1Q-2

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

神田 尚 † 尾形 哲也 ‡ 駒谷 和範 ‡ 奥乃 博 ‡

† 京都大学工学部情報学科 ‡ 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

人間の学習において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でありながら, 親の発声する音声を模倣することができる. このような模倣学習の枠組みを, ロボティクスの観点から構成論的手法によって解明を試みる研究が多く成されている. また, 現在の音声処理のほとんどは音響信号のみに着目しているが, 人間は音声を認識する際, 音響信号のみを認識しているのではなく, それを発声する自らの声道挙動を同時に連想しているのではないかと考えられる. 身体的拘束である声道のなめらかな動作から生み出される音声のダイナミクスを考慮することで, 音素間を自然につなげた音声を扱えることが期待できる.

本研究では, 音と身体のダイナミクスの関係に着目し, 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の実現を目的とする. 実験には, 聴覚空間（音）・構音空間（身体）を有する DIVA モデル [1], 学習器にパラメータ付再帰結合神経回路モデル Recurrent Neural Network with Parametric Bias (RNNPB)[2] を用いた. 複数の音声に対する聴覚・構音ダイナミクスを RNNPB に学習させ, 横矢らの RNNPB による模倣手法 [3] を利用し, 未学習の音声信号に対して連想・生成のシミュレーションを行った.

2. 人間型声道モデル

声道モデルには聴覚空間 (音) と構音空間 (身体) を有する DIVA モデル [1] を用いる. 本来, このモデルの聴覚空間は Miller が提唱した 3 次元座標空間 [4] であるが, 本研究では, 人間の感覚上の音の高さを表す尺度である mel 周波数をケプストラムに変換した Mel-Frequency Cepstrum Coefficient (MFCC) を聴覚空間として用いる. これは, Miller の聴覚空間が音響信号の第 1, 2, 3 フォルマントを利用して定義されており, より多くの音響情報を含む MFCC を聴覚空間として用いることで, Miller の聴覚空間では扱えない音声に対処するためである. 構音空間は, Maeda により提案された声道モデル [5] である. 母音生成時の構音器官の正中矢状面を撮影，形状の主成分分析を行い，7 つの主成分で声道構音器官の形状を表現している. このため, 解剖学的知見に基づく Maeda モデルが身体拘束としての利用に適すると考えられる．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. 神経回路モデル

3.1 Recurrent Neural Network with Parametric Bias

我々は, 谷ら [2] によって提唱された図 1 のような Parametric Bias (PB) を持つ RNN を神経回路モデルとして用いる. 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 値の変更によって 1 つの RNNPB に複数のパターンを埋めこむことが容易になり, 各パターンに対応した PB 値の入力によって, 希望のパターンを出力させることができる.

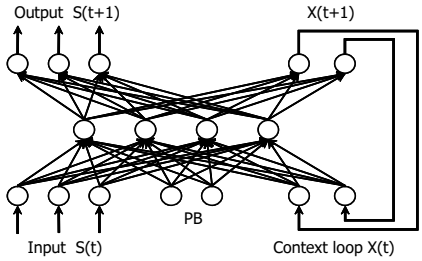

図 1: Recurrent Neural Network with Parametric Bias

3.2 PB 値の学習方法

PB の内部値はニューロンの重み・閾値と同様に, 時刻 t 毎に出力誤差から学習信号 δ_t を求め, それらを Back-Propagation Through Time を行うことにより計算される. また, 通常のニューロン同様, 式 (1) のシグモイド関数を通して出力される. 本研究では PB 値を全ステップ共通にするという制約を加えるため, パラメータの修正量を式 (1) によって与える. ε は学習定数である. このようにして学習されたパラメータ値は各パターンのダイナミクスを保持し, それらを自己組織化させることができる [2].

$$p_i = sigmoid(\rho_i) \quad (1)$$

$$\Delta\rho_i = \varepsilon \cdot \sum_t \delta_{i,t} \quad (2)$$

4. 音声模倣手順

本研究では, 音と身体のダイナミクスの関係に着目することで模倣音声を生成する手法を設計する. 模倣手順の概観を図 2 に示す. 入力データは, sampling 周波数 16kHz の音響信号を 24 チャネルのフィルタバンク分析を行い 5 次元の MFCC に変換した結果得られた特徴量と, 声道モデルとして用いた Maeda モデルの 7 次元パラメータのうち表 1 の 1 番から 6 番のパラメータを各データの最小値・最大値により正規化した値を用いる. これらの各特徴量を同期させ，RNNPB への入力信号として 1step が 20msec の 11 次元ベクトルを得る．本手法は学習・連想・生成の 3 フェーズからなり，概要は以下の通りである.

学習 声道モデルに構音動作を行わせ, 発声した音声データにより RNNPB の重みを学習させる. ここで, 自己の構音動作と聴覚ダイナミクスを結び付ける.

連想 音声ダイナミクスのみを入力として与え, 構音ダイナミクスの連想を行う. 得られたデータにより, 重みを固定した RNNPB を用いて PB 値のみを計算する.

生成 最後に, 連想時に得られた PB 値を学習済の RNNPB に入力することで音声を生成する.

Vowel Imitation using Vocal Tract Model and Neural Network : *Hisashi Kanda (Kyoto Univ.), Tetsuya Ogata (Kyoto Univ.), Kazunori Komatani (Kyoto Univ.), and Hiroshi G. Okuno (Kyoto Univ.)*

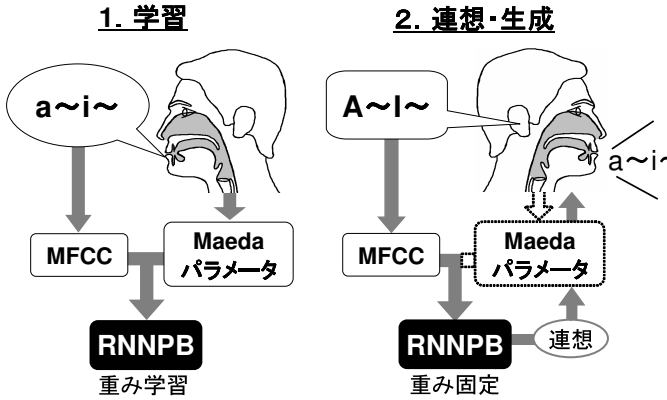

図 2: 模倣手順の概要

5. 音声模倣実験

上記の模倣手順に従って, 学習した母音列の音声ダイナミクスのみから構音ダイナミクスが連想されるかどうかを確認する. また, 未学習の母音列を含む音声についても検討を行う.

5.1 実験条件

三連続母音 *iue*, *ueo*, *oai* の聴覚・構音ダイナミクスを作成し, RNNPB に学習させた. RNNPB の構成は, 入出力層数 : 11, 中間層数 : 20, コンテキスト層数 : 15, PB 層数 : 2 である. 次に, 1 話者が発声した, 三連続母音 1 サンプル (*iue*, *ueo*, *oai*, *iuo*, *iae*, *ioe*, *iua*) から聴覚ダイナミクスを作成し, 連想・生成を行った.

5.2 実験結果・考察

学習・連想の結果得られた 2 次元 PB 空間を図 3 に示す. また, 音声 *iue* について, 原音声及び音声生成時における模倣音声の第 1・第 2・第 3 フォルマント (F1, F2, F3) を図 4 に示す. 図 3 より, 連想により得られた PB 値は, 学習時と同じ音声に対する PB 値付近にプロットされており, 図 4 から模倣音声が原音声のフローをおよそ再現していることが確認できる.

また, 未学習の母音列を含む音声 *iuo*, *iae*, *ioe*, *iua* に対する PB 値は, *iue* 付近にプロットされた. また, 図 5 から, ある程度フローとしては近い音声を再現することはできたが, 人が弁別できるようなはっきりとした音を生成することができなかった.

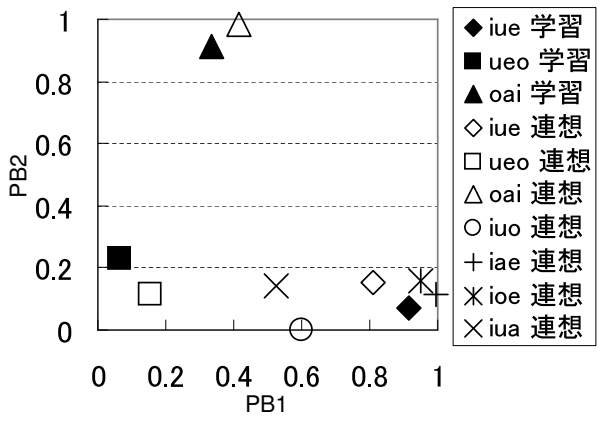

図 3: 三連続母音に対する PB 空間

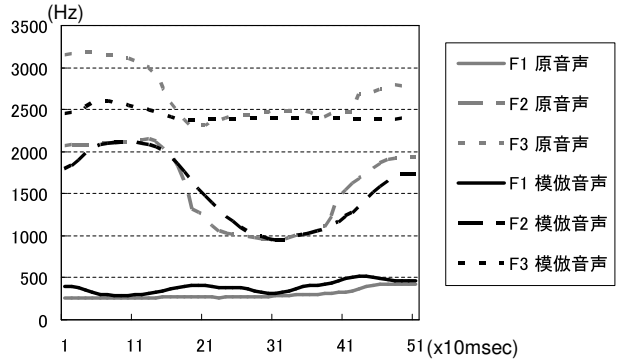

図 4: *iue* の原音声と模倣音声のフォルマント比較

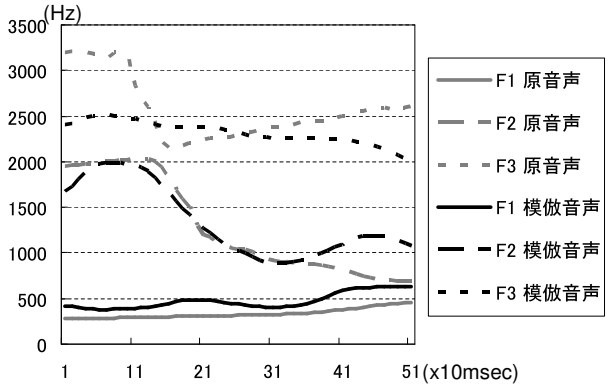

図 5: *iuo* の原音声と模倣音声のフォルマント比較

6. おわりに

本稿では, 聴覚・構音ダイナミクスに着目し, 声道モデルと神経回路モデルを用いた音声模倣モデルの設計と検証を行った. 検証の結果, 学習されたものと同じ母音列に対しては, 精度良く模倣ができた. 未学習の母音列を含む音声については, ある程度似た音声が生成されたものの, 人がはっきりと判別できるまでには至らなかった.

この問題を解決するには, 今回時不変とした PB 値を時系列化する必要がある. そこで, 今後は, 文献 [2, 6] を応用し, PB 値を時系列化することで, 上記の問題に取り組んでいきたい.

謝辞 本研究は科研費萌芽研究 (No.17650051), 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] F. H. Guenther, “A neural network model of speech acquisition and motor equivalent speech production”, Biological Cybernetics, Vol.72, pp.43–53, 1994.

[2] 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.

[3] R. Yokoya and T. Ogata and J. Tani and K. Komatani and H. G. Okuno, “Experience Based Imitation Using RNNPB”, IROS-2006.

[4] J. D. Miller, “Auditory-perceptual interpretation of the vowel”, J. ASA., Vol.85, pp.2114–2134, 1990.

[5] 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.

[6] 松本他, “RNNPB を用いて獲得した疑似シンボルによる人間とロボットの協調の実現”, 情報処理学会第 68 回全国大会, 7L-6, 2006.