Speech Recognition by Neural Net Pattern Recognition Equations with Self-organization

Sung-Ill Kim*, Hyun-Yeol Chung**
*Division of Electrical and Electronic Engineering, Kyungnam University
**School of Electrical Engineering and Computer Science, Yeungnam University

(Received November 6 2002; revised June 2 2003; accepted June 19 2003)

Abstract

The modified neural net pattern recognition equations were attempted to apply to speech recognition. The proposed method has a dynamic process of self-organization that has been proved to be successful in recognizing a depth perception in stereoscopic vision. This study has shown that the process has also been useful in recognizing human speech. In the processing, input vocal signals are first compared with standard models to measure similarities that are then given to a process of self-organization in neural net equations. The competitive and cooperative processes are conducted among neighboring input similarities, so that only one winner neuron is finally detected. In a comparative study, it showed that the proposed neural networks outperformed the conventional HMM speech recognizer under the same conditions.

Keywords: *Speech recognition, Coupled pattern recognition, Neural network, HMM, Depth perception*

I. Introduction

Hitherto, many researches on hidden Markov model (HMM)[1-3] or artificial neural networks (ANN)[4-6] have been conducted in the field of speech recognition. The HMM approach, particularly, has appeared a main stream to tackle the problem by giving accurate probabilistic acoustic models. However, it is still difficult to give a satisfactory explanation of humanlike speech understanding, because it is originally based on the probabilistic modeling concepts. As the alternative approach, therefore, ANN, such as multilayer perceptron [4], time-delay neural network[5], or hidden control neural network[6] etc., have been developed by modeling information processing mechanism of physiological human brain. One of their major strength is in the fact that there is no need for any mathematical assumptions about statistical distributions or independence among input frames. However, there are still demerits of dealing with too many parameters in both training and recognition processes.

In stereoscopic vision, a human brain recognizes three-dimensional depth[7-9] by fusing the vast information coming through left and right eyes. The depth perception phenomenon has been successfully simulated by using the recently modified algorithms[10-14] of neural networks such as two[13] or three layered[13,14] neural networks. It is assumed that the human brain system has complicated neural networks fusing disparities between the different images by a self-organizing process of competition and cooperation. The neural networks have even simpler architecture than ordinary ANN. The network parameters are always fixed and not revised at any time. Moreover, it has a characteristic process of identifying most likely neuron among confusable candidates,

Corresponding author: Hyun-Yeol Chung (hychung@yu.ac.kr)
School of Electrical Engineering and Computer Science, Yeungnam University, 214-1, Gyungsom, Gyungbuk, 712-749, Korea

resulting in a clear depth perception of a specific object. Namely, the depth perception is conducted by a dynamic process of self-organization between two kinds of input similarities from left and right retinas.

In a similar way, it is assumed that speech recognition is conducted by self-organization of neurons handling similarities between the input speech signals and the memorized patterns in our brain. The memorizing process is regarded as a learning process something like Kohonen self-organizing maps[15]. In its application to speech recognition, the input similarities are supposed to be competed and cooperated among neurons through the dynamic process. The neural networks then trigger competition among similarities of hypothetic speech as well as cooperation among neighboring similarities of temporal frames. As a result, the so-called winner-take-all process plucks only one winner neuron out of candidate ones.

This study details the new mathematical algorithms of stereoscopic vision neural networks, namely, coupled pattern recognition (CPR) equations with self-organization mentioned above. Moreover, the new approach investigates how well the neural net equations work in identifying a specific speech among confusable hypotheses. In addition, the comparative study with the existing HMM would be made under the same condition.

II. Neural Net Pattern Recognition Equations

2.1. Similarity Measure

If two objects are separated in depth from the viewer, the relative positions of their images will differ in two eyes. Our brains are capable of measuring this disparity by fusing information of the binocular difference and process it to estimate depth. For the similarity measure between the both features, it is necessary to take the absolute value of the difference of two corresponding feature points and sum over all points in the feature area.

When the above-mentioned concepts are applied to speech recognition, the similarity λ_u^a at u-th temporal frame to a certain vocal signal /a/ can be defined as

$$\lambda_u^a = \frac{\log N(o_u;\mu_a,\Sigma_a) - <\log N>}{<\log N>} \qquad (1)$$

where N is Gaussian probability density function with input data o_u, mean μ_a, and covariance $\sum_a$. <log N> means an average over temporal frames. As shown in this equation, the similarity between two features can be obtained after normalization. From the equation, it is noticed that λ_u^a is a deviation of the similarity measure from its mean or standard value.

2.2. Coupled Pattern Recognition Equations

In the course of the process, the activity of a neuron with the highest similarity excels others that would vanish. Such a winner-take-all system is given by a coupled pattern recognition (CPR) process[15-17] as following:

$$\xi_u^a(t) = -\frac{dU}{d\xi_u^a(t)} \qquad (2)$$

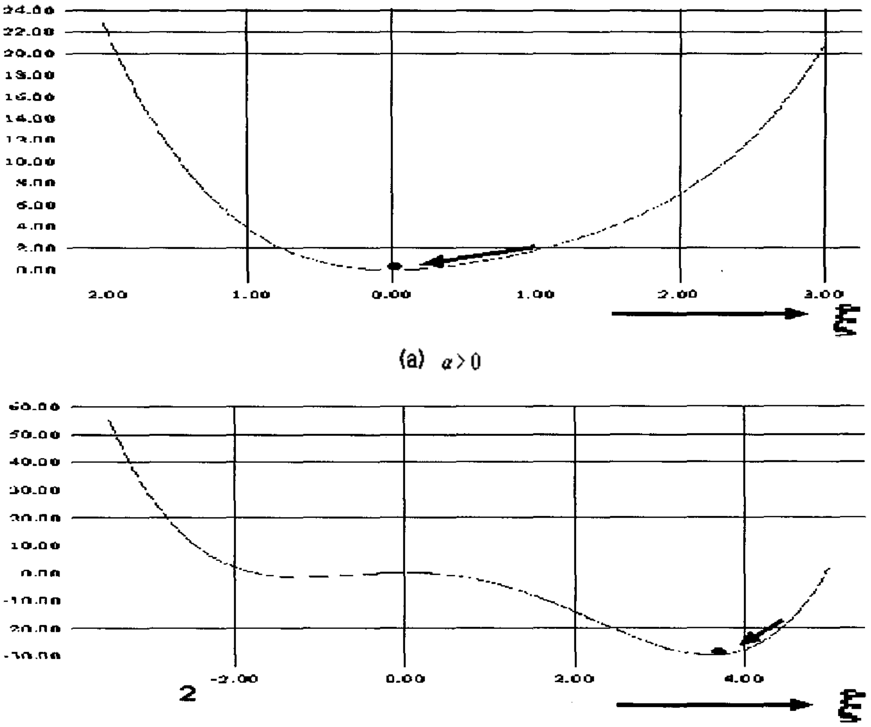
where ξ_u^a is a time-dependent activity of neuron in which the amplitude of ξ_u^a means a projection of test patterns.

$$U(\xi_u^a(t)) = \frac{\alpha}{2}\xi_u^a(t)^2 - \frac{E}{3}\xi_u^a(t)^3 + \frac{C}{4}\xi_u^a(t)^4 \qquad (3)$$

where $U(\xi_u^a(t))$ is a non-vanishing part of U under the derivative of equation (2). In this equation, particularly, the second term plays a role in breaking a left-right symmetry in the potential, while the third term restricts the activities. α_u^a is given by

$$\alpha_u^a(t) = -\lambda_u^a + (B+C)\sum_{\substack{a'=a-a_s \\ a'\neq a}}^{a+a_s} \xi_u^{a'}(t)^2 - D\sum_{\substack{u'=n-l \\ u'\neq u}}^{n+l} \xi_{u'}^a(t)^2 \qquad (4)$$

where B, C, D are positive constants which are chosen appropriately. In equation of α_u^a, the first term is a similarity of input data at an u-th temporal frame of an arbitrary hypothesis /a/. The second term means a competitive coupling among neighboring neural activities of all candidates, while the third term represents a cooperative coupling among neighboring frames. The



(a) $a > 0$

(b) $a < -E^2/4.5C$

Figure 1. Potential functions for loser neural activity (a) and winner neural activity (b).

summation index of competitive coupling runs over the disparity search area defined as $a - a_s \leq a' \leq a + a_s$ with a restriction of $a' \neq a$. The summation index of cooperative coupling, on the other hand, runs over the cooperation area defined as $u - l \leq u' \leq u + l$ with a restriction of $u' \neq u$. Accordingly, it is noticed that $\alpha_u^a(t)$ depends on neural activities of both $\xi_u^{a'}(t)$ and $\xi_{u'}^a(t)$ as well as input similarity λ_u^a.

In actual applications, it is dealt with only two cases of potential functions in which ξ_u^a converges to an absolute minimum value. Figure 1 shows typical potential functions where a stable solution is decided in each case by the minima of the potential depending on the value of α_u^a.

In case of $\alpha > 0$, ξ_u^a converges to a global minimum. In case of $\alpha < -\dfrac{E^2}{4.5C}$, on the other hand, ξ_u^a saturates in a certain positive value corresponding to an absolute minimum in potential function. Therefore, it is noticed that only one neuron among candidates would win the competition through a dynamic process of self-organization in potential function.

2.3. Self-organizing Process with Competition and Cooperation

The similarity values of all possible candidates are obtained as a result of similarity measure. Figure 2 shows an example of similarity map where each candidate phoneme has a similarity value in each frame. In the next step, the similarity map is given to CPR equations with self-organization in which the dynamic process with

INPUT

frame	/n/	/m/	/o/	/g/	/w/
1	0,172669	0,007747	-0,179798	0.068170	-0,317374
2	0,047739	0,021844	0,012022	0,106935	-0,377080
3	-0,053958	-0,254189	0,174484	0,140137	-0,321096
4	-0,020677	-0,345811	0,166542	0,152011	-0,270617
5	0,071875	-0,109546	0,026478	0,047362	-0,181884
6	0,164128	-0,066376	-0,075502	0,000766	-0,187911
7	0,074848	0,021229	0,011177	-0,173780	-0,040727
8	0,075048	-0,128097	0,029788	-0.138120	0,028273
9	0,151001	-0,058196	-0,134349	-0,094952	-0,014505
10	0.181342	-0,005437	-0,214245	-0,072309	-0,070694
11	0,132347	0,084662	-0.163194	-0,224362	-0,046461
12	0,052027	0,157427	0,039553	-0,173396	-0,324618
13	0,112184	0,316814	0,044812	-0,315088	-0,632532
14	0,088750	0,277316	0,008593	-0,229108	-0,520211
15	0,064446	0,061100	0,028512	-0,384859	0,038372

Figure 2. Example of input normalized similarity map of candidate phonemes.

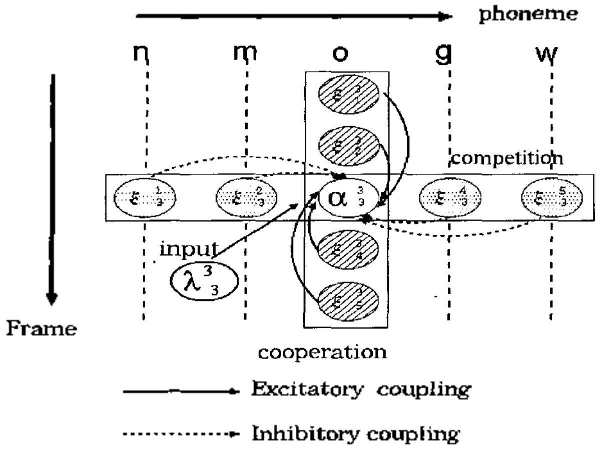


Figure 3. Competitive process among the similarity values in different hypotheses and cooperative process among the values in neighboring frames.

competitive and cooperative coupling is then performed among similarities as shown in figure 3.

The value of α_u^a is influenced by not only input similarity but also neighboring neural activity of $\xi_{u'}^{a'}$. The neurons compete one another over the competitive search area and cooperate over the cooperative search area. Namely, α_u^a has an inhibitory coupling among neural activities in all candidates as well as an excitatory coupling in neighboring temporal frames. Through the dynamic process of self-organization, they approach an arbitrary specific value in the long run, independent of the initial values of ξ_u^a.

It is noticed in the above-mentioned process that the self-organization process exerts a great influence on α_u^a as well as ξ_u^a. Figure 4 and 5 show an examples of time dependent behaviors of α_u^a and ξ_u^a, respectively, at the fifth frame of each candidate phoneme.

In a parameter setting, α_u^a starts with positive value because all initial values of ξ_u^a are given as 1. If the value of ξ_u^a begins to drop owing to a dynamic movement of potential form particularly described in figure 1(a) for α_u^a >0, the value of α_u^a in phoneme /n/ also begins to fall down while the values of other phonemes rise. In this case, it is revealed that the excitatory coupling in /n/ gets more activated than inhibitory one, so that it accelerates ξ_u^a to increase. Accordingly, the value of ξ_u^a turns to grow rapidly, depending on the potential function in figure 1(b). On the contrary, the values of ξ_u^a in other phonemes fall down to 0. In this case, inhibitory coupling becomes more activated than excitatory one. As a result, when ξ_u^a reaches a certain saturated point through the cycles of recurrent networks, it is called a winner neuron, while it is called a loser neuron when it loses the whole activity to become close to 0. Figure 6 shows an example of winner and loser neurons.

In this example, winner neurons are given to /n/ between 1-th and 11-th frame and to /m/ between 12-th and 15-th

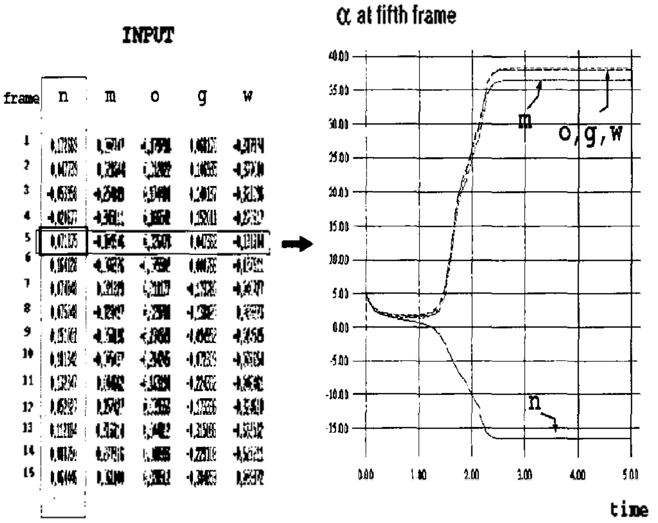


Figure 4. Time dependent behaviors of α_u^a at the fifth frame of each candidate.

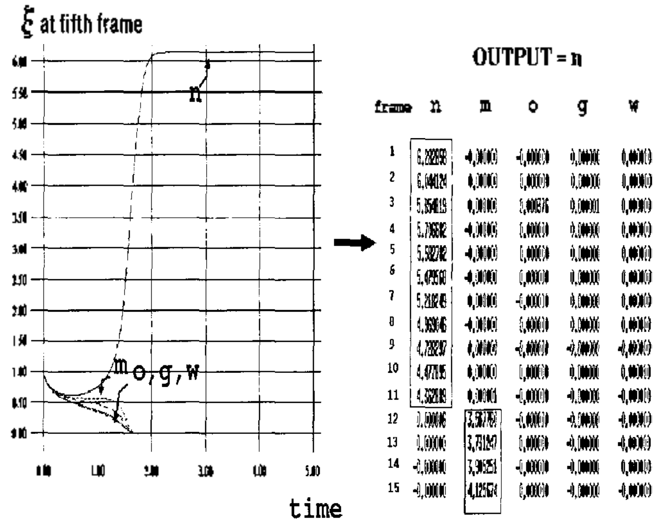
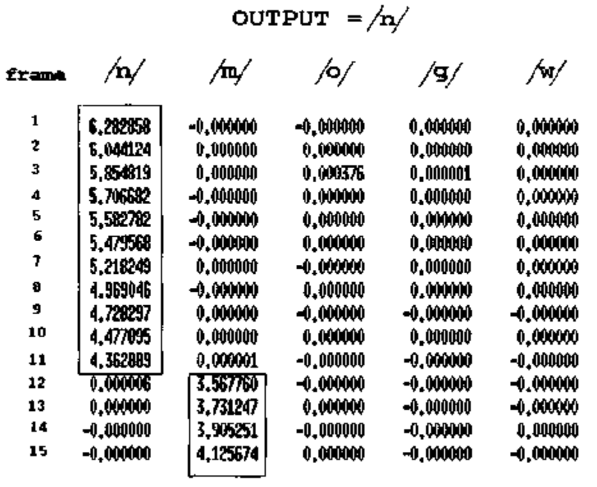


Figure 5. Time dependent behaviors of ξ_u^a at the fifth frame of each candidate.



```
                    OUTPUT  = /n/

frame    /n/        /m/        /o/        /g/        /w/

  1    6,282858  -0,000000  -0,000000   0,000000   0,000000
  2    6,040124   0,000000   0,000000   0,000000   0,000000
  3    5,854819   0,000000   0,000376   0.000001   0,000000
  4    5,706682  -0,000000   0,000000   0,000000   0,000000
  5    5,582782  -0,000000   0,000000   0,000000   0,000000
  6    5,479568  -0,000000   0,000000   0,000000   0,000000
  7    5,218249   0,000000  -0,000000   0,000000   0,000000
  8    4,969046  -0,000000   0,000000   0,000000   0,000000
  9    4,728297   0,000000  -0,000000  -0,000000  -0,000000
 10    4,477095   0,000000   0,000000   0,000000   0,000000
 11    4,362889   0,000001  -0,000000  -0,000000  -0,000000
 12    0.000006   3.567760  -0,000000  -0,000000  -0,000000
 13    0,000000   3,731247   0,000000  -0,000000  -0,000000
 14   -0,000000   3,905251  -0,000000  -0,000000   0,000000
 15   -0,000000   4,125674   0,000000  -0,000000  -0,000000
```

Figure 6. Winner and loser neurons as a result of dynamic process of self-organization with B=0.25, C=1.25, D=0.60, E=3.0, I=4.0.

frame, while other phonemes have only loser neurons. Consequently, /n/ is recognized as the most likely candidate to input speech, as it has even more winner neurons than others.

2.4. Application to Speech Recognition

The procedure of constructing speech recognition system using CPR neural net equations is as followings.

(1) Using a number of speech corpuses that was hand-segmented and hand-labeled to phonemes, the standard models are built for each phoneme that is supposed to be memorized in our brains.

(2) For recognition, input phonemes are first parameterized into a sequence of acoustic feature vectors. The extracted acoustic feature signals are referred to each standard model to obtain a similarity λ_u^a at a certain frame /u/ of an arbitrary phoneme /a/.

(3) The similarity λ_u^a is then given to CPR neural net equations to trigger neural activities such as ξ_u^a or α_u^a, each of which means membrane potential of neurons in the terminology of neural networks.

(4) Self-organization with competition and cooperation is then activated to carry out on the similarity map. The process makes the neural activities move toward a stable state. As a consequence, only one winner neuron would be chosen, beating others down.

III. Experimental Evaluation

The effectiveness of speech recognition by CPR neural net equations has been assessed on two kinds of test sets and compared with the conventional HMM recognizer under the same experimental condition. For the training of acoustic phoneme models, the feature parameters were first extracted from two different databases that consist of hand-segmented and -labeled phonemes. The one is ATR speech database of 4000 words spoken by 10 male speakers. The other is ASJ continuous speech database of 500 sentences by 6 male speakers. For recognition, the test sets were composed of two kinds, one from database of 216 words and the other from 240 words, each of which was spoken by 3 male speakers, respectively.

Table 1 shows an analysis condition of speech signals, in which 10 dimensional mel-frequency cepstrum coefficients (MFCC) and their derivatives were parameterized from the original speech signals.

For a comparative study, the proposed system was compared with the conventional HMM recognizer with a single mixture and three states.

Table 2 shows the average recognition accuracies of 3 speakers in speaker independent experiments, in which CPR recognizer was compared with HMM on two kinds of test sets. The recognition accuracies using CPR were 82.77 % and 79.53 % on 216 and 240 test sets, which were compared with 71.56% and 72.37% by HMM, respectively.

Table 3 shows the improvement rates of overall performance by CPR compared with HMM recognizer. The average recognition accuracies of CPR were 11.2 %

Table 1. Analysis of speech signal.

Sampling rate	16 Khz, 16 Bit
Pre-emphasis	0.97
Window	16 msec Hamming window
Frame period	5 ms
Feature parameters	10 order MFCC + 10 order delta MFCC

Table 2. Comparison of CPR with HMM on two test sets.

Phoneme	216 set		240 set	
	HMM	CPR	HMM	CPR
NG	53.46	85.44	59.62	89.03
A	92.55	95.24	93.85	96.67
B	76.62	81.01	86.79	86.79
CH	84.62	87.69	100	75
D	69.84	84.38	74.07	59.26
E	64.77	89.77	80.86	96.3
G	57.14	63.64	45.71	44.44
H	63.46	59.62	53.33	60
I	69.16	92.21	84.18	97.98
J	97.01	97.01	93.1	93.1
K	55.25	69.86	67.02	61.7
M	61.9	54.72	86.67	66.67
N	44.3	46.25	50	50
O	70.58	95.27	66.67	90.79
P	64	44	100	77.78
R	62.34	62.34	42.3	32
S	89.01	85.71	76.4	80.9
SH	96.05	82.89	91.11	95.56
T	4.35	39.13	15.38	48.72
TS	65.22	86.96	89.74	89.74
U	94.78	67.24	59.8	68.67
W	84.38	75.76	91.03	78.91
Y	61.36	65.91	87.3	91.67
Z	87.76	85.71	93.1	93.33
Total (%)	71.56	82.77	72.37	79.53

Table 3. Overall comparison of CPR with HMM in performance.

Database	Recognizer	Recognition Accuracy	Improvement in CPR
216 data	HMM	71.56%	
	CPR	82.77%	11.20%
240 data	HMM	72.37%	
	CPR	79.53%	7.20%

and 7.2% higher than HMM on 216 and 240 test sets, respectively. As shown in this table, it was shown that CPR gave better performance than the existing HMM recognizer.

However, as shown in phoneme 'P' for example, the accuracies based on CPR do not always show better performance in every phoneme than HMM, because CPR still has a lack of exact modeling of the inner change of acoustic features like HMM. Since this study is restricted to phoneme recognition, moreover, we should make further experiments to word or continuous speech recognition as future works.

IV. Conclusions

This study presented a new approach of speech recognition using CPR neural net equations with self-organization. From the comparative study, it was shown that the proposed method outperformed the conventional HMM recognizer. In recognition process, the self-organization of competition and cooperation plays a crucial role in determining the most likely neuron in a similar way as a depth perception in stereoscopic vision. Therefore, it is noticed that the visual cognitive mechanism might be also useful and beneficial in picking up specific speech signals out of confusable candidates. Moreover, we can see that the proposed neural net equations are able to give even simpler architecture than other ordinary ANN.

References

1. P. C. Woodland, C. J. Leggestter, J. J. Odell, V. Valtchev, and S. J. Young. "The 1994 HTK large vocabulary speech recognition system," *Proc. IEEE Int. Conf. On Acoustics, Speech, and Signal Processing*, 1, 73-76, Detroit, 1995.
2. X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Makov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, U. K., 1990.
3. L. Rabiner, A Tutorial "Hidden Markov models and selected applications in speech recognition," A. Waibel and K.-F. Lee, editors, Readings in Speech Recognition, 267-296, Morgan Kaufmann, San Mateo, 1990.
4. C. J. Bourlard, and Wellekens, "Links between Markov models and multi-layer perceptrons," *IEEE Trans. Patt. Anal. Machine Intell.*, 12, 1167-1178, 1990.
5. T. Waibel, G. Hanazawa, K. Hinton, Shikano et al., "Phoneme recognition using time-delay neural networks," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37 (3), 329-339, 1989.
6. Martinelli, "Hidden control neural network," *IEEE Trans. on Circuits and Systems, Analog and Signal Processing*, 41 (3), 245-247, 1994.
7. D. Reimann, T. Ditzinger, E. Fischer, and H. Haken, "Vergence eye movement control and multivalent perception of Autostereograms," *Biol. Cybern*, 73, 123-

128, 1995.

8. D. Reinmann, H. Haken, "Stereo vision by self-organization," Biol. Cybern., **71**, 17-26, 1994.
9. S. Amari, and M. A. Arbib, "Competition and cooperation in neural nets," *Systems Neuroscience*, 119-165, Academic Press, 1977.
10. Y. Yoshitomi, T. Kanda, and T. Kitazoe, "Neural nets pattern recognition equation for stereo vision," *Trans. IPS*, 29-38, 1998.
11. Y. Yoshitomi, T. Kitazoe, J. Tomiyama and Y. Tatebe, "Sequential stereo vision and phase transition," *Proc. of Third Int. Symp. On Artificial Life, and Robotics*, 318-323, 1998.
12. T. Kitazoe, J. Tomiyama, Y. Yoshitomi, et al., "Sequential stereoscopic vision and hysteresis," *Proc. Neural Information Processing*, 391-396, 1998.
13. T. Kitazoe, S.-I. Kim, and T. Ichiki, "Acoustic speech recognition by two and three layered neural networks with competition and cooperation," *Proceeding of International Workshop on Speech and computer*, 111-114, 1999.
14. T. Kitazoe, S-I Kim, and T. Ichiki, "Speech Recognition using Stereovision Neural Network Model," *Proc. International Symposium on Artificial Life and Robotics*, **2**, 576-579, 1999.
15. T. Kohonen, "Self-organizing map," *Proc. IEEE*, **78** (9), 1464-1480, 1990.

[Profile]

● Sung-Ill Kim

Sung-Ill Kim was born in Kyungbuk, Korea, 1968. He received his B. S. and M. S. degrees in the Department of Electronics Engineering from Yeungnam University, in 1997, and Ph.D. degree in the Department of Computer Science & Systems Engineering from Miyazaki University, Japan, in 2000. During 2000 to 2001, he was a postdoctoral researcher in the National Institute for Longevity Sciences, Japan. He worked in the Center of Speech Technology, Tsinghua University, China during 2001 to 2003. Currently, he is full-time lecturer in the Division of Electrical & Electronic Engineering, Kyungnam University since 2003. His research interests include speech/emotion recognition, neural networks, and multimedia signal processing. E-mail: kimstar@kyungnam.ac.kr

● Hyun-Yeol Chung

Hyun-Yeol Chung was born in Kyungnam, Korea, 1951. He received his B. S. and M. S. degrees in the Department of Electronics Engineering from Yeungnam University, in 1975 and 1981, respectively, and the Ph.D. degree in the Information Sciences from Tohoku University, Japan, in 1989. He was a professor from 1989 to 1997 at the School of Electrical and Electronic Engineering, Yeungnam University. Since 1998 he is a professor in the School of Electrical Engineering and Computer Science, Yeungnam University. During 1992 to 1993, he was a visiting scientist in the Department of Computer Science, Carnegie Mellon University, Pittsburgh, USA. He was a visiting scientist in the Department of Information and Computer Sciences, Toyohashi University, Japan, in 1994. He was a principle engineer, Qualcomm Inc., USA, in 2000. His research interests include speech analysis, speech/speaker recognition, multimedia and digital signal processing application. E-mail: hychung@yu.ac.kr