音声言語情報処理 35−11
（2001. 2. 3）

# 視覚パターン認知方程式を応用した連続音声（単語）認識

北添　徹郎　舟森　信
宮崎大学工学部
889 - 2192　宮崎県宮崎市学園木花台西1 - 1　宮崎大学工学部
*e-mail:* kitazoe@cs.miyazaki-u.ac.jp

ステレオビジョン神経回路方程式が音声認識に適用された。この方程式はこれまでの神経回路モデルと異なり学習機構をもたず少ないパラメータで高速に処理される。学習機構は音素毎にガウス分布として表現され記憶されているものとして、入力信号はそれらの分布と比較され類似度が計算された後、神経回路に入力される。これらの時系列データの流れは一定区間のフレームから成る窓に入りここで神経回路によって競合と協調が行われた後、特定の音素が認識されて出力される。これらの出力された音素の記号列はDP法によって参照単語と比較される。こうして9人の男性話者100単語に対して96．7％の認識結果を得、通常のHMMの97．9％に接近した。さらにDP法の代わりに、離散HMMを用いたハイブリッドなアルゴリズムについても論じられている。

# Application of Visual Pattern Processing Equations to Continuous Speech Recognition

Tetsuro Kitazoe, Makoto Funamori
*Department of Computer Science and Systems Engineering*
*Faculty of Engineering, Miyazaki University*
*1-1, Gakuen Kibanadai Nishi, Miyazaki, 889-2192 Japan*
*e-mail:* kitazoe@cs.miyazaki-u.ac.jp

## ABSTRACT

The two or three layered networks 2LNN, 3LNN which originate from stereovision neural network are applied to speech recognition. To accommodate sequential data flow, we consider a window to which new acoustic data enter and from which final neural activities are output. Inside the window recurrent neural network develops neural activity toward a stable point. The process is called Winner-Take-All (WTA) with cooperation and competition. The resulting neural activities clearly showed recognition of a continuous speech of a word. The string of phonemes obtained is compared with reference words by using dynamical programming (DP) method. The resulting recognition rate amounts to 96.7% for 100 words spoken by 9 male speakers, which is compared to 97.9% by hidden markov model (HMM) with three states and single gaussian distribution. The present results which are close to those of HMM seem noticeable because the architecture of the neural network is very simple and parameters in the neural net equations are small numbered and always fixed. It is also discussed how to construct a hybrid recognition system of discrete HMM and 2(3)LNN model.

## 1. INTRODUCTION

Since we recognize speech through neural network in the brain, many works on this line have been conducted for speech recognition. Though probabilistic acoustic models represented by Hidden Markov Model (HMM) has been widely used recognizers, it has been long standing go to let machine enable human abilities of speech recognition in the brain. Various kind of neural networks have been proposed for speech recognition such as multilayer perceptrons (MLP)[1,2], time delayed neural network (TDNN)[3], hidden control neural network (HCMM)[4], hybrid system combining HMM and MLP (HMM/MLP)[1] and fully recurrent neural network (FRNN)[5,6], notable things are that these models use more or less learning algorithms of back propagation of error and need many parameters to be adjusted.

In the previous works, we employed a new approach to the problem by applying stereo vision neural network to hearing system [7,8,9,10,11]. The neural networks are two or three layered (2LNN, 3LNN) and the parameters in the equations are fixed and not changed at any time. The learning processes are considered as that the feature parameters characteristic of each phoneme are stored or memorized in the brain in the form of probability density functions. We consider recognition processes as that the neural network equations employed from visual system process the similarities between the characteristic phonetic features stored in our memory and the input acoustic data from our ears, eventually giving stable neural activities. The resulting phoneme recognition rate was fairly good, resulting 7-9% higher than HMM model. In the present paper we are going to give an algorithms for the continuous speech recognition. The major problem in this case are how to introduce an algorithms to the real time acoustic data flows and how to employ the neural network to process the data flows, giving the recognition of continuous speech. Two kind of algorithms are introduced for continuous speech recognition. One is DP method and the other is a hybrid system of discrete HMM and 2(3)LNN model, both of which show a fairly good performance.

## 2. APPLICATION OF NEURAL NETWORK TO SPEECH RECOGNITION

The speech(phoneme) recognition system using stereo vision neural net equations is divided into four main processes;

(1) A number of training speech data are classified and parameterized into sequences of feature vectors for each phonemes. The feature vectors are used to form standard Gaussian PDFs which are supposed to be memorized in our brain for each phoneme.

(2) An input phonemes are referred to these memorized phoneme data and a similarity measure is obtained by comparing the input phoneme data with the memorized PDF of each phoneme.

(3) Suppose that there is a neuron activity $\xi_u^a$ according to the similarity measure $\lambda_a^u$ to a certain phoneme /a/ at the frame number u.

(4) The stereo vision neural net equations are performed to make an activity $\xi_a^u$ move toward a stable point after the equations receive the similarity measure as an input and a recognition results are achieved when it reaches to a stable state.

The memorized standard acoustic models for each phonemes are expressed in terms of Gaussian PDF for input $o$.

$$\mathrm{N}(o;\mu_a,\Sigma_a)=\frac{1}{\sqrt{(2\pi)^n|\Sigma_a|}}e^{-\frac{1}{2}(o-\mu_a)^t\Sigma^{-1}(o-\mu_a)} \qquad (1)$$

where $\mu_a$ is a mean value and $\Sigma_a$ is covariance matrix, of feature vectors for training data of a phoneme /a/. The normalized similarity $\lambda_a^u$ of input data $o_u$ at u-th frame to a certain phoneme /a/ is defined as

$$\lambda_u^a = \frac{\log N(o_u;\mu_a,\Sigma_a)-<\log N>}{<\log N>} \qquad (2)$$

where <logN> means an average over phonemes at the same frame.

## 3. TWO LAYERED NEURAL NET EQUATIONS

Since 3LNN has a similar property with 2LNN, we discuss about 2LNN which is given as

$$\tau_1\dot{\xi}_u^a(t) = -\xi_u^a(t) + f(\alpha_u^a) \qquad (3)$$

$$\tau_2\,\dot{\alpha}_u^a = -\alpha_u^a + A\lambda_u^a - B\sum_{a'\neq a} g(\xi_u^{a'}(t)) + D\sum_{u'=n-l1}^{u+l2} g(\xi_{u'}^a(t)) \qquad (4)$$

f(x) is a well known as sigmoid function and g(u) is a function given by

$$f(x) = (\tanh(w(x-h))+1)/2 \qquad (5)$$

$$g(u) = u^+ = (u+|u|)/2 \qquad (6)$$

where A,B,D,w,h are positive constants which are to be chosen appropriately .

Figure 1 shows three layered structure of the stereo vision neural network.

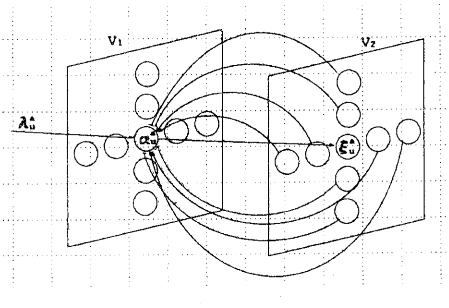

**FIGURE 1**. Three layered neural network (2LNN)

To understand the qualitative feature of the equations consider an equilibrium for $\alpha(\dot{\alpha} = 0)$. We obtain

$$\tau_1 \dot{\xi}_u^a(t) = -\xi_u^a + f(\alpha_u^a) \tag{7}$$

$$\alpha_u^a = A\lambda_u^a - B\sum_{a' \neq a} g(\xi_u^{a'}(t)) + D\sum_{u'=u-l}^{u+l} g(\xi_{u'}^a(t)) \tag{8}$$

Notice that the equations (7),(8) give the same solution as (3),(4) if the stable solution is unique. And simulations show this is the case. Equations (7) and (8) are understood as that the similarity measures $\lambda_u^a$ are inputted into the $\alpha$ layer and the outputted $\alpha's$ are fed to the $\xi$ layer. In the $\xi$ layer, it is noticed that large (small) $\alpha$ gives large (small) $\xi$ due to the sigmoid function. The new $\xi's$ thus obtained are again brought back into the $\alpha$ layer and the same procedure is repeated. In the $\alpha$ layer Winner-Take-All processes take place with competition (the second term of (8)) and cooperation (the third term of (8)). The typical dynamics is shown in figure 2 where cooperation works in the neighboring frames for the same phoneme, while competition does against other phonemes at the same frame. The Winner-Take-All processes accelerate the neuron activities toward stable points where we will get a speech recognition.

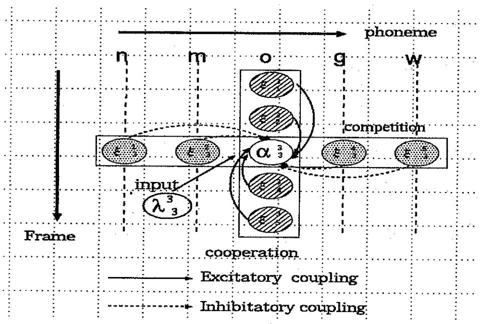

**FIGURE 2.** Dynamical process of neural activities with cooperation and competition.

## 4. CONTINUOUS SPEECH RECOGNITION

For continuous speech, it is considered that input data $\lambda_u^a$ are fed to the neural network (3), (4) and the activities $\alpha$, $\xi$ in the network develop their value toward a stable point. We judge as that a phoneme /a/ is recognized at a flame u if $\xi_u^a$ goes close to 1, while it is not recognized if $\xi_u^a$ goes close to 0. In the present study, we take equations (7),(8) to save calculation time. Differential equations (7),(8) are described as a loop of N steps in numerical calculation where time is divided by discrete span $\Delta t$. Thus, $\xi$ develops as $\xi(t)$, $\xi(t+\Delta t), \cdots$ to $\xi(t+N\Delta t)$.

To treat data flow $\lambda_u^a$ sequentially, we consider that the data enter into a window with L frames and the neural network (7), (8) processes the L frame data for N steps. Then, the data are passed one frame forward through the window, where the initial values of $\xi$ are set to zero before entering to the window. The procedure is stated more in detail as follows; When input data $\lambda_{u+L-1}^a, \lambda_{u+L-2}^a, \ldots, \lambda_u^a$ are entered inside the window, equations (7), (8) develop activities $\xi(t)$ until it arrives at $\xi_{u'}^a(t+N\Delta t)$ for $u' = u, u+1, \ldots, u+L-1$. Then, new data $\lambda_{u+L}^a$ enter into the window from the left and the old data $\lambda_u^a$ get out from the window. At the same time whole $\xi_{u'}^a$ are replaced by $\xi_{u'-1}^a$, setting the initial values of $\xi_{u'-1}^a$ for the next N step loop calculations. At this time we have new comers $\xi_{u+L}^a$ entering into the window and the final ones $\xi_u^a$ outputted from the window. We eventually judge if phoneme /a/ at u-th frame is recognized or not according to whether the final $\xi$ outputted from the window is close to 1 or 0, respectively. A sequence of the same processes continues until whole input data go through the window completely and whole values of final $\xi$ are obtained.

## 5. EXPERIMENTAL RESULTS

We extracted a total 24 labeled phonemes from ATR Japanese speech database composed of 4000 words spoken by 10 male speakers and from ASJ speech database of 500 sentences by 6 male speakers to make Gaussian PDFs for each phoneme. For recognition test which is independent, we take 100 words of the training data. The experimental conditions are as follows

| | |
|---|---|
| *Sampling rate* | *16kHz,16bit* |
| *Pre-emphasis* | *0.97* |
| *Window function* | *16ms Hamming window* |
| *Frame period* | *5ms* |
| *Feature parameters* | *10-order MFCC*<br>*+10-order ΔMFCC* |

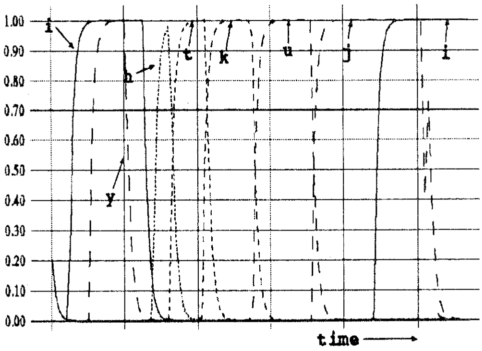

**FIGURE 3.** Best two of $\xi$ are selected and plotted against frame (time). Here parameters are set as window size L=10, Δt=0.01, N=100, A=3.0, B=3.0, D=0.5, w=2.0, h=1.0, l1=6, l2=0, $\tau_1$=1

In figure 3 the typical result for a word pronounced /i/ /k/ /u/ /j/ /i/ is shown for the best two $\xi_u^a$ outputted from the window. In the figure the best $\xi$ are read sequentially as /i/ /h/ /t/ /k/ /u/ /j/ /i/. It is noticed that /y/ has rather high values following to /i/, because /y/ resembles to /i/. /h/ and /t/ do not have correspondence in the reference word and may be recognized as context dependent effects between /y/ and /k.

| NG | a | b | ch | d | e | f | g | h | i | j | k | m | n | o | p | r | s | sh | t | ts | u | w | y | z | _ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .0 | .2 | .0 | .1 | .0 | .0 | .4 | .0 | .0 | .2 | .0 | .8 | .9 | .0 | .0 | .2 | .0 | .0 | .0 | .0 | .5 | .0 | .0 | .3 | .0 |
| | 1 | .0 | .0 | .2 | .2 | .0 | .0 | .9 | .0 | .0 | .2 | .0 | .0 | .1 | .0 | .3 | .0 | .0 | .0 | .0 | .1 | .8 | .0 | .0 | .0 |
| | | 1 | .0 | .4 | .0 | .0 | .5 | .0 | .0 | .1 | .0 | .3 | .3 | .0 | .0 | .4 | .1 | .2 | .0 | .0 | .3 | .3 | .0 | .1 | .0 |
| | | | 1 | .0 | .0 | .0 | .0 | .2 | .1 | .0 | .4 | .0 | .0 | .0 | .1 | .0 | .2 | .4 | .2 | .2 | .0 | .0 | .0 | .0 | .0 |
| | | | | 1 | .0 | .0 | .4 | .0 | .0 | .1 | .0 | .1 | .1 | .0 | .0 | .2 | .0 | .0 | .1 | .1 | .0 | .0 | .0 | .3 | .0 |
| | | | | | 1 | .0 | .0 | .1 | .1 | .0 | .0 | .0 | .0 | .0 | .0 | .3 | .0 | .0 | .0 | .0 | .2 | .0 | .5 | .0 | .0 |
| | | | | | | 1 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 |
| | | | | | | | 1 | .0 | .0 | .3 | .0 | .3 | .3 | .0 | .0 | .4 | .0 | .0 | .0 | .0 | .2 | .0 | .0 | .1 | .0 |
| | | | | | | | | 1 | .0 | .0 | .6 | .0 | .0 | .4 | .6 | .0 | .0 | .0 | .6 | .2 | .0 | .3 | .0 | .2 | .0 |
| | | | | | | | | | 1 | .6 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .8 | .0 | .0 |
| | | | | | | | | | | 1 | .0 | .0 | .0 | .0 | .0 | .2 | .0 | .0 | .0 | .0 | .0 | .0 | .3 | .0 | .0 |
| | | | | | | | | | | | 1 | .0 | .0 | .0 | .6 | .0 | .0 | .0 | .6 | .2 | .0 | .0 | .0 | .2 | .0 |
| | | | | | | | | | | | | 1 | .9 | .0 | .0 | .5 | .0 | .0 | .0 | .0 | .1 | .0 | .0 | .0 | .0 |
| | | | | | | | | | | | | | 1 | .0 | .0 | .5 | .0 | .0 | .0 | .0 | .4 | .0 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | 1 | .0 | .0 | .0 | .0 | .0 | .0 | .4 | .5 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | | 1 | .1 | .0 | .0 | .6 | .2 | .0 | .0 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | | | 1 | .0 | .0 | .0 | .0 | .3 | .0 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | | | | 1 | .3 | .0 | .7 | .0 | .0 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | | | | | 1 | .2 | .3 | .0 | .0 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | | | | | | 1 | .7 | .0 | .0 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | | | | | | | 1 | .0 | .0 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | | | | | | | | 1 | .5 | .2 | .4 | .0 |
| | | | | | | | | | | | | | | | | | | | | | | 1 | .0 | .0 | .0 |
| | | | | | | | | | | | | | | | | | | | | | | | 1 | .0 | .0 |
| | | | | | | | | | | | | | | | | | | | | | | | | 1 | .0 |

**TABLE 1.** Likelihood table among 24 phonemes for the dynamical programming method./-/ shows missing phonemes in inputted word against a reference word.

```
input word:_aaaaaaaaaaaaaaaaaaaayyiiiiiiiiiiiiiiiiiiii
           jjjjjjjjjjjjjjjjjjCjjjjjjjjjjjjoooowooowwwoooooooooo
           oooooooooooooooooooooooooooooooooooooooooohoh
```

```
reference words

 1 aaaaaaaaaaaaaaaaaaaaaaiiiiiiiiiiiiiiiiiiiiiiiiii
   jjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjoooooooooooooooooooooo
   oooooooooooooooooooooooooooooooooooooooooo   185.20

40 kkkkkkkkkaaaaaaaaaaaaaaaaiiiiiiiiiiiiiiiiiSSSSSSSS
   SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSooooooooooooooooooo
   oooooooooooooooooooooooooooooooooooooooooooooo   93.30

57 gggggggggggggggeeeeeeeeeeeeeeeeeeeeeeeekkkkkkkkkk
   kkkiiiiiiiiiiiiiiiiiiiijjjjjjjjjjjjjjjjjjjjooooooo
   oooooooooooooooooooooooooooooooooooooooooooooooo
   oooooo   93.00

31 ooooooooooooooooooooooooookkkkkkkkkkkkkkkuuuuuuuuu
   uuuuuuuuuuuuuujjjjjjjjjjjjjjjjjjjjjjjjjjjoooooooo
   oooooooooooooooooooooooooooooooooooooooooooooooo
   oooooooo   84.50

 6 aaaaaaaaaaaaaaaaaaaSSSSSSSSSSSSSSSSSSSSSSSSSSiiiii
   iiiiiimmmmmmmmmmmmmmmooooooooooooooooooooooootttttt
   tttttttttoooooooooooooooooooooooooooooooooooo   73.30
```

**FIGURE 4.** An example of dynamical programming where a string of phonemes recognized by the neural networks is compared with reference words and the best five words are selected. The best scored /a/ /i/ /j/ /o/ is correctly recognized in this case.

To get the recognition rate, a sequence of phonemes for a word obtained from the window are compared with a list of reference words. We make the list of reference words composed of phonemes with the mean length which is estimated from the data of the same word spoken by many different people. We take dynamical programming (DP) method to match an inputted words with the reference word, where a likelihood table between different phonemes is estimated from phoneme recognition results given by neural networks reported previously [8-11](TABLE1). An example for DP is given in FIGURE 4, where higher scores for reference words show better likelihood for inputted data.

| | 2LNN | HMM |
|---|---|---|
| mau | 98 | 98 |
| mht | 100 | 98 |
| mms | 93 | 98 |
| mmy | 94 | 97 |
| mnm | 95 | 97 |
| msh | 94 | 98 |
| mtk | 99 | 98 |
| mtm | 97 | 99 |
| mtt | 100 | 98 |
| average | 96.7 | 97.9 |

**TABLE 2.** The recognition results for 100 words spoken by 9 male persons

Experiment was performed in this way and 100 words uttered by 9 male speakers were recognized with the rate of 96.7%, where HMM model with 3 states and single gaussian distribution gave 97.9% for the same sample data, as shown in Table 2.

## 6. APPLICATIN OF DISCRETE HMM TO 2(3)LNN

Since a discrete HMM(DHMM) is a probabilistic automaton, it emits or accepts a sequence of discrete symbols in a probabilistic way. When a sequence of phonemes are produced through the neural network as stated in the previous sections, it is possible to apply discrete HMM for the sequence of phonemes in order to recognize continuous speech instead of applying DP algorithm stated in section 4. Before going to continuous speech recognition, we have to train DHMM for data of each phoneme with discrete data which are obtained from 2LNN. 200 words by 9 male speakers are used to train DHMM for data preprocessed by 2LNN. The DHMM thus trained is applied to recognize 100 words of the same 9 speakers which is independent of training data. The results show 2% better recognition rate for 24 phonemes than the original 2LNN model decided by majority as shown in Table 3. However, it is found that the above training data are still too small to train DHMM for continuous speech recognition.

| | 2LNN | HMM |
|---|---|---|
| NG | 76.8 | 81.1 |
| a | 94.2 | 97.5 |
| b | 48.1 | 57.4 |
| ch | 20.0 | 51.1 |
| d | 41.0 | 51.3 |
| e | 81.3 | 88.4 |
| f | 94.4 | 94.4 |
| g | 20.5 | 22.6 |
| h | 63.0 | 63.0 |
| i | 89.0 | 91.1 |
| j | 86.2 | 89.0 |
| k | 52.2 | 50.8 |
| m | 62.0 | 62.7 |
| n | 49.4 | 55.6 |
| o | 90.0 | 94.1 |
| p | 55.6 | 55.6 |
| r | 61.9 | 41.9 |
| s | 87.7 | 87.7 |
| sh | 91.6 | 96.3 |
| t | 35.6 | 42.2 |
| ts | 91.1 | 100.0 |
| u | 68.6 | 72.2 |
| w | 83.3 | 80.6 |
| y | 92.6 | 92.6 |
| z | 85.1 | 87.0 |
| average | 74.99 | 77.43 |

**TABLE 3.** The recognition results for 24 phonemes of 100 words by 9 male speakers

## 7. CONCLUSIONS AND DISCUSSIONS

The two or three layered networks 2LNN, 3LNN which originate from stereovision neural network are applied to speech recognition. To accommodate sequential data flow, we consider a window to which new outputted acoustic data enter and from which final neural activities are outputted. Inside the window recurrent neural network develops neural activity toward a stable point. The process is called

Winner-Take-All (WTA) with cooperation and competition. The resulting neural activities clearly showed recognition of a continuous speech of a word. The string of phonemes obtained is compared with reference words by using DP matching. The recognition results are 96.7%, compared with 97.9% by HMM. To apply DHMM to the NN model is discussed. It is shown that the hybrid model gives 2% better performance than the single NN model. Though it is straight forward to apply DHMM for continuous speech recognition, it seems necessary to use much more data for training than those given in the present experiment. A simple step forward will be to use multi-gaussian distribution to obtain more accurate similarity measures and further improvement of recognition is expected together with the study of more input data. The nice feature of our model is that it does not have many parameters to be adjusted and the algorithm for recognition is simple.

## Reference

[1] Bourlard, C.J.Wellekens. "Link between Markov Models and Multi-layer Perceptoron" IEEE Trans. Patt. Anal. Machine Intell., Vol.12, pp.1167-1178, 1990

[2] Hung, A.Kuh. "A Combined Self-Organizing Feature Map and Multilayer Perceptron for Isolated Word Recognition" IEEE Trans. on Signal Processing, Vol.40,pp.2651-2657, 1992

[3] J.Lang, A.Waibel, G.E.Hinton. "A Time-Delay Neural Network Architecture for Isolated Word Recognition" Artificial Neural Networks, Paradigms, Applications and Hardware, 1992

[4] Martinelli. "Hidden Control Neural Network" IEEE Trans. on Circuits and Systems, Analog and Signal Processing 41(3):245-247,1994

[5] T.Robinson(1992) Recurrent Nets for Phone Probability Estimation. Proceedings of the ARPA Continuous Speech Recognition Workshop, Stanford, Sept.

[6] Williams,R.J.,Zipser.D.(1990)Gradient based Learning Algorithms for Recurrent Connectionist Networks. Tech. Rep. NU=CCS-90-9,Northeastern University, College of Computer Science, Boston

[7] T.Kitazoe, J.Tomiyama, Y.Yoshitomi, and T.Shii "Sequential Stereoscopic Vision and Hysteresis" Proc. of Fifth Int.Conf. on Neural Information Processing, pp. 391-396, 1998

[8] T.Kitazoe,S-I.Kim,T.Ichiki. Speech recognition using Stereovision Neural Network Model. Fourth International Symposium on Artificial Life and Robotics, pp.576-579,Vol2, January, Beppu, Oita, Japan, 1999.

[9] Acoustic Speech Recognition Model by Neural Net Equation with Competition and Cooperation(Tetsuro Kitazoe,Tomoyuki Ichiki,Sung-Ill-Kim) ICSLP'98(The 5th International Conference on spoken Language Processing,Vol 7.pp3281-3284,30th November-4th December, Sydney,Australia,1998

[10]T.Kitazoe,S-I.Kim,T.Ichiki,M.Funamori.Acoustic Models in Speech Recognition by Stereo Vision Neural Nets. International Conference on Speech Processing,pp81-86,Vol1,August,Seoul,Korea,1999

[11] T.Kitazoe, S-I.Kim, T.Ichiki, M.Funamori. Acoustic Speech Recognition by Two and Three Layered Neural Networks with Competition and Cooperation. International Workshop SPEECH AND COMPUTER, pp.111-114, October, Moscow, Russia, 1999.