ECRONICON OPEN ACCESS

EC NEUROLOGY
Research Article

Phoneme Acquisition Modelling Study for Standard-Chinese Children

Mengxue Cao*

School of Chinese Language and Literature, Beijing Normal University, China

***Corresponding Author**: Mengxue Cao, School of Chinese Language and Literature, Beijing Normal University, No. 19, Xin Jie Kou Wai Da Jie, Beijing, China.

Received: May 15, 2017; **Published:** June 02, 2017

Abstract

This paper simulates the early phoneme acquisition process for children whose mother-tongue language is Standard Chinese. A neural-computational model, based on the growing self-organizing map algorithm, is used to simulate the learning child, and the optimized growing strategy is implemented into the learning process. The experimental results show that, at early stages of language acquisition, by processing the acoustic features of the syllables in the speech signal, young children have the ability, on the one hand, to acquire vowel phonemes and establish the acoustic vowel space, on the other hand, to acquire the manner-of-articulation features of consonants.

Keywords: *Neural-Computational Model; Growing Self-Organizing Map; Phoneme Category; Child Language Acquisition; Vowels and Consonants*

Abbreviations

SOM: Self-Organizing Map; GSOM: Growing Self-Organizing Map; CV: Consonant–Vowel; SCSC: Syllable Corpus of Standard Chinese; BMU: Best Matching Unit; RMS: Root Mean Square

Introduction

Language acquisition is a complex learning process which involves both statistical learning and communicative learning [1,2] mechanisms. On the one hand, by exposing to the mother-tongue language environment, children are able to acquire linguistic knowledge from the statistical patterns of language stimuli by utilizing abilities such as pattern recognition and computational processing [3,4]. On the other hand, during their interactions with the care-takers, children have the ability to improve their learning results by communicative feedbacks [1]. As a starting point, this study focuses on the statistical learning process only, and simulates the acquisition of vowel and consonant phonemes of Standard Chinese for young children in the early stage of language acquisition.

From the perspective of computational modeling, language acquisition can be abstracted as a knowledge learning process. During a common knowledge learning process, not only the knowledge itself but also the fields it covers continually increase as the learning develops. Therefore, the dynamic scalability of knowledge is an essential feature during this process. Among various kinds of knowledge learning algorithms, the Self-Organizing Map (SOM) [5,6] can simulate the topographic structure of knowledge and the self-organizing process of knowledge learning. The learning strategy of SOM is consistent with the "perceptual magnet effect" [7,8] of language learning. Therefore, SOM is widely applied in linguistic related studies [9-15]. However, SOM has its own difficulties in simulating the incremental nature of knowledge [16], so it is not appropriate to apply SOM directly to the language acquisition modeling task.

Inspired by the SOM algorithm, many extendable self-organizing algorithms have been proposed. Ideas of Li's and Alahakoon's are of the most representative. Li and his colleagues proposed two extendable models, the DevLex [13] and DevLex-II [14], and conducted modeling studies on children's early lexicon development [13,14]. Their research suggests that extendable self-organizing neural networks

work better in the modeling of language acquisition related tasks. Taking the data mining task as a starting point, Alahakoon and his colleagues proposed the Growing Self-Organizing Map (GSOM) algorithm [17,18]. Comparing with DevLex and DevLex-II, GSOM has simpler structures and more flexible growing mechanisms that are more suitable for complex modeling tasks such as language acquisition [19,20].

In this study, we will take the GSOM algorithm as the modeling basis, and simulate the phoneme acquisition process with neurocomputational networks.

Materials and Methods

Materials

Audio data are used as the language stimuli in this study. As a simplification, here we only focus on limited vowel and consonant phonemes, and we only consider the CV syllable type. In addition, as required by the model's training procedure, those audio data are converted into model-friendly representations. Details are described as the followings.

Audio Data

The audio data used in this study is adopted from the Syllable Corpus of Standard Chinese (SCSC). The corpus is recorded by 15 male speakers from Beijing area with a sampling rate of 16 kHz. In this study, we do not consider the multi-speaker situation, so we only extract the data of one speaker from the 15 speakers as the audio database for our modeling experiment. As a simplification, we only focus on the monosyllabic words of CV syllable type, which consists of the combinations of 6 plosive consonants ([p], [p^h], [t], [t^h], [k] and [k^h]) and 5 vowels ([i], [e], [a], [u] and [o]). Based on those phoneme and syllable type limitations, we selected 94 meaningful monosyllable words as the training set of our modeling experiment.

Feature Representation

Foster-Cohen [21] points out that children treat the whole word or syllable as a perceptual unit at the early stages of language acquisition, and only with their language development, can they gradually perceive separate phonemes and treat each phoneme as a perceptual unit. Therefore, in the present study, we use the whole monosyllabic word as a perceptual unit, and code the acoustic features of the syllable as a whole.

Spectrogram and duration are the most important acoustic features of a speech signal. However, the coding method in either SOM or GSOM is not capable of representing features from frequency and time domains simultaneously. In this study, we use the spectral state feature map [20] to represent the acoustic features of each audio stimulus. For each speech signal in the training set, we use 22 units to represent its frequency domain features (each unit represents 1 Bark), and 57 units to represent its time domain features (each unit represents 10 ms). Therefore, the acoustic features of each speech signal are represented by the spectral state feature sequence of 22 × 57=1,254 units. Figure 1 shows the acoustic feature representation of the CV syllable [t^hi]. In the frequency domain, the formant features of consonant [t^h] and vowel [i] and their formant transition features are clearly represented; in the time domain, the duration information of [t^h], [i] and the whole syllable are clearly represented as well. Since we do not focus on the acquisition of monosyllable tones in this study, we do not code the tonal features into the acoustic representation of the speech signals.

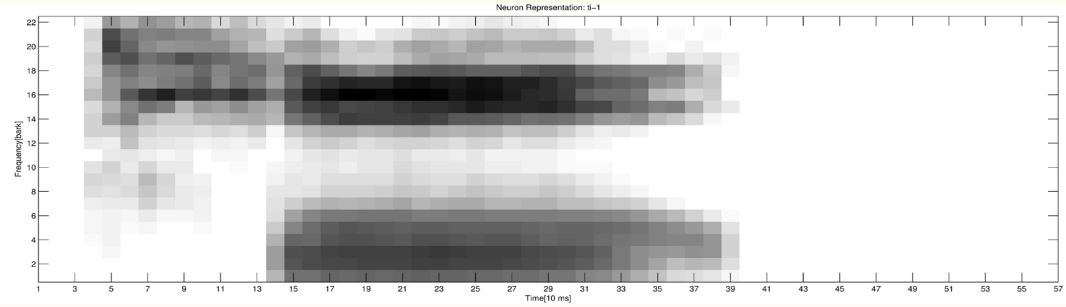

Figure 1: *The acoustic feature representation of the CV syllable [thi]. The x-axis represents the time domain and y-axis represents the frequency domain. The activation state of each unit is represented by gray-scale marker, where 0 (white) represents non-activated, and 1 (black) represents fully activated.*

Methods

The GSOM Structure

Comparing with SOM, GSOM has a simpler structure. At the initial state, there are only four nodes in the GSOM neural network. Since all of the four nodes are edge nodes, they are all extendable at the beginning (see Figure 2). As the learning process develops, the network structure extends automatically by adding in new nodes. The model initializes the feature vectors of the newly-added nodes according to the feature vectors of the Best Matching Unit (BMU) and BMU's neighboring nodes. This enables the new nodes to be added smoothly into the present neural network.

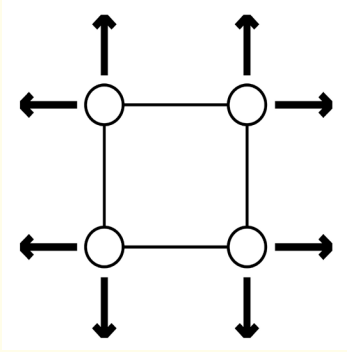

Figure 2: *At the initial state, the network can extend its structure towards any direction at any of the four initial neuron nodes.*

The Optimized Growing Strategy

In the GSOM algorithm, if a BMU reaches growing threshold [17], the network will extend itself by adding new nodes to all available positions among the direct neighbors of the BMU node. This growing strategy lacks of clear direction, and leads to large number of redundant nodes. To overcome this drawback, we adopted the "Cross-Insert" algorithm [22,23] and proposed an optimized growing strategy.

In order to let the newly-added node fit into the present network more effectively, the position for node-adding should stay as close as possible to the BMU and BMU's related nodes. If more than one available position is of the closest to the BMU node (i.e., BMU_1), then their distance to the second best matching unit (BMU_2) is checked; if they again have equal distance to the BMU_2, their distance to the third best matching unit (BMU_3) is checked. This process continues until an optimized position is found (as shown in Equation 1). In Equation 1, n

represents the position of the newly-added node, n_{BMU1} represents all the direct neighboring positions of the BMU_1, and N_{BMU} represents the BMU sequence ranked by their feature distance to the training token.

$$n=\text{argmin}_{i,j} \{||n_i\text{-}BMU_j||\}, i \in n_{BMU1}, j \in N_{BMU}$$

Training procedures

The training process consists of the initializing phase and the growing phase. At the initializing phase, there are only four nodes in the network. As an advantage of this kind of structure, every node is the edge node at the beginning, and, therefore, is able to grow the network to any direction when necessary. This structure increases the flexibility of the network at early stages and enables the network to acquire new knowledge more rapidly.

At the growing phase, each training token is inputted into the network sequentially, so that one token is trained for multiple iterations and then the next token enters. The learning rate and neighborhood size is reset to their initial value when a new token enters. During the training process, the network will update the feature vectors of the BMU node and the nodes within BMU's neighborhood by the rule described in Equation 2 [19,20]. In Equation 2, $R_{learn}(t)$ represents the learning rate function that decrease the learning rate by each iteration; h(t) is the Gaussian neighborhood function that determines the activated area within BMU's neighborhood; x(t) represents the feature vector of the input training token, and ω(t) represents the feature vector of the BMU node. The feature updating rule described in Equation 2 is consistent with the perceptual magnet effect.

$$\omega_i(t+1)=\omega_i(t)+R_{learn}(t)\times h(t)\times(x(t)-\omega_i(t)), i \in N$$

If a BMU node in the network reaches its growing threshold, and that node is an edge node, then the model will add a new neuron node at the optimized position using the optimized growing strategy, and therefore grows the network to an optimized direction.

Results and Discussion

A complete simulation in our experiment involves 61 training steps, including 1 initializing training and 60 growing trainings. After each training step, we examined the learning result in details. At the examining stage, the training tokens are inputted into the trained network, and BMUs are identified as the learned phonemes by calculating the Euclidean distance between the feature vectors of the input token and nodes in the trained network. Both of the vowel and consonant phoneme features are stored in the same trained knowledge network. Five simulations were performed with identical training data and modeling parameters. From the root-mean-square (RMS) based and network-structure based analyses, each of the five simulations reflects similar results. Therefore, the following RMS analysis is demonstrated by the average result of the five simulations, and the network structure analysis is demonstrated by one of the five simulations.

Root Mean Square Analysis

The root mean square (RMS) (see Equation 3 [23]) is a common measure for competitive learning tasks [24]. In Equation 3, N represents the number of input tokens, K represents the number of neuron nodes within the network, μ_{ki} represents the correlation factor between the input token x and the network node c. If the node c is the BMU of the token x, $\mu_{ki}=1$, otherwise, $\mu_{ki}=0$. The RMS value is determined by the feature differences between the input token and the BMU node. The less is the difference, the smaller the RMS value is, and the better the learning results are.

$$MSE=\frac{1}{N}\sum_{i=1}^{N} E_i \ , E_i=\sum_{k=1}^{K} \mu_{ki} \ \| x_i - c_k \|$$

We have examined the RMS values of each training step for the five simulations. Figure 3 shows the averaged RMS values of the five simulations at each training step. As the training develops, the RMS value of the network decreases. This result indicates that, with a series of training steps, the network can acquire the acoustic features of the training data well, and guarantee a high standard learning result.

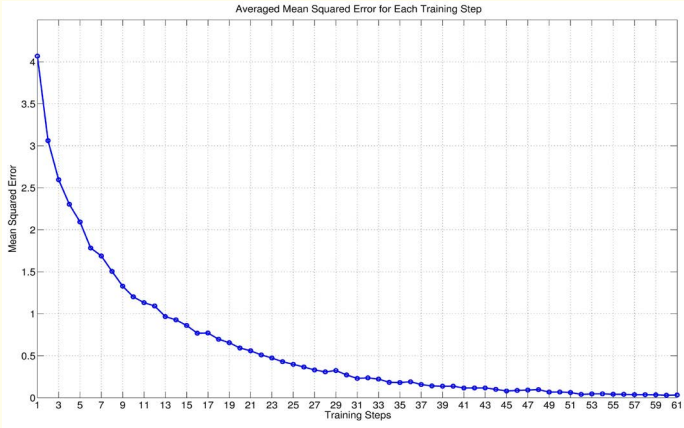

Figure 3: *The averaged RMS values for each training step of the five simulations. The x-axis represents the training steps and y-axis represents the RMS value.*

Network Structure Analysis

After training, each neuron node in the knowledge network represents the acoustic features of both vowels and consonants of a monosyllabic word. We examined the network structure and the acoustic features represented by the nodes in the network. By analyzing the representation of vowel features, we can observe the acquisition result of vowel categories; and by examining the representation of manner-of-articulation features, we can observe the acquisition result of consonants.

Vowel Cluster Analysis

As shown in Figure 4, we can clearly observe different vowel clusters. Syllables contain vowel [a] are located at the right area of the network; syllables contain vowel [i] are located at the upper area; syllables contain vowel [u] are located at the left area; and syllables contain vowel [e] and [o] are located at the central area of the network. This distribution of vowel clusters is consistent with the "vowel triangle" in the acoustic space, which forms into the "front–center–back" and "low–high" spatial relations. The distribution of vowels [e] and [o] has some overlapped areas. This phenomenon indicates that, when processing the acoustic features of [e] and [o], young children cannot clearly perceive the differences between those two vowels, and therefore have difficulties in distinguishing those two phoneme categories.

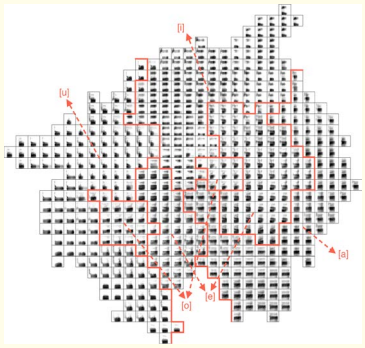

Figure 4: *The trained knowledge network and the vowel features represented by the neuron nodes in the network. Each square in the network represents a neuron node (510 in total), and the spectrogram in each square represents the acoustic feature representation of the node; read lines indicates the borders between each vowel clusters that are marked by red texts.*

The above results show that the modeling algorithm can acquire the acoustic features of vowels in the syllable and clearly distinguish different vowel categories. Therefore, the vowel distribution pattern corresponding to the vowels' features in the acoustic space is established. Consequently, we predict that, during language acquisition, young children have the ability to acquire vowel categories and build acoustic-related spatial structures by processing the acoustic features of the speech signals.

Consonant Cluster Analysis

As shown in Figure 5, we can observe a clear distinction between the syllables contain aspirated plosive and unaspirated plosive. In the knowledge network, different from the distribution of vowel categories, clusters of different manners of articulation distribute scattered across the network. Comparing Figure 5 and Figure 4, however, we observe that the distribution of aspirated and unaspirated features relies on the distribution of vowel categories. From right to left, as the vowel moves from low to high, aspirated and unaspirated plosives are reasonably distributed within each vowel category.

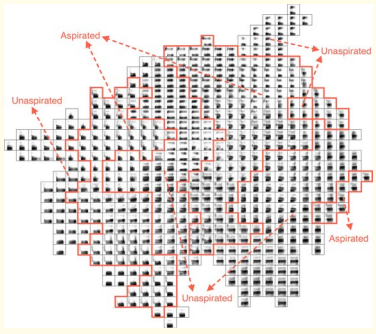

Figure 5: *The trained knowledge network and the aspirated/unaspirated features represented by the neuron nodes in the network. Each square in the network represents a neuron node (510 in total), and the spectrogram in each square represents the acoustic feature of the node; read lines indicates the borders between different manner-of-articulation clusters that are marked by red texts.*

From the perspective of acoustic representation, spectrogram can clearly reflect the differences between aspirated and unaspirated plosives. However, we cannot observe a clear distinction between aspiration and non-aspiration in the trained knowledge network. This situation can be interpreted as the followings. As stated above, in this study we simulate the early stage of phoneme acquisition where young children treat the whole syllable as a perceptual unit. At this stage, the model (i.e., the simulated child) does not have the ability to process vowels and consonants independently. As a result, vowels dominant the perception of a syllable with their significantly longer duration and higher amplitude. However, this does not mean that the model (i.e., the simulated child) does not have the ability to distinguish aspirated and unaspirated plosives. In fact, the model can form the aspiration and non-aspiration feature clusters within each vowel category, and establish the distinction with the help of the perception of vowels.

The above results together indicate that the modeling algorithm proposed in this study has the ability to distinguish the aspiration and non-aspiration feature and establish the manner-of-articulation clusters based on the distribution of vowel categories. Consequently, we predict that, during language acquisition, young children have the ability to acquire the manner-of-articulation of consonants by processing the acoustic features of CV syllables.

Conclusion

In this study, based on the GSOM algorithm and the optimized growing strategy, we simulated the early phoneme acquisition of Standard-Chinese children. As pointed out by Foster-Cohen [21], the whole syllable is treated as a perceptual unit for children at this stage. By analyzing the RMS results of the network at each training step, we conclude that the proposed modeling algorithm can achieve good

learning results. By analyzing the learned network structure, we observe that the model (i.e., the simulated child) can distinguish different vowel categories and manner-of-articulation of aspiration and non-aspiration, by processing the acoustic features of the whole syllable. Since vowels dominant the acoustic feature representation of a syllable, the knowledge network perceives acoustic features mainly based on the distribution of vowel categories. Within each vowel cluster, the network can further establish vowel-based distinctions between aspirated and unaspirated plosives.

The modeling results predict that, at early stages of language acquisition, by processing the acoustic features of the whole syllable, children have the ability to acquire vowel categories and their spatial relations, and build manner-of-articulation distinctions based on vowel categories.

The GSOM-based self-organizing neural network model proposed in this study can well simulates the knowledge topographic structure, the incremental nature of knowledge, and the self-organizing process of knowledge learning. Its learning pattern is consistent with the perceptual magnet effect, and can simulate the statistical learning mechanism. In future studies, we will focus on the transition process from the syllable-based perception to phoneme-based perception, and examine the mechanisms behind.

Acknowledgements

This study is funded by China Postdoctoral Science Foundation (NO. 2016M590057) and Beijing Normal University Young Teachers Grant (NO. SKXJS2015012).

Conflict of Interest

We declare no conflict of interest exists.

Bibliography

1. Kuhl Patricia K. "Early language acquisition: cracking the speech code". *Nature Reviews Neuroscience* 5.11 (2004): 831-843.
2. Kuhl Patricia K. "Brain mechanisms in early language acquisition". *Neuron* 67.5 (2010): 713-727.
3. Kuhl Patricia K. "Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception". *Journal of Phonetics* 21.1-2 (1993): 125-139.
4. Maye Jessica., *et al*. "Infant sensitivity to distributional information can affect phonetic discrimination". *Cognition* 82.3 (2002): B101-B111.
5. Kohonen, Teuvo. "The self-organizing map". *Neurocomputing* 21.1 (1998): 1-6.
6. Kohonen Teuvo. "Essentials of the self-organizing map". *Neural Networks* 37 (2013): 52-65.
7. Kuhl Patricia K. "Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not". *Attention, Perception and Psychophysics* 50.2 (1991): 93-107.
8. Kuhl P., *et al*. "Linguistic experience alters phonetic perception in infants by 6 months of age". *Science* 255.5044 (1992): 606-608.
9. Ritter Helge and Teuvo Kohonen. "Self-organizing semantic maps". *Biological Cybernetics* 61.4 (1989): 241-254.
10. Kröger Bernd J., *et al*. "Learning to associate speech-like sensory and motor states during babbling". *Proceedings of the 7th International Seminar on Speech Production* (2006): 67-74.
11. Kröger Bernd J., *et al*. "Towards a neurocomputational model of speech production and perception". *Speech Communication* 51.9 (2009): 793-809.

12. Gauthier Bruno., *et al*. "Simulating the acquisition of lexical tones from continuous dynamic input". *The Journal of the Acoustical Society of America* 121.5 (2007): EL190-EL195.

13. Li Ping., *et al*. "Early lexical development in a self-organizing neural network". *Neural Networks* 17.8 (2004): 1345-1362.

14. Li Ping., *et al*. "Dynamic Self-Organization and Early Lexical Development in Children". *Cognitive Science* 31.4 (2007): 581-612.

15. Li Ping and Xiaowei Zhao. "Self-organizing map models of language acquisition". *Frontiers in Psychology* 4 (2013): 828.

16. French Robert M. "Catastrophic forgetting in connectionist networks". *Trends in Cognitive Sciences* 3.4 (1999): 128-135.

17. Alahakoon Damminda., *et al*. "Dynamic self-organizing maps with controlled growth for knowledge discovery". *IEEE Transactions on Neural Networks* 11.3 (2000): 601-614.

18. Matharage Sumith and Damminda Alahakoon. "Growing Self Organising Map Based Exploratory Analysis of Text Data". *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* 8.4 (2014): 639-646.

19. Cao Mengxue., *et al*. "Growing self-organizing map approach for semantic acquisition modeling". Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on. IEEE (2013): 33-38.

20. Cao Mengxue., *et al*. "Interconnected growing self-organizing maps for auditory and semantic acquisition modeling". *Frontiers in Psychology* 5 (2014): 236.

21. Foster-Cohen and Susan H. "An introduction to child language development". Pearson Education (1999).

22. Tai Wei-Shen and Chung-Chian Hsu. "A growing mixed self-organizing map". Natural Computation (ICNC), 2010 Sixth International Conference on. IEEE (2010): 986-990.

23. Tai Wei-Shen and Chung-Chian Hsu. "Growing Self-Organizing Map with cross insert for mixed-type data clustering". *Applied Soft Computing* 12.9 (2012): 2856-2866.

24. Du K-L. "Clustering: A neural network approach". *Neural Networks* 23.1 (2010): 89-107.

Volume 6 Issue 5 June 2017