SPEECH RECOGNITION USING MODULAR ORGANIZATIONS BASED ON MULTIPLE HOPFIELD NEURAL NETWORKS

Yohji Fukuda and Haruya Matsumoto

Faculty of Engineering, Kobe University
1-1, Rokkodai, Nada, Kobe 657, JAPAN

ABSTRACT - In this paper, we describe the speech recognition that introduces the modular organizations based on Multiple Hopfield Neural Network (MHNN). MHNN is composed of several Hopfield Networks that are connected to each other. Each network owns the individual energy function but when connected, the total energy of the whole network can be minimized because MHNN interacts between the networks. Our recognition architecture has two phases. In the first phase, two mapping-type networks extract features from a spectrum data and a pitch data as the modular organization's inputs. In the second phase, MHNN recognizes the speech signal by the interactions between the networks. We perform Japanese name recognition by using MHNN.

INTRODUCTION

Recently, neural networks for the speech recognition have been studied in the hope of achieving human-like performance. Especially, Kohonen Network (Kohonen, 1988) and Time Delay Neural Network (Waibel, Hanzawa, Hinton, Shikano & Lang, 1989) obtain good performance for the speech recognition. However, their networks do not achieve a human-like parallel processing. The reason of this is that they process only one kind of input such as a cepstrum, a bank filter, and so on. Because the speech signal includes the information of a spectrum, a pitch, and time progress, the speech recognition is difficult to treat such as the complex information. The problem of speech recognition is that we treat only one kind of data instead of the complex information.

It is assumed that biological systems consist of several modular organizations and their interactions act effectively on the dynamical behavior of the total system. We suppose that each modular organization acts feature extraction of spectrum, accent, and so on at lower order in the brain. At higher order, man understands the words and the contexts from the obtained features at lower order. The expected advantage for such a modular organization is the mutual coupling of some simple modules at large scale architecture.

The methodology for integrating some modular organizations, therefore, is important to construct more complex and higher order neural systems. Some recent modular organizations can be interpreted as the integrated systems through this attempt. For instance, such systems are Adaptive Resonance Theory (ART) (Carpenter & Grossberg, 1988), Bidirectional Associative Memory (BAM) (Kosko, 1988), Cross-Coupled Hopfield Net (CCHN) (Tsutsumi, 1990), and so on.

Multiple Hopfield Neural Network (MHNN) is composed of several Hopfield Networks (HN) (Hopfield & Tank, 1988) that are connected to each other. Each network owns the individual energy function but when connected, the total energy of the whole network can be minimized because MHNN interacts between HNs, just like BAM does.

Our recognition architecture has two phases. In the first phase, two mapping-type neural networks extract features from a spectrum data and a pitch data respectively. MHNN has two HNs whose inputs are the mapping-type network's outputs. In the second phase, MHNN recognizes the speech signal from obtained features at each network.

In the simulation, we perform Japanese name recognition that uttered by a Japanese male speaker.

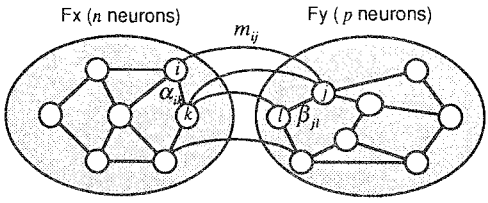


Figure 1. The architecture of MHNN.

Lyapunov function

MHNN is composed of several HNs that are connected to each other. The architecture of MHNN is shown in Figure 1. In this paper, we assume that MHNN is composed of two HNs. We denote by $\mathbf{M}$ the *n-by-p* weight matrix that interconnects two HNs of neurons. One HN Fx has n neurons, and the other HN Fy has p neurons. The weight m_{ij} connects from the ith neuron in Fx to the jth neuron in Fy. $\mathbf{A}$ (α_{ik}) and $\mathbf{B}$ (β_{jl}) are within-network weight matrix of Fx and Fy, respectively. The activation of neuron is described by a real-valued potential $x_i(t)$ (the ith neuron in Fx) and $y_j(t)$ (the jth neuron in Fy). The neuron transforms this activation into a signal $S_i = S_i(x_i)$ and $V_j = V_j(y_j)$. S_i and V_j can be any monotone-increasing function: $S_i', V_j' > 0$. We define the following Lyapunov function L for MHNN.

$$L = -\sum_{i=1}^{n}\sum_{j=1}^{p} S_i(x_i)V_j(y_j)m_{ij} + \sum_{i=1}^{n}\int_0^{x_i} S_i'(\theta_i)b_i(\theta_i)d\theta_i + \sum_{j=1}^{p}\int_0^{y_j} V_j'(\varepsilon_j)b_j(\varepsilon_j)d\varepsilon_j + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p} m_{ij}^2$$
$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} S_i(x_i)S_k(x_k)\alpha_{ik} + \frac{1}{4}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_{ik}^2 - \frac{1}{2}\sum_{j=1}^{p}\sum_{l=1}^{p} V_j(y_j)V_l(y_l)\beta_{jl} + \frac{1}{4}\sum_{j=1}^{p}\sum_{l=1}^{p}\beta_{jl}^2 \quad (1)$$

where b_i is arbitrary. The time differentiation of the Lyapunov function L given in the equation (1) is represented as follows:

$$\dot{L} = -\sum_{i=1}^{n} S_i'\dot{x}_i\{\sum_{j=1}^{p} V_j(y_j)m_{ij} - b_i + \sum_{k=1}^{n} S_k(x_k)\alpha_{ik}\} - \sum_{j=1}^{p} V_j'\dot{y}_j\{\sum_{i=1}^{n} S_i(x_i)m_{ij} - b_j + \sum_{l=1}^{p} V_l(y_l)\beta_{jl}\}$$
$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\dot{\alpha}_{ik}\{S_i(x_i)S_k(x_k) - \alpha_{ik}\} - \frac{1}{2}\sum_{j=1}^{p}\sum_{l=1}^{p}\dot{\beta}_{jl}\{V_j(y_j)V_l(y_l) - \beta_{jl}\} \quad (2)$$
$$-\sum_{i=1}^{n}\sum_{j=1}^{p}\dot{m}_{ij}\{S_i(x_i)V_j(y_j) - m_{ij}\}$$

$$\dot{x}_i = \sum_{j=1}^{p} V_j(y_j)m_{ij} - b_i + \sum_{k=1}^{n} S_k(x_k)\alpha_{ik} \quad (3)$$

$$\dot{y}_j = \sum_{i=1}^{n} S_i(x_i)m_{ij} - b_j + \sum_{l=1}^{p} V_l(y_l)\beta_{jl} \quad (4)$$

The dynamical system x and y are defined by the equation (3) and (4). If $\mathbf{M}$, $\mathbf{A}$, and $\mathbf{B}$ are fixed, we can deduce the following equation:

$$\dot{L} \le 0 \quad (5)$$

Thus the dynamical system by the equation (1) is globally stable, just like BAM does. It is assume that

MHNN has non-interconnection between Fx and Fy, the equation (3) is the following:

$$\dot{x}_i = \sum_{k=1}^{n} S_k(x_k)\alpha_{ik} - b_i \quad (6)$$

Replacing $S_k(x_k) = V_k, \alpha_{ik} = T_{ik} / C_i$ and $b_i = x_i / R_i C_i - I_i / C_i$, we obtain the following:

$$C_i \dot{x}_i = -x_i / R_i + \sum_{k=1}^{n} T_{ik} V_k + I_i \quad (7)$$

This equation is the same as Hopfield circuit. Therefore, MHNN is that each HN owns the individual energy function but when connected, the total energy of the whole network can be minimized because MHNN interacts between networks.

Row & column problem

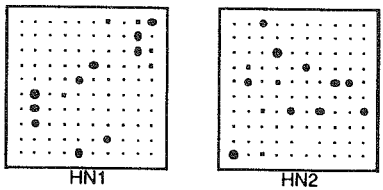
To confirm the efficiency, we perform a simulation called "Row & Column Problem." This problem must be only one unit "1" output in each row and each column in the *n-by-n* square, all other unit being "0" output. HN1 is defined so that the energy function such as only one "1" output is in each row and HN2 is defined in each column.

$$E_{HN1} = \frac{A}{2} \sum_{X=1}^{n} (\sum_{i=1}^{n} S_{Xi} - 1)^2 \quad (8)$$

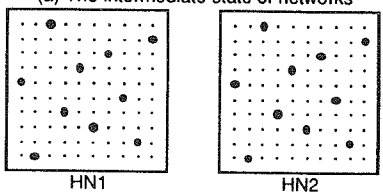
$$E_{HN2} = \frac{B}{2} \sum_{i=1}^{n} (\sum_{X=1}^{n} V_{Xi} - 1)^2 \quad (9)$$

The total energy function E is defined the following:

$$E = \frac{A}{2} \sum_{X=1}^{n} (\sum_{i=1}^{n} S_{Xi} - 1)^2 + \frac{B}{2} \sum_{i=1}^{n} (\sum_{X=1}^{n} V_{Xi} - 1)^2 + \frac{C}{2} \sum_{X=1}^{n} \sum_{i=1}^{n} (S_{Xi} - V_{Xi})^2 \quad (10)$$



(a) The intermediate state of networks



(b) The final state of networks

Figure 2. The typical convergence of "Row & Column Problem".

The last term is zero if and only if the final state of HN1 corresponds to the final state of HN2. Figure 2 shows the results of a simulation that illustrates the typical convergence of a state such as an intermediate

state to a final state. Each network behaves as the individual energy function to be minimized. After a while, the total energy of the whole network is minimized because MHNN interacts between HN1 and HN2. Finally, the final state of HN1 corresponds to the final state of HN2.

JAPANESE NAME RECOGNITION

Recognition architecture

We perform Japanese name recognition that uttered by a Japanese male speaker. We use the following five Japanese names; "okue", "kamei", "tamio", "kikuta", and "matuo". 256-point FFT with Hanning window is computed every 10 ms from the input words sampled at 8 KHz. The training data has 25 words (the number of each word is 5) and the testing data has also 25 words.

Our recognition architecture has two mapping-type neural networks and one MHNN. The proposed network is shown in Figure 3. The mapping-type networks use the following two types; Recurrent Neural Network (RNN) (Fukuda & Matsumoto, 1991) and Non-recurrent Neural Network (NNN). NNN and RNN extract features from a spectrum data and a pitch data respectively. The spectrum data represents phoneme information at every time and the pitch data represents a relationship between phoneme and time progress. We use RNN for the feature extraction from the pitch data because RNN has advantage of treating the temporal sequences independent of time length. Each network classifies the inputs data into vowels, "a", "i", "u", "e", "o"; consonants, "k", "t", "m", and nothing as "N". MHNN has two HNs that are called HN1 and HN2. The inputs of HN1 and HN2 are the outputs of NNN and RNN respectively. HN has *8-by-time length* units. Each energy function of HN is defined so that the only one "1" output is in the each phoneme column. Furthermore, the energy function of HN2 is defined so that the neighboring row units that belong to the same class intensify each other. Because the pitch data has the relationship between phoneme and time progress, the energy function of HN2 must have such relationship. If the final state of HN1 corresponds to the final states of HN2, MHNN can recognize words from the obtained features at each network.

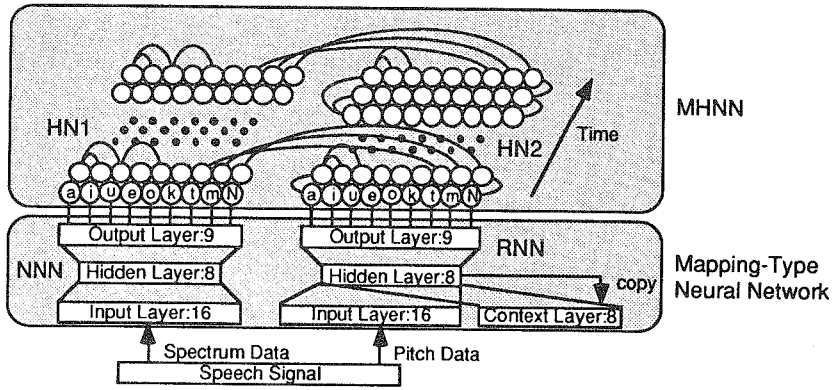


Figure 3. The proposed network.

In the simulation, the spectrum data is 16 order melscale spectrum coefficients and the pitch data is 16 order cepstrum coefficients from the 43rd point to the 58th point. RNN consists of 16 input units, 8 hidden units, 9 output units and 8 context units. NNN consists of 16 input units, 8 hidden units, and 9 output units. RNN and NNN are trained by Back Propagation algorithm. For comparison, NNN whose input is the spectrum data and the pitch data without MHNN is also trained to perform the same task. This network is called CompNNN, consists of 32 input units, 16 hidden units, and 9 output units.

The results of simulation

Figure 4 shows the change of error during a typical learning. Ordinate in the figure indicates the mean

squared error of output layer, averaged over the output units. In the figure, the solid line and the dot line show the learning in NNN whose input is the spectrum and CompNNN whose input is the spectrum and the pitch. As it is clear from the figure, the final error of CompNNN is the same as NNN though CompNNN has many parameters and inputs data. Because the architecture of NNN is more simple than CompNNN, the learning time of NNN is faster than CompNNN in the absolute time. We suppose that the modular organization needs to act feature extraction respectively for several information of speech signal .

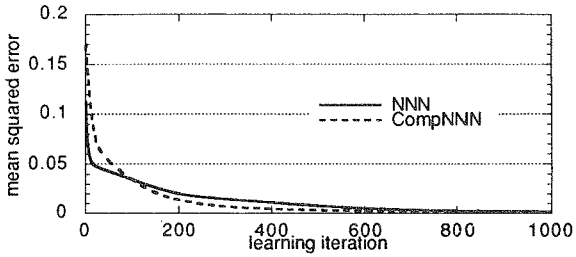
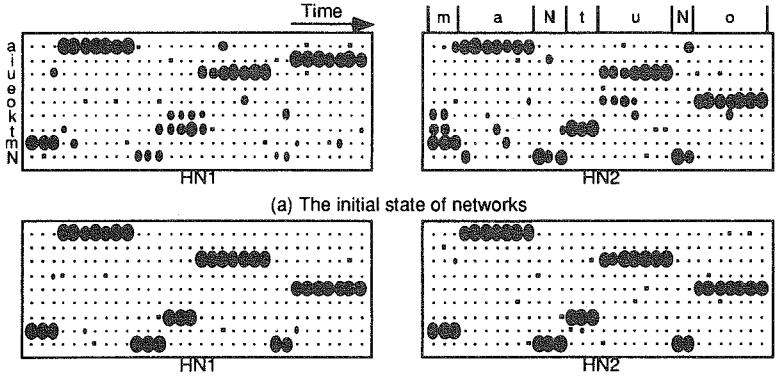


Figure 4. The change of error during typical learning.

Figure 5 shows the results of a simulation that illustrates the typical convergence of a state such as an initial state to a final state with MHNN when "matuo" was uttered. Figure 5 (a) shows the initial state of HN1 and HN2. As it is clear from the figure, HN1 and HN2 recognize the phoneme incorrectly at the several points. Figure 5 (b), at the final state, shows that MHNN interacts between HN1 and HN2. HN1 and HN2 recognize correctly at the point of incorrect recognition in the figure (a). Table 1 shows the results of the word recognition experiments with testing data. MHNN obtains better performance than NNN and CompNNN.



(a) The initial state of networks

(b) The final state of networks

Figure 5. The word recognition with MHNN when "matuo" was uttered.

Table 1. The result of the Japanese name recognition.

network	NNN	CompNNN	MHNN
input data	spectrum	spectrum, pitch	spectrum, pitch
correct recognition	18/25	20/25	22/25

Generally, the input of spectrum data correctly recognizes vowels and consonants that appear in the early time. The input of pitch data recognizes correctly consonants that appear in the intermediate

time. The reason of this is that each data has the different information. The speech recognition system should be capable of integrating several information.

CONCLUSIONS

We discussed the approach for the speech recognition that introduces the modular organizations based on MHNN. After training phase, each network may be the incomplete recognition network because of trapping the local minimum or less training data. However, we can obtain the better recognition results by the interactions between their networks. MHNN may be capable of the speech recognition system integrating each modular organization network, such as the phoneme recognition network, the context recognition network, and so on. It is shown that the modular organizations based on MHNN is useful for the speech recognition system.

MHNN has a capability to integrate several information. This is important for the large scale neural systems because the systems can be composed of some simple modular organizations.

REFERENCES

Carpenter, G.A. & Grossberg, S. (1988) *The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network*, IEEE Computer Vol.21, No.3, 77-88

Fukuda, Y. & Matsumoto, H. (1991) *Phoneme Recognition using Recurrent Neural Networks*, Proc. of 2nd European Conference on Speech Communication and Technology, Vol.3, 1419-1422

Hopfield, J.J. & Tank, D.W. (1985) *Neural Computation Decisions in Optimization Problems*, Biological Cybernetics, Vol.52, 141-152

Kohonen, T. (1988) *The Neural Phonetic Typewriter*, IEEE Computer Magazine, 11-22

Kosko, B. (1988) *Bidirectional Associative Memories*, IEEE Trans. on System, Man and Cybernetics, Vol.18, No.1, 49-60

Tsutsumi, K. (1990) *Cross-Coupled Hopfield Nets via Generalized-Delta-Rule-Based Internetworks*, Proc. of International Joint Conference on Neural Network. Vol.I, 259-265

Waibel, A.H., Hanzawa, T., Hinton, G., Shikano, K. & Lang, K. (1989) *Phoneme Recognition using Time-Delay Neural Networks*, IEEE Trans. on Acoustic, Speech, and Signal Processing, Vol.37, No.3, 328-339