K. Bloom. Accoustic properties that determine adult's preference for 3-month-old infant vocalizat. *Infant behavior and development*, Vol. 17, pp. 461–464, 1994.

[6] M. Peläez-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.

[7] Carre R. Prediction of vowel systems using a deductive approach. In *In Proceedings of the International Conference on Spoken Language Processing 96*, pp. 434–437, 1996.

[8] M. L. Patterson and J. F. Werker. Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, Vol. 6, No. 2, pp. 191–196, 2003.

[9] Andrew N. Meltzoff and M. Keith Moore. Explaining facial imitation: A theoretical model. *Early Development and Parenting*, pp. 179–192, 1997.