Segmenting Sound Signals and Articulatory Movement using Recurrent Neural Network toward Phoneme Acquisition

○協 神田 尚（京都大学） 正 尾形 哲也（京都大学）
協 駒谷 和範（京都大学） 協 奥乃 博（京都大学）

Hisashi KANDA, Graduate School of Informatics, Kyoto University, hkanda@kuis.kyoto-u.ac.jp
Tetsuya OGATA, ogata@kuis.kyoto-u.ac.jp
Kazunori KOMATANI, komatani@kuis.kyoto-u.ac.jp
Hiroshi G. OKUNO, okuno@kuis.kyoto-u.ac.jp

This paper proposes a computational model for phoneme acquisition by infants. Infants perceive speech 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. The key idea to solve this problem is that articulatory mechanisms such as the vocal tract help human beings to perceive sound units corresponding to phonemes. To segment acoustic signal with articulatory movement, our system was implemented by using a physical vocal tract model, called the Maeda model, and applying a segmenting method using Recurrent Neural Network with Parametric Bias (RNNPB). This method determines segmentation boundaries in a sequence using the prediction error of the RNNPB model, and the PB values obtained by the method can be encoded as kind of phonemes. Experimental results demonstrated that our system could self-organize the same phonemes in different continuous sounds. This suggests that our model reflects the process of phoneme acquisition.

Key word: Phoneme Acquisition, Physical Vocal Tract Model, Recurrent Neural Network.

1 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 by generating those sounds repeatedly by trial and error. This ability is closely related to the language acquisition.

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’ vocal formation through vocal imitation [1, 2]. 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 [3], and recent neuroscience studies seem to show the idea to be an active process involving motor cognition [4].

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) [5] to segment a continuous temporal sequence consisting of acoustic signal with articulatory movement. From the viewpoint of considering sounds as temporal sequences, we have already developed a vocal imitation system [6], 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 [7]. 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.

2 LEARNING ALGORITHM

This section describes the method to learn and segment temporal sequence dynamics. We apply the RNNPB model, which was first proposed by Tani [5] as the forward forward model. It generates complex movement sequences, which are encoded as the limit-cycling dynamics and/or the fixed-point dynamics of the RNN.

2.1 RNNPB model

The RNNPB model has the same architecture as the conventional Jordan-type RNN model [8], 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 1 shows the network configuration of the RNNPB model.

![RNNPB model](image)

Fig. 1 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 [9], as in a con-
The given sequence is divided into $N$ sections. Each section has the same length. The boundary step $s_i$ ($i = 0, 1, \ldots, N$) is set as follows.

$$s_i \leftarrow i \cdot T / N$$

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 $E_i$ ($i = 0, 1, \ldots, N - 1$) is obtained as follows.

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

Step 4: Update the length of each section

The boundary step $s_i$ ($i = 1, \ldots, N - 1$) is updated by using the following rules:

$$s_i \leftarrow \begin{cases} s_i - d_s & \text{if } E_{i+1} \geq E_i \\ s_i + d_s & \text{if } E_{i+1} < E_i, \end{cases}$$

where $d_s$ 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. The algorithm can decrease the error by modifying the position of each boundary.

2.3 Learning of PB Vectors

The learning algorithm for the PB vectors is a variant of the BPTT algorithm. The step length of $i$th 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 $k$th unit of the parametric bias at the section $S_i$ in the sequence are as follows:

$$\delta p_{i,k} = \epsilon \sum_{t=i}^{s_{i+1}} \delta_{i,k}(t),$$

$$p_{i,k} = \text{sigmoid}(p_{i,k} + \delta p_{i,k}),$$

where $\epsilon$ is a coefficient. In Eq. 4, the $\delta$ force for updating the internal values of the PB $p_{i,k}$ 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.

3 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.

We used the vocal tract model proposed by Maeda [10]. This model has seven parameters determining the vocal tract shape: Jaw position (JP), Tongue dorsal position (TDP), Tongue dorsal shape (TDS), Tongue tip position (TTP), Lip opening (LO), Lip protrusion (LPR), Larynx position (LP). The parameters were derived by principal component analysis of cineradiographic and labiolingual data from French speakers. 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.

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 mid sagittal 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.

4 EXPERIMENTAL SYSTEM AND RESULTS

4.1 Experimental system

Our experimental 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 then normalized and synchronized. In this phase, the parameter $d_s$ was set at 0.1. The size of the RNNPB model and the time interval of the sequence data differed according to the experiment.

4.1.1 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 Hanning 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.

4.1.2 Vocal Tract Parameter

We applied the Maeda model with the six parameters except LP. We used the Maeda model, which captures the phonetically important characteristics of speech.
Maeda model produces vowel sounds, the seventh parameter \( \text{LP} \) and 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.

4.2 Experimental results

4.2.1 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, 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/, /eoe/, /iue/, /iao/, /oai/, /ueo/, /uo/, /uia/, /eiu/, /oai/, and /oeu/ (1380 ms and 30 ms/step), produced by the Maeda model.

Figure 2 shows the PB space for \( N = 8 \). In Fig. 3, 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 the category of the 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. The PB values for the same vowel, including the learning data, were mapped with sufficient dispersion.

Figure 4 shows the transition of the PB values for the input data /eiu/ and /uia/ in the learning phase. In Fig. 5, the PB values of section \( S_i \) for input data /uia/ were close to those of the sections \( S_{i,1}, S_{i,2}, S_{i,3} \) for input data /eiu/. When comparing Fig. 4 and 5, we confirmed that the category of the phoneme \( /i/ \) in Fig. 4 corresponded to the transitions of the PB values in Fig. 5.

5 DISCUSSION

5.1 Segmentation ability by RNNPB

Our segmenting method determines the segmentation boundaries using the prediction error of the RNNPB model. In the experiment 4.2.1, 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. 2(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.
Even if the numbers and categories of phoneme in the given acoustic signal is unknown, our system can self-organize the position of phoneme transitions, and the phoneme categories by preparing enough number of boundaries N. The twice numbers of given phonemes are enough as the N.

5.2 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. 4, it is confirmed that each phoneme category /i/, /u/ and /e/ formed a plot but a small cluster consisted of multiple plots. In Fig. 5, 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 [13]. 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.

6 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.

Fig. 5 Changing PB values for input sequence data /eui/ and /uia/ in the PB space.

Our future work includes to imitate speech using automatically extracted PB values corresponding to phonemes from speech 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.

References