



Unsupervised Learning of Vowels from Continuous Speech based on Self-organized Phoneme Acquisition Model

MIYAZAWA Kouki¹, KIKUCHI Hideaki¹, MAZUKA Reiko²

¹ Graduate School of Human Sciences, Waseda University, Japan

² RIKEN Brain Science Institute, Japan

m-kouki@moegi.waseda.jp, kikuchi@waseda.jp, mazuka@brain.riken.jp

Abstract

All normal humans can acquire their native phoneme systems simply by living in their native language environment. However, 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 read speech that has a limited vocabulary as input and do not handle a continuous speech that is almost comparable to a natural environment. Therefore, in this study, we use 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 vowel system. Our model is designed to learn values of the acoustic characteristic of a natural continuous speech and to estimate the number and boundaries of the vowel categories without using explicit instructions. In the simulation trial, we investigate the relationship between the quantity of learning and the accuracy for the vowels in a single Japanese speaker's natural speech. As a result, it is found that the vowel recognition accuracy of our model is comparable to that of an adult.

Index Terms: language acquisition, vowels, neural network

1. Introduction

The major current speech-processing technologies use a statistical model that expresses an acoustic characteristic of the 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 can realize a high recognition accuracy in a specific speech environment, but it has certain disadvantages: it costs considerable time and money to acquire the training data, and it cannot flexibly adapt itself to speaker and noise changes.

On the other hand, infants do not need explicit instructions to acquire phonemic systems. Furthermore, the phonemes acquired by humans are robust against environmental changes. We aim at obtaining a cue to improve speech processing technology from the mechanism of this ability for superior phoneme acquisition. Consequently, we focus on the process through which humans acquire the vowel systems of their languages and examine it computationally. In the next section, we introduce the recent studies and explain our new trial.

2. Background

2.1. Acquisition of phonemic systems

The development of phoneme perceptions is examined to test

an infant's sensitivity to discriminate between two phoneme stimuli. According to [1], newborns can discriminate between any of the phonemic categories. For example, English four-month olds are sensitive to the German vowel contrasts /ɜ/ and /o/[1]. The native vowels are acquired in the first six months, and the native consonants are acquired within the first one year. In this process, infants lose their ability to discriminate between nonnative phonemic categories. In addition, 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 phonemic 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 /r/ and /l/, which assumed the second and third formant frequency with competitive Hebbian learning[3] or discrimination learning of English /l/, /i/, /e/, and /e/ and Japanese /i/, /i:/, /e/, and /e:/, which assumed the first and second formant and vowel duration times with a Gaussian mixture model [4]. These studies succeeded in optimizing specific phonemic systems by using unsupervised learning; therefore, their results support the fact that language experience plays an important role in language acquisition.

However, the input data used in these studies is read speech of isolated words or generated stimulus sets based on the result of the sound analysis. According to [5], as for the cepstrum distance between phonemes, read speech is larger than the natural speech; therefore, the learning experiment of the recent studies may be easy in comparison with the learning of the infants in the real environment. In addition, these studies selected specific acoustic properties that are important for discriminating phonemes in order to acquire learning data. However, there is a question as to which specific features should be analyzed to learn because it is doubtful that infants can selectively extract formant frequency values. Furthermore, a natural speech that infants hear in a real environment includes many types of phonemes and nonspeech sounds, but it has not been elaborated as to how they acquire the ability of

discriminating a specific phoneme pair from others in a real speech. Hence, we attempt to simulate learning of a vowel system with the input that comprises natural sounds and features and propose a learning model that can faithfully reproduce human auditory expressions.

3. Our Model

3.1. Analysis of spectral time series

We approximated the preprocess of the input speech sound to the human auditory system as following.

3.1.1. 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. This process used the *Hidden Markov Model Toolkit* (HTK). The summary is as follows.

- ① Calculate the frequency-amplitude spectrum by FFT. This process imitates the function of the inner ear cells.
- ② Calculate cepstrum by log conversion and the *Discrete Cosine Transform* (DCT), and extract the low-dimension cepstrums. By this process, the vocal tract spectrum, which is effective for the identification of the phoneme, is extracted.
- ③ Using mel filter bank analysis, calculate the cepstrum into 12 dimension coefficients. The mel scale expresses the frequency response characteristics of an acoustic cell.

3.1.2. Dynamic features

Δ MFCC is the feature that was suggested to express the dynamic change in the speech. Δ MFCC is the regression coefficients that are extracted for every MFCC frame over an approximately 50[ms][6]. t , Θ , and C denote the frame number, the total number of frames, and the MFCC respectively.

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(C_{t+\theta} - C_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \quad (1)$$

Our model supposes ' $t = 2$ ', which is default of HTK. [7] shows that the auditory system may analyze the dynamic characteristics of speech signals, which is based on the reply frequency of the acoustic cell. Consequently, it has biological validity to introduce Δ MFCC. We used 26 dimension features, which are MFCC12, log power (C0), Δ MFCC12, and Δ C0.

3.2. Unsupervised Clustering

This section explains our unsupervised learning algorithm of vowel system acquisition from continuous speech.

3.2.1. Self-organizing Maps

As an algorithm of the unsupervised class, we use the *Self-organizing Map* (SOM) model[8]. A SOM is a type of a neural network model, which reflects the fact that characteristic 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 phonemic system. A basic SOM consists of an input layer and a competition layer. 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:

$$m_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{iR}) \quad (2)$$

These vectors are normalized so that their sum equals 1. The reference vector is copied from the strength of the synapse combination between nerve cells. 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."

$$c = \arg \max_i \{m_i \cdot p\} \quad (3)$$

Physiologically, the mechanism that determines the winner node is realized by inhibitory connections from neurons. Finally, the reference vectors of the winner node and the neighbor node are approximated by the input vector.

$$m_i := \frac{m_i + h_{ci}p}{\|m_i + h_{ci}p\|} \quad (4)$$

The value h denotes the degree of update:

$$h_{ci} = \alpha \exp\left(-\frac{\|r_i - r_c\|^2}{2\sigma^2}\right) \quad (5)$$

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. It is known that the distinction performance improves when the values of α and σ are reduced in proportion to the number of learning times. 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 .

3.2.2. Unification of categories

A SOM model classifies input data into same numbers of classes as the number of competing layer's nodes. However, the number of vowel categories is different in each language, and an infant can learn the number of vowel 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 nodes were arranged in a line, and the following technique was suggested[9]; 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:

$$L_i = V_i / dM_i \quad (6)$$

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

$$dM_i = |m_i - m_{i-1}| + |m_i - m_{i+1}| \quad (7)$$

Hence, we assumed that a virtual node was next to the node at each of both the ends of the SOM and had a reference vector of 0.

- 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. (This process has been originally incorporated in this study because it is simple and easy.)

We provide examples to classify many types of normal distributions. The number and boundaries of these categories are unknown. The upper panel in Figure 1 shows an example of the results of learning by a SOM. The lower panel shows a result of unifying categories by using the data density histogram on the SOM. The categories are unified and three Gaussian distributions are extracted.

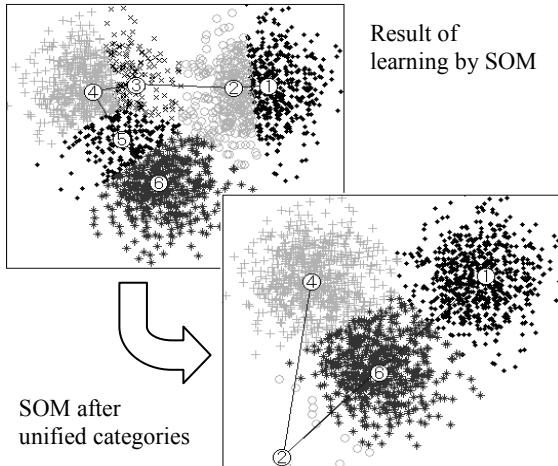


Figure 1: Unification of SOM clusters[9].

In addition, a phenomenon identical to the unification of categories based on the appearance frequency of the input is confirmed in the auditory cortex of the human[3].

4. Simulation

4.1. Training data

In this investigation, our model learns and organizes vowel categories by natural utterances. Input data are a single Japanese speaker's speech because we assumed that the newborn first heard only the sound of the parents.

For evaluating the learning results, we used *the Corpus of Spontaneous Japanese* (CSJ)[10] that has time information of each phoneme. The details of the used data are listed in Table 1. We chose two men and two women in order to vary speech styles, so 12 speakers were chosen. *Academic Presentation Speech* (APS) is the live recording of academic presentations. *Simulated Public Speech* (SPS) contains speeches given by general speakers. The re-reading speeches are re-read transcription of the APS or SPS by the same speaker. In addition, the audio file of CSJ is recorded in 16 bit at 16 KHz. We use the HTK to calculate MFCC and Δ MFCC 26 dimensions, the frame rate is 20[ms], and the frame shift length is 10[ms].

4.2. Methods

We used the continuous 100[s] (10000 frames) of the speech features data mentioned above for evaluation data and used the rest for training data. We extracted continuous 300[s] (30000 frames) from a random time point of training data and used it for learning. The training of the SOM model was performed with MATLAB Neural Network Toolbox. We used 36 nodes in a one-dimension SOM.

Table 1. Details of the speech used in experiment.

Speaker	CSJ ID	Speech Style	sex	age
1	A01M0015	APS	male	35~40
2	A03M0018	APS	male	35~40
3	A01F0143	APS	female	35~40
4	A02F0082	APS	female	30~35
5	R00M0036	re-reading	male	30~35
6	R00M0187	re-reading	male	35~40
7	R00F0028	re-reading	female	30~35
8	R00F0178	re-reading	female	35~40
9	S00M0025	SPS	male	40~45
10	S00M0199	SPS	male	30~35
11	S00F0041	SPS	female	45~50
12	S00F0088	SPS	female	30~35

In the initial state, all the reference vectors of the SOM are initialized to 0. By one learning time step, one frame is chosen from among the beginning of the training data and is input to the SOM. 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 (the distance to the most far-off node) to 1. From 1000 to 30000 steps belong to the tuning phase; the value of α changes from 0.2 to 0.0. We decided these parameters based on a pilot run, so that the acquisition efficiency of a vowel sound became higher.

4.3. Evaluation

For the SOM for which learning was over, we estimated the vowel categories which were acquired by the SOM. First, we extracted the vowels /i/, /e/, /o/, /a/, and /u/ from the evaluation data based on a phoneme label of CSJ. In Japanese, the long vowels /i:/, /e:/, /o:/, /a:/, /u:/ do not qualitatively differ from the corresponding short vowels, and hence, we do not distinguish between long and short vowels. It is one central frame of each vowel that we used for evaluating the model. If the length of a vowel was less than 80[ms], we judged that it was inarticulate and excluded it. The evaluation data were chosen to become the same number about each vowel category.

For the SOM that finished learning, we use the method described in 3.2.2 so as to unify the categories. Next, we input the set of evaluation data into the SOM and estimated the number and boundaries of each vowel categories. In addition, we define the vowel identification rate of these trials as the average of the recognition accuracy of each vowel.

4.4. Results

Figure 2 shows the vowel recognition accuracy of each vowel category uttered by every speaker after the training finished. Each value and each vertical bar in the graph indicates the mean and standard deviation of six times of cross-validation training, respectively.

The vowel recognition accuracy was totally 63.5%. Seeing it in speakers, it was between 47.3% and 75.9%. Seeing it in speech-styles, APS, re-reading, and SPS were 62.9%, 62.1%, and 65.4% respectively, however there was no significant difference ($F(2,9) = 4.26, p = .9$). It was significantly different between vowels ($F(4,55) = 2.54, p < .0001$), with the recognition of /e/ being the highest and that of /u/ being the lowest. Further, there were no significant differences between male and female ($t(10) = 0.96, p = .4$). On average, a SOM obtained 10.6 clusters. Consonants, which have stationary characteristics, tend to form independent categories. For example, there are semivowels and fricatives.

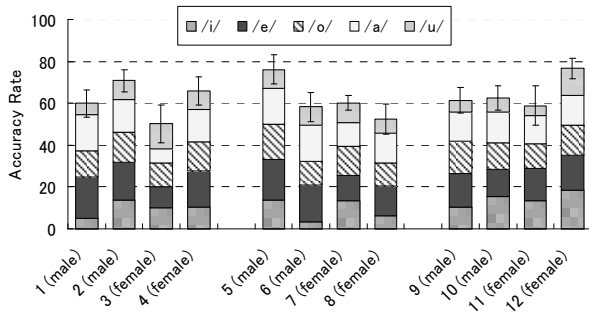


Figure 2: Learning result of each speaker.

In addition, the vowel identification rate of the learning results did not change when the quantity of training data is 100[s]. Therefore, by this simulation, our model arrives at the convergence state by learning for 100[s].

Table 2 lists the correct error identification rates of the learning results. Each value is the average of all trials. “in” implies input vowels (correct label), and “out” implies the recognized phoneme. “others” indicates the rate of classifying sounds into categories, except the vowel categories. There were many cases that were classified /i/ in /e/, and /u/ in /e/ and other categories by mistake.

Table 2. Confusion matrix of the vowel identification rate of model [%]

in\out	i	e	o	a	u	others
i	56.1	23.7	4.5	0.2	6.1	9.4
e	9.2	79.0	0.4	0.8	4.2	6.4
o	0.9	1.6	66.1	8.5	5.4	17.5
a	0.3	0.9	6.7	71.9	2.0	18.2
u	9.7	14.8	4.8	2.7	41.2	26.8

5. Discussion

Concerning Japanese, our model could acquire the detailed vowel categories of the input language by consecutive speech. In a natural speech, vowels have the stationary feature, which is a temporarily unchanging part on the spectrum; hence, the SOM model is suitable for learning such patterns.

To evaluate the validity of the vowel systems that our model has learned, we show the identification rates of a human hearing independent vowels in Table 3[11]. This experiment used a continuous reading speech uttered by a male announcer. Four women heard the vowels, which were picked up from this utterance, and they answered which phoneme they heard.

Table 3. Confusion matrix of the vowel identification rate of Human [%][11]

in\out	i	e	o	a	u	others
i	52.0	0.0	0.0	0.0	2.0	46.0
e	4.0	70.0	3.0	12.0	5.0	6.0
o	0.0	1.0	58.0	11.0	25.0	3.0
a	0.0	1.0	0.0	57.0	2.0	40.0
u	10.0	3.0	1.0	1.0	53.0	32.0

The average of correct identification percentage was 58%. Comparing Table 3 and Table 2, it is observed that the recognition accuracy of /e/ was high and that there was a mistake in identifying /u/ besides the vowels that were common to our model and the human. However, it is rare that a human commits an error and so mishears a vowel as another vowel. Because our model could not categorize sufficient number of consonants, there is a possibility that the case of a human mishearing a consonant was replaced by an error for vowels. We are considering this result in reference to an

infant's phoneme perception ability within the first half year which consonant system is not acquired.

6. Conclusions

We performed an experiment using a neural network model so as to elucidate the process through which humans acquire the phoneme system of their native language. 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, even if it is a short time of around 100[s], and a self-organizing learning process can learn the vowels at the same level as do humans. In other words, a little quantity of speech includes the information that is necessary for estimating vowel categories. This result supports the hypothesis that infants learn the statistical frequency of specific sound properties of an adult speech.

We are improving our model in reference to the human auditory system to learn consonants; it has an acoustic characteristic of an unsteady short-time spectrum. In addition, it is said that at the initial stage of language acquisition, an infant hears the *infant-directed speech* (IDS) of its mother, and we are analyzing the difference between the acoustic features of an IDS and SPS[12]. For the future studies, we will use an IDS for input so as to examine its role.

7. Acknowledgements

This work was supported by KAKENHI 21610028.

8. References

- [1] Polka, L. and Werker, J. F., “Developmental changes in perception of non-native vowel contrasts”, *J Exp Psychol Hum Percept Perform.*, 20(2):421–435, 1994.
- [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”, *Proc. Natl. Acad. Sci.*, 104:13273–13278, 2007.
- [5] Nakamura, M., Iwano, K. and Furui, S., “Analysis of acoustic characteristics in spontaneous speech using Corpus of Spontaneous Japanese”, *IPSP SIG Notes SLP.*, 103:7–12, 2004.
- [6] Furui, S., “Speaker-independent isolated word recognition using dynamic features of speech spectrum”, *IEEE transactions on acoustics, speech, and signal processing.*, 34:52, 1986.
- [7] 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.
- [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 Trans. Inf.*, J79-D-II(7):1280–1290, 1996.
- [10] Maekawa, K., “Corpus of Spontaneous Japanese: Its Design and Evaluation”, *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition.*, 7–12, 2003.
- [11] Kuwahara, H. and Sasaki, H., “Perception of Vowels and C-V Syllables Segmented from Connected Speech”, *The Acoustical Society of Japan.*, 28(5):225–234, 1972.
- [12] Miyazawa, K., Kikuchi, H., Shinya, T. and Mazuka, R., “The dynamic structure of vowels in Infant-directed speech: RIKEN Japanese Mother-Infant Conversation Corpus”, *IEICE technical report. Speech.*, 109(308):67–72, 2009.