音声言語情報処理 25－5
（1999. 2. 5）

Neural Net Equations with Competition and Cooperation for Speech Recognition

Tetsuro Kitazoe , Sung-Ill Kim , Tomoyuki Ichiki
Department of Computer Science and Systems Engineering
Faculty of Engineering , Miyazaki University
1-1 , Gakuen Kibanadai Nishi , Miyazaki , 889-2192 Japan

Abstract—— The stereo vision neural-net equations, known to process a depth perception, are applied to speech recognition . We use a recently developed three layered neural net (TLNN) equations with competition and cooperation for stereo vision . We use a Gaussian PDF to represent memorized data of each phoneme in our memory, and the similarities of an input phoneme with respect to the memorized ones were calculated . The TLNN equations are applied to the similarities with best 5 hypotheses among 24 kinds of phonemes . The average rates for speaker independent recognition are 78.05 % for 216 word database and 78.94 % for 240 word database by TLNN equations which are compared to 71.56 % and 72.37 % by Hidden Markov Model(HMM), respectively .
Key Words—— Speech Recognition, Stero Vision, Nueral Network, Hidden Markov Model

神経回路方程式を用いた競合と協調による音声認識

北添徹郎 、金星一 、市来知幸
〒 889-2192, 宮崎市学園木花台西 1-1
宮崎大学工学部情報工学科

あらまし—— 奥行き知覚のメカニズムとして知られているステレオビジョン神経回路方程式を用いて音声認識への適用を試みる。数年前ラインマンおよびハーケンにより提案され、最近本研究室で発展させられた競合と協調によるステレオビジョン神経回路を用いる。我々は、各音韻の特徴量が脳に蓄積されており、入力音声がそれらと比較され、音声認識を行なっていると考える。本研究では 24 セットの音韻のシミラリティを求め、その上位 5 つの音韻のみを最終的な候補として、ステレオビジョン神経回路方程式にかける。その結果、216 単語データベースにおいては 78.05 %、240 単語データベースにおいては 78.94 %という音韻の認識率が得られ、HMM による認識率、各 71.56 %と 72.37 %を上回る結果を得た。
キーワード—— 音声認識、ステレオビジョン、神経回路、隠れマルコフモデル

1. INTRODUCTION

Recently, many studies have been conducted for improvement of large vocabulary continuous speech recognition . There are two main trends in the art. One is to develop good acoustic models, such as triphone models with mixed Gaussian distribution and/or with tree based clustering models . The other is to reduce perplexity by using language models, such as n-gram and context free grammar . It was reported recently[1] that improving the acoustic models is much more effective than reducing perplexity; the improvement in word recognition rate achieved by increasing phoneme recognition by merely 1-2 % corresponds to the improvement achieved by decreasing perplexity as much as 10 - 20 % .

We applied stereo vision neural-net equations, known to process a depth perception, to speech recognition . In the stereo vision, the two-dimensional images of a 3-dimensional object are captured at two different points, left and right eyes . The local disparities(or similarities) between the two 2-dimensional images are input to the neural-net equations with competition and cooperation, whereby a clear depth perception can be obtained[2,3,4,5] . In real life speech recognition, it is considered that characteristic features of each phoneme are stored in our memory through daily life training and input speech data are compared with the memorized data to estimate their similarities(or disparities) for each phoneme .

We use a recently developed three layered neural net(TLNN) equations for stereo vision[4,5] to process similarities among phonemes . When the equations are applied to phoneme recognition, it develops competition among similarities of different phonemes and cooperation among neighboring frame variables, and a so-called winner-take-all process selects a specific phoneme as a recognized one, beating others down to zero . We used a Gaussian probability density function(PDF) to represent memorized data of each phoneme in our brain, and the similarities of an input phoneme with respect to the memorized ones were calculated . The simulation was performed for Japanese phoneme database[6,7] . It was found that the neural-net equations gave a clear recognition to each phoneme .

As training data, we used 4000 words recorded by 10 different speakers and 500 sentences recorded by 6 different speakers . We run recognition tests by using 10 dimensional MFCC coefficients and their derivatives . The recognition rates was greatly raised when each phoneme was divided into two parts, before and after the mid flame position of a phoneme, and the respective similarities were calculated separately. The TLNN equations are applied to the similarities with best 5 hypotheses among 24 kinds of phonemes . The average rates for speaker independent recognition were 78.05 % for 216 word database and 78.94 % for 240 word database by TLNN equations which were compared to 71.56 % and 72.37 % by Hidden Markov Model(HMM), respectively .

2. TLNN EQUATIONS WITH COMPETITION AND COOPERATION

The recently developed algorithm by TLNN equations has been successful for stereoscopic depth perception . Figure 1 shows that depth perception is successfully achieved with a dynamic process of competition and cooperation to treat the images coming from left and right eyes .

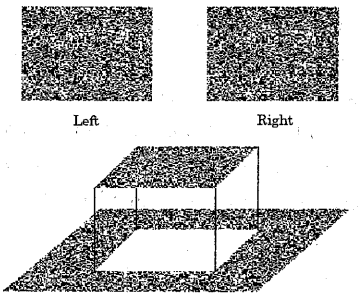

Figure 1: (Top) Pair of random-dot stereograms presented to the left and right eyes. (Bottom) 3-dimensional image of the random-dot stereograms viewed by the present neural net model

In the present paper, as a nueral net equations we employ TLNN equations for speech recognition to process the similarities between the characteristic features stored in our memory and the input speech from our ears . The TLNN equations are given as

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

where $\xi_u^a(t)$ is a time-dependent neuron activity and f(x) is a well known as sigmoid fuction given by

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

$$\dot{\beta}_u^a = -\beta_u^a + g(\alpha_u^a) + g(\xi_u^a) \tag{3}$$

where g(u) is a function given by

$$g(u) = u^+ = \begin{cases} u, u > 0 \\ 0, u \le 0 \end{cases} \tag{4}$$

$$\dot{\alpha}_u^a = -\alpha_u^a + A\lambda_u^a - B\sum_{a\prime \neq a} g(\xi_u^{a\prime}(t)) + D\sum_{u\prime=n-l}^{u+l} g(\xi_{u\prime}^a(t)) \quad (5)$$

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

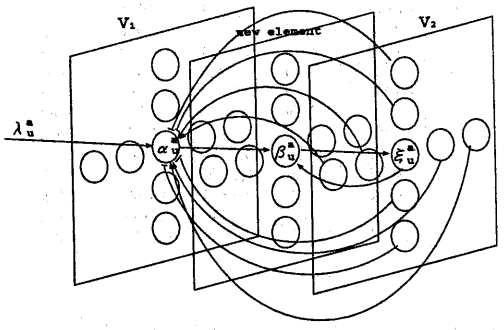

Figure 2: Three layered structure for TLNN equations

α_u^a receives not only similarity λ_u^a but also influence of neighboring neuron activities $\xi_{u\prime}^{a\prime}$. In equation (5), the second, third and forth terms in α_u^a are referred as the input, competitive and cooperative terms, respectively . The second term describs an similarity of input data at u-th frame to a certain phoneme /a/, the third term in α_u^a represents a competition with activities $\xi_u^{a\prime}(a\prime \neq a)$ of other phonemes and the forth term means a cooperation of the neighboring frames in the same phonemes . In $\sum_{a\prime \neq a}$ the summation indices run over the disparity search area(DSA) defined as $a - a_s <= a\prime <= a - a_s$ with the restriction $a\prime \neq a$. In $\sum_{u\prime=n-l}^{u+l}$ the summation indices run over the cooperation area(CA) defined as $u - l <= u\prime <= u + l$ with the restriction $u\prime \neq u$. The neural network for these equations has a three layered structure as shown in figure 2 .

To understand the qualitative feature of the equations, consider equilibrium solution $\dot{\alpha}_u^a = \dot{\beta}_u^a = \dot{\xi}_u^a = 0$. Equations (4)(5)(6) are written as

$$\xi_u^a = f(g(\alpha_u^a) + g(\xi_u^a)) \quad (6)$$

In figure 3, the curves of y = ξ and y = f(g(α) + g(ξ)) are shown for changing value of α_u^a from positive large one (a) down to small positive value (e) . The solutions are given by the intersection of the two curves . If α_u^a decreases from (a) to (e), the solution maintains approximately $\xi_u^a = 1$ until it reaches to the value of (d) . On the contrary, if α_u^a increases from (e) to (a), approximately $\xi_u^a = 0$ solution is maintained until it reaches to the value of (b) . From this fact we obtain two conclusions .

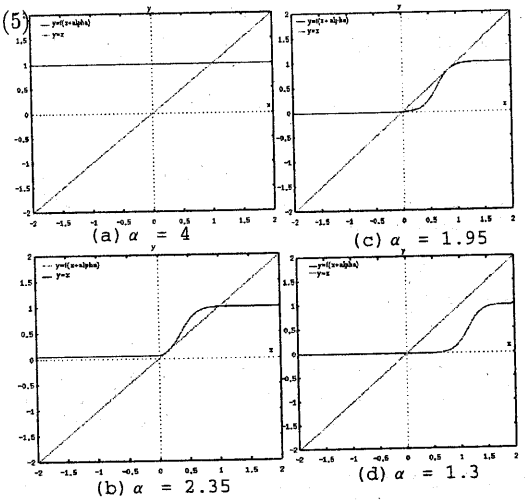

Figure 3: Curves of y = ξ and y = f(g(α) + g(ξ))

(1) For large α, ξ has high value(approximately 1) while it has low value(approximately 0) for small α .

(2) The solution ξ has different path according to whether α is increased or decreased, suggesting that there exists a hysteresis phenomenon .

The third solution which exists between (b) and (c) is not stable if we assume w > 1 .

3. APPLICATION TO SPEECH RECOGNITION

The speech(phoneme) recognition system by TLNN equations is divided into three main processes;

(1) A number of training speech data are classfied and parameterised 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 ξ_u^a in accordance with the similarity measure λ_u^a to a certain phoneme /a/ at the frame member u .

(4) The TLNN equations are performed to make an activity ξ_u^a 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 models for each phonemes are expressed in terms of Gaussian PDF for input o .

$$\mathrm{N}(o;\mu_a,\Sigma_a) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_a|}} e^{-\frac{1}{2}(o-\mu_a)^t\Sigma^{-1}(o-\mu_a)} \quad (7)$$

where μ_a is a mean value of feature vectors for training data of a phoneme /a/ . The covariance matrix Σ_a is given by

$$\Sigma_a = \frac{1}{N}\sum_{n=1}^{N}(o_n-\mu_a)(o_n-\mu_a)^t \quad (8)$$

where o_n is a training data of a phoneme /a/ .

The normalized similarity λ_u^a of input data o_u at u-th frame to a certain phoneme /a/ is defined as

$$\lambda_u^a = \frac{N(o_u;\mu_a,\Sigma_a) - < N >}{< N >} \quad (9)$$

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

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	-0,532532
14	0,088750	0,277316	0,008593	-0,229108	-0,520211
15	0,064446	0,061100	0,028512	-0,394859	0,038372

Figure 4: Input similarity map of best 5 hypotheses for TLNN equations

We apply TLNN equations to a speech recognition by using the similarity measures between memorized PDF's and input phoneme data . α is influenced additively by the input λ and the competition-cooperation terms accerelate winner-take-all processes . We call a winner neuron when ξ gets a positive value finally and a loser neuron when it approaches zero, losing whole activity . In actual time dependence, the situation is more complicated because α depends on neighboring ξ's and thus varies with time .

Figure 4 shows an example of a similarity map of best 5 hypotheses selected by similarity measures . Figure 5 shows results of recognition after TLNN

OUTPUT = n

frame	n	m	o	g	w
1	6,282958	-0,000000	-0,000000	0,000000	0,000000
2	6,044124	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	4,477895	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 5: Recognition results using TLNN equations

equations are applied . In this example, /n/is a winner for the frames 1-11, while /m/ is a winner for the frames 12-15 . Thus, we conclude /n/ is correctly recognized in average . To get an understanding for the processes dynamically, we notice the sigmoid form which depends on the typical values of α . Since the stable solution for the equations is decided by the equation (6), giving either high value(approximately 1) or low value(approximately 0), we set all ξ's=0.5 for the initial values in the following discussions .

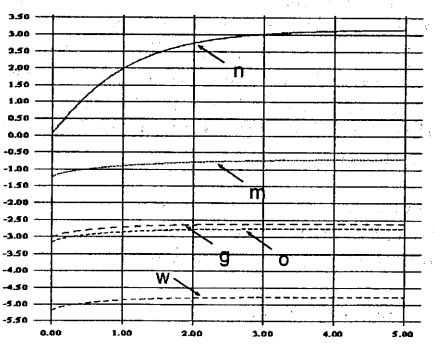

Figure 6: Time dependent behaviors for α

Figure 6,7 show the time dependent behaviors for α's and ξ's, at the 5-th frame for phoneme /n/,/m/,/o/,/g/,/w/ when the similarity map of figure 4 is input . Initially, only the difference among phonemes comes from α's in TLNN equations, where input λ's are different . ξ for

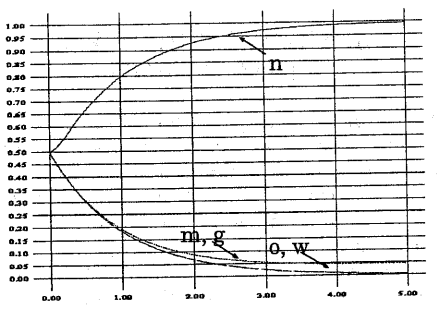

Figure 7: Time dependent behaviors for ξ

/m/,/o/,/g/,/w/ begin to decrease due to the sigmoid form for α <0 . On the contrary, α for /n/ with the biggest λ begins to take positive values as the competition term increases . When α^n becomes positive, the activity ξ^n turns to increase according to the sigmoid form . Figure 7 shows that ξ^n of phoneme /n/ turns to increase . At this stage, it is noticed that the cooperation term in α^n helps to rise α^n and accelerate ξ^n to increase . α's for other phonemes, on the other hand, begin to decrease because of the increase of their competition terms due to increasing ξ^n .

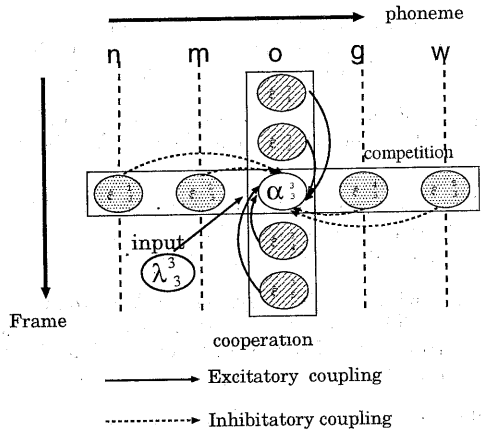

Figure 8: Process of competition among similarities of different phonemes and cooperation among neighboring framing data

Therefore, ξ's of other phonemes continue to decrease and finally goes to zero . Thus, the phoneme /n/ is recognized through TLNN equations . The applied parts of the processes of competition and cooperation are illustrated in figure 8 . In conclusion, the processes due to competition and cooperation in TLNN equations play a good role to make a definite recognition . A block diagram of the overall speech recognition system using TLNN equations is shown in figure 9 .

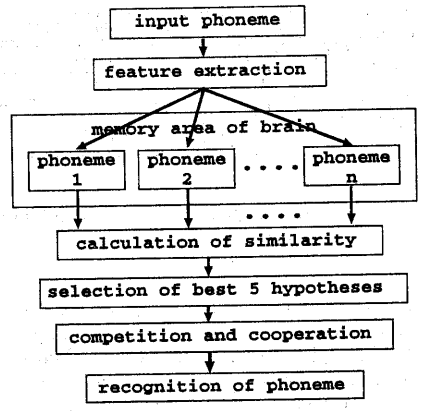

Figure 9: Overall block diagram of speech recognition system using TLNN equations

4. TECHNICAL IMPROVEMENTS AND EXPERIMENTAL RESULTS

To make Gaussian PDF for each phoneme from training data, we extracted labeled phonemes from ATR data[10] composed of 4000 words spoken by 10 male speakers and from ASJ data of 500 sentences[11] by 6 male speakers . The input data for recognition experiment were composed of two kinds, one from database of 216 words and the other from one of 240 words, spoken by 3 different male speakers respectively .

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

Table 1: Analysis of speech signal

The speech data were analyzed as follows; To compare our neuron model with the conventional model, the phoneme recognition experiment was performed for HMMs with single mixture and three states, by using the same database . We run recognition tests by using 10 dimensional MFCC coefficients and their derivatives as illustrated in table 1 . The cepstrum data of each phoneme was divided into two parts, before and after the mid flame position, and used to form two Gaussian PDF's respectively . The input data divided into two parts

Data	Model	Recognition rates
216 data	HMM	71.56
〃	TLNN	78.05
240 data	HMM	72.37
〃	TLNN	78.94

Table 2: Recognition results of speaker independent for two types of data

were compared to corresponding part of the Gaussian PDF's separately and a similarity map was obtained . The TLNN equations were applied to these similarity maps with best 5 hypotheses among 24 kinds of phonemes .

Pho	216 data		240 data	
	HMM	TLNN	HMM	TLNN
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
ALL	71.56	78.05	72.37	78.94

Table 3: Comparison between HMMs and TLNN models

The recognition results were shown in table 2 and 3, where the average rates for speaker independent recognition were 78.05 % for 216 word database and 78.94 % for 240 word database by TLNN equations which were compared to 71.56 % and 72.37 % by HMMs, respectively . We can see that our neuron model gave approximately 7 % higher in average recognition rates than the performance of HMMs for each database .

5. CONCLUSION AND DISCUSSION

We applied recently developed TLNN equations for stereo vision, known to process a depth perception, to speech recognition . In real life speech recognition, it is considered that characteristic features of each phoneme are stored in our memory through daily life training and input speech data are compared with the memorized data to estimate their similarity (or disparity) for each phoneme . When the equations are applied to phoneme recognition, it develops competition among similarities of different phonemes and cooperation among neighboring frame data, and selects a specific phoneme as a recognized one, beating others down to zero . The TLNN equations were applied to the similarities with best 5 hypotheses among 24 kinds of phonemes . The average rates for speaker independent recognition were 78.05 % for 216 word database and 78.94 % for 240 word database by TLNN equations which were compared to 71.56 % and 72.37 % by HMMs, respectively .

References

[1] S. Nakagawa "Ability and Limitation of Statistical Language Model" Proc.of ASJ:23-26,spring, 1998

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

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

[4] Y.Yoshitomi, T.kanda, T.kitazoe "Neural Nets Pattern Recognition Equation for Stereo Vision" Trans.IPS.Japan:29-38, 1998

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

[6] ATR Japanese Speech Database and Technical Report, Japan, 1988

[7] ASJ Continuous Speech Corpus for Research, Japan, 1991