Proceedings of the 2007 IEEE/RSJ International
Conference on Intelligent Robots and Systems
San Diego, CA, USA, Oct 29 - Nov 2, 2007

Vocal Imitation Using Physical Vocal Tract Model

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

Abstract— **A vocal imitation system was developed using a computational model that supports the *motor theory of speech perception*. 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.**

I. INTRODUCTION

Our final goal is to clarify the development process in the early-speech period of human infants. In this paper, we mainly focus on their vowel imitation using computational model that supports the *motor theory of speech perception*. The target are primitive utterances such as cooing[1] or babbling[2] before infants utter first words.

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 is closely related to the cognitive development of language. Recently, many researchers have designed robots that duplicate the imitation process of human infants in terms of the constructive approach.

Typical methods of vocal imitation using vocal tract models first segment speech signals into multiple units of phonemes and then learn the corresponding vocal tracts shapes. To imitate a target speech signal, these fixed units are concatenated in an appropriate order so that a generated speech signal resembles the target signal. Therefore, it is necessary to interpolate adjacent units that are individually learned. This does not, however, reflect the articulatory mechanism of humans. Articulatory motions for the same phoneme are dynamically changed according to the context of continuous speech, (e.g. coarticulation). This effect derives from a physical constraint that articulatory motions should be continuous in sound generation. Therefore, we should reflect this constraint in vocal imitation.

In this study, we propose a speech imitation model based on the *motor theory of speech perception* [1], which was developed to explain why speech sound (in the form of phonemes) is characterized by motor articulation information. This model is based on the observation that human infants actively imitate the speech sounds of their parents by distinguishing between imitable and unimitable features. The model captures sounds not as a set of phonemes but as temporal dynamics. To apply this model, we use Recurrent Neural Network with Parametric Bias (RNNPB) [2] and an anatomic vocal tract model, called the Maeda model, to recreate physical constraints.

In the remainder of this paper, section II introduces the *motor theory of speech perception*. Section III describes the vocal tract model and RNN model used as the learning method. Section IV describes our imitation model and system. Section V gives the results of some experiments with our proposed method. Section VI discusses the adequacy and generalization capabilities of our system as an imitation model, and section VII concludes the paper.

II. MOTOR THEORY OF SPEECH PERCEPTION

In this section, we describe the *motor theory of speech perception* with consideration of the association between speech perception and production in speech communication.

Speech is formed by complex cooperative action of the articulatory organs transforming a sequence of discrete phonetic units into continuous sounds. As a result, speech has a complicated configuration, and no acoustic invariants corresponding with phonemes have ever been found [3]. Nevertheless, human beings can hear the intended phonetic gestures of a speaker. The *motor theory of speech perception* was proposed as an answer to this question. This theory basically insists on the following two propositions.

1) Speech perception is active processing for the listener, and there is a special sensory mechanism for speech sound, called "speech mode."
2) Speech perception is executed through the speech production process.

In other words, we can make sense out of what we hear because we guess how the sounds are produced. Although this motor theory has been controversial, recent neuroimaging studies seem to support the idea of perception as an active process involving motor cognition [4]–[6].

Starting from the *motor theory of speech perception*, we propose that the motor information in speech, which enables

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

[1]The murmuring sound of a dove or a sound resembling it.

[2]A meaningless confusion of words or sounds.

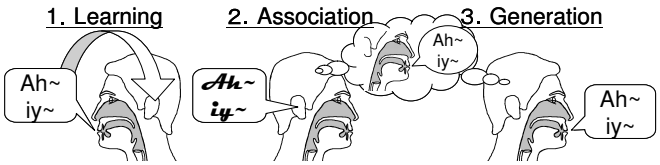


Fig. 1. Imitation process.

the recovery of articulatory motions, enables the vocal imitation required for infants to learn spoken vocabulary. This function is essential for subsequent processes such as word identification.

III. VOCAL IMITATION SYSTEM

A. Overview of Our Imitation Process

In this section, we present an overview of our system imitating the sound of a human voice. As illustrated in Fig. 1, our imitation process consists of three phases: learning, association, and generation. The system executes the following tasks.

1) Learning (Babbling)
   The vowel imitation system make articulatory motions to produce sounds, and it acquires the mapping between motions and sounds. This phase corresponds to babbling in infants.
2) Association (Hearing parents' speech sounds)
   In this phase, a speech sound is input to the system. The system associates the sounds with an articulation producing the same dynamics as the heard sound.
3) Generation (Vocally imitating heard sounds)
   Finally, the system use the articulatory motion to produce a imitation speech sound.

In this process, one problem is how to get an appropriate articulation from a speech sound input. We need a method of connecting an articulatory motion with the corresponding sound dynamics. To solve this problem, we use the method proposed by Yokoya et al. [7], which connects a robot motion with an object motion via RNNPB, to connect articulatory motions with sound dynamics.

B. Physical Vocal Tract Model

A speech production model simulating the human vocal tract system incorporates the physical constraints of the vocal tract mechanism. 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.

In this study, we use the vocal tract model proposed by Maeda [8]. This model has seven parameters determining the vocal tract shape, which 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 [9] and STRAIGHT [10], we think that Maeda model, with physical constraints based on anatomical findings, is the most appropriate, because of our aim to simulate the development process of infant's speech.

TABLE I
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

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, which defines the outline of the vocal tract shape. The resulting vocal tract shape is transformed into a vocal tract 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 a method that enables our imitation model to learn temporal sequence dynamics. For this method, we apply the FF-model (forwarding forward model) proposed by Tani [2], which is also called RNN with Parametric Bias (RNNPB) model.

1) RNNPB model: The RNNPB model has the same architecture as the conventional Jordan-type RNN model [11], except for the PB nodes in the input layer. Unlike the other input nodes, these PB nodes take a constant value throughout each time sequence Figure 2 shows the network configuration of the RNNPB model. The RNNPB model works as a prediction system: its input data is current sensory state S(t) and its output data is predicted sensory state S(t+1) in the next step. The context layer has a loop that inputs current output as input data in the next step.

After learning time sequences using the back propagation through time (BPTT) algorithm [12], the RNNPB model self-organizes the PB values at which the specific properties of each individual time sequence are encoded. As a result, the RNNPB model self-organizes a mapping between the PB values and the time sequences. In our study, input data S(t) are articulatory and sound parameters in time t, and one pair of the PB values represents a time sequence of an articulatory motion and sound by the motion.

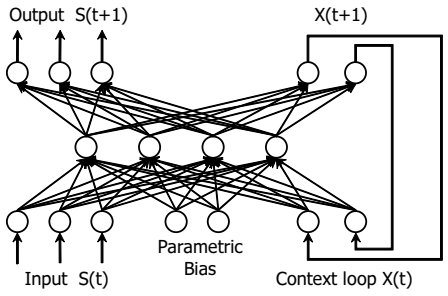


Fig. 2. RNNPB model.

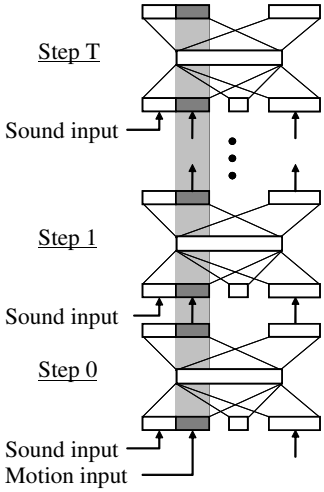


Fig. 3. Forward calculation of PB values.

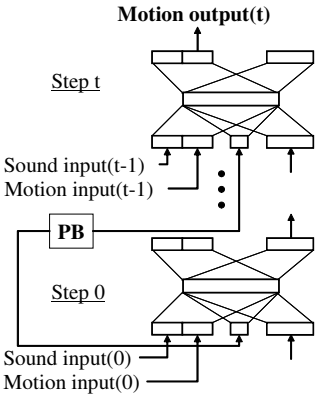


Fig. 4. Articulatory motion generation.

2) Learning of PB Vectors: The learning algorithm for the PB vectors is a variant of the BPTT algorithm. The length of each sequence is denoted by T. For each of the articulatory parameters outputs, the backpropagated errors with respect to the PB nodes are accumulated and used to update the PB values. The update equations for the ith unit of the parametric bias at t in the sequence are as follows:

$$\delta\rho_i = \varepsilon \cdot \sum_{t=0}^{T} \delta_i(t), \quad (1)$$

$$p_i = sigmoid(\rho_i), \quad (2)$$

where ε is a coefficient. In Eq. 1, the δ force for updating the internal values of the PB p_i is obtained from the summation of the delta error δ_i. The delta error δ_i is backpropagated from the output nodes to the PB nodes: it is integrated over the period from the 0 to T steps. Then, the current PB values are obtained from the sigmoidal outputs of the internal values.

D. Calculation in Association and Generation Phases

After the RNNPB model is organized via the BPTT and the PB values are calculated in the learning phase, the RNNPB model is used in the association and generation phases. This subsection describes how the RNNPB model is used in the latter two phases.

The association phase corresponds to how infants recognize the sound presented by parents, i.e., to how the PB values are obtained. The PB values are calculated from Eq. 1 and 2 by the organized RNNPB without updating the connection weights. At this point, 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 the motion 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, as illustrated in Fig.3. Put simply, the motion input layer plays the same role as the context layer does.

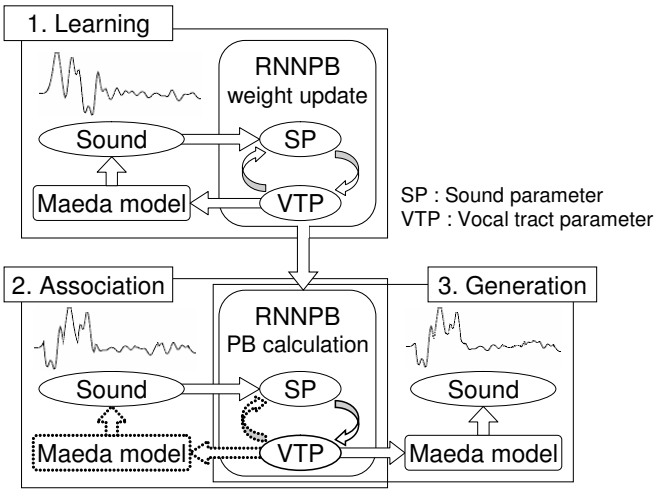


Fig. 5. Diagram of the experimental system.

The sound generation phase corresponds to what articulation values are calculated, as shown in Fig 4. The motion output of the RNNPB model is obtained in a forward calculation. The PB values obtained in the association phase are input to the RNNPB in each step.

IV. MODEL AND SYSTEM

A. Experimental System

In this subsection, we describe our experimental system, which is illustrated Fig.5. This system model was used to verify the relation between vocal imitation and the phoneme acquisition process according to the *motor theory of speech perception*. To simplify the system, we purposely used a simple vocal tract model and target vowel sound imitation.

In the learning phase, several articulatory motions are put into Maeda model, and learn temporal sequence dynamics of an articulatory motion and the speech sound for the motion by RNNPB. We first decide arbitrarily motion parameters: initial values of each motion parameters are all zero, and we produce sequences of vocal tract parameters by interpolating some vowel parameters, which are already known. 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 set of the vocal tract and sound parameters, which are normalized and synchronized. The size of the RNNPB model and the time interval of the sequence data differed according to the experiment. In the association phase, we put speech sound data into the system. The corresponding PB values are calculated for the given sequence by the organized RNNPB to associate the articulatory motion for the sound data. In the generation phase, the system generates these imitation sounds by inputting the PB values obtained in the association phase into the organized RNNPB.

B. Sound Parameters

To convert a speech waveform into feature parameters, we use the Mel-Frequency Cepstrum Coefficient (MFCC), which is based on the known frequency variation of the human ear's critical bandwidths. 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 10kHz. They were analyzed using a Hamming window with a 40-ms frame length and a 17-ms frame shift, forming five-dimensional MFCC feature vectors. The number of mel filterbanks was 24. In addition, Cepstrum Mean Subtraction (CMS) [13] was applied to reduce linear channel effects.

C. Vocal Tract Parameter

In the experiments, we applied the Maeda model - with the first six parameters listed in Table I. When Maeda model produces vowel sounds, the seventh parameter 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 articulation for an unexperienced sound. Therefore, to help prevent extraordinary articulation, we execute temporal smoothing of the vocal tract parameters produced by the RNNPB. Concretely, the vocal tract parameters in each step are calculated by averaging those of the adjacent steps.

V. EXPERIMENTS

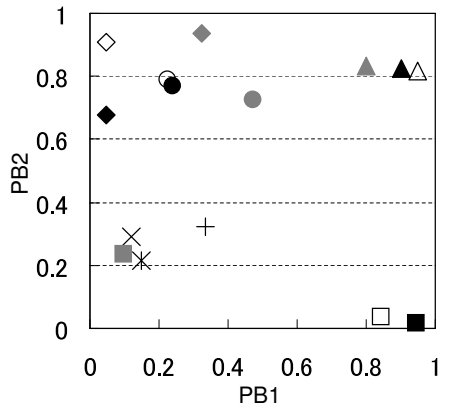
A. Model Verification by Two Continuous Vowels

We carried out this experiment to verify the adequacy of our system by comparing the use of sound and articulatory information with the use of only sound information.

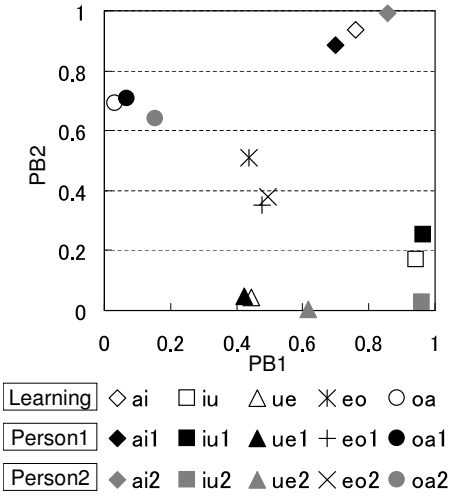
For the experiment, we organized two RNNPBs. One, called RNNPB–1, learned only the MFCC parameters as sound information. The input and output layers had five units, the hidden layer had 20 units, the context layer had 10 units, and the PB layer had two units. The other, called RNNPB–2, learned both the MFCC and vocal tract parameters as sound and articulatory information. The input and output layers had 11 units, the hidden layer had 20 units, the context layer had 15 units, and the PB layer had two units. The learning data consisted of the following vowels: $/ai/$, $/iu/$, $/ue/$, $/eo/$, and $/oa/$ (380 ms, 20 ms/step), produced by the Maeda model. In the association phase, We inputted MFCC parameters, which were produced by recording the speech sounds of two speakers, into each organized RNNPB. Each RNNPB obtained the PB values from each set of sound data. The recording data used the same vowels as those in the learning data. In the following, we describe the association data of one person with the additional character '1', e.g., $/ai_1/$, and that of the other person with the additional character '2', e.g., $/ai_2/$. Figure 6 shows the PB space acquired by each organized RNNPB. The two parametric values in the RNNPBs correspond to the X–Y axes.

Figure 6(a) shows the PB space when only sound information was used. Although some of the PB values for the same vowel sounds were closely mapped, $/ai/$ and $/oa/$ was not clearly classified, and $/iu_2/$ had been confused with $/eo/$.

Meanwhile, Fig. 6(b) shows the PB space when both sound and articulatory information was used. The PB values



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



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

Fig. 6. PB space.

for the same vowel sounds, including the learning data, were mapped with sufficient dispersion. We confirmed that RNNPB–2 could recognize the vowel sounds correctly. As we can see from table II, there are sharp differences between vocal tract parameters of $/a/$ and $/o/$, which are acoustically similar. In fact, it is said that articulation information helps human beings to recognize speech sounds.

TABLE II
PARAMETERS OF VOWEL $/a/, /o/$ FOR THE MAEDA MODEL.

Parameter number	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

B. Vocal Imitation

We next carried out an experiment to verify the adequacy of our imitation model by having it imitate both experienced and unexperienced sounds.

1) Imitation of Two Continuous Vowels: In the learning phase, we organized the following RNNPB: the input and output layers had 11 units, the hidden layer had 20 units, the context layer had 15 units, and the PB layer had two units. The RNNPB learned the MFCC and vocal tract parameters

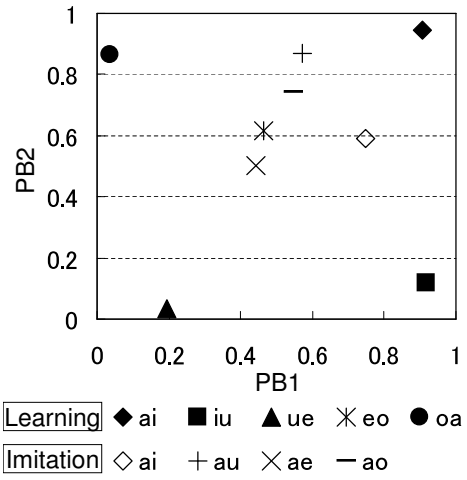


Fig. 7. PB space for two continuous vowels: five learned sounds and the four associated sounds, where the first vowel was /a/.

of the learning data (/ai/, /iu/, /ue/, /eo/, and /oa/, 320 ms and 20 ms/step), produced by the Maeda model. In the association phase, we inputted the MFCC parameters, generated by recording the speech sounds of a person, into the organized RNNPB, and we obtained the PB values for each of the sounds. Table III summarizes the recording two continuous vowels sounds. In the generation phase, we used the PB values to reproduce each of the recording sounds.

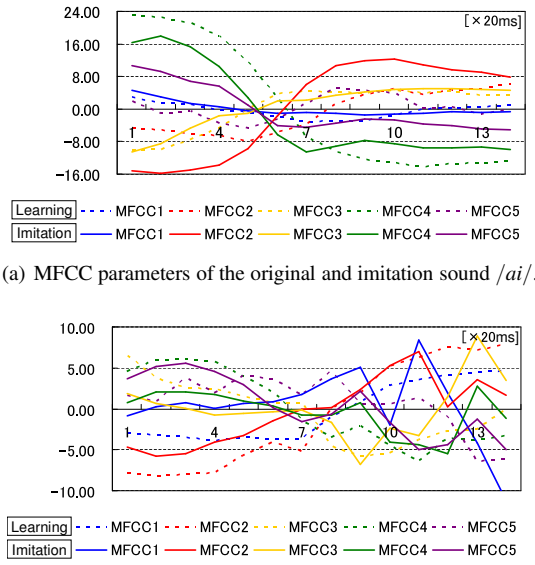
Figure 7 shows the resulting PB space, consisting of five learned sounds and four associated sounds, where the first vowel was /a/. Figure 8 shows the time series variation of the MFCC parameters for the original and imitation sounds /ai/ and /au/, as examples of an experienced sound and an unexperienced sound, respectively. The vertical axis represents the MFCC value, and the horizontal axis represents time [x 20 ms]. We could confirm that the imitation sound /ai/ reproduced the original sound. On the other hand, although the imitation sound /au/ differed from the original sound in the last part, the sound was reproduced to a differentiable extent. Most of the imitation sounds were similar to the original ones.

2) Imitation of Three Continuous Vowels: In the learning phase, we organized the following RNNPB: the input and output layers had 11 units, the hidden layer had 25 units, the context layer had 15 units, and the PB layer had two units. The RNNPB learned the MFCC and vocal tract parameters of the learning data (/iue/, /ueo/, and /oai/, 580 ms, and 30 ms/step), produced by the Maeda model. In the association phase, we inputted the MFCC parameters generated by recording the speech sounds of a person into the organized

TABLE III

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/



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

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

Fig. 8. MFCC parameters.

RNNPB, and we obtained the PB values for each of the sounds. Table IV summarizes the recording of the three continuous vowels sounds. In the generation phase, we used the PB values from the association phase to produce imitation sounds for the data.

Figure 9 shows the resulting PB space, which consisted of three learned sounds and five associated sounds, where the first vowel was /i/. The PB values for /iue/, an experienced sound, were mapped close to those of the learning phase. Figure 10 shows the time series variation of the MFCC parameters for the original and imitation sound /iuo/, as an example of an unexperienced sound. The vertical axis represents the MFCC value, and the horizontal axis represents time [x 30 ms]. Although /iuo/ was reproduced exactly, to an extent, some of the unexperienced sounds were reproduced either as undifferentiable, unclear sounds, or the same as the experienced sounds.

VI. DISCUSSION

A. Adequacy of Proposed Model

As we can see from Fig 6, RNNPB–1, which used only sound information, acquired PB values that were affected by acoustic similarities in the sound data, and it made mistakes in recognizing the sounds. On the other hand, despite of the differences between the two speakers, RNNPB–2, which used both sound and articulation information, acquired PB values that were mapped closely to the same sounds, and it robustly recognized the sounds. These results show that articulation information helps human beings to recognize speech sounds, thus supporting the *motor theory of speech*

TABLE IV

RECORDING OF THREE CONTINUOUS VOWELS.

Experienced	Unexperienced			
/iue/	/iuo/	/iae/	/ioe/	/iua/
/ueo/	/uea/	/uei/	/uao/	/uio/
/oai/	/oau/	/oae/	/oui/	/oei/

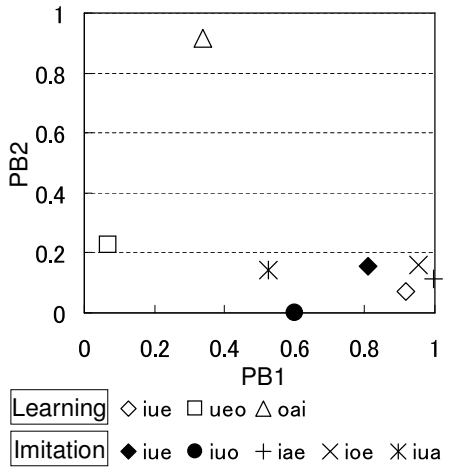


Fig. 9. PB space of three continuous vowels: three learned sounds and five associated sounds, where the first vowel was /i/.

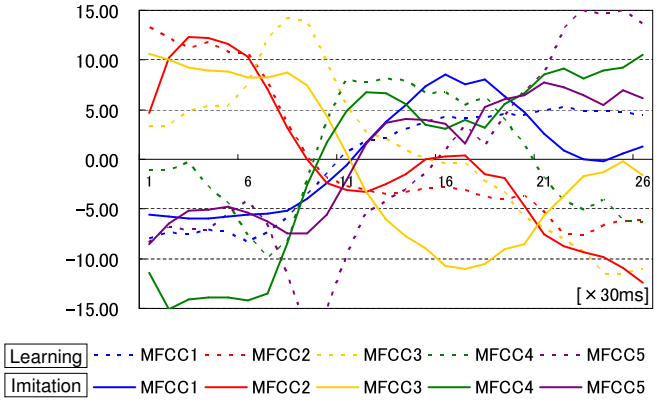


Fig. 10. MFCC parameters of the original and imitation sound /*iuo*/.

perception. We have thus confirmed the adequacy of our imitation model for targeting language acquisition in infants.

B. Imitation Capability

In the case of imitating two continuous vowels, our system could accurately reproduce, to an extent, most of the heard sounds that were experienced or unexperienced. Meanwhile, in the case of imitating three continuous vowels, although the system could imitate experienced sounds, it had difficulty in imitating unexperienced sounds. The reason is that adding a vowel to the target sounds for imitation caused increased diversity in the combination of sounds.

The limitation in our model implies the requirement of acquiring phonemes. The RNN used in our model encoded a vowel sequenceas a dynamics: a PB vector. However, the result showed that two or more PB vectors were required to imitate a sequence with three vowels. This requirement could relate to the form of phoneme. Tani has already proposed for automatic extraction of the PB vectors from a sequence [2]. It would be interesting to investigate the correspondence between the extracted PB vectors and phonemes.

VII. CONCLUSIONS

We have proposed a vocal imitation system focused on the physical constraints of the human vocal tract and on treating speech sounds as dynamic sequences. Through experiments, we have verified the properties and the imitation capability of the system. The results show that the system could robustly recognize speech sounds without exhibiting the effects of differences between two speakers, and it could imitate experienced sounds accurately. In the case of imitating unexperienced sounds, two continuous vowels could reproduced accurately , to an extent, whereas three continuous vowels posed difficulties in accurate generation. These results imply the possibility that the PB values for the RNNPB used in this model correspond directly to phonemes.

Our future work include extracting phonemes from speech sounds through an automatic tuning method for the RNNPB parameters.

VIII. ACKNOWLEDGMENTS

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (A) (No. 17680017, 2005-2007), Grant-in-Aid for Exploratory Research (No. 17650051, 2005-2006), and Kayamori Foundation of Informational Science Advancement.

REFERENCES

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

[2] 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.

[3] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, "Theorem of the invariant structure and its derivation of speech gestalt," in *Proc. Int. Workshop on Speech Recognition and Intrinsic Variations*, 5 2006, pp. 47–52.

[4] D. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annu. Rev. Neurosci*, vol. 27, pp. 169–192, 2004.

[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] R. Yokoya, T. Ogata, J. Tani, K. Komatani, and H. G. Okuno, "Experience based imitation using RNNPB," in *IEEE/RSJ IROS2006*, 2006.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, 1972.