6X-2

RNNPBによる音響模倣・分節化を用いた音素獲得モデルの提案

神田 尚 † 尾形 哲也 † 駒谷 和範 † 奥乃 博 †

† 京都大学大学院情報学研究科知能情報学専攻

1. はじめに

言語の学習において音声模倣は重要な役割を占めている. 例えば, 人の乳児は身体が未発達ながら, 親の発する音声を模倣することができ, 次第に音素を獲得していく.

従来, 音声模倣・音素獲得過程の解明を目的とする認知発達研究において, 言語を認識するためには声道という身体拘束が必要条件としてモデル化を行っている [1, 2]. しかし, これらの研究で扱う音声の単位は音素に限られていた. 我々は「幼児は音素を獲得する以前に連続音響信号の模倣を行う」という仮説に基づき, 神経回路モデル Recurrent Newral Network with Parametric Bias(RNNPB) [3] による音声模倣モデル [4] を提案している. このモデルは, 計算機上の声道モデルが発する音響信号を音素に区切らずに RNNPB で学習し, 人間音声の模倣を行う.

本研究では, この音声模倣モデルに RNNPB の時系列データ分節化手法 [5] を新たに適用し, 連続音響信号の分節化を試みた.

2. 声道モデル

声道モデルには Maeda モデル [6] を用いた. このモデルは, 母音生成時の構音器官の正中矢状面を撮影し, 形状の主成分分析を行った結果得られた 7 つの声道パラメータによって声道器官の形状を表現している. Maeda モデルの 7 次元パラメータを表 1 に示す.

表 1: Maeda モデルのパラメータ

パラメータ番号	パラメータ名
1	Jaw position
2	Tongue dorsal position
3	Tongue dorsal shape
4	Tongue tip position
5	Lip-opening
6	Lip-protrusion
7	Larynx position

3. RNNPB による時系列データの分節化

3.1 RNNPB

RNNPB は, 現状態を入力とし, 次状態を出力とする予測器である (図 1 参照). RNNPB は再帰結合を持ち, 非線形な時系列パターンを学習することができる. さらに, PB 層と呼ばれる入力層を持ち, PB 層の内部値 (PB 値) の変更によって複数の時系列データを生成可能である. また, RNNPB を認識器として用い, 希望する時系列データを生成するような PB 値を得ることが出来る.

3.2 複数シーケンスへの分節化手法

本研究では, 安定した予測が可能な区間を単一のシーケンスとみなして分節化を行う. そこで, 次のような分節化手法を行った.

i) **初期化** : 与えられた時系列データを, 与えられた分節数に対して均等な区間に分割する.

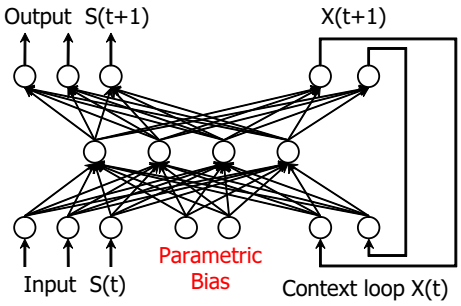

図 1: Recurrent Neural Network with Parametric Bias

ii) **学習** : 各区間について RNNPB の学習を行い, 結合重みと PB 値を更新する.

iii) **誤差算出** : 各区間について予測値を求め, その区間における予測誤差の平均値を求める.

iv) **区間更新** : iii) で誤差が隣接区間のものより大きければ区間幅を減少, 小さければ区間幅を増加させる.

v) 全体の誤差が小さくなるまで ii) から iv) を繰り返す.

単一の PB 値で予測可能なシーケンス (=単一のダイナミクス) では予測誤差が小くなる. 一方, 複数のダイナミクスから成るシーケンスでは予測に失敗し, 誤差が大きくなる. そのため, 上記の手順により, 区間の境界がダイナミクスの境界に近付き, 力学構造に基づいた分節化が可能となる.

4. 音素獲得モデル

本研究では, 音声・声道のダイナミクスの関係に着目し, 音響信号の分節化による音素獲得モデル化した.

本手法は, 既に提案した音声模倣モデル [4] の学習フェーズを基に, 声道モデルに与えた声道動作と, それにより発声した音声データを RNNPB で学習させるとともに, 第 3 節の分節化手法を適用し, 各区間での声道と音声ダイナミクスを結び付け, 力学構造を抽出する. また, これにより抽出した力学構造を音声を認識する際に利用し, 音声模倣が可能になると考えられる.

5. 音声分節化実験

5.1 実験目的

本実験では, 音素獲得手順に従って, 音声・声道のダイナミクスの分節化を行い, 求めた分節位置の検証を行った.

5.2 実験条件

RNNPB への入力には, 三母音 /*aiu*/, /*iue*/, /*ueo*/, /*eoa*/, /*oai*/(各 1380msec) の 5 種類を用いた. 各入力データは, Maeda モデルが生成した音響信号に MFCC 分析 (フィルタバンク:24 次元, 窓幅:250msec, シフト幅:100msec) を行った結果得られた 5 次元の音響特徴量と, 各音響信号に対する 7 次元声道パラメータのうち 6 次元 (表 1 の 1 番から 6 番) を使用し, これらを同期させた 11 次元データである. 各データは, 最小値・最大値により 0 から 1 に正規化を行い, 30msec/step とした. RNNPB の各層数は, 入出力層数 : 11, 中間層数 : 40, 文脈層数 : 25, PB 層数 : 2

Phoneme Acquisition Model by Imitating and Segmenting Sound Signals on The Basis of RNNPB: *Hisashi Kanda, Tetsuya Ogata, Kazunori Komatani, and Hiroshi G. Okuno (Kyoto Univ.)*

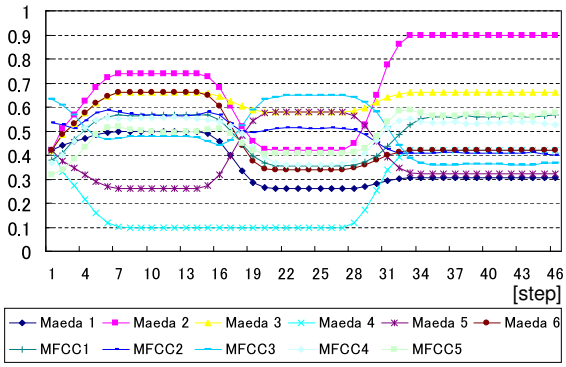

図 2: RNNPB への入力データ /*ueo*/

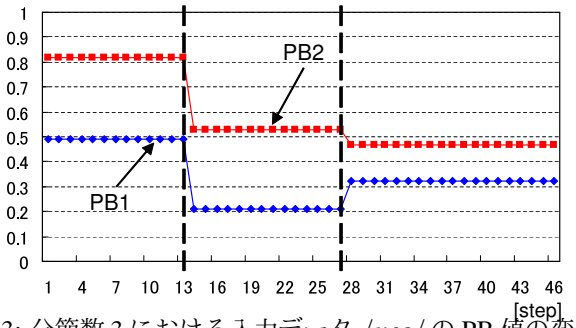

図 3: 分節数 3 における入力データ /*ueo*/ の PB 値の変化

に設定し, 分節数 3 と 8 についてそれぞれ同じ入力データ・学習法で分節化を行った.

5.3 実験結果・考察

5.3.1 音響信号の分節位置と実音素の対応

RNNPB への入力例として, 11 次元データ /*ueo*/ を図 2 に示し, 学習後に得られた PB 値の変化を分節数 3 と 8 についてそれぞれ図 3, 4 に示した. 図 3, 4 の破線は各分節位置を表している.

図 2 より, 入力データ /*ueo*/ の定常母音区間は前から順に 7~14step, 22~27step, 34~46step である. 図 3, 4 をそれぞれ入力データと比較すると, 母音変化に伴って PB 値が変化していることを確認できる. 得られた分節位置は, 分節数 3 のとき 13, 27step, 分節数 8 のとき 5, 13, 15, 24, 27, 29, 44 であり, 分節数 3 の場合は母音が遷移する前に分節位置を決定し, 分節数 8 の場合は母音定常区間と母音遷移区間に分けるように分節位置を決定している. 他の入力データに対しても, ほぼ同様の結果が確認された.

5.3.2 PB 値による音素クラスタリング

図 5 に, 分節数 8 における分節幅の大きい順に 3 つの区間の PB 値を 2 次元 PB 空間として示している. この 3 区間は, 入力データの母音の定常部分に相当する区間であると言える. 各区間の PB2 の値について, 母音 /*i*/ の区間はほぼ 1 に近い値をとり, 母音 /*a*/ と /*o*/ の区間はそれぞれ 0 に近い値, 母音 /*u*/ と /*e*/ の区間はそれぞれ 0.2~0.4 の値をとっており, 同じ母音に対する PB2 の値に対し分布が形成されていることが確認できる. このことは, 分節数 3 の場合には現れず, 各 PB 値が母音を特徴を表すような分布にはならなかった.

以上から, 本手法の分節化により連続音響信号と音素の対応関係を抽出できる可能性を示した. しかし, 実際に音素獲得を目指すには, 各母音と母音間の遷移に対してロバストな拘束条件が必要である. 本研究では, 声道を

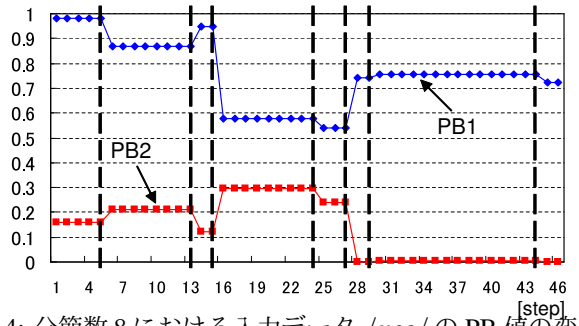

図 4: 分節数 8 における入力データ /*ueo*/ の PB 値の変化

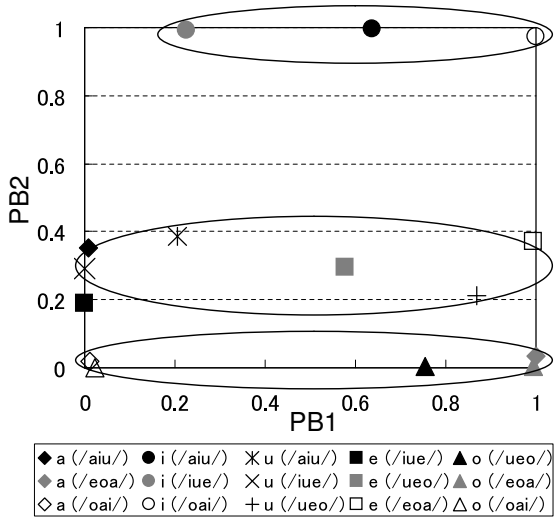

図 5: 分節数 8 における母音定常区間の PB 空間

その拘束条件としているが, さらに, 他者の教示による修正・追加学習によって音素カテゴリを形成していく枠組みが必要であると考えられる.

6. おわりに

本稿では, 声道モデルと神経回路モデルを用いて声道・音声ダイナミクスを分節化を行い, その検証を行った.

今回の実験では, 提案手法による音声・声道ダイナミクスの分節化について検証するのみにとどまっている. 今後の課題として, 分節化による音素獲得の可能性を示し, 3 母音以上の長いシーケンスに対する音声模倣を行う予定である.

謝辞 本研究は, 科研費, GCOE, 栢森情報科学振興財団設立 10 周年記念特別研究助成の支援を受けた.

参考文献

[1] K. Miura, M. Asada and Y. Yoshikawa, “教示者の無意識的引き込み模倣に基づく母音カテゴリの発見”, ロボティクス・メカトロニクス講演会 2007, Vol. CD-ROM, 1A2-L07, 2007.

[2] J. Hörnstein and J. S. Victor, “A Unified Approach to Speech Production and Recognition Based on Articulatory Motor Representations”, IROS 2007, San Diego, Oct., 2007.

[3] J. Tani and M. Ito, “Self-Organization of Behavioral Primitives as Multiple Attractor Dynamics A Robot Experiment”, IEEE Trans SMC Part A: Systems and Humans, Vol.33, No.4, pp.481–488, 2003.

[4] 神田 尚, 尾形 哲也, 駒谷 和範, 奥乃 博, “人間型声道モデルと神経回路モデルを利用した母音模倣”, 第 69 回情処全大, 2007.

[5] 村瀬 昌満, 尾形 哲也, 谷 淳, 駒谷 和範, 奥乃 博, “RNNPB による自然言語列と動作列の意味的結合と人間ロボットインタラクション”, 第 69 回情処全大, 2007.

[6] 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, Kluwer Academic Publishers, pp.131–149, 1990.