声道物理モデルとリカレントニューラルネットワークによる母音模倣

○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)

Vowel Imitation with Physical Vocal Tract Model and Recurrent Neural Network

*Hisashi KANDA†, Tetsuya OGATA†, Kazunori KOMATANI†, Hiroshi G. OKUNO†
(†Graduate School of Informatics, Kyoto University)

Abstract— A vocal imitation system was developed using a computational model that explains the process of language acquisition by infants. A critical problem in vocal imitation is how to generate speech sounds produced by adults, whose vocal tracts have physical properties (i.e., articulatory motions) differing from those of infants' vocal tracts. To solve this problem, a model based on the *motor theory of speech perception*, was constructed. This model suggests that infants simulate the speech generation by estimating their own articulatory motions in order to interpret the speech sounds of adults. Applying this model enables the vocal imitation system to estimate articulatory motions for unexperienced speech sounds that have not actually been generated by the system. The system was implemented by using Recurrent Neural Network with Parametric Bias (RNNPB) and a physical vocal tract model, called the Maeda model. Experimental results demonstrated that the system was sufficiently robust with respect to individual differences in speech sounds and could imitate unexperienced vowel sounds.

Key Words: vocal tract, articulation, imitation, cognitive

1. はじめに

本論文の目的は, 幼児の言語獲得プロセスにおける初期音声模倣モデルの構築とそのモデル検証である. 人の言語獲得過程は, 今だ解明には至っていない. この問題に対する答えの探究は, 言語の起源や構造を理解するための一つのアプローチとして, 様々な学問領域において議論されてきた [1, 2]. 主に, 言語の獲得とは, 語彙や文法の習得を焦点とされているが, それよりずっと以前の段階から, 幼児は発声活動 (クーイング[1] や喃語[2]) を行っている. 実際に, 幼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おり [3], 幼児期の発声発達過程の詳細を知ることは, 言語獲得の基礎を考える上で重要であろう. 近年, このような幼児の模倣学習における認知発達過程を, 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 [4].

音声模倣の従来研究では, 始めに各音素パターンを用意している. 模倣を行う際には, 入力される音声をセグメントに分けて音素パターンと比較し, 最も近いパターンを適応することで原音声の音素系列を求めている. しかし, 連続音声においては, 同時調音の影響から各音素に対して一定のパターンをとらない. これは, 調音器官である声道の連続的な動作によって引き起こされるものである. つまり, 音声模倣を行う際には, 声道という身体拘束を考慮する必要があると考えられる. さらに, 発達心理学によれば, 「幼児の音声言語獲得は, 分節音の獲得の前に語全体の音形の獲得から始まる」[3] とされている. つまり, 幼児が音素単位で音声模倣を行っているとは考えにくい. また, 音声模倣モデルを構築する上で注目すべき知見として, "音声知覚の運動理論" [5] がある. この理論は, 音声を生成する声道が言語知覚の拘束条件であると仮定している.

本研究では, 身体拘束である声道と発声される音声の関連性に着目した "音声知覚の運動理論" をベースとし, 音声を音素に分節化せず一つの時系列ダイナミクスとして扱う音声模倣モデルを提案する. この際, 人間を模した声道モデルを用いることで, 上記の模倣時における身体拘束をシミュレートできると期待できる. 具体的には, 身体拘束である声道の物理モデルとして Maeda モデル [6] を用い, 複数の音声を時系列データとして同時に扱うために, Recurrent Neural Network with Parametric Bias (RNNPB)[7] と呼ばれる神経回路モデルを採用した. RNNPB は, 学習した複数種類の時系列データをパラメータにより同定・生成する機能を有していると同時に極めて優れた汎化能力を有している. 実験としては, 母音音声を対象として提案モデルの基本特性を明らかにする.

以下, 第 2. 章では, "音声知覚の運動理論" を説明し, 第 3. 章では, 使用する声道モデルとリカレントニューラルネットワークモデルについて説明する. 第 4. 章では, システムの設計・実装について詳細に述べる. 第 5. 章で, 提案モデルの検証実験とその結果を示し, 最後に第 6. 章で本稿をまとめる.

2. 音声知覚の運動理論

音声の知覚と生成の関係を重視した理論である「音声知覚の運動理論 (*motor theory of speech perception*, 以下では運動理論)」[8] を以下に説明する.

[1] 生後 2〜3ヶ月の幼児が発する「クー」又は「アー」などの音声.
[2] 乳児のまだ言葉にならない発声. 子音＋母音の構造を持ち, クーイングと区別される.

Table 1 Parameters of the Maeda model

Parameter number	Parameter name
1	Jaw position
2	Tongue dorsal position
3	Tongue dorsal shape
4	Tongue tip position
5	Lip-opening
6	Lip-protrusion
7	Larynx position

音声は, 離散的な記号である音素列が, 調音器官の協調動作により連続的な音に変換されたものである. その結果, 音声は極めて複雑な形をとり, 音素に一対一に対応するような音響的な特徴 (音響的不変量) は見つかっていない [9]. それにもかかわらず, なぜ意図した音素が聞き取れるのかは, 今も未解決の音声知覚の中心的問題である. この問題に対する一つの解答として運動理論が提唱された. 運動理論の基本的な主張は以下の 2 つである.

1. 音声知覚は, 聞き手が積極的に音声を聞き取るための能動的処理であり, 音声のための特別な知覚機構, 「音声モード (speech mode)」が存在する.
2. 音声知覚は音声生成過程を介して行われる.

つまり, 聞き手が「聴取した音を自分が発話するとしたら, どのような調音運動を行うであろうか」という自分の調音器官の運動感覚を作り出すことで, 音声知覚における比較や判断を行っているとする理論である.

運動理論が提唱されて以来, 音声モードの存在に対する疑問も出されている [10]. しかし近年, 認知神経科学の急速な発展によって, 運動理論の神経基盤につながる発見が相次ぎ, 運動理論を支持する知見が多く得られている [11]. 音声模倣モデルを構築する上で, この音声知覚と音声生成の関係は注目すべきものと考えられる.

3. 声道モデルとリカレントニューラルネットワークモデル

3·1 声道物理モデル

声道物理モデルには, Maeda により提案された声道モデル [6] を用いる. 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 つの主成分で声道構音器官の形状を表している.

声道モデルには他にも，PARCOR[12] や極めて高精度な音響再現機能を有する STRAIGHT[13] などがあるが，第 1 章で述べたように，声道モデルを “身体の拘束条件” として利用するため，人間の解剖学的知見に基づく Maeda モデルが適切であると考えられる．Maeda モデルの 7 次元パラメータを表 1 に示す. 各 Maeda パラメータは-3~3 の実数値により表現される.

3·2 リカレントニューラルネットワークモデル

我々は, 谷ら [7] 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. また, 学習に時間がかかる反面，少量の学習データから汎化できるといった利点がある.

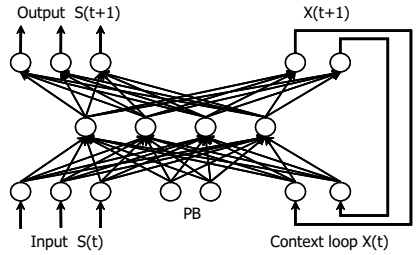


Fig.1 Recurrent Neural Network with Parametric Bias

3·3 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 BackPropagation 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 [7].

$$p_i = sigmoid(\rho_i) \tag{1}$$

$$\Delta\rho_i = \varepsilon \cdot \sum_t \delta_{i,t} \tag{2}$$

4. 音声模倣システム

模倣手順の概観を図 2 に示す. 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, 自己の構音動作と発声音声を結び付ける. これは, 幼児の音声バブリングに相当する.

連想 システムに音声データのみを入力として与え, それに対応する構音動作の連想を行う.

生成 最後に, 連想時に得た構音動作によって模倣音声を生成する.

ここで, 各音声入力に対する適切な構音動作をどのように連想するかが問題となる. そこで我々は, RNNPB を用いて物体動作とロボットの挙動の結びつけを行った Yokoya et al. [14] の手法を適用し, 音声ダイナミクスに対応する構音動作の結びつけを行った.

入力データは, サンプリング周波数 10kHz 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た. 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.

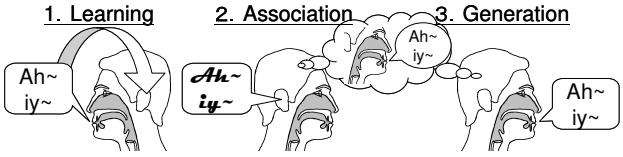


Fig.2 Imitation process.

5. 実験

第 4 章のシミュレーション設計に従って以下の実験を行い, 本手法の妥当性を検証した.

5·1 実験 1：身体拘束を反映した模倣モデルの妥当性評価実験

本実験では, 聴取した音声を同定する際に, 音声情報のみを用いた場合と, 音声 · 声道情報の両方を用いた場合とで比較検証し, 本システムの妥当性を示す. 音声データは, 連続音声として最も短い二連続母音音声を扱った.

5·1.1 実験条件

学習フェーズにおいて, 音声情報 (5 次元 MFCC パラメータ) のみを入力データとして学習した RNNPB (RNNPB–1：入出力層数 5, 中間層数 20, 文脈層数 10, PB 層数 2) と, 音声 (5 次元 MFCC)· 声道 (6 次元 Maeda パラメータ) 情報を入力データとして学習した RNNPB(RNNPB–2：入出力層数 11, 中間層数 20, 文脈層数 10, PB 層数 2) を構成した. 学習に用いた音声は Maeda モデルにより生成した 5 つの二連続母音 $/ai/, /iu/, /ue/, /eo/, /oa/$(各 380msec) である. 次に, 連想フェーズにおいて, 構成済みの RNNPB に, 2 話者の発声した実音声から作成した MFCC パラメータを入力し, 各音声に対する PB 値を求めた. 録音した実音声は, 学習済みの二連続母音パターンと同じものであり, 以下では話者 1, 2 の音声を $/ai_1/, /ai_2/$ のように表す. 録音した音声は 350∼400msec である.

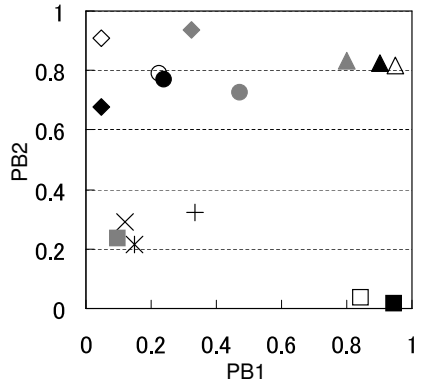
5·1.2 結果 · 考察

RNNPB–1, 2 により得られた二次元 PB 空間を図 3(a), 3(b) に示す. 図 3(a) を見ると, 各 $/ai/, /oa/$ の PB 値が近い位置にプロットされた. また, $/iu_2/$ は $/iu/$ と離れて $/eo/$ に非常に近い位置にプロットされた. これらは, RNNPB–1 が音声パターンを音響的な近さのみによって同定したため, 誤った認識結果になったと考えられる.

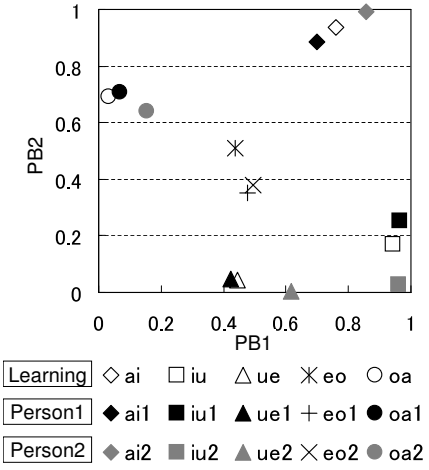
対して, 図 3(b) を見ると, 話者 1, 2 に対する PB 値は, 学習時に得た PB 値とほぼ同じ位置にプロットされており, 実音声入力時のノイズや話者性による音響歪みの影響を吸収していることが確認できる. これは, RNNPB–2 が音響信号のみではなく, それに対応する声道挙動のダイナミクスを同時に学習しているため, 入力音声パターンの変動に対してロバストな連想を行った結果であると解釈できる. 実際に, 例として音響的に近い $/a/$ と $/o/$ を Maeda モデルが発声する際の声道パラメータを表 2(PN：Parameter number) で比較すると, 明確な差が現れている.

Table 2 Comparison between vowel $/a/$ and $/o/$ of Maeda parameters.

PN	1	2	3	4	5	6
$/a/$	-1.5	2.0	0.0	-0.5	0.5	-0.5
$/o/$	-0.7	3.0	1.5	0.0	-0.6	0.0



(a) PB space of RNNPB–1, using only sound information.



(b) PB space of RNNPB–2, using both sound and articulatory information.

Fig.3 PB space.

5·2 実験 2：音声模倣実験

本実験では, 既知 · 未知音声パターンに対する音声模倣をタスクとし, 提案モデルの妥当性を検証する.

5·2.1 実験条件

学習フェーズにおいて, 音声 (5 次元 MFCC)· 声道 (6 次元 Maeda パラメータ) 情報を入力データとして学習した RNNPB (入出力層数 11, 中間層数 20, 文脈層数 10, PB 層数 2) を構成した. 学習に用いた音声は Maeda モデルにより生成した 5 つの二連続母音 $/ai/, /iu/, /ue/, /eo/, /oa/$(各 380msec) である. 次に, 連想フェーズにおいて, 表 3 の 1 話者による実音声 20 パターンから作成した MFCC パラメータを構成済みの RNNPB に入力し各音声パターンに対する PB 値を求め, 生成フェーズで模倣音声を生成した.

Table 3 Recording of two continuous vowels.

Experienced	Unexperienced		
/ai/	/au/	/ae/	/ao/
/iu/	/ia/	/ie/	/io/
/ue/	/ua/	/ui/	/uo/
/eo/	/ea/	/ei/	/eu/
/oa/	/oi/	/ou/	/oe/

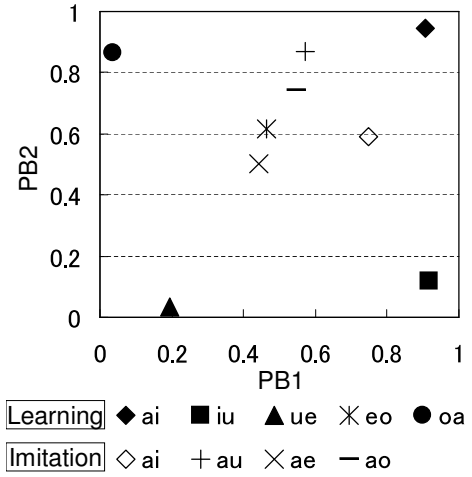


Fig.4 PB space for two continuous vowels: five learned sounds and the four associated sounds, where the first vowel was /a/.

5·2.2 結果 · 考察

図 4 は, 既学習音声と先頭母音が /a/ の未知音声パターンの PB 値をプロットしている. 図 5 は, 既知 · 未知の音声パターン /ai/, /au/ に対する原音声 · 模倣音声の MFCC パラメータを示したものである. 縦軸が MFCC 値, 横軸が時間 (×20 msec) である.

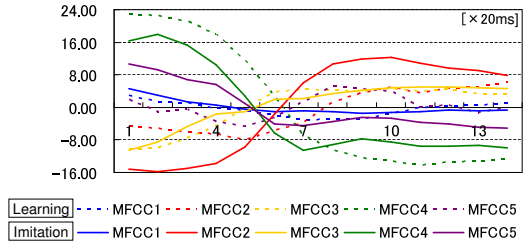
図 5(a) を見ると, 模倣音声が原音声を再現していることが確認できる. 一方, 図 5(b) を見ると, 模倣音声の後半の MFCC パラメータが原音声と異なっているが, 音声は弁別可能な程度に再現されていることが確認できた. また, 模倣音声のほとんどが原音声に似ていることを確認した.

6. まとめ

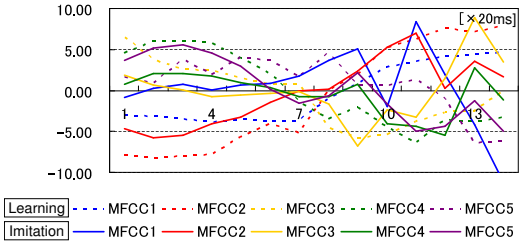
本稿では, 音声をダイナミクスとして扱うことに着目し, 音声ダイナミクスを生成する声道の身体拘束を反映した音声模倣生成モデルの構築と検証を行った. 検証の結果, 身体情報を用いることで, 話者性の音響的な差異に関わらず, 同じ音声パターンを同定することが可能となり, 音声知覚の運動理論を指示する結果が得られた.

参考文献

[1] 須藤珠水, 茂木健一郎. 言語獲得期における語彙学習とカテゴリー認知のメカニズム. 信学技報, SIS2004-4, pp. 17–22, 2004.



(a) MFCC parameters of the original and imitation sound /ai/.



(b) MFCC parameters of the original and imitation sound /au/.

Fig.5 MFCC parameters.

[2] 正高信男. ヒトはいかにヒトになったか. 岩波書店, 2006.
[3] 早川勝廣. 月刊言語. 大修館, 9 2006.
[4] 浅田稔. 模倣発達の構成論的展開. 数理科学, Vol. 40, No. 9, pp. 61–69, 2002.
[5] A. M. Liberman, F. S. Cooper, and et al. A motor theory of speech perception. In *Proc. Speech Communication Seminar, Paper-D3*, Stockholm, 1962.
[6] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In *Speech production and speech modeling*, pp. 131–149. Kluwer Academic Publishers, 1990.
[7] 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.
[8] A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, Vol. 21, pp. 1–36, 1985.
[9] 峯松信明, 西村多寿子, 西成活裕, 櫻庭京子. 構造不変の定理とそれに基づく音声ゲシュタルトの導出. 電子情報通信学会技術研究報告. SP, 音声, Vol. 105, No. 98, pp. 1–8, 2005.
[10] M. E. H. Schouten. The case against a speech mode of perception. *Acta Psychologica*, Vol. 44, pp. 71–98, 1980.
[11] 柏野牧夫. 音声知覚の運動理論をめぐって. 日本音響学会講演論文集, 1-2-10, pp. 243–246, 9 2004.
[12] 東倉洋一, 板倉文忠. Parcor 帯域圧縮方式における音声品質向上. 電子通信学会論文誌, J61-A, No. 3, pp. 254–261, 3 1978.
[13] 河原英紀. 聴覚の情景分析が生んだ高品質 vocoder: Straight. 日本音響学会誌, Vol. 54, No. 7, pp. 521–526, 7 1998.
[14] R. Yokoya, T. Ogata, J. Tani, K. Komatani, and H. G. Okuno. Experience based imitation using rnnpb. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, May 2006.