Phoneme Recognition based on Two-Layered Stereo Vision Neural Network

Sung-Il Kim[†] and Nag-Cheol Kim[††]

ABSTRACT

The present study describes neural networks for stereoscopic vision, which are applied to identifying human speech. In speech recognition based on 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, the two-layered SVNN was 7.7 % higher in recognition accuracies than the hidden Markov model (HMM). From the evaluation results, it was noticed that SVNN outperformed the existing HMM recognizer.

2층 구조의 입체 시각형 신경망 기반 음소인식

김성일[†] · 김낙철[††]

요 약

본 연구는 입체 시각을 위한 신경망에 대한 연구 결과로서 인간의 음성을 인식하는데 적용된다. 입체 시각 신경망(SVNN)에 기반한 음성인식에서, 먼저 입력된 음성 신호를 표준 모델과 비교함으로써 유사성이 얻어진다. 이 값들은 다이나믹한 처리 과정으로 주어지고 이웃한 신경소자들 사이에서 경쟁적이고 협력적인 처리를 거치게 된다. 이러한 다이나믹한 처리과정을 통해 단 하나의 가장 우수한 신경세포(winner neuron)만이 최후에 검출된다. 비교연구에서, 2층 구조의 SVNN은 HMM 인식기보다 인식정확도 측면에서 7.7% 더 높았다. 평가 결과, SVNN은 기존의 HMM 인식기 성능을 능가하는 것으로 나타났다.

Key words: Speech Recognition, Neural Network, Stereoscopic Vision Neural Network, HMM

1. 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 (ANN) [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 human-like speech understanding modeling. As the alternative approach, therefore, ANNs such as multilayer perceptron[4], time-delay neural network[5], hidden control neural network[6] etc., have been introduced by modeling and processing mechanism of physiological human brain. One of the major strengths 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. One is that it

접수일 : 2002년 5월 23일, 완료일 : 2002년 8월 16일

[†] 중국 청화대학 음성기술센터 연구원
[††] 정회원, 대구기능대학 전자과 조교수

has much simpler architecture because network parameters are always fixed and not revised at any time. The other is that it has a powerful capability in information process of identifying the most likely neuron among confusable candidates. The process is made by both cooperative and competitive process among their similarities. The input streams for stereoscopic vision are similarity or disparity between two different data from left and right retinas. These stereo vision neural networks (SVNN)[7-9] process input visual data, yielding a depth perception of a specific object.

In a similar way, it is assumed that speech recognition can be performed by the same process between the vocal features as input data into a human auditory organ and the memorized ones as standard models in human brain. In this process, 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 that 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 for stereoscopic vision to speech recognition, on the viewpoint of information processing based on the neural networks.

The recently modified algorithms using SVNN, which have been optimized through preliminary investigations based on the coupled pattern recognition equations[10,11] and three-layered neural network[12-14], were successful in both stereoscopic depth perception and speech recognition. In a new approach, the proposed algorithms have much simpler architecture for recognition process than the past investigations mentioned above. In addition, it would be explored how well the new type of neural networks can be adopted to human speech identification. In addition, a comparative study with the existing HMM speech recognizer would be made under the same condition.

2. Stereoscopic Vision Neural Networks

2.1 Stereoscopic Depth Perception

When we look at a specific object, visual information from left and right eyes, which is seen differently as if it is slightly shifted horizontally, is delivered to our binocular neurons. By fusing left and right incoming information, therefore, we are able to perceive its depth. This is known as a depth perception phenomenon that can be realized by estimating visible differences between left and right retinas. It has been proved by assuming that our brain has neural networks fusing a disparity between two kinds of different visible images.

In stereoscopic vision, our brain system recognizes a three-dimensional depth by dealing with the vast information incoming through left and right retinas, which is then processed chiefly in the visual area 2(V2) of the neural networks of the brain. If two objects are separated in depth from the viewer, the relative positions of their images will differ in two eyes. Accordingly, our brains are capable of measuring the disparity and processing it by a dynamic process, so that depth is finally perceived. Figure 1 shows that the depth perception is successfully achieved by neural net equations with a dynamic process with both competitive

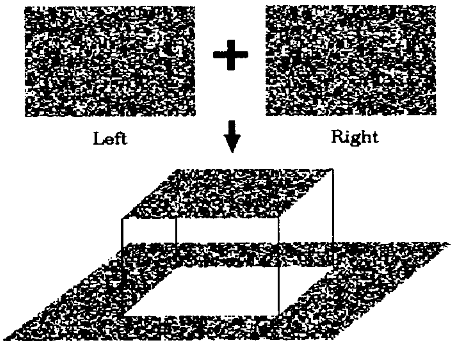


Figure 1. (Top) Pair of random-dot stereograms through left and right eyes. (Bottom) three-dimensional image of the stereograms obtained by a dynamic process of SVNN.

and cooperative process on the random dot stereograms. It was produced by displacing the square area in the random dot image horizontally by a certain amount, and by taking both image with original square area and image with horizontally displaced square area. If stereoscopically fused, the central square can be seen as if it is floated over the image plane, through a competitive process among input similarities of stereoscopic neurons and a cooperative process among them as well.

We will describe the recently developed two-layered SVNN equations with dynamic process of competitive and cooperative coupling among input similarities, which have been successful in perceiving depth for stereoscopic vision. It would be then explored how well the dynamic process of SVNN equations works in speech recognition.

2.2 Two-Layered SVNN Equations

The two-layered SVNN equations are given as

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

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

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

α_u^a is given as following:

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

where the second, third, and forth terms are referred as the input similarity, competitive and cooperative coupling, respectively. Therefore α_u^a is always influenced by input similarity, λ_u^a, as well as neighboring neural activities, $\xi_{u'}^{a'}$. The summation indices of competitive coupling run over the search area, for instance, all available hypotheses within the range of $a - a_s \le a' \le a + a_s$ with a restriction of $a' \neq a$. On the other hand, those of cooperative coupling run over the search area, for instance, the temporal frames within the range of $u - l \le u' \le u + l$ with a restriction $a' \neq a$. λ_u^a is a normalized similarity represented as

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

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

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

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 α_u^a determines a certain point on the curve of sigmoid function. Therefore, the output value of ξ_u^a depends on what values α_u^a takes.

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

$$\xi_u^a(t) = f(\alpha_u^a) \qquad (6)$$

The solution of the equation means that has an

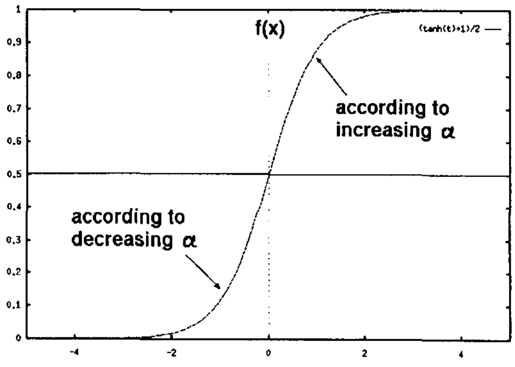


Figure 2. Sigmoid function with a coefficient α_u^a

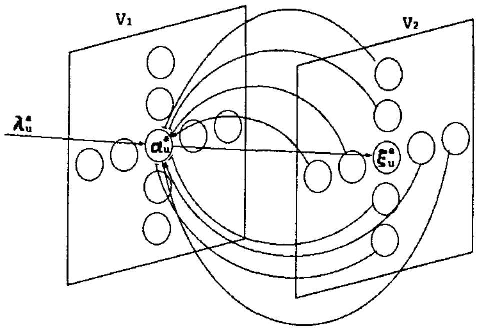
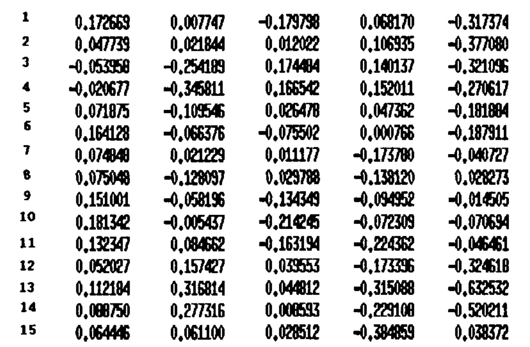


Figure 3. Two-layered SVNN with a dynamic process of competitive and cooperative coupling between α_u^a ξ_u^a and

identical movement of sigmoid function in proportion to the coefficient value of 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.

3. Application of SVNN to Speech Recognition

In speech recognition based on SVNN, similarities are first obtained by comparing incoming vocal features with trained standard models. Figure 4 shows an example of normalized input similarity values in five different candidate phonemes.



INPUT

frame	/n/	/m/	/o/	/g/	/w/
1	0,172663	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 4. Example of normalized input similarity values in each candidate phoneme

The similarity map is then given to dynamic process with competitive and cooperative coupling in SVNN. Figure 5 shows the dynamic process among input similarities.

As shown in this figure, the first layer, α_u^a, is influenced by not only input similarities but neighboring neural activities. Namely, it is activated by an inhibitory coupling among hypotheses and by an excitatory coupling among neighboring frames as well. The dynamic process ultimately makes each neuron converge to a certain final value, independent of initial conditions of parameters in SVNN.

Figure 6 shows an example of time-dependent behaviors of α_u^a at the fifth frame of each candidate phoneme in which the similarity value of /n/ becomes even bigger in the process of recursive processing than other candidates. Since the excitatory coupling is more activated than inhibitory one, the

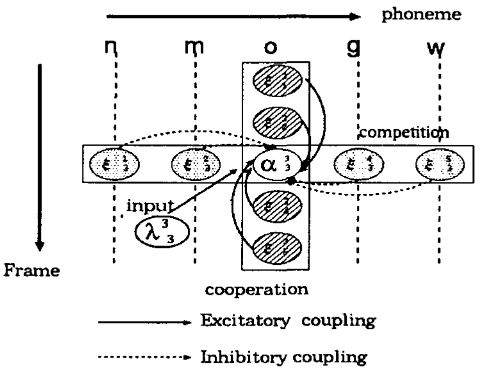


Figure 5. Competitive coupling among similarities in different phonemes and cooperative coupling in temporal frames

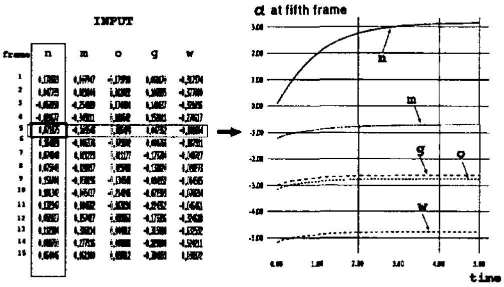


Figure 6. Time-dependent behaviors of α_u^a at the fifth frame of each candidate phoneme

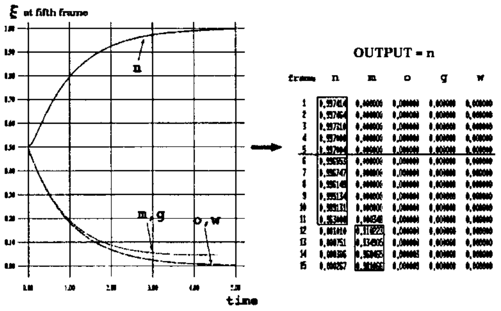


Figure 7. Time-dependent behaviors of ξ_u^a at the fifth frame of each candidate phonemeigure

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 8. Output values as a result of the dynamic process with A=3.0, B=3.5, D=2.0, w=1.0, h=0.5 in two-layered SVNN equations

cooperative process causes α_u^a to grow more than the others.

Figure 7 shows an example of time-dependent behaviors of ξ_u^a which is influenced by α_u^a. It starts with the preset initial values in the neural net equations. First of all, α_u^a takes values corresponding to λ_u^a. Then, ξ_u^a updates its value through competitive and cooperative process among neighboring neural activities. In this figure, for example, the value of ξ_u^a in /n/ grows to converge to a maximum point, while the 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 recursive dynamic process, therefore, only one specific neuron wins over the other neurons whose activities are damped to minimum points.

Figure 8 shows an example of the output values through the dynamic process. 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.

4. Experimental Conditions

Japanese phoneme recognition based on SVNN was conducted, which was also compared with the performance of HMM speech recognizer with a structure of a single mixture and three states. For training standard models, first of all, each recognition system 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, the test data consisted of two kinds, one from database of 216 words and the other from 240 words, 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.

Table 1. Analysis of speech signals

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

5. Experiments

The speaker independent recognition accuracies based on two-layered SVNN were shown in table 2. When using two-layered SVNN, the average recognition accuracies were 77.41% and 82.87% for 216 and 240 test set, which were compared with 71.56% and 72.37% by HMM, respectively.

Table 3 shows the overall recognition accuracies on the performance in SVNN compared with HMM. On 216 test set, the accuracies for two-layered SVNN were about 5.9 % higher than HMM. On 240 test set, on the other hand, the accuracies for two-layered SVNN were 9.5 % higher than HMM. As shown in this table, therefore, the two-layered SVNN was 7.7 % higher in average than HMM. As a result, it was noticed that SVNN outperformed the existing HMM recognizer.

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

Phoneme	216 test set		240 test set	
	HMM	SVNN	HMM	SVNN
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

6. Conclusion

The focus of the present study is on enhancing the discriminative capability in detecting the most likely candidate out of confused sounds. In this respect, the proposed neural networks were proved to be successful in performing them. Particularly, it was noticed 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 human speech identification. Besides, totally new type of neural network is able to yield much simpler architecture than other ordinary ANNs. From the experimental results, moreover, we could see that the proposed approach with the unique characteristics in recognizing speech had better recognition performance than the existing HMM recognizer. However, the accuracies based on SVNN do not show always better performance than those based on HMM in every phoneme. Since this study is restricted to phoneme recognition, as future works, we should make further experiments to word or continuous speech for a real-world application.

References

[1] K.F. Lee, H.W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models", IEEE Transaction on Acoustic, Speech and Signal Processing, pp.1641-1648, 1989.

[2] 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, Vol.1, pp.73-76, 1995.

[3] Huang,X.D., Ariki,Y., Jack,M.A, "Hidden Makov Models for Speech Recognition", Book: Edinburgh University Press, Edinburgh, U.K., 1990.

[4] H. Bourlard, C.J. Wellekens, "Links between

Markov Models and Multi-layer Perceptrons", IEEE Transaction Pattern Analysis Machine Intelligence, Vol.12, pp.1167-1178, 1990.

[5] J. Lang, A. Waibel, 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, pp.388-408, 1992.

[6] G. Martinelli, "Hidden Control Neural Network", IEEE Transaction on Circuits and Systems, Analog and Signal Processing 41(3), pp.245-247, 1994.

[7] D. Reimann, T. Ditzinger, E. Fischer, H. Haken, "Vergence eye movement control and multivalent perception of Autostereograms", Biol. Cybern., Vol.73, pp123-128, 1995.

[8] D. Reinmann, H. Haken, "Stereo Vision by Self-organization", Biol. Cybern., Vol.71, pp.17-26, 1994.

[9] S. Amari and M.A. Arbib, "Competition and Cooperation in Neural Nets", Systems Neuroscience, Academic Press, pp.119-165, 1977.

[10] Y. Yoshitomi, T. Kanda, T. Kitazoe, "Neural Nets Pattern Recognition Equation for Stereo Vision", Trans. IPS, pp.29-38, 1998.

[11] T. Kitazoe,T. Ichiki, S-Ill Kim, "Acoustic Speech Recognition Model by Neural Net Equation with Competition and Cooperation", Proc. International Conference on Spoken Language Processing, Vol.7, pp.3281-3284, 1998.

[12] T. Kitazoe, J. Tomiyama, Y. Yoshitomi et al., "Sequential Stereoscopic Vision and Hysteresis", Proc. Neural Information Processing, pp.391-396, 1998.

[13] Y. Yoshitomi, T. Kitazoe, J. Tomiyama, Y. Tatebe, "Sequential stereo Vision and Phase Transition", Proc. Third International Symposium on Artificial Life and Robotics, pp.318-323, 1998.

[14] T. Kitazoe, S-Ill Kim, T. Ichiki, "Speech Recognition using Stereovision Neural Network Model", Proc. International Symposium on Artificial Life and Robotics, Vol.2, pp.576-579, 1999.

Sung-Ill Kim

Sung-Ill Kim was born in Kyungbuk, Korea, in 1968. He received the 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. Currently, he has been working in the Center of Speech Technology, Tsinghua University, China. His research interests include speech/emotion recognition, neural network, and multimedia signal processing. E-mail; ksistar@hotmail.com

Nag-Cheol Kim

Nag-Cheol Kim was born in Kyungbuk, Korea, in 1963. He received the B.S., M.S. and Ph.D. degrees in the Department of Electronics Engineering from Yeungnam University, in 1985, 1987, 2002, respectively. From 1987 to 1996, he was a researcher in the Korea Electrotechnology Research Institute (KERI). Currently, he has been working in the Electronics Department of Daegu Polytechnic College. His research interests include Speech Signal Processing and EMI/EMC. E-mail; nckim@tgpc.ac.kr