2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Segmenting Acoustic Signal with Articulatory Movement using Recurrent Neural Network for Phoneme Acquisition

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

Abstract— **This paper proposes a computational model for phoneme acquisition by infants. Human infants perceive speech sounds not as discrete phoneme sequences but as continuous acoustic signals. One of critical problems in phoneme acquisition is the design for segmenting these continuous speech sounds. The key idea to solve this problem is that articulatory mechanisms such as the vocal tract help human beings to perceive speech sound units corresponding to phonemes. That is, the ability to distinguish phonemes is learned by recognizing unstable points in the dynamics of continuous sound with articulatory movement. We have developed a vocal imitation system embodying the relationship between articulatory movements and sounds produced by the movements. To segment acoustic signal with articulatory movement, we apply the segmenting method to our system by Recurrent Neural Network with Parametric Bias (RNNPB). This method determines the multiple segmentation boundaries in a temporal sequence using the prediction error of the RNNPB model, and the PB values obtained by the method can be encoded as kind of phonemes. Our system was implemented by using a physical vocal tract model, called the Maeda model. Experimental results demonstrated that our system can self-organize the same phonemes in different continuous sounds. This suggests that our model reflects the process of phoneme acquisition.**

I. INTRODUCTION

Our goal is to clarify how to acquire the ability to distinguish phonemes in the early period of human infants. Human infants can acquire spoken language through vocal imitation of their parents. Despite their immature bodies, they can imitate their parents' speech sounds by generating those sounds repeatedly by trial and error. This ability is closely related to the cognitive development of language.

Many researchers took notice of the relationship between articulatory movements and sounds produced by the movements. They have designed simulations and robots that duplicate the developmental process of infants' vowel acquisition through vocal imitation [1], [2], [3]. These studies were based on the idea that articulatory mechanisms such as the vocal tract enable us to acquire phonemes, i.e. speech sound in the form of phonemes is characterized by motor articulatory information. This idea has been advocated as the *motor theory of speech perception* [4], and recent neuroscience studies seem to show the idea to be an active process involving motor cognition [5], [6].

Segmenting acoustic signals with articulatory movements is essential for phoneme acquisition; the reason is that human infants do not know the given phonetic distinction inherently. The human development studies described above assume that acoustic signals consist of discrete phoneme sequences in advance, and they search for vocal tract shapes corresponding to phonemes. However, articulatory movements for the same phoneme dynamically change according to the context of continuous speech (e.g. coarticulation). This effect derives from a physical constraint that articulatory movements should be continuous in sound generation. We assume that human infants regard phoneme sequences as continuous acoustic signals. As they grow, infants will acquire the ability to discover phoneme units in a continuous speech sound by prosody, rhythm, stress and whether they can imitate the sound or not.

We use Recurrent Neural Network with Parametric Bias (RNNPB) [7] to segment a continuous temporal sequence consisting of acoustic signal with articulatory movement. From the view point of considering sounds as temporal sequences, we have already developed a vocal imitation system [8], which used the RNNPB model and a physical vocal tract model, called the Maeda model, to simulate the physical constraints. We, furthermore, apply to our system the segmenting method by RNNPB [9]. This method can segment several kinds of sequences into primitive sections using the prediction error of the RNNPB model and encode the segmented sections as a set of parameters, called PB values. It is assumed that the method enables to encode the position of phoneme transition as the segmented sections.

Section II gives an overview of our imitation process, and it describes the vocal tract model and the RNN model. Section III describes our imitation model and the system. Section IV gives the results of experiments with our system. Section V discusses the adequacy of our system as a phoneme acquisition model, and Section VI concludes the paper.

II. VOCAL IMITATION PROCESS AND MODEL

A. Overview of Our Imitation Process

In this section, we present an overview of our system for imitating speech sounds. As illustrated in Fig. 1, our imita-

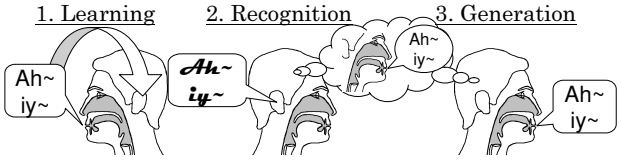


Fig. 1. Imitation process.

H. Kanda, T. Ogata, K. Komatani and H. G. Okuno are with the Graduate School of Informatics, Kyoto University, Kyoto, Japan {hkanda, ogata, komatani, okuno}@kuis.kyoto-u.ac.jp

tion process consists of three phases: learning, recognition, and generation.

1) Learning (Babbling)
   The vowel imitation system makes articulatory movements to produce sounds, and it makes a connection between an articulatory movement and the sound produced by the movement. This phase corresponds to babbling in infants.
2) Recognition (Hearing parents' speech sounds)
   In this phase, we put a speech sound into the system. The system recognizes the sounds with an articulation producing the same dynamics as the heard sound.
3) Generation (Vocally imitating heard sounds)
   Finally, the system uses the articulatory movement to imitate a speech sound.

The learning phase uses the RNNPB method of segmenting temporal sequences. Our imitation model can self-organize so as to connect an articulatory movement with the corresponding sound dynamics. Additionally, in the recognition and generation phases, the connection is available for our model to imitate speech sounds.

B. Physical Vocal Tract Model

A speech production model simulating the human vocal tract system incorporates the physical constraints of the vocal tract mechanism and the acoustic constraints of speech production. The parameters of the vocal tract with physical constraints are better for continuous speech synthesis than acoustic parameters such as the sound spectrum. This is because the temporal change of the vocal tract parameters is continuous and smooth, while that of the acoustic parameters is complex, and it is difficult to interpolate the latter parameters between phonemes.

We used the vocal tract model proposed by Maeda [10]. This model has seven parameters determining the vocal tract shape, and they were derived by principal components analysis of cineradiographic and labiofilm data from French speakers. Table I lists the seven shape parameters. Although there are other speech production models, such as PARCOR [11] and STRAIGHT [12], we think that the Maeda model, with its physical constraints based on anatomical findings, is the most appropriate, because of our aim to simulate the development process of infant's speech. This model for generating acoustic signals is a very simplified articulatory model, and the sound units corresponding to phonemes are expressed in these articulatory terms.

TABLE I
PARAMETERS OF THE MAEDA MODEL.

Parameter number	Parameter name
1	Jaw position (JP)
2	Tongue dorsal position (TDP)
3	Tongue dorsal shape (TDS)
4	Tongue tip position (TTP)
5	Lip opening (LO)
6	Lip protrusion (LPR)
7	Larynx position (LP)

Each Maeda parameter takes on a real value between -3 and 3 and may be regarded as a coefficient weighting an eigenvector. The sum of these weighted eigenvectors is a vector of points in the midsagittal plane that defines the outline of the vocal tract shape. The resulting vocal tract shape is transformed into an area function, which is then processed to obtain the acoustic output and spectral properties of the vocal tract during speech.

C. Learning Algorithm

This subsection describes the method to learn and segment temporal sequence dynamics. We apply the RNNPB model, which was first proposed by Tani [7] as the forwarding forward model. It generates complex movement sequences, which are encoded as the limit-cycling dynamics and/or the fixed-point dynamics of the RNN.

1) RNNPB model: The RNNPB model has the same architecture as the conventional Jordan-type RNN model [13], except for the PB nodes in the input layer. Unlike the other input nodes, these PB nodes take a constant value throughout each temporal sequence and are used to implement a mapping between fixed-length values and temporal sequences. Figure 2 shows the network configuration of the RNNPB model.

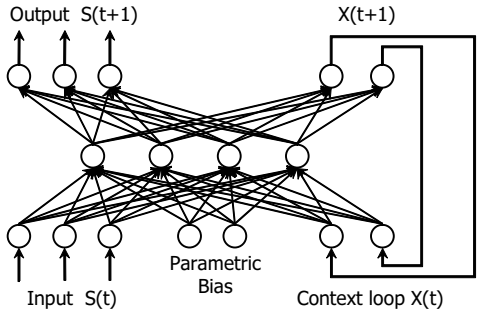


Fig. 2. RNNPB model.

Unlike the Jordan-type RNN model, the RNNPB self-organizes the values in the PB nodes that encode the sequence during the learning process. The common structural properties of the training data sequences are acquired as connection weights by using the back propagation through time (BPTT) algorithm [14], as in a conventional RNN. Meanwhile, the specific properties of each individual temporal sequence are simultaneously encoded as PB values. As a result, the RNNPB model self-organizes a mapping between the PB values and the temporal sequences.

2) Segmenting Temporal Sequence Data: Our segmenting method determines the segmentation boundaries using the prediction error of the RNNPB model. Systems using this approach usually consist of dynamic recognizers that predict the target sequences. The dynamic sequence is articulated based on the predictability of the recognizer. The method we used to segment acoustic signals with articulatory movements uses the prediction error of RNNPB model and the number of segmentations. Its description is as follows: Consider the

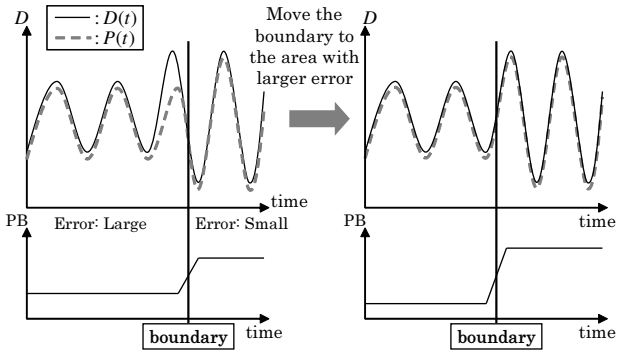


Fig. 3. Segmenting multiple dynamics.

problem of segmenting a dynamic sequence, $D(t)$, whose length is T into N sections, which are represented as S_0, S_1, ..., S_{N-1}. The boundary step between S_{i-1} and S_i is represented by $t = s_i$, that is, S_i is defined as $[s_i, s_{i+1}]$. The segmenting process consists of five steps.

Step 1: Initialization
The given sequence is divided into N sections. Each section has the same length. The boundary step s_i $(i = 0, \cdots, N)$ is set as follows.

$$s_i \longleftarrow i \cdot T/N \tag{1}$$

Step 2: RNNPB training
The connection weights and PB values of the RNNPB model are updated with the given sequence, while the PB values are kept constant in each section, S_i.

Step 3: Calculate prediction errors
In each S_i, the prediction errors of the RNNPB model, P(t), are calculated, and the average error of the section E_i $(i = 0, \cdots, N-1)$ is obtained as follows.

$$E_i \longleftarrow \frac{1}{s_{i+1} - s_i} \cdot \sum_{t \in S_i} ||D(t) - P(t)|| \tag{2}$$

Step 4: Update the length of each section
The boundary step s_i $(i = 1, \cdots, N-1)$ is updated by using the following rules:

$$s_i \longleftarrow \begin{cases} s_i - ds & if\ E_{i-1} \geq E_i \\ s_i + ds & if\ E_{i-1} \leq E_i, \end{cases} \tag{3}$$

where ds is a parameter used to update the section length.

Step 5: Repeat Steps 2 to 4 until the whole error is less than the threshold.

If a sequence is generated by using simple dynamics, the prediction error of the RNNPB will be small, even when the PB values are fixed. However, if a sequence is generated by using multiple dynamics, the prediction error at the boundary between dynamics will increase as shown in Fig. 3. The algorithm can decrease the error by modifying the position of each boundary.

3) Learning of PB Vectors: The learning algorithm for the PB vectors is a variant of the BPTT algorithm. The step length of ith section S_i in a sequence is denoted by $s_{i+1} - s_i$. For each of the articulatory and sound parameters outputs, the back-propagated errors with respect to the PB nodes are accumulated and used to update the PB values. The update equations for the kth unit of the parametric bias at the section S_i in the sequence are as follows:

$$\delta\rho_{i,k} = \varepsilon \cdot \sum_{t=s_i}^{s_{i+1}} \delta_{i,k}(t), \tag{4}$$

$$p_{i,k} = sigmoid(\rho_{i,k} + \delta\rho_{i,k}), \tag{5}$$

where ε is a coefficient. In Eq. 4, the δ force for updating the internal values of the PB $\rho_{i,k}$ is obtained from the sum of the delta errors $\delta_{i,k}$. The delta error $\delta_{i,k}$ is backpropagated from the output nodes to the PB nodes: it is integrated over the period from s_i to s_{i+1} steps. Then, the current PB values $p_{i,k}$ are obtained from the sigmoidal outputs of the updated internal values.

D. Calculation in Recognition and Generation Phases

After the RNNPB model is organized in the learning phase, it is used in the recognition and generation phases.

The recognition phase corresponds to how infants recognize sounds presented by parents, i.e. to how the PB values are obtained. The PB values of each section are calculated from Eq. 4 and 5 by using the organized RNNPB without updating the connection weights. However, there is no vocal tract data because the system is only hearing sounds without articulating them, unlike in the learning phase. The initial vocal tract values are input to vocal tract units of the input layer in step 0, and the outputs are calculated forward in the closed-loop mode from step 1. More generally, the outputs in the motion output layer in step $t-1$ are the input data in the motion input layer in step t.

The generation phase corresponds to what articulation values are calculated. The motion output of the RNNPB model is obtained in a forward calculation. The PB values obtained in the recognition phase are input to the RNNPB in each step.

III. VOCAL IMITATION SYSTEM

A. Experimental System

Our experimental system is illustrated Fig. 4. This system was used to verify the relation between vocal imitation and the phoneme acquisition process. To simplify the system, we purposely used a simple vocal tract model and target vowel sound segmentation.

In the learning phase, we first use a cubic interpolation method to produce sequences of vocal tract parameters for the Maeda model as articulatory movements. Second, the sequences are put into the Maeda model to produce the corresponding sounds, which are then transformed into temporal sound parameters. Finally, the RNNPB learns each the sound and the vocal tract parameters, which are normalized and synchronized. In this phase, the parameter ds was set at 0.1.

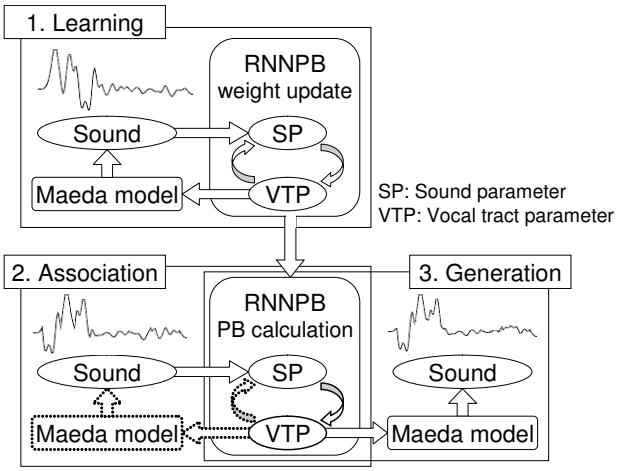


Fig. 4. Diagram of the experimental system.

The size of the RNNPB model and the time interval of the sequence data differed according to the experiment.

In the recognition phase, speech sound data is put into the system. The corresponding PB values are calculated for the given sequence by the organized RNNPB in order to associate the articulatory movement with the sound data.

In the generation phase, the system generates imitation sounds by inputting the PB values obtained in the recognition phase into the organized RNNPB.

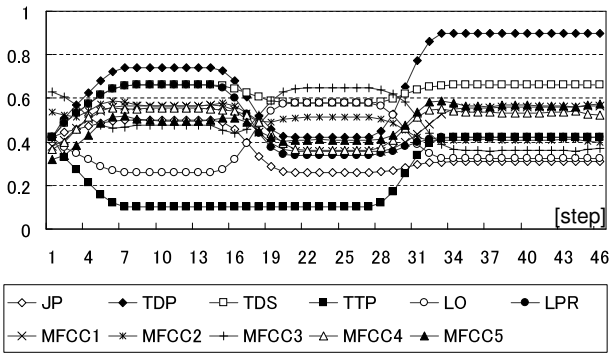
B. Sound Parameters

To convert a speech waveform into feature parameters, we use the Mel-Frequency Cepstrum Coefficient (MFCC). Filters spaced linearly at low frequencies and logarithmically at high frequencies capture the phonetically important characteristics of speech.

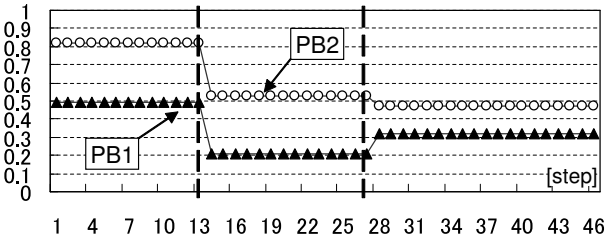
In the experiments, the speech signals were single channel with a sampling frequency 10 kHz. They were analyzed using a Hamming window with a 25-ms frame length and a 10-ms frame shift, forming five-dimensional MFCC feature vectors. The number of mel filterbanks was 24. In addition, a Cepstrum Mean Subtraction was applied to reduce linear channel effects.

C. Vocal Tract Parameter

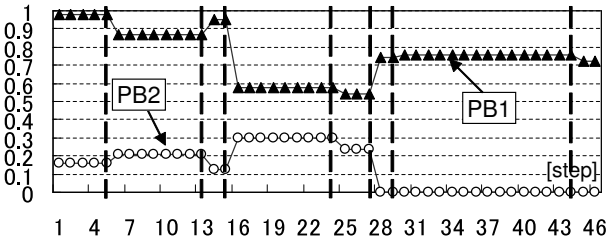
We applied the Maeda model with the first six parameters listed in Table I. The reason for choosing only these six parameters is that when the Maeda model produces vowel sounds, the seventh parameter LP has a steady value. In the generation phase, it is possible for the vocal tract parameters produced by the RNNPB to temporally fluctuate without human physical constraints. This occurs if the system does not easily associate the articulatory movements of an unexperienced sound. Therefore, to help prevent extraordinary articulation, we temporally smoothed the vocal tract parameters produced by the RNNPB. Concretely, the vocal tract parameters in each step were calculated by averaging those of the adjacent steps.



(a) Input data: /*ueo*/.



(b) The sequence of PB values for input data /*ueo*/ for $N = 3$.



(c) The sequence of PB values for input data /*ueo*/ for $N = 8$.

Fig. 5. Input data and obtained PB values in the learning phase.

IV. EXPERIMENTS

A. Model Verification by Segmenting Three-Vowel Data

We verified the capability of the segmenting method based on an experiment altering the number of segmentations N from three to eight. We assumed that our system did not know the number of phonemes in the input data. The organization of RNNPB for each N is as follows: 11 input/output nodes, 40 hidden nodes, 25 context nodes, and 2 PB nodes. The learning data consisted of the following five patterns of three-vowel data: /*aiu*/, /*iue*/, /*ueo*/, /*eoa*/, and /*oae*/ (1380 ms, 30 ms/step), produced by the Maeda model.

Figure 5(a) shows the learning data /*ueo*/. Figure 5(b), 5(c) shows the sequence of PB values for the leaning data /*ueo*/ obtained by organized RNNPB. The vertical dotted line represents the boundary step s_i segmented by RNNPB in the learning phase. Figure 5(b) shows PB values of /*ueo*/ for $N = 3$, and Fig. 5(c) shows those for $N = 8$. The boundary steps, representing the steps just before the transitions in the

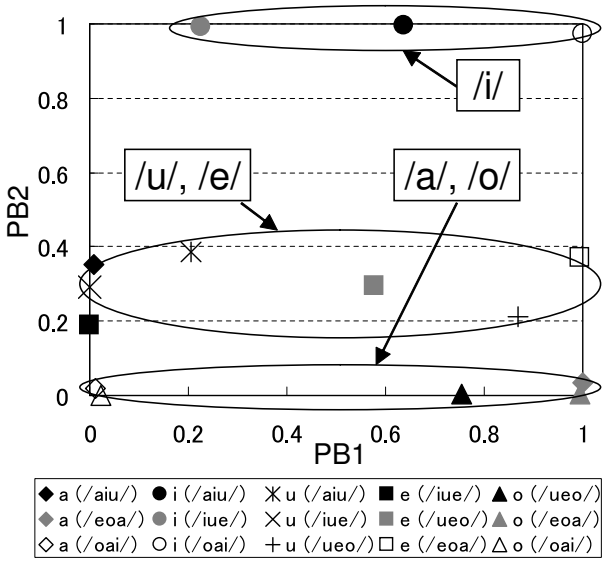


Fig. 6. The PB space for each flat part in input sequences for N = 8.

input sequence, in Fig. 5(b) were $s_1 = 13$ and $s_2 = 27$. Those in Fig. 5(c) were $s_1 = 5$, $s_2 = 13$, $s_3 = 15$, $s_4 = 24$, $s_5 = 27$, $s_6 = 29$ and $s_7 = 44$ dividing the input sequence /*ueo*/ into flat and transition segments. We confirmed that as the size of N increases, the boundary steps become more stable in the learning phase. Similar results were also acquired for the other input data.

Figure 6 shows the PB space for $N = 8$. In Fig. 6, the PB values represent the phonemes of a set of three-vowel data aligned according to the length of the three longest sections of an input sequence. The PB space has a tendency to classify the PB values according to PB2 in the following three categories: /*a*/ and /*o*/, /*u*/ and /*e*/, and /*i*/. A comparable result had not been acquired for other sizes of N.

B. Segmentation for Phoneme Acquisition

Next, we carried out an experiment to verify phoneme acquisition using our imitation model in the learning phase. The organization of RNNPB is as follows: 11 input/output nodes, 50 hidden nodes, 10 context nodes, and 2 PB nodes. In this experiment, the parameter ds was set to 0.1. RNNPB learned the MFCC and vocal tract parameters of ten patterns of three-vowel data: /*aiu*/, /*aoe*/, /*iue*/, /*iao*/, /*ueo*/, /*uia*/, /*eoa*/, /*eui*/, /*oai*/, and /*oeu*/ (1380 ms and 30 ms/step), produced by the Maeda model.

Figure 7 shows the PB space after learning. In Fig. 7, the PB values represent the phonemes of a set of three-vowel data aligned according to the length of the three longest sections of a learning sequence. The PB values for the same vowel, including the learning data, were mapped with sufficient dispersion.

Figure 8 shows the transition of the PB values for the input data /*eui*/ and /*uia*/ in the learning phase. In Fig. 8, the PB values of section S_2 for input data /*uia*/ were close to those

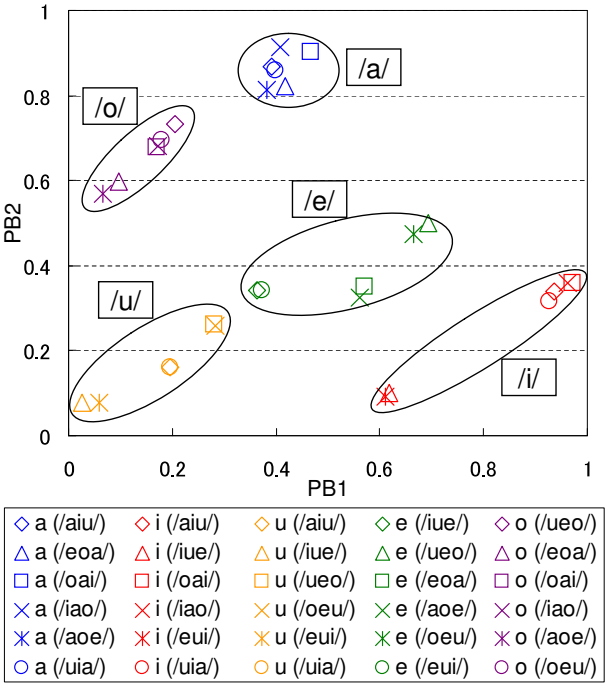


Fig. 7. The PB space in the learning phase.

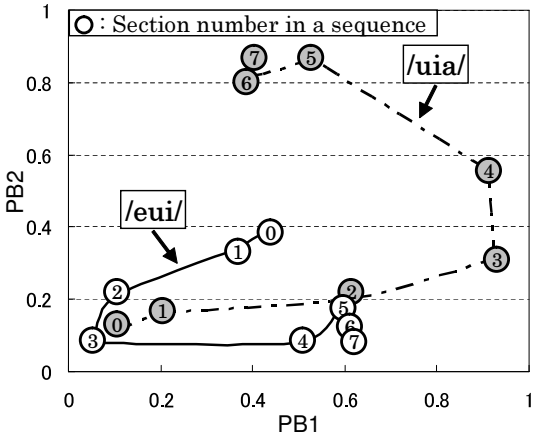


Fig. 8. Changing PB values for input sequence data /*eui*/ and /*uia*/ in the PB space.

of the sections $S_{5,6,7}$ for input data /*eui*/. When comparing Fig. 7 and 8, we confirmed that the category of the phoneme /*i*/ in Fig. 7 corresponded to the transitions of the PB values in Fig. 8.

V. DISCUSSION

A. Segmentation ability by RNNPB

In the experiment IV-A, the number of segmentation boundaries was set arbitrarily for given continuous sounds including unknown number of phoneme. As a result of learning phase where the positions of the segmentation

boundaries self-organized, the almost boundaries tended to gather to the transition phases of the phonemes given in an input sequence (see Fig. 5(c)). Several sections are defined by these boundaries, and it was confirmed that length of some sections of phoneme parts are significantly longer than that of transition parts between phonemes. It is also confirmed that the obtained PB values of the sections corresponds to given phonemes.

HMM is one of the representative methods to segment sounds into phonemes. However, the phoneme categories should be given in advance, and a large amount of learning data is required. Our method determines the segmentation boundaries using the prediction error of the RNNPB model. This enables our method to obtain the position of phoneme transitions and the phoneme categories, without the information of the numbers and categories of phoneme in the input acoustic signals. Furthermore, it needs only a small amount of learning data to organize phoneme system in PB space.

B. Context dependency for each sound

Our system could encode the same phonemes in acoustic signals as the near PB values in the PB space. In this sense, each phoneme category is defined independently from the other phonemes. However, in Fig. 7, it is confirmed that each phoneme category $/i/$, $/u/$ and $/e/$ formed a plot but a small cluster consisted of multiple plots. In Fig. 8, the transitions of PB values pass through different points in the same phoneme categories. This means that the PB values representing the same phoneme are changed by the adjacent phonemes in a given phoneme sequences. It is assumed that this represents coarticulation designed in general speech recognition systems. In this sense, each phoneme is determined context dependently on the other phonemes.

Tani et al. showed that the internal symbolic process, being embedded in the dynamical attractor in a mobile robot system [15]. In his experiment, the robot acquired the attractors representing the observed objects as the activities in RNN nodes. These attractors were also represented by complex clusters, and the positions of the active points were fluctuated by the context, i.e. trajectory of the mobile robot. This bilateral characteristic, that is context dependency or independency, is one of the interesting and essential properties in dynamical systems representation.

It is confirmed that Fig. 7 of the obtained PB space corresponds to the map of "vowel triangle" shown in [16]. Concretely, PB1 and PB2 corresponds to 1st and 2nd formant frequency respectively. Our model uses Maeda model of which vocal tract parameters include the property of PARCOR [11]. The property could extract formant peak frequency in a sound and contribute to organize the relation between PB values and formant frequencies.

VI. CONCLUSIONS

This paper proposed a phoneme acquisition system focusing on segmentation of the dynamic sequences of acoustic signals with the articulatory movements generated by the Maeda model. Concretely, our model uses a RNNPB model trained with several acoustic sequences and articulatory movements including unknown numbers and kinds of phonemes. The experimental results demonstrated that our system with RNNPB model automatically found the segmentation boundary of the phonemes and found phonemes were encoded as the PB values.

Our future work includes to imitate speech sounds using automatically extracted PB values corresponding to phonemes from speech sounds through simulating mother and child interaction. The acoustic babbling should be introduced into our model as the exploring and learning phase of corresponding between generated acoustic signal and articulatory movements.

VII. ACKNOWLEDGMENTS

This research was partially supported by Global COE, the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (S), Grant-in-Aid for Young Scientists (A), and Kayamori Foundation of Informational Science Advancement.

REFERENCES

[1] B. de Boer, "Self-organization in vowel systems," *Journal of Phonetics*, vol. 28, no. 4, pp. 441–465, October 2000.

[2] P. Y. Oudeyer, "The self-organization of speech sounds," *Journal of Theoretical Biology*, vol. 233, no. 3, pp. 435–449, 2005.

[3] K. Miura, M. Asada, K. Hosoda, and Y. Yoshikawa, "Vowel acquisition based on visual and auditory mutual imitation in mother-infant interaction," in *ICDL2006*, 2006.

[4] A. M. Liberman, F. S. Cooper, and et al., "A motor theory of speech perception," in *Proc. Speech Communication Seminar, Paper-D3*, Stockholm, 1962.

[5] L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolatti, "Speech listening specifically modulates the excitability of tongue muscles: a tms study," *European Journal of Cognitive Neuroscience*, vol. 15, pp. 399–402, 2002.

[6] G. Hickok, B. Buchsbaum, C. Humphries, and T. Muftuler, "Auditory-motor interaction revealed by fmri," *Area Spt. Journal of Cognitive Neuroscience*, vol. 15, no. 5, pp. 673–682, 2003.

[7] J. Tani and M. Ito, "Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment," *IEEE Transactions on SMC Part A*, vol. 33, no. 4, pp. 481–488, 2003.

[8] H. Kanda, T. Ogata, K. Komatani, and H. G. Okuno, "Vocal imitation using physical vocal tract model," in *IEEE/RSJ IROS2007*, 2007.

[9] 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 IROS2007*, 2007.

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

[11] N. Kitawaki, F. Itakura, and S. Saito, "Optimum coding of transmission parameters in parcor speech analysis synthesis system," *Transactions of the Institute of Electronics and Communication Engineers of Japan (IEICE)*, vol. J61-A, no. 2, pp. 119–126, 1978.

[12] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 1997, pp. 1303–1306.

[13] M. Jordan, "Attractor dynamics and parallelism in a connectionist sequential machine," in *Eighth Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale, NJ, 1986, pp. 513–546.

[14] D. Rumelhart, G. Hinton, and R. Williams, *Learning internal representation by error propagation*. Cambridge, MA, USA: MIT Press, 1986.

[15] J. Tani, "Model-based learning for mobile robot navigation from the dynamical systems perspective," *IEEE Trans. on Systems, Man, and Cybernetics Part B: Cybernetics*, vol. 26, no. 3, pp. 421–436, 1996.

[16] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-hall, 1978.