A Computational Model on the Emergence of Vowel Transition Perception

Ching-Pong Au[†] **and Christophe Coupé**[‡]

† Language Engineering Laboratory, City University of Hong Kong, Hong Kong SAR, China

‡ Laboratoire Dynamique du Langage, Lyon, France

E-mail: bong.au@plink.cityu.edu.hk, ccoupe@ens-lyon.fr

ABSTRACT

The speech generalizing abilities of infants in their native language improve as they grow up throughout their first half-year [1]. The emergence of this ability is a self-organizing process because the access of word meaning of the young infants is still very limited. Guenther and Gjaja simulated this phenomenon successfully by using self-organizing maps [2]. In their models, however, only static features are considered as inputs, and internal temporal relationships within speech sounds are not considered. In the present study, we simulate the same phenomenon by modifying a word recognition model that takes temporal features into account [3,4]. A new self-organizing algorithm, relying on the competition among dendrites or neurons, is used in our model.

1. INTRODUCTION

Experiments show that two acoustic sounds are perceived more similar if they are acoustically closer to their prototype. Kuhl named this phenomenon, "perceptual magnet effect" (PME) because the prototype of sounds functions like a magnet for the other members in the category and assimilates neighboring stimuli, effectively pulling them toward the prototype in the perceptual space. [5]. The formation of this effect has been found in very young infants whose abilities in generalizing the speech in their native language improve as they grow up throughout their first half-year [1]. Since infants' access to the meaning of speech is still very limited around one year old, the emergence of the magnet effect is likely to be a self-organizing process. Guenther and Gjaja (G&G) hypothesized that the effect emerges because of the non-uniformity of the auditory map due to the auditory experience of the native language, and simulated this effect successfully by using self-organizing maps [2]. However, this pioneering model simulating PME abstracts away two important characteristics of human speech: (1) The representation of the inputs are not very likely to be realistic because a few single neurons are assumed to represent the formant frequencies of the speech; (2) Temporal relationships within the speech sound are ignored in the model. In this paper, we adapt G&G's hypothesis and intend to improve the two aspects that their model lacks.

In a recent model [6], instead of using the former kind of idealistic representation of formants, a number of neurons representing signals from different frequencies and ordered tonotopically are used as the input representation in agreement with the fact that tonotopical representation were found in various auditory nuclei and auditory cortex. In the model, selection of neurons due to the input signals is used as the development algorithm of the model, called 'survive-and-spread' algorithm: only neurons get enough activation can survive; and then, if further activated, their dendritic connections spread to the nearby input units. The simulation shows the emergence of PME due to different linguistic environment [6]. The idea is based on the fact that a large number of neurons die during the early brain development (see review in [7]) and the philosophical account that the neuronal death is not just random, but a strategy of development [8]. In the present study, tonotopical representation of inputs and a similar selectionistic development algorithm are used.

In G&G's model, although using formant values to represent single vowels may be sufficient, speech sounds with internal temporal relationships such as diphthongs cannot be satisfactorily represented. In the present study, we modify a biologically plausible word recognition model recently proposed by Hopfield and Brody (H&B), which takes temporal features into account, to fit our purpose [3,4].

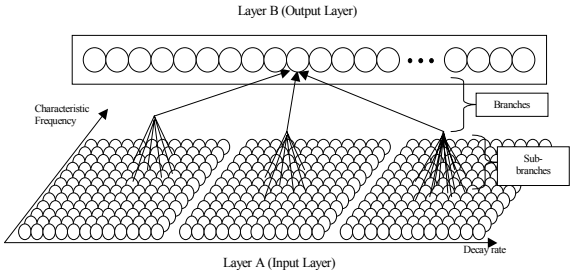
Our model here was tested with artificial formants. The mechanism of the model will be introduced in section 2 and its simulation results will be reported in section 3. Finally, conclusions and future development will be discussed.

2. THE MODEL

The model has 2 layers: layer A (input layer) and B (output layer). In layer A, there are three distinct bi-dimensional maps of inputs (The number of the input maps is not restricted to 3, but this is the minimum required number to show the convergence of the synchronized firing patterns). These 3 maps can be conceptually located at any level of the auditory pathway. Each of these maps has a fixed number of neuronal units organized in a two dimensional manner. Neuronal units are organized tonotopically from the lowest to the highest frequency in one dimension while the neuronal units are organized from the lowest to the highest decay rate (will be explained later) in another

ISBN 1-876346-48-5

(fig.1). Layer B contains a varying number of neurons. Each neuron has 3 branches and each of these branches has



a varying number of dendrites that collects signals from the units in layer A.

Fig.1 Structure of our model

The tonotopical organization of the frequency in the auditory nuclei and auditory cortex has been well attested. However, the existence of the ordering of the decay rate is relatively hypothetical. The assumption is made primarily because of the functional requirement of the system: since human has the ability of generalizing the length of speech, there must be some way to represent the continuous relationship of time in the auditory pathway. Relying on a bi-dimensional tonotopical organization is reasonable because bi-dimensionally organized cortex can be found in other animals such as mustached bat [9]. However, claims on the reality of the network structure cannot be firmly defended since fine details about the organization of many parts of the human auditory pathway are still unknown.

The learning algorithm (or development algorithm, a term that may be more appropriate) includes a few steps for the iteration, as detailed below:

1. A large number of developing neurons are added to layer B with the 3 branches attached to the 3 input maps accordingly (See fig. 1). Each branch initially attaches to five input units (in a cross shape), but locates randomly on the map.
2. A number of signals simulating artificial formants are fed into the system. Neurons with sufficient activation from the input are kept in the layer while those are not activated frequently die immediately and are taken away from the layer. A 'survival' index is used to measure how often the neuron is activated. Whenever the neurons fire, the index will be increased by a small value. If the value of this index is over a threshold by the end of this iteration, the neuron can survive.
3. At the same time, the surviving neurons from the previous iterations spread their dendrites further by adding sub-branches to the four sides of the existing sub-branches if the 'spreading' index of the sub-branches is higher than a threshold. Those sub-branches that play a role in synchronizing and firing the neuron increase their 'spreading' index by a certain value, and at the same time, the 'spreading' index of other sub-branches in the same neuron is increased by a smaller value. However, each unit in all the input maps can allow only a limited number of sub-branches to connect to. Once the units have reached the limit, even if a sub-branch nearby is activated, no new sub-branch can grow into this fully occupied unit.

The simulation can go on until the system becomes stable (i.e. most of the popular input sites are being blocked). At this time, no significant change in the responses can be observed.

Since we consider the temporal relationship features within the speech patterns, the recalling mechanism is largely based on the concept of the H&B model. The model relies on the transient synchronization of integrate-and-fire spiking neurons with convergent firing rates. There are three layers in their model. The first layer is a tonotopically organized layer in which there are three types of neurons responding to the onsets, peaks and offsets of the incoming signals; the second layer contains both excitatory (α cells) and inhibitory (β cells) neurons. The neuron outputs of the first layer are linked to the neurons in the second layer. The α and β neurons in the second layer are interconnected and some of them are connected to the third layer, i.e. the output layer.

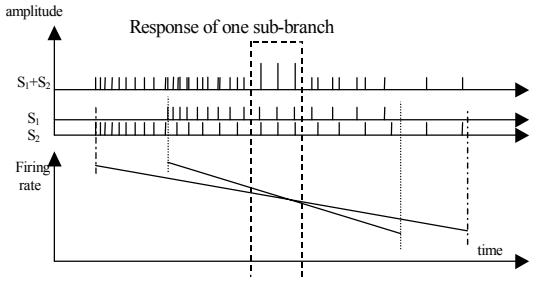


Fig. 2 Simplification of the model

In this model, the response of each neuron is a pattern of decreasing firing rate (see signals, S_1 and S_2 in fig. 2; decay rate is the slope of the lines in the lower graph). In the neurons, with a large number of incoming links, all the signals received are summed up. Only those with similar frequencies and phases can add up to a larger value of amplitude as shown in the upper panel (S_1+S_2) of fig. 2. If a large number of inputs of a neuron have the same frequencies and phases, the neuron fires if the amplitude is over a certain threshold.

In the model we develop here, we introduce a learning algorithm that requires a demanding computation power. Therefore, we simplify H&B's recalling mechanism to let us see the interesting results obtained by using this innovative learning algorithm, as our interest is not the detailed behavior in the neurons. The simplification includes the following: (1) the patterns for pulses with decreasing firing rate, as in the upper panel of fig 2, are only represented by straight lines, as shown in the lower panel. (2) The second layer of the H&B model is not explicitly implemented in our model. We assume that patterns can be synchronized if the frequencies of the inputs alone match at a certain time; The phases are assumed to be then always synchronized as they are in H&B's model by the α and β neurons, but which are ignored in our model. (3) To keep our model simple and easy to be analyzed, all the neurons in the output layers have only 3 input links (the 3 branches) from the input layer as shown in fig. 1. Neurons in layer B can be activated only when all three branches

have at least one sub-branch containing a suitable pattern such that the three patterns can be synchronized successfully. Recall that the whole system is simply feeding in an input signal to layer A and then observing the responses of all the neurons in layer B.

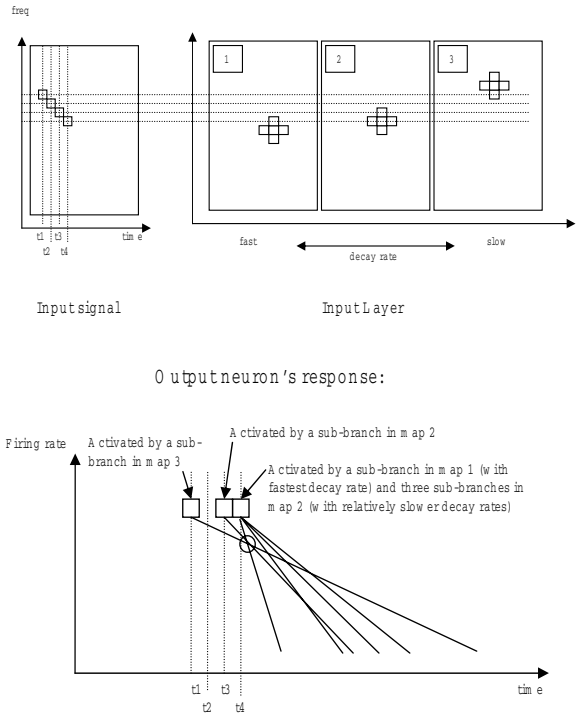


Fig. 3 An illustration of the mechanism of neuron firing

With the above simplification, we end up with a simple recall mechanism that can capture the temporal properties in speech. The recall mechanism is illustrated with an example in fig. 3. Assuming there is one neuron which sub-branches of each branch connect to the 3 input maps in cross shapes (initially, the connections are in cross shapes, but when the algorithm goes on, it can be spread and form other shapes). Shown is a pattern detected from a formant of a diphthong (The 4 squares in the upper left graph). The squares activate the input neurons in 4 different characteristic frequencies at 4 different time. At time t1, only one sub-branch in map 3 is activated. No sub-branch is activated at time t2. Then, at time, t3, one sub-branch in map 2 is active, and finally 1 sub-branch in map 1 and 3 sub-branches in map 2 are simultaneously activated at t4 (fig. 3, the upper right graph). Consequently, since the 3 signals (3 straight lines, one from each map) are synchronized (at the intersection point marked with a circle, fig. 3, the lower graph), this neuron fires.

Based on the system described above, PME can be shown. The results and the evaluation of the system performance will be reported in the next section.

3. RESULTS

In the simulations, an artificial linguistic environment is provided to the system in the learning stage (development stage). Formants of an artificial vowel transition with variations in length, starting and ending frequencies are used as the linguistic environment. Since only onset detectors are simulated, only the black squares of the patterns in fig. 4 can be detected through the input units.

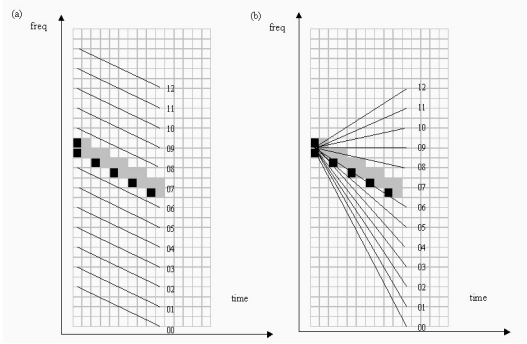


Fig. 4 Testing Stimuli from two different dimensions

Two sets of testing patterns that vary in two different dimensions are used to test the system in different stages of training. The first testing set contains formants with variation of average frequency but constant slope fixed at the mean value in the learning environment (fig. 4a). The second set varies in formant slopes realized by fixing the starting frequency but varying the ending frequency (fig. 4b).

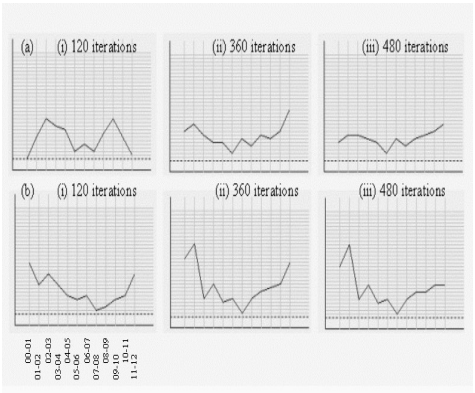


Fig. 5 Results of two runs with testing stimuli in fig. 4a.

By feeding in the 13 signals in fig. 4a, firing patterns of the neurons in layer B are obtained. The number of neurons having different responses to two neighboring signals is counted and plotted in fig. 5 (comparison between signal 00 and 01, between 01 and 02, and so on). Different 'infants' show different developmental pathways. Fig. 5a and 5b are the ontogenetic changes for two 'infants' (two trials) of the model. Fig. 5aiii and 5biii (stabilized stage) shows that the closer the signals to the mean of the training sets (around signal 07), the higher the similarity of the two neighboring patterns. This explains the V-shape of the curves.

The general patterns fit the phenomenon of PME, although some parts are not very smooth. The wrinkled portions can have two possible explanations. An optimistic explanation may be that the model can realistically show the real fact

ISBN 1-876346-48-5

that each individual has such a non-smooth response and the wrinkles were hidden by averaging the responses of a number of subjects in the empirical experiments [1]. Another possibility may be due to the unrealistic characteristics of our model. The population of neurons is limited by the present technology. Only a limited number of neurons are allowed in the model. The curves may become smoother if a larger population can be considered in the model. Further investigation in the future is necessary.

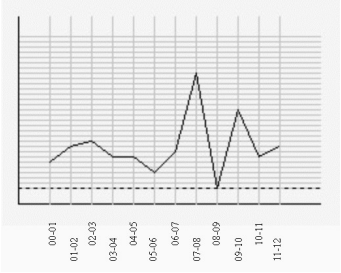


Fig. 6 Result with testing stimuli in fig. 4b. (480 iterations)

Stimuli in fig. 4b are used to test the trained system and the result is shown in fig. 6. Left hand side shows the comparison of the signals with greater slope. PME can be shown partly on the left hand side. Right hand side is not consistent with the PME. The comparisons of the responses drop/raise sharply at the right hand side of the pattern because signals with little or no frequency variation cannot be detected through the onset detectors. The weaknesses should be able to be improved if other detectors (peak and offset) are also implemented in future.

5. CONCLUSIONS

In this study, we have built a model that tries to capture the internal temporal information of signals so that the emergence of the PME on vowel transitions can be realized. The model can preliminarily show the PME quite well for transitions but not for steady vowels. Not responding to a steady vowel seems very odd as a human behavior. However, this is, to a certain extent, consistent with the experimental results suggested by some early perceptual studies that vowel identification is more accurate for vowels cued only by formant transitions than for the formants of the steady portions heard in isolation [10].

We have also tried to apply the model on real speech data. Some specialized settings are added to the present model in order to make learning of real speech possible. The main challenge is the computational resources required to run models with a large number of highly connected neurons and correctly capture the diversity of real signals. Our first attempts with a small number of connections have provided preliminary but encouraging results: with either real speech or artificial noisy vowels organized in clusters as input signals, some of the clusters of sounds show the emergence of PME. In order to deal with the computational burden due to the large number of connections in our model, we are currently developing some efficient computing algorithms by modifying some algorithms used in image processing. We hope these faster algorithms can help in shortening the computation time so that we can see some interesting results in the future models.

ACKNOWLEDGEMENTS: This work was supported in part by grants from the City University of Hong Kong [#9010001], the Hong Kong Research Grants Council [#9040781], and Academia Sinica [Taiwan].

REFERENCES

[1] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science*, vol. 255, pp. 606-608, 1992.

[2] F. H. Guether and M. N. Gjaja, "The Perceptual Magnet Effect as an Emergent Property of Neural Map Formation," *JASA*, vol. 100, pp. 1111-1121, 1996.

[3] J. J. Hopfield and C. D. Brody, "What is a moment? "Cortical" sensory integration over a brief interval", *PNAS*, vol. 97, no. 25, pp. 13919-13924, 2000.

[4] J. J. Hopfield and C. D. Brody, "What is a moment? Transient synchrony as a collective mechanism for spatiotemporal integration", *PNAS*, vol. 98, no. 3, pp. 1282-1287, 2001.

[5] P. K. Kuhl, "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not," *Perception & Psychophysics*, vol. 50, pp. 93-107, 1991.

[6] C-P. Au, "Emergence of Speech Perception by Plausible Neuronal Phenomena" in *Language Acquisition, Change and Emergence*, J. W. Minett and W. S-Y. Wang, Ed., forthcoming.

[7] R. W. Oppenheim, "Programmed Cell Death," in *Fundamental Neuroscience*, M. J. Zigmond, Ed., pp581-609, 1999.

[8] R. Dawkins, "Selective Neurone Death as a Possible Memory Mechanism", *Nature*, vol. 229, pp. 118-119, 1971.

[9] N. Suga, "Auditory Neuroethology and Speech Processing: Complex-Sound Processing by Combination-Sensitive Neurons," *Auditory Function*, G. M. Edelman, W. E. Gall and W. M. Cowan, Ed., pp. 679-720, 1988.

[10] R. R. Verbrugge, W. Strange, D. P. Shankweiler and T. R. Edman, "What Information Enables a Listener to Map a Talker's Vowel Space?," *JASA*, vol. 60, pp. 198-212, 1976.