# The Multi Timescale Phoneme Acquisition Model of the Self-Organizing Based on the Dynamic Features

*MIYAZAWA Kouki* [1], *MIURA Hideaki*[1], *KIKUCHI Hideaki* [1], *MAZUKA Reiko* [2]

[1] Graduate School of Human Sciences, Waseda University, Japan
[2] RIKEN Brain Science Institute, Japan
m-kouki@moegi.waseda.jp, miura_hideaki@asagi.waseda.jp, kikuchi@waseda.jp, mazuka@brain.riken.jp

## Abstract

It is unclear as to how infants learn the acoustic expression of each phoneme of their native languages. In recent studies, researchers have inspected phoneme acquisition by using a computational model. However, these studies have used a limited vocabulary as input and do not handle a continuous speech that is almost comparable to a natural environment. Therefore, we use a natural continuous speech and build a self-organization model that simulates the cognitive ability of the humans, and we analyze the quality and quantity of the speech information that is necessary for the acquisition of the native phoneme system. Our model is designed to learn values of the acoustic features of a continuous speech and to estimate the number and boundaries of the phoneme categories without using explicit instructions. In a recent study, our model could acquire the detailed vowels of the input language. In this study, we examined the mechanism necessary for an infant to acquire all the phonemes of a language, including consonants. In natural speech, vowels have a stationary feature; hence, our recent model is suitable for learning them. However, learning consonants through the past model is difficult because most consonants have more dynamic features than vowels. To solve this problem, we designed a method to separate "stable" and "dynamic" speech patterns using a feature-extraction method based on the auditory expressions used by human beings. Using this method, we showed that the acquisition of an unstable phoneme was possible without the use of instructions.

**Index Terms**: language acquisition, dynamic features, neural network, consonants

## 1. Introduction

The major current speech-processing technologies use a statistical model that expresses an acoustic characteristic of a speech. This acoustic model is built by employing supervised learning that uses a large quantity of speech and transcription data that prescribe detailed information about the language to learn (the number/type/meaning of a word/phoneme). Such a model has certain disadvantages: it costs considerable time and money to acquire the training data. On the other hand, infants do not need explicit instructions to acquire phonetic systems. We aim at obtaining a cue to improve speech processing technology from the mechanism of this ability for superior phoneme acquisition. We build a self-organization model that simulates the cognitive ability of the humans, and we examine computationally the process through which humans acquire the phoneme systems of their languages.

In a natural speech, vowels have the stationary feature, which is a temporarily unchanging part on the spectrum; hence, our recent SOM model[1] is suitable for learning such patterns. However, learning is difficult in the past model because many consonants have the more dynamic features than vowels. Therefore, in this study, we improved upon our recent model and examined the mechanism involved in as well as the quality and quantity of speech data that is required for the acquisition of the complete phoneme system of a native language. To handle vowels and consonants that have varying acoustic constancy, we herein propose a model that uses a "dynamic filter" based on the processing that occurs in the human auditory system. In the next section, we introduce the recent studies and explain our new trial.

## 2. Background

### 2.1. Acquisition of native phonetic systems

In the human, native vowels are acquired in the first six months, and native consonants are acquired within the first one year. According to [2], infants show superior ability in short-term learning too. In [2], infants were familiarized for two minutes with continuum stimulus that changed gradually from /ta/ to /da/ and exhibited either a bimodal or unimodal distribution. After the trial, only infants in the bimodal condition were able to discriminate between the /ta/ and /da/ tokens[2]. Based on this knowledge, there is a hypothesis that infants have the ability to learn the statistical frequency of specific sound properties and that they use input from adults to separate native phonetic boundaries. Nevertheless, there is also a hypothesis that humans have specific linguistic abilities. Therefore, a problem arises whether a language system is acquired by an innate mechanism or by language experience.

### 2.2. Computational model of phoneme learning

One of the approaches to clarify this discussion is to build a computational model of language acquisition and to estimate how and how much information about speech signals can simulate the human ability of speech perception. For example, there are recent studies such as discrimination learning of English and Japanese liquids, which assumed F2 and F3 with competitive Hebbian learning[3] or discrimination learning of English and Japanese vowels, which assumed F1 and F2 and duration times with a Gaussian mixture model[4].

However, the input data used in these studies is the isolate word or the generated stimuli. Therefore, the learning experiment of the recent studies may be easy in comparison with the learning of the infants in the real environment. Hence, we attempt to simulate learning of a vowel system with the input that comprises natural sounds and features and propose a learning model (chapter 3.1 in details) that can faithfully reproduce human auditory expressions. As a result, it was shown that a natural speech includes the sufficient information about the vowel system peculiar to a language, without depending on the speech-style and sex, and a self-organizing learning process can learn the vowels at the same level as do humans[1]. These studies succeeded in optimizing specific phonetic systems by using unsupervised learning; therefore, these results support the fact that language experience plays an important role in language acquisition.

In the learning results observed in [1], phonemes that had stationary features tended to form independent categories (e.g., vowels and fricatives). However consonants such as explosives were not acquired through the model. The result obtained in [1] had the acoustic characteristics of an unsteady short-time spectrum. In the model, input data was classified according to the appearance frequency and distribution of phonemes; however, consonants had unsteady characteristics with many short-term spectral changes and a short distribution time unlike vowels and appeared rarely. Therefore, it was difficult for them to form an independent cluster. To overcome this problem, we improved upon our model by using the same technique as that used in the human auditory system for handling the unstable characteristics of speech.

### 2.3. Dynamic feature extraction in auditory system

According to [5], the auditory system may analyze the dynamic characteristics of speech signals, which is based on the reply frequency of the acoustic cell. Figure 1 shows the reply of the primary auditory nerve to the syllable /pu/. The figure to the left shows a spectrogram of the input. The figure in the center shows a "high-frequency cell" group, and the large change in the spectrum represents the reply. The figure to the right shows a "low-frequency cell" group, and the stable spectrum represents the reply (e.g., a vowel) [5].

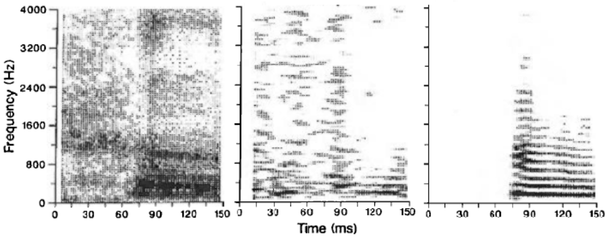

Figure 1: *Reply of auditory nerve to /pu/ [5]*

Moreover, it is said that the processing of speech in the auditory system is multi-timescale. For example, one pathway handles the phoneme on a 10-ms timescale and another pathway handles the syllable line on a 100-ms timescale [6].

We assume that the structure of the auditory nervous system of an infant is similar to that of an adult because a newborn can discriminate syllable. Consequently, the design of our model on the basis of this knowledge is biologically valid. Therefore, in this study, we separated the speech input according to its dynamic or stable spectral features as a part of the preprocessing step during acquisition learning. The separated speech data was input to separate models to overcome the failure of the unification model, in which the speech input was not separated. The reason the unification model failed is that the appearance frequency of each phoneme is different. Specifically, we consider it biologically valid to use the ΔMFCC in the filter described in section 3.2.2 and to separate speech data on the basis of dynamic features according to the ΔMFCC. A similar technique was suggested by [8] to extract vowels from continuous speech using the power of the spectrum.

## 3. Our Model

This section explains the functioning of our unsupervised learning algorithm for the acquisition of a phoneme system[1] and this method's problems, and the preprocessing by which the input is separated on the basis of its stable or unstable characteristics.

### 3.1. Unsupervised Clustering

As an algorithm of the unsupervised class, we use the *Self-organizing Map* (SOM)[9]. A SOM is a type of a neural network model, which reflects the fact that expression in the cerebrum sensory area is organized by perceptual experience. A SOM can classify high-dimensional input signals without instruction and estimate the categories; therefore, we assume that it is adequate as a language acquisition model of the phonetic system. A basic SOM consists of an input layer and a competition layer. We used in a one-dimension SOM, the nodes were arranged in a line. Each unit of the competition layer (the neural node) is connected with the input layer with different weight values (the reference vector). Let *R* denote the dimensions of the input vector; then, the reference vector of node *i* is $m_i$, which is as follows:

$$m_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{iR}) \tag{1}$$

When the input vector *p* is given, node *c* for which the inner product of its reference vector with the input vector is largest becomes the "winner."

$$c = \arg\max_i \{m_i \cdot p\} \tag{2}$$

Finally, the reference vectors of the winner node and the neighbor node are approximated by the input vector.

$$m_i := \frac{m_i + h_{ci} p}{\| m_i + h_{ci} p \|} \tag{3}$$

The value *h* denotes the degree of update:

$$h_{ci} = \alpha \exp\left( - \frac{\| \mathrm{r_i} - \mathrm{r_c} \|^2}{2\sigma^2} \right) \tag{4}$$

*α* represents the learning constant; $r_i$, the coordinate of node *i*; and $r_c$, the coordinate of the winner node. *σ* expresses the size of the neighborhood. The learning process of a SOM proceeds by repeating these steps. First, the ordering phase is learning in big values of *α, σ*. Second, the tuning phase is learning relations between inputs in small *α, σ*. In this study, the training of the SOM model was performed with MATLAB.

A SOM model classifies input data into same numbers of classes as the number of competing layer's nodes. However, an infant can learn the number of phoneme categories of its native language. In consideration of this point, we introduce a framework to integrate similar categories into a single category. In a one-dimension SOM, the following technique was suggested[10]; our model was based on this technique and evaluated the number of categories.

① Construct a histogram expressing the degree of integration of the categories in the SOM that finished learning. The frequency value is calculated as follows:

$$L_i = V_i / dM_i \tag{5}$$

*Vi* indicates the number of input data having node *i* as the winner node, and *dMi* denotes similarity between the reference vector of neighboring nodes, that is as follows:

$$dM_i = | m_i - m_{i-1} | + | m_i - m_{i+1} | \qquad (6)$$

② Examine the node corresponding to the peak of the histogram. Replace the weights of nodes, except for the peak node with a 0 vector, and calculate correspondence with the input data and the winner node again.

### 3.2. Analysis of spectral time series

Section 3.2.1 describes the preprocessing similar to [1]. Section 3.2.2 explains the dynamic filter that we used on the basis of the dynamic feature extraction described in 2.3.

### *3.2.1. Mel Frequency Cepstral Coefficients*

First, we calculate the raw speech into *mel frequency cepstral coefficients* (MFCCs), that is, acoustic features often used in speech-processing technologies. We used 26 dimension features, which are MFCC12, the logarithm power (C0), ΔMFCC12, and ΔC0. This process used the *Hidden Markov Model Toolkit* (HTK). The frame rate is 25[ms], and the frame shift length is 10[ms].

### *3.2.2. Dynamic features Filter*

As discussed in 2.2, our recent model [1] is difficult to learn unstable phonemes because such phonemes have fewer appearances and their distributions are unstable. On the other hand, acoustic cells in human auditory systems can separate the stable features of speech from the unstable ones. One of the methods for separating the stable features of speech from the unstable ones for engineering purposes is to calculate the square sum of the ΔMFCC [11]. ΔMFCC is the feature that was suggested to express the dynamic change in a speech. It is the regression coefficients that are extracted for every MFCC frame over an approximately 50[ms]. *t*, Θ, and *C* denote the frame number, the total number of frames, and the MFCC respectively. We supposes '*t* = 2', which is default of HTK.

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta \left( \mathrm{C}_{t+\theta} - C_{t-\theta} \right)}{2 \sum_{\theta=1}^{\Theta} \theta^2} \qquad (7)$$

In this study, we calculated the value of the square sum of 1–12 dimensions of the ΔMFCC. If it was less than a specific threshold, we judged the speech data as stable because such a value indicated very few changes in the spectrum (MFCC). If it was above the threshold, we judged the speech data as unstable. We refer to the part of the model that performs this function as the "dynamic filter."

### *3.2.3. Evaluation experiment of the dynamic filter*

We performed a preliminary evaluation experiment using *the Corpus of Spontaneous Japanese* (CSJ) [12] to determine an appropriate threshold for the dynamic filter. The audio file of CSJ is recorded in 16 bit at 16 KHz. We chose two men and two women. First, we analyzed their speeches by the method described in 3.1.2 and applied a dynamic filter to each frame, extracting only the frame determined to have unstable features. Next, according to the phoneme labels provided in the CSJ, we ascertained the phoneme applicable to each frame.

Figure 2 shows the result of the experiment. Each bar in the graph indicates the ratio of the frames of each phoneme category for all the frames separated by the dynamic filter. Each value printed on the bars indicates the number of extracted frames. The horizontal axis shows the threshold values, and the vertical axis shows the ratios.

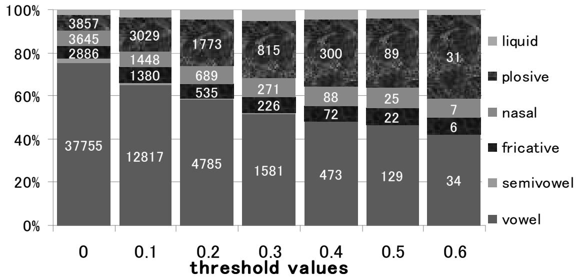

Figure 2: *Ratios of the number of frames*

All the data shown are averages. Overall, as the threshold increases, the ratio of phonemes with unstable characteristics increases, whereas the ratio of phonemes with stable characteristics decreases. Therefore, it may be said that phonemes unstable characteristics were successfully extracted by the dynamic filter. Considering the rate of decline of the number of frames, we set the value of the appropriate threshold for the dynamic filter to 0.2 and performed experiments on the model with the dynamic filter, as described in following section.

# 4. Simulation

### 4.1. Methods

In this investigation, our model learns and organizes phoneme categories by natural utterances. The same as [1], input data are a single Japanese speaker's speech because we assumed that the newborn first heard only the sound of the parents. We separated speech data into stable and unstable speech using a dynamic filter prior to all processing, and then, we input the separated data to different SOM models. In the following text, the term "unstable SOM" refers to a model to which unstable speech was input (square sum is above the threshold), whereas "stable SOM" refers to a model to which stable speech was input (square sum is below the threshold).

We chose five men and five women. The details of the used data are listed in Table 1. All these speeches-style is Simulated Public Speech that contains speeches given by general speakers. We analyze these speeches by 3.1.2 method.

Table 1. *Details of the speech used in experiment.*

| Speaker ID | Sex | Age | Speaker ID | Sex | Age |
|---|---|---|---|---|---|
| S00F0031 | F | 25to29 | S00M0065 | M | 20to24 |
| S00F0041 | F | 35to39 | S00M0199 | M | 30to34 |
| S00F0082 | F | 20to24 | S01F0006 | F | 20to24 |
| S00F0088 | F | 30to34 | S01M0101 | M | 20to24 |
| S00M0025 | M | 20to24 | S01M0205 | M | 20to24 |

We used the continuous 103.2[s] (10320 frames) of the speech features data mentioned above for training data and used the rest for evaluation data. We applied the dynamic filter to the training data and performed five times of cross-validation experiment for speech learning using the stable SOM and the unstable SOM. The learning parameter for both the SOMs was set to the same value as in [1]. We used 28 nodes in a one-dimension SOM. By one learning time step, one frame is chosen from among the beginning of the training data and is input to the SOM. The data were input twice repeatedly. The first 1000 steps belong to the ordering phase; the value of *α* changes from 0.9 to 0.2 and that of *σ* changes from maximum to 1. From 1000 to 20640 steps belong to the tuning phase; the value of *α* changes from 0.2 to 0.0.

### 4.2. Evaluation

For the SOM for which learning was over, we estimated the phoneme categories which were acquired by the SOM. First, we extracted the phonemes from the evaluation data based on a phoneme label of CSJ. It is one central frame of each vowel that we used for evaluating the model. The evaluation data were chosen to become the same number about each vowel category. Table 2 is the list of phonemes for the evaluation.

Table 2. *The details of the phoneme to evaluate*

| Group | Phoneme | Group | Phoneme |
|---|---|---|---|
| vowel | /i/,/e/,/o/,/a/,/u/ | voiceless plosive | /k/,/t/ |
| semivowel | /y/ | voiced plosive | /g/,/d/ |
| fricative | /s/ | liquid | /r/ |
| nasal | /m/,/n/ | | |

For the SOM that finished learning, we use the method described in 3.1 so as to unify the categories. Next, we input the set of evaluation data into the stable SOM and unstable SOM, and estimated the number and boundaries of each phoneme categories as follows.

(1) For each SOM cluster, we calculated the share rate for each phoneme category (the percentage of evaluation data that belonged to the same phoneme category and ranked in the same cluster)

(2) For each cluster, we selected a phoneme with the highest share rate as the phoneme label of the cluster.

(3) When there was a phoneme without a matching cluster, we searched for a cluster with a high share rate in a given order. If there was a cluster that did not match any phoneme, we matched it. When a suitable cluster did not exist, the share rate of the phoneme was assumed to be 0%.

(4) For each phoneme label, the share rate of its matching cluster was its identification rate.

Further, we define the phoneme identification rate for these trials as the average recognition accuracy of each phoneme. If the recognition accuracy of a phoneme was found to be 0.00, we concluded that there was a failure in assigning a category to the phoneme and ignored it.

### 4.3. Results

Figure 3 shows the recognition accuracy for each phoneme category occurring in every phoneme group after the completion of training. Each value and vertical bar in the graph indicates the mean and standard deviation for the five speech durations and ten speakers during the cross-validation training, respectively. For each phoneme group, the left-hand-side bar shows the result of the unstable SOM, the bar in the center shows the result of the conventional model [1] that does not use a dynamic filter (threshold = 0), and the right-hand-side bar shows the result of a stable SOM.

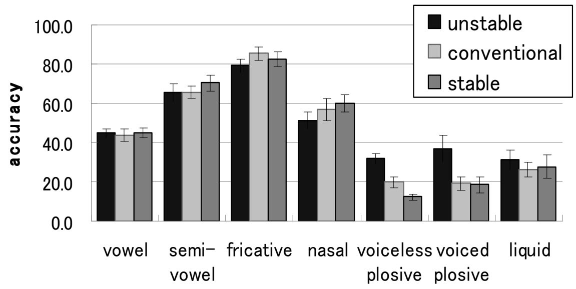

Figure 3: *Learning result of each SOM model.*

As a result of ANOVA, it was significantly different between voiced plosive (F(2,27) = 3.35, p<.001) and voiceless plosive (F(2,27) = 3.35, p <.05),. Further, there were no significant differences between vowels, semivowels, fricative, nasal, and liquid. Learning unstable consonants was difficult in our recent model [1]. However, it was effectively shown that by using the unstable SOM, clustering was possible.

## 5. Discussion

We performed an experiment using a neural network model to elucidate the process through which humans acquire the phoneme system of their native language. We introduced a dynamic filter, similar in function to the auditory nerve in human beings, and input the obtained stable/unstable speech data separately to different models. As a result of experiment, the identification rate of plosives improved by approximately 10%-15% in comparison with recent model. It was shown that natural speech contains sufficient information about the phoneme system of a particular language even when the speech is as short as about 100 [s]. Further, speech resources and a learning process that infants can use to learn phonemes were presented. This result supports the hypothesis that infants learn using the statistical frequency of specific sound features found in an adult's speech.

In addition, it is said that at the initial stage of language acquisition, an infant hears the *infant-directed speech* (IDS), and we will use an IDS for input so as to examine its role.

## 6. Acknowledgements

This work was supported by KAKENHI 21610028.

## 7. References

[1] Miyazawa, K., Kikuchi, H., Mazuka, R., "Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model", Interspeech, 2914-2917, 2010.

[2] Maye, J., Werker, J. F., and Gerken, L., "Infant sensitivity to distributional information can affect phonetic discrimination", Cognition., 82(3):B101-B111, 2002.

[3] Guenther., F. H., and Gjaja., M. N., "The perceptual magnet effect as an emergent property of neural map formation", J. Acoust. Soc. Am., 100:1111-1121, 1996.

[4] Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., and Amano, S., "Unsupervised learning of vowel categories from infant-directed speech", PNAS, 104:13273-13278, 2007.

[5] Carney, L. H., and Geisler, C. D., "A temporal analysis of auditory-nerve fiber responses to spoken stop consonant-vowel syllables", J. Acoust. Soc. Am., 79(6):1896-1914, 1986.

[6] Hickok, G., and Poeppel, D., "The cortical organization of speech processing", Nat Rev Neurosci, 8:393-402, 2007.

[7] Takara, T., Fujita, Y., Sunagawa, T., Oshiro, T., and Iwai, S., "A functional model for acquisition of spoken language based on the clustering method - in case of words and vowel phonemes", IEICE Tech. Rep., 109(308):61-66, 2009. (in Japanese)

[8] Kohonen, T., "The self-organizing map", Proceedings of the IEEE., 78(9):1464-1480, 1990.

[9] Terashima, M., Shiratani, F., and Yamamoto, K., "Unsupervised cluster segmentation method using data density histogram on self-organizing feature map", IEICE, J79-D-II(7):1280-1290, 1996.

[10] Hodoshima, N., Arai, T., Kusumoto, A., and Kinoshita, K., "Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments," J. Acoust. Soc. Am., 119(6), 4055-4064, 2006.

[11] Maekawa, K., "Corpus of spontaneous Japanese : its design and evaluation", SSPR2003, 7-12, 2003. (in Japanese)