THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.1E 2003. 3. pp. 3~10

Stereo Vision Neural Networks with Competition and Cooperation for Phoneme Recognition

Sung-Ill Kim*, Hyun-Yeol Chung**
*Division of Electrical and Electronic Engineering, Kyungnam University
**School of Electrical Engineering and Computer Science, Yeungnam University
(Received September 9 2002; revised November 22 2002; accepted December 20 2002)

Abstract

This paper describes two kinds of neural networks for stereoscopic vision, which have been applied to an identification of human speech. In speech recognition based on the stereoscopic vision neural networks (SVNN), the similarities are first obtained by comparing input vocal signals with standard models. They are then given to a dynamic process in which both competitive and cooperative processes are conducted among neighboring similarities. Through the dynamic processes, only one winner neuron is finally detected. In a comparative study, with, the average phoneme recognition accuracy on the two-layered SVNN was 7.7% higher than the Hidden Markov Model (HMM) recognizer with the structure of a single mixture and three states, and the three-layered was 6.6% higher. Therefore, it was noticed that SVNN outperformed the existing HMM recognizer in phoneme recognition.

Keywords: *Speech recognition, Neural network, Stereoscopic vision, HMM, Depth perception*

Ⅰ. Introduction

In the field of speech recognition or speech understanding, many studies have been conducted on the basis of Hidden Markov Model (HMM)[1-3] and several kinds of artificial neural networks (ANNs)[4-6]. Though HMM has been regarded as a useful recognizer by producing relatively accurate probabilistic acoustic models, it still has a weakness in the viewpoint of the modeling with human-like speech understanding. As the alternative approach, therefore, ANNs such as multi-layer perceptron[4], time-delay neural network[5], or hidden control neural network[6] etc., have been introduced by modeling an information processing mechanism of physiological human brain. The one of major strength of them 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 as well as structural complexity.

In the neural networks for stereoscopic vision, there are two beneficial features compared with the above-mentioned neural networks. The one thing is that it has much more simple architecture because the network parameters are always fixed and not revised at any time. The other is that it has a powerful information processing capability of identifying the most likely neuron among confusable candidates. The process is made by both cooperative and competitive process among their similarities. These stereo vision neural networks (SVNN)[7-9] process those input

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

visual data, yielding a depth perception of a specific object in stereoscopic vision.

In the same way, it is assumed that speech recognition can be performed by the same process between vocal features as input data and memorized ones as standard models in human brain. In the processing, SVNN triggers not only competition among similarities in all possible speech candidates but cooperation among ones in temporal frames of the candidates, and finally so-called winner-take-all process plucks only one neuron from the candidates. Though it has not been found if a visual processing mechanism for depth perception is compatible with an actual hearing system for speech recognition, it is worth to apply the cognitive architecture in stereoscopic vision to speech recognition, on the viewpoint of information processing based on the neural networks.

In this new approach, the recently modified algorithms of SVNN, which have been optimized through preliminary investigations[10-12], were successful in stereoscopic depth perception. We will describe the recently developed two- and three-layered SVNN equations with dynamic process of competitive and cooperative coupling among input similarities. It would be then explored if the dynamic process of SVNN works well in speech recognition.

II. Stereo Vision Neural Networks

2.1. Depth Perception in Stereoscopic Vision

If two objects are separated in depth from the viewer, the relative positions of their images will differ in the two eyes. Our brains are capable of measuring this disparity and using it to estimate depth, which is processed on the chiefly visual area 2(V2) of the neural networks of the brain.

Figure 1 shows the depth perception phenomenon that is given by fusing the disparity of two different images. It was produced by displacing the square area in the random-dot image horizontally by a certain amount, where we took one image with an original square area and the other one with a horizontally displaced area.

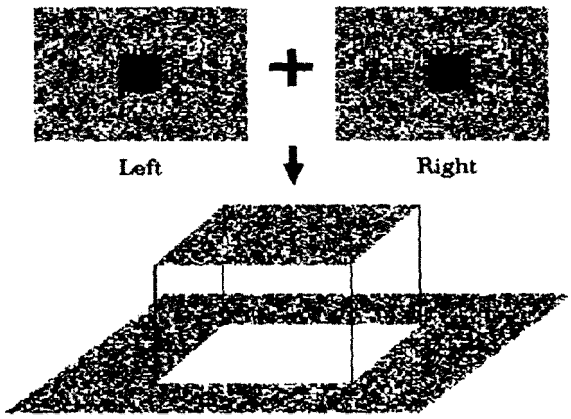

Figure 1. Depth perception using binocular difference (Top) Pair of random-dot stereograms presented to the left and right eyes. (Bottom) 3-dimensional image of the stereograms viewed by the present neural net models.

If stereoscopically fused, the central square can be seen as if it is floated over the image plane, through competitive process among input similarities and cooperative process among them as well. The membrane potentials of binocular neurons corresponding to each possible disparity interact each other through a neural network and specific neurons win through a process of competition and cooperation, resulting in a success of depth perception.

2.2. Two-layered SVNN Equations

The two-layered SVNN equations are given as

$$\tau_1 \dot{\xi}_u^a(t) = -\xi_u^a(t) + f(\alpha_u^a) \tag{1}$$

where $\xi_u^a(t)$ is a time-dependent neural activity, for example at an u-th temporal frame of an arbitrary vocal sound /a/, in which f(x) is a sigmoid function, that is

$$f(x) = \frac{\tanh(w(x-h))+1}{2} \tag{2}$$

$\alpha_u^a(t)$ is given as following:

$$\tau_2 \dot{\alpha}_u^a(t) = -\alpha_u^a + A\lambda_u^a - B \sum_{\substack{a'=a-a_1 \\ a'\neq a}}^{a+a_2} g(\xi_u^{a'}(t)) + D \sum_{\substack{u'=u-l \\ u'\neq u}}^{u+l} g(\xi_{u'}^a(t)) \tag{3}$$

where the second, third, and forth terms are referred as the input similarity, competitive and cooperative coupling,

respectively. Therefore $\alpha_u^a(t)$ is always influenced by input similarity, $\lambda_u^a(t)$, as well as neighboring neural activities, $\xi_{u'}^{a'}(t)$. The summation indices of competitive coupling run over the search area of all available candidates, within the range of $a-a_s \le a' \le a+a_s$ with a restriction of $a' \neq a$. On the other hand, the summation indices of cooperative coupling run over the search area of all frames, within the range of $u-l \le u' \le u+l$ with a restriction of $u' \neq u$. $\lambda_u^a(t)$ is a normalized similarity represented as following

$$\lambda_u^a = \frac{\log N(o_u;\mu_a,\Sigma_a) - <\log N>}{<\log N>} \tag{4}$$

where N is a Gaussian probability density function (PDF) with input data o_u, mean μ_a, and covariance $\sum_a$. <log N> is an average value over temporal frames. On the other hand, g(u) is a function given by

$$g(u) = u^+ = \frac{u+|u|}{2} \tag{5}$$

A, B, D, w, h, and τ_1, τ_2 used in the above equations are all positive constants which are to be chosen appropriately.

Figure 2 shows that the value of $\alpha_u^a(t)$ determines a certain point on the curve of sigmoid function. It is noticed that the output value of $\xi_u^a(t)$ depends on what values $\alpha_u^a(t)$ takes. The neural net equations make $\alpha_u^a(t)$ and $\xi_u^a(t)$ move toward a stable point (0 or 1) in the Sigmoid function ultimately.

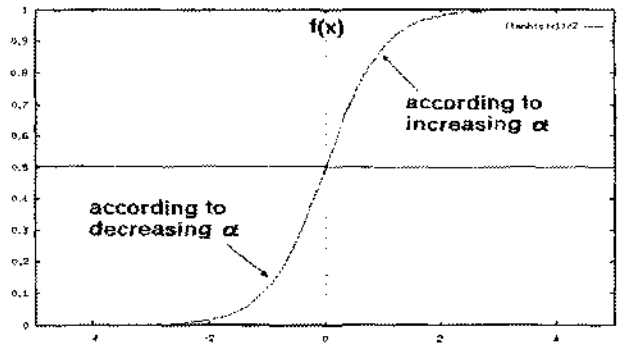

Figure 2. Sigmoid function with a coefficient $\alpha_u^a(t)$.

Figure 3 shows the two-layered SVNN with a process of competitive and cooperative coupling between two layers such as $\alpha_u^a(t)$ and $\xi_u^a(t)$. At equilibrium of $\dot{\xi}_u^a = \dot{\alpha}_u^a = 0$, the equation (1) may be written as

$$\xi_u^a(t) = f(\alpha_u^a) \tag{6}$$

The solution of the equation means that $\xi_u^a(t)$ has an identical movement in proportion to the coefficient value of $\alpha_u^a(t)$ that is greatly affected by both competitive and cooperative process among input similarities. As a result, the most stable state in the equation would be obtained through dynamic process of SVNN equations.

2.3. Three-layered SVNN Equations

The three-layered SVNN equations are given as

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

where $\xi_u^a(t)$ is a time-dependent neural activity and f(x) is the sigmoid function defined in the two-layered, in which $\beta_u^a(t)$ is a coefficient represented as a middle layer, that is

$$\tau_2 \dot{\beta}_u^a(t) = -\beta_u^a(t) + g(\alpha_u^a(t)) + g(\xi_u^a(t)) \tag{8}$$

where g(u) is a linear function given by

$$g(u) = u^+ = \begin{cases} u, u > 0 \\ 0, u \le 0 \end{cases} \tag{9}$$

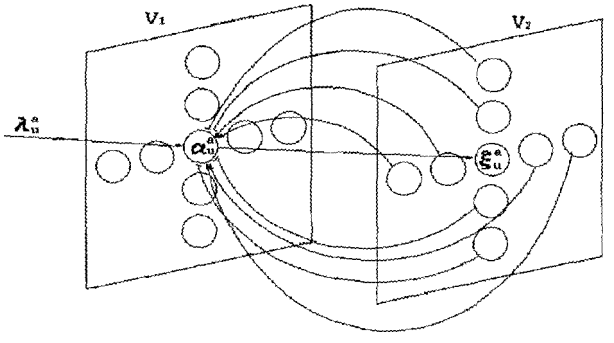

Figure 3. Two-layered SVNN with a dynamic process of competitive and cooperative coupling between $\alpha_u^a(t)$ and $\xi_u^a(t)$.

$\alpha_u^a(t)$ is given as following

$$\tau_3 \dot{\alpha}_u^a(t) = -\alpha_u^a + A\lambda_u^a - B \sum_{\substack{a'=a-\alpha_1 \\ a' \neq a}}^{a+\alpha_2} g(\xi_u^{a'}(t)) + D \sum_{\substack{v'=v-l \\ v' \neq u}}^{v+l} g(\xi_{v'}^a(t)) \quad (10)$$

A, B, D, and, τ_1, τ_2, τ_3 used in the above equations are all positive constants. Figure 4 shows the three-layered SVNN with a process of competitive and cooperative coupling among three layers such as $\alpha_u^a(t)$, $\beta_u^a(t)$, and $\xi_u^a(t)$.

The equilibrium solution can be considered by assuming that $\dot{\xi}_u^a = \dot{\alpha}_u^a = \dot{\beta}_u^a = 0$. Therefore, the equations of (7), (8), (10) can be represented as following

$$\xi_v^a(t) = f(g(\alpha_u^a) + g(\xi_u^a)) \quad (11)$$

The solution of the equation can be an intersection between y = ξ and y = $f(g(\alpha) + g(\xi))$. Figure 5 shows a simulation on movement of the intersection point between two equations, in which the intersection value varies from 0 to 1.

In this simulation, it is noticed that the value of intersection rises in proportion to an increase of $\alpha_u^a(t)$. As a result of the dynamic process, $\xi_u^a(t)$ will approach a certain stable point, so that only one neuron would be determined regardless of initial conditions of SVNN equations.

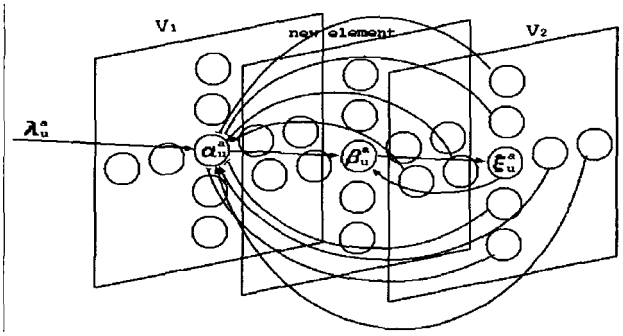

Figure 4. Three-layered SVNN with a dynamic process of competitive and cooperative coupling among $\alpha_u^a(t)$, $\beta_u^a(t)$, and $\xi_u^a(t)$.

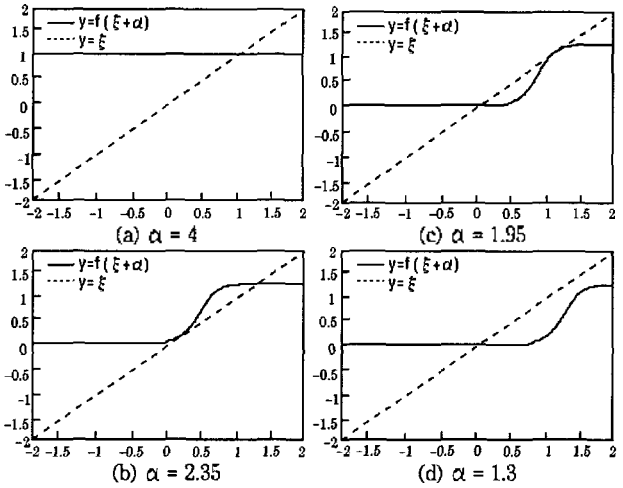

Figure 5. The curves of $y = \xi$ and $y = f(g(\alpha) + g(\xi))$.

III. Application of SVNN to Speech Recognition

The two kinds of SVNN equations mentioned above have a common process in dealing with input similarities in spite of each different mechanism. Namely, two different neural net equations feature a dynamic process with competition and cooperation. In speech recognition based on SVNN, the similarities are first obtained by comparing the input vocal signals with the trained standard models.

The similarity map is then given to the dynamic process with competitive and cooperative coupling. Figure 6 shows the dynamic process among input similarities.

As shown in this figure, the first layer, $\alpha_u^a(t)$, is influenced by not only input similarities but neighboring

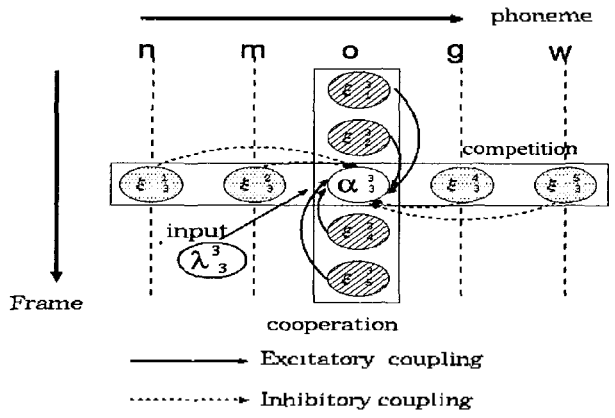

Figure 6. Competitive coupling among similarities in all possible candidates and cooperative coupling in temporal frames.

neural activities. Namely, it is activated by an inhibitory coupling among candidates and by an excitatory coupling among neighboring frames as well. The dynamic process ultimately makes each neuron to converge to a certain final value independent of the initial conditions of parameters in neural net equations.

Figure 7 shows an example of time-dependent behaviors of $\alpha_u^a(t)$ at the fifth frame of every candidate phonemes in which the similarity value of /n/ becomes even bigger through a recursive processing than other candidates. Since the excitatory coupling is more activated than the inhibitory one, the cooperative coupling causes $\alpha_u^a(t)$ to grow while others fall down.

Figure 8 shows an example of time-dependent behaviors of $\xi_u^a(t)$ influenced by $\alpha_u^a(t)$. It starts with the initial value

OUTPUT = /n/

frame	/n/	/m/	/o/	/g/	/w/
1	0.997414	0.000000	0.000000	0.000000	0.000000
2	0.997464	0.000000	0.000000	0.000000	0.000000
3	0.997310	0.000000	0.000000	0.000000	0.000000
4	0.997000	0.000000	0.000000	0.000000	0.000000
5	0.997004	0.000000	0.000000	0.000000	0.000000
6	0.996553	0.000000	0.000000	0.000000	0.000000
7	0.996747	0.000000	0.000000	0.000000	0.000000
8	0.996149	0.000000	0.000000	0.000000	0.000000
9	0.995134	0.000000	0.000000	0.000000	0.000000
10	0.989131	0.000000	0.000000	0.000000	0.000000
11	0.963000	0.000348	0.000000	0.000000	0.000000
12	0.001010	0.910223	0.000000	0.000000	0.000000
13	0.000751	0.934905	0.000000	0.000000	0.000000
14	0.000306	0.968465	0.000003	0.000000	0.000000
15	0.000267	0.981066	0.000009	0.000000	0.000000

Figure 9. Output values as a result of the dynamic process using 2 and 3LNN equations with A=3.0 (3.0), B=3.5 (3.5), D=2.0 (1.5), w=1.0 (2.5), h=0.5 (0.5) in 2LNN (3LNN).

preset in the neural net equations. $\alpha_u^a(t)$ first takes values corresponding to $\lambda_u^a(t)$. $\xi_u^a(t)$ then updates its value

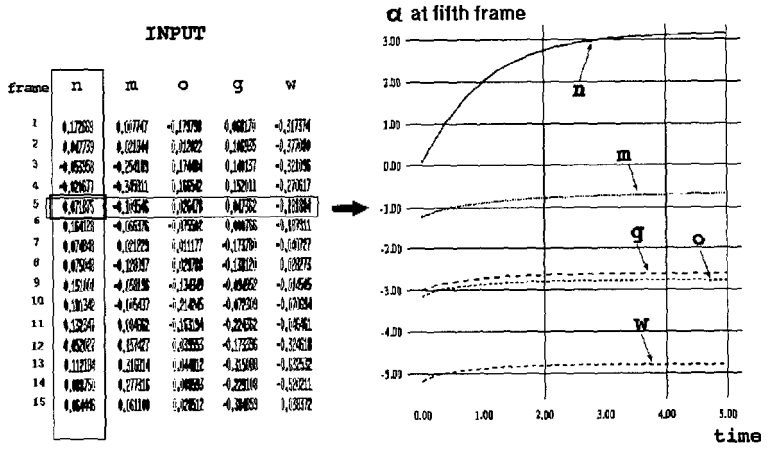

Figure 7. Time-dependent behaviors of $\alpha_u^a(t)$ at the fifth frame of every candidate phonemes.

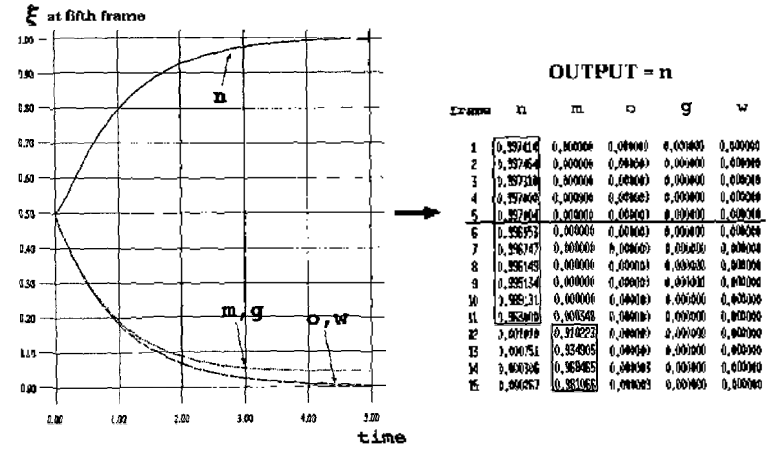

Figure 8. Time-dependent behaviors of $\xi_u^a(t)$ at the fifth frame of every candidate phonemes.

through competition-cooperation process among neighboring neural activities. In this figure, for example, the value of $\xi_{ij}^{a}(t)$ in /n/ grows to converge to a maximum point, while others fall down to approach minimum values.

The binocular neurons compete over the inhibitory coupling area and simultaneously cooperate over the excitatory area. Through the dynamic process, therefore, only one specific neuron wins over the other neurons whose activities are damped to minimum points. As shown in figure 2(2LNN) and 5(3LNN), the dynamic process ultimately makes $\alpha_{ij}^{a}(t)$ and $\xi_{ij}^{a}(t)$ to move toward a stable point, namely 0 or 1, regardless of the initial conditions of parameters.

Figure 9 shows an example of output values through dynamic process in which parameters were determined experimentally. The neuron with value of near 1 is called a winner neuron, whereas one with value of 0 is called a loser neuron. Since /n/ has more winner neurons than others, it is finally recognized as the most likely candidate to input speech.

Table 1. Analysis of speech signals.

[illegible]	16 Khz, 16 Bit
[illegible]	0.97
[illegible]	16 msec Hamming window
[illegible]	5 ms
[illegible]	10 order MFCC + 10 order delta MFCC

IV. Experiments and Discussion

The Japanese phoneme recognition based on SVNN was conducted, which was also compared with the performance

Table 2. Comparison of HMM with two-layered SVNN on two test sets

	[illegible]		[illegible]	
NG	53.46	84.81	59.62	89.10
A	92.55	95.03	93.85	99.23
B	76.62	74.68	86.79	88.68
CH	84.62	78.46	100.00	91.67
D	69.84	64.06	74.07	70.37
E	64.77	86.36	80.86	98.77
G	57.14	46.75	45.71	36.11
H	63.46	50.00	53.33	60.00
I	69.16	85.71	84.18	97.64
J	97.01	94.02	93.10	89.66
K	55.25	61.18	67.02	63.12
M	61.90	44.33	86.67	76.67
N	44.30	40.00	50.00	45.83
O	70.58	92.18	66.67	96.48
P	64.00	61.53	100.00	100.00
R	62.34	28.94	42.30	22.36
S	89.01	90.10	76.40	91.01
SH	96.05	84.21	91.11	95.56
T	4.35	33.33	15.38	56.41
TS	65.22	86.95	89.74	89.74
U	94.78	61.66	59.80	85.33
W	84.38	54.54	91.03	91.15
Y	61.36	63.63	87.30	94.84
Z	87.76	66.67	93.10	92.59
Total(%)	**71.56**	**77.41**	**72.37**	**82.87**

Table 3. Comparison of HMM with three-layered SVNN on two test sets.

	[illegible]		[illegible]	
NG	53.46	83.54	59.62	89.03
A	92.55	94.62	93.85	96.41
B	76.62	77.22	86.79	86.79
CH	84.62	83.08	100.00	83.33
D	69.84	71.88	74.07	62.96
E	64.77	86.74	80.86	96.30
G	57.14	48.05	45.71	38.89
H	63.46	51.92	53.33	60.00
I	69.16	86.04	84.18	96.97
J	97.01	94.03	93.10	93.10
K	55.25	67.58	67.02	58.16
M	61.90	47.17	86.67	60.00
N	44.30	38.75	50.00	50.00
O	70.58	91.56	66.67	89.45
P	64.00	40.00	100.00	77.78
R	62.34	25.97	42.30	35.65
S	89.01	86.81	76.40	78.65
SH	96.05	84.21	91.11	95.56
T	4.35	28.99	15.38	48.72
TS	65.22	86.96	89.74	89.74
U	94.78	62.07	59.80	68.00
W	84.38	69.70	91.03	74.15
Y	61.36	72.73	87.30	92.86
Z	87.76	87.76	93.10	93.33
Total(%)	**71.56**	**78.05**	**72.37**	**78.94**

of HMM speech recognizer with a structure of a single mixture and three states. For training standard models, fist of all, each of recognition systems used two kinds of the phoneme-labeled training database. The labeled phonemes were extracted from ATR Japanese word speech database which was composed of 4000 words spoken by 10 male speakers, and from ASJ Japanese continuous speech database which was composed of 500 sentences by 6 male speakers. For evaluation, test data consisted of two kinds, one from database of 216 words set and the other from 240 words set, each of which is the phoneme balanced ATR database spoken by 3 male speakers, respectively.

Table 1 shows the analysis of speech signals in which 10 dimensional Mel-frequency Cepstrum coefficients (MFCC) and their derivatives were used for feature parameters. In order to absorb the changes of the characteristic features in the same phoneme, the feature data is divided into two parts, the former half and the latter half in each phoneme. The input data, which is divided into two parts, is compared with the corresponding part of the Gaussian PDF's separately and a similarity map is then obtained for a dynamic process of SVNN.

The speaker independent recognition accuracies based on two- and three-layered SVNN were shown in Table 2 and 3, respectively.

When using two-layered SVNN, the average recognition accuracies of 3 speakers were 77.41% and 82.87% for 216 and 240 test sets, which were compared with 71.56% and 72.37% by HMM, respectively. When using three-layered SVNN, on the other hand, the accuracies were 78.05% 78.94% for 216 and 240 test set, respectively.

Table 4 shows the overall recognition accuracies on the performance based on SVNN compared with HMM. On 216-test set, the accuracies for two- and three-layered SVNN were about 5.9% and 6.5% higher than HMM, respectively. On 240-test set, on the other hand, the accuracies for two- and three-layered SVNN were 9.5% and 6.6% higher than HMM, respectively. As shown in this table, two-layered SVNN was 7.7% higher in average and three-layered was 6.6% higher. It is presumed that the differences of performance between two- and three-layered SVNN occur because of the difference of mechanism of choosing winner neuron.

Table 4. Overall recognition accuracies for two- and three-layered SVNN in comparison with HMM.

Test set	Recognizer	Recognition Accuracy	Improvement
216 set	HMM	71.56%	
	2LNN	77.41%	5.9%
	3LNN	78.05%	6.5%
240 set	HMM	72.37%	
	2LNN	82.87%	9.5%
	3LNN	78.94%	6.6%

As a result, it was found that SVNN outperformed the existing HMM recognizer. However, as shown in phoneme "R" for example, the accuracies based on SVNN do not show always better performance in every phoneme than HMM. In order to reduce the error rate in performance, the first thing to be considered is an exact modeling of the inner change of phonetic features. Since this study is restricted to phoneme recognition, in addition, we should make further experiments to word or continuous speech recognition as future works.

VI. Conclusion

The present study focuses on enhancing the discriminative capability in detecting the most likely candidate out of confused sounds. The proposed neural networks were proved to be successful in performing them in this respect. Particularly, it was revealed that the mechanism of dynamic process for stereoscopic vision, which played a crucial role in selecting the best candidate as winner neuron, might be compatible with the underlying principle of speech identification. In addition, we could see that the totally new types of the neural networks for speech recognition were able to yield much simpler architecture than the other ordinary artificial neural networks. From the experimental results, moreover, it was shown that the proposed approach with the unique characteristics in recognizing speech had better recognition performance than the existing HMM recognizer.

References

1. P. C. Woodland, C. J. Leggestter, J. J. Odell, et.al., "The 1994 HTK Large Vocabulary Speech Recognition System," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1, 73-76, 1995.
2. X. D. Huang, Y. Ariki, M. A. Jack, *Hidden Makov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, U.K., 1990.
3. K. F. Lee, H. W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Transaction on Acoustic, Speech and Signal Processing*, **37** (11), 641-1648, 1989.
4. H. Bourlard and C. J. Wellekens, "Links between Markov Models and Multi-layer Perceptrons," *IEEE Transaction Patt. Anal. Machine Intell.*, **12**, 1167-1178, 1990.
5. J. Lang, A. Waibel and G. E. Hinton, "A Time-Delay Neural Network Architecture for Isolated Word Recognition," Artificial Neural Networks, Paradigms, Applications and Hardware Implementations, IEEE press, New York, 388-408, 1992.
6. G. Martinelli, "Hidden Control Neural Network," *IEEE Transaction 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 and 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, Academic Press, 119-165, 1977.
10. Y. Yoshitomi, T. Kanda, T. Kitazoe, "Neural Nets Pattern Recognition Equation for Stereo Vision," Trans. IPS, 29-38, 1998.
11. Y. Yoshitomi, T. Kitazoe, J. Tomiyama, Y. Tatebe, "Sequential stereo Vision and Phase Transition," *Proc. Third International Symposium 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.

[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