声道物理モデルの母音列繰り返し模倣による音素獲得シミュレーション

○尾形哲也　神田尚　高橋徹　駒谷和範　奥乃博

京都大学大学院情報学研究科

Phoneme Acquisition by Iterative Imitations of Vowel Sequences using Vocal Tract Model

Tetsuya Ogata, Hisashi Kanda, Toru Takahashi, Kazunori Komatani, and Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University

Abstract － We hypothesized that infants use self-vocalization babbling to explore imitatable and unimitatable elements in their mothers' voices. We constructed a phoneme acquisition model using continuous sound imitation between a human and an infant model. We applied Recurrent Neural Network with Parametric Bias (RNNPB) to learn the experience of self-vocalization, to recognize the human voice, and to produce the sound imitated by the infant model. The experimental results revealed that as imitation interactions were repeated, the formants of sounds of our system moved closer to those of human voices, and our system could self-organize the same vowels in different continuous sounds.

***Key Words*:** Imitation, Forward model, and Recurrent Neural Network with Parametric Bias

1. はじめに

乳児は言語獲得の基礎として，初語を発する以前から発声活動（バブリング）を行っている．乳児のバブリングは音素で表現できないが，音素獲得を経て初語発声に至る過程で必ず観測される活動である．

乳児は親の声の模倣を通して音声言語を獲得すると言われている[1, 2]．乳児はその成長過程で模倣を試行錯誤することで，音素単位の発見が可能になると考えられる．ここで，「乳児が音素の事前知識を持たない状態から如何にして母国語の音素体系を獲得するか」という課題がある．この問題に対して，音声模倣を通じた乳児の構音発達過程モデルを提案する．本研究で対象とする乳児の発達時期は，人間の大人と同程度の母音生成が可能となる3ヶ月頃から，母音模倣を通じて母国語特有の母音知覚が始まる6ヶ月の約3ヶ月である．

本モデルは次の4つのフェーズからなる．(1) 乳児の自己発声音とその構音動作の学習, (2) 乳児による親から発せられた音声の認識, (3) 認識音声に対する乳児の模倣音声生成, (4) 模倣精度の高い音声を自己選択．以上の音素獲得モデルに対し，親を人間，子を乳児モデルとして音素獲得シミュレーションの実現を行う．

2. 本研究における音素獲得と課題

発声は声道という身体により生成される行為であり，この身体拘束が連続音声の構造に影響していると考えられる．本研究では，音素獲得を「自分の身体（構音動作）により親の音声が模倣可能となること」と定義する．

ここで具体的な課題は以下の 3 点である．

課題 1: 音声生成過程における身体拘束利用

音声の発達には，その生成機構である声道発達の拘束が有効に働いていると考えられ，モデルに声道（身体）を組み込むことは本質であると考えられる．

課題 2: 音声 構音動作の時系列ダイナミクス処理

乳児は音素に関する事前知識を持たない．音素の種類，クラス数が未知の場合，隠れマルコフモデル (HMM) などの統計的確率モデルを用いた音響信号処理は必ずしも効率的ではない．また大量学習データを必要とする手法は，乳児の発達過程のモデルには適切でないといえる．

課題 3: 自己組織的な母音構造抽出

従来，音声模倣インタラクションによる母音獲得研究が行われている[3, 4, 5]．これらの研究では課題 1 を考慮した音素獲得を実現しているが，インタラクションにおける発話は 1 母音と定義されていた．つまり，音響信号が離散的な音素列であることを前提としている．しかし，親の子供への語りかけを考えた場合，一母音であるケースは稀である．より自然な発達過程のモデル化のためには，特に「離散から連続へ問題の拡張」が必要といえる．

3. 音声模倣による母音獲得手法

3.1 提案する母音獲得プロセス

本モデルは，既に提案した音声模倣モデル[6]を基に，声道モデルによる構音動作と音声データを RNNPB で学習させると共に時系列ダイナミクス分節化手法を適用し，各分節区間での力学構造を抽出する．さらに，抽出した力学構造を利用することで，既知・未知の聴取音声が認識し，連続音声の模倣を行う．提案する模倣プロセスを図1に示す．学習・認識・生成の4フェーズがある．

1. 学習（音声バブリング）

模倣システムに構音動作を入力し，複数母音を含む音声を生成させ，構音動作と生成音声を関連付ける．このフェーズは，乳児であるシステムが「どのよ

うに声道を動かせばいかなる音声が生成されるか」を学習する”音声バブリング”に相当する．

2. 認識（親の発声音を聴取）
　本システムに人間が発話した音声を入力し，入力音声に近い構音動作をRNNPBで計算する．

3. 生成（聴取音に対する音声模倣）
　本システムは認識フェーズで計算した構音動作を用いて模倣音声を生成する．

4. 選択（学習音声の選択）
　模倣音声のうち，入力音声に近いものを選択し，これを新たな学習データとする．その後，1に戻る．

　このプロセスを繰り返すことで模倣可能な音声種類が広がっていく．また上記の各フェーズにおいて，時系列ダイナミクス分節化手法[10]を適用する．

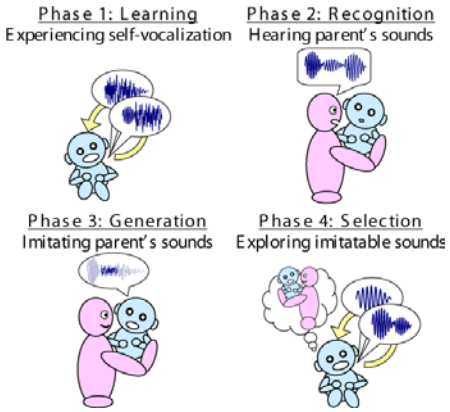

Figure 1　Proposed phoneme acquisition process.

3.2　音声合成器: 声道物理モデル

　まず2章で示した課題1に対応するため，Maeda により開発された声道物理モデルを利用する[7]．このモデルでは，構音パラメータを設定することで声道形状を決定し，それに対応する音声を生成することができる．各パラメータは，JP: 顎位置の上下，TDP: 舌背位置の前後，TDS: 舌背位置の上下，TTP: 舌尖の上下，LO: 唇の開閉，LPR: 唇位置の前後，LP: 喉頭位置の上下，に対応する．

　音声合成器としての声道モデルには，他にもPARCOR [8]や極めて高精度の音響再現機能を持つSTRAIGHT [9]などがある．しかし，本研究では，人間の認知発達過程をシミュレートするため，解剖学的知見に基づくMaeda モデルを採用した．

3.3　RNNPBモデルの学習と時系列データの分節化

　課題2に対応するため，RNNPBを利用した．本モデルは現状態ベクトルを入力とし，次状態ベクトルを出力とする予測器である[11]．Jordan型RNN[12]と同様にRNNPBモデルはBack Propagation Through Time (BPTT) 学習法[13]を用いる．RNNPBモデルの大きな特徴は，きわめて少数の学習時系列データから，大域的な汎化構造を自己組織化する点になる．つまり， RNNPBは，PB値と時系列パターンの汎化写像を自己組織化する．

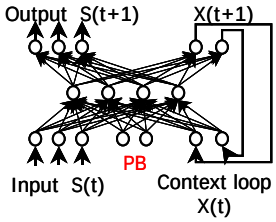

Figure 2　RNNPB

　RNNPBは学習・認識・生成の3つの動作モードを持つ．学習モードでは，RNNPBはセンサーやモータ出力の予測誤差からBPTT法を使用して，結合重みとPB値を更新する．シーケンスのステップ長をlとしたときPB値p_iの更新式は以下のようになる．

$$\delta\rho_t = k_{bp} \cdot \sum_{0}^{L_s} \delta_t^{bp} \qquad (1)$$

$$p_t = sigmoid(\rho_t / \zeta). \qquad (2)$$

(1)式のδ_t^{bp}は各時刻のデルター誤差で，これをTステップで移動平均をとる．内部値ρ_t をシグモイド関数で[1, 0]に正規化して PB 値としている．

　認識モードでは，入力シーケンスに対応する PB 値を求める．具体的には入力シーケンスを RNNPB に予測させ，その予測誤差からネットワーク重みの更新はせずに，PB 値のみを式(1)(2)で求める．生成モードでは，出力希望のシーケンスに対応する PB 値を PB ノードにセットし，RNN の閉ループモード（出力$S(t+1)$を再度入力する）で計算を行い，シーケンスを生成する．

　RNNPBはその入力時系列データの予測状態に基づいて分節化する．具体的には，単一のダイナミクスから生成されたシーケンスは安定した予測が可能であるという「予測可能性」を利用する．RNNPBは，ある一定のPB 値が与えられている際のダイナミクス特性は同一である．単一のPB 値で予測可能なシーケンス，すなわち単一のダイナミクスから生成されたシーケンスでは，予測が成功するためエラーは小なくなる．一方，複数のダイナミクスから生成されたシーケンスでは，ダイナミクスが切り替わる際に予測エラーが大きくなる．この予測エラーを分節化に利用する．概念図を図2に示す．

　長さTの時系列データ$X(t)$をN個のセクション$S_0, S_1, \ldots, S_{N-1}$に分割する問題を考える．$S_{i-1}$と$S_i$の間の境界時間を$t = s_i$によって表す． S_iは$[s_i, s_{i+1}]$と定義される．

Step 1: 初期化
　入力時系列データを同じ長さのN個のセクションに分割する．

$$s_i \leftarrow T \cdot i / N \qquad (i = 0, \ldots, N) \qquad (3)$$

Step 2: RNNPB学習
　各セクションS_i内でPB値は固定し，RNNPBの重みを学習する．

Step 3: 予測誤差の計算
　各セクションでの予測誤差$P(t)$を計算しその最大値E_iを求める．

$$E_i \leftarrow \max_{t \in S_i} \| X(t) - P(t) \| \qquad (i = 0, ..., N) \qquad (4)$$

Step 4: セクション長の調整

各セクションの境界s_iを以下の式により更新する．

$$s_{i+1} \leftarrow \begin{cases} s_{i+1} - ds & if\ E_i \geq E_{i+1} \\ s_{i+1} + ds & if\ E_i < E_{i+1} \end{cases} \qquad (5)$$

Step 5: 最大予測誤差E_iが閾値以下になるまで，Step 1～4を繰り返す．

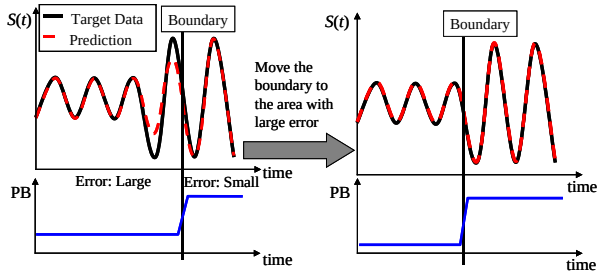

Figure 3　Articulating into multiple sequences

提案手法は，境界位置を調整して，全体の予測誤差最小となる境界位置を自動的に求める．

4. システム実装

本章では，本音素獲得モデルのシステム設計について述べる．音素獲得システムの概観を図4に示す．

4.1 音素獲得システムの概要

本システムにおける音素獲得の手順は以下の通りに行う．

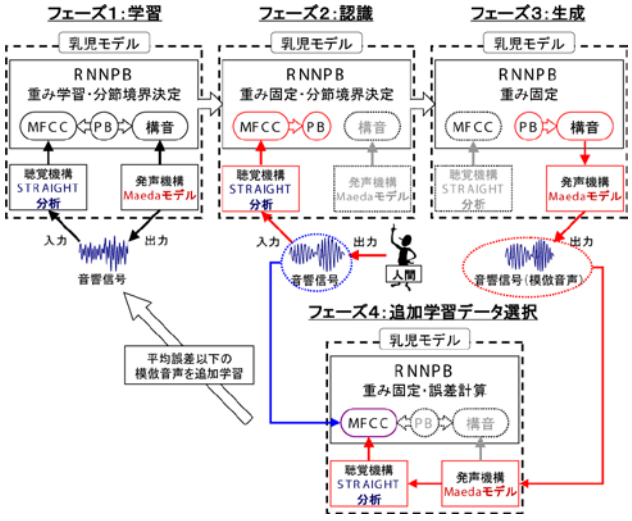

Figure 4　Diagram of developed system.

フェーズ1: 学習Maeda モデルに時系列構音パラメータを入力する．初回のフェーズ1では，ランダムバブリングによる構音パラメータを入力する．そのほかの場合は，フェーズ4で選択した学習音声の構音パラメータを入力する．Maedaモデルから生成された音響信号をSTRAIGHT分析し，MFCC 特徴量に変換，これと構音パラメータを1組として正規化する．これを用いてRNNPBの重み学習を行い，各入力データに対するPB 値と分節境界を求める．

フェーズ2: 認識人間の発する音声をSTRAIGHT分析によりMFCC特徴量に変換する．これをRNNPBの結合重みを固定した状態で予測エラーを求め，PB値と分節境界を求める．

フェーズ3: 生成フェーズ2で求めたPB 値と分節境界位置を利用し，RNNPBのフィードフォワード計算により構音パラメータを求める．求めた構音パラメータをMaedaモデルに入力し，模倣音声を生成する．

フェーズ4: フェーズ3で生成したすべての模倣音声と人間が発した音声から，模倣音声のMFCCs特徴量の2乗誤差を計算する．全体の平均予測誤差以下となる模倣音声を,フェーズ1の追加学習データとして選択する．

5. 実験

5.1 実験条件

提案音素獲得プロセスの検証実験結果を示す．各実験の条件として，システムには音素数や音素クラスについての事前知識は与えていない．

提案プロセスのフェーズ1～3を用いて，乳児モデルがフェーズ1でランダムな発声を行うバブリング学習により，人間の音声に対して模倣可能かを検証することである．

フェーズ1において，構音特徴量をランダムに変化させたものの中から構音可能な5種類の単音を選び学習データを作成した．選出した各単音をv_i ($i = 1 \ldots 5$) で表現する．学習音素列パターンを以下に示す．（各1350-msec, 30-msec/step）

/$v_1v_2v_3$/, /$v_2v_3v_4$/, /$v_3v_4v_5$/, /$v_4v_5v_1$/, /$v_5v_1v_2$/,
/$v_2v_1v_5$/, /$v_1v_5v_4$/, /$v_5v_4v_3$/, /$v_4v_3v_2$/, /$v_3v_2v_1$/,

一単音の発音時間は約400-msecであり，合計発音時間が1350-msec となるよう単音間の遷移時間を決定した．フェーズ1において学習パターンをRNNPBに学習させた．RNNPBの構成は，入出力層5次元，中間層40次元，文脈層5次元，PB層2次元である．

5.2 実験結果

図5にフェーズ2における話者2名の認識音声に対するバブリング直後（1st），追加学習一回目（2nd），追加学習二回目（3rd）模倣平均誤差を示す．各話者を認識するRNNPBが追加学習を行うたびに，誤差が減少することが確認できる．

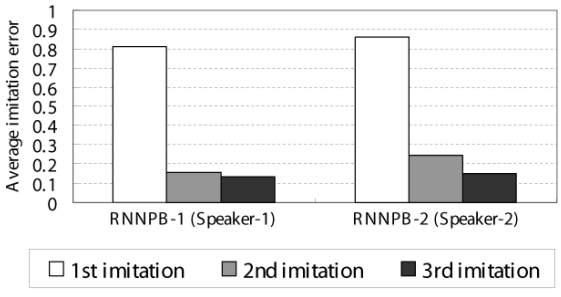

Figure 5　Average imitation error in 1st, 2nd, and 3rd generation phases.

図6は，バブリング学習直後のRNNPBにより自己組織化したPB空間とフォルマント空間の解析結果である．PB空間の解析手順は次のとおりである．

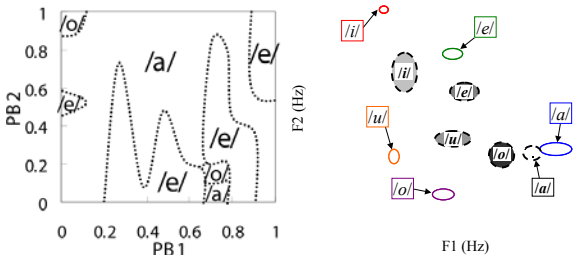

Figure 6　PB space and Formant space after babbling learning

1. PB 空間を10　10 の格子に分割.
2. 分節数N = 1 として，各格子のPB値から300-msecの音声を生成.
3. 生成音の第1，2フォルマントの平均値を計算.
4. Maedaモデル母音と生成音のフォルマント二乗誤差を算出.
5. 二乗誤差最小となるMaedaモデル母音を各格子点の代表母音に設定.

各音素の境界は二乗誤差に応じて曖昧となるが，ここでは便宜上明確な境界を点線であえて示した．この図から/a/，/e/が空間に広く形成されていることが確認できる．これは実際の幼児が生後一年間において発話回数の多い母音と対応している[14].

図7，8に一回目，二回目の追加学習後のPB空間解析結果，フォルマント空間を示す．RNNPBの追加学習一回目では，母音/a/, /e/が広く，バブリング直後より単純な空間が構成される．追加学習二回目で，母音/o/の面積が広がり，母音/u/も新たに出現する．

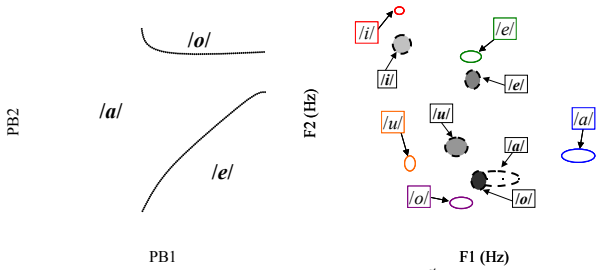

Figure 7　PB space and Formant space after 1st incremental learning

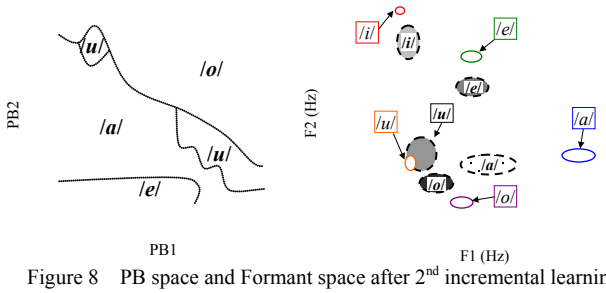

Figure 8　PB space and Formant space after 2nd incremental learning

また各フォルマント空間から，ランダムバブリング直後の模倣音声は，/a/以外の母音について認識音声のフォルマント分布との差が大きい．追加学習一回目後では，母音/i/，/u/，/e/，/o/のフォルマント分布が，認識音声に近付いていることが確認できる．さらに追加学習二回目においては，母音/u/と/o/の模倣音声に対してF1-F2空間上での改善が大きいことが確認できる．

6. まとめ

本稿では，バブリング学習に基づく追加学習母音獲得モデルを構築し，シミュレーションによる母音カテゴリ自己組織化の検証を行った．本実験では，1回の追加学習についての検証のみにとどまっている．今後の課題として，追加学習の反復による母音カテゴリ収束の可能性を示し，音声模倣能力の発達過程の再現を目指す．

謝辞　本研究は，学術創成研究，科研費基盤 B，科研費基盤 S の支援を受けた．

参考文献

[1] N. Masataka and K. Bloom, “Acoustic properties that determine adult's preference for 3-month-old infant vocalization,” Infant behavior and development, vol. 17, pp. 461-464, 1994.
[2] M. Pel´aez-Nogueras, J. L. Gewirtz, and M. M. Markham, “Infant vocalizations are conditioned both by maternal imitation and motherese speech, Infant behavior and development,” vol. 19, p. 670, 1996.
[3] B. de Boer, “Self-organization in vowel systems,” J. Phonetics, vol. 28, no. 4, pp. 441-465, 2000.
[4] P. Y. Oudeyer, “The self-organization of speech sounds,” J. Theoretical Biology, vol. 233, no. 3, pp. 435-449, 2005.
[5] K. Miura and et al., “Vowel acquisition based on visual and auditory mutual imitation in mother-infant interaction,” in proc. of ICDL2006, 2006.
[6] H. Kanda, T. Ogata, K. Komatani, and H. G. Okuno, “Segmenting acousitic signal with articulatory movement using recurrent neural network for phoneme acquisition,” in IEEE/RSJ IROS 2008, 2008.
[7] S. Maeda, “Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model,” Speech production and speech modeling, pp. 131-149, 1990.
[8] N. Kitawaki, F. Itakura, and S. Saito, “Optimum coding of transmission parameters in PARCOR speech analysis synthesis system,” Trans. IEICE Japan, vol. J61-A, no. 2, pp. 119-126, 1978.
[9] H. Kawahara, K. Masuda, and A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction,” Speech Communication, vol. 27, pp. 187-207, 1999.
[10] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. G. Okuno, “Two way translation of compound sentences and arm motions by recurrent neural networks,” in IEEE/RSJ IROS-2007, 2007.
[11] J. Tani and M. Ito, “Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment,” IEEE Trans. On SMC Part A, vol. 33, no. 4, pp. 481-488, 2003.
[12] M. Jordan, “Attractor dynamics and parallelism in a connectionist sequential machine,” in Annu. Conf. Cog. Sci. Soc., 1986, pp. 513-546.
[13] D. Rumelhart, G. Hinton, and R. Williams, “Learning internal representation by error propagation.” MIT Press, 1986.
[14] K. Ishizuka and et al., “Longitudinal developmental changes in spectral peaks of vowels produced by Japanese infants,” J. Acoust. Soc. Am., vol. 121, no. 4, pp. 2272-2282, 2007.