Acoustic Speech Recognition Model by Neural Net Equation with Competition and Cooperation

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

ABSTRACT

The equation of neural nets for stereo vision is applied to speech recognion. We use Coupled Pattern Recognition (CPR) equation which has been shown to organize depth perception very well through competition and cooperation. We construct Gaussian probability density function(pdf) for each phoneme from a number of training data. The input data to be recognized are compared to the pdf's and the similarity measures are obtained for each phoneme. The CPR equation develops neuron activities by receiving the similarity measures as input. A recognition is achieved when the activities arrive at a stable states. The recognition rates for 25 Japanese phoneme are 74.75 % in average which is compared to 71.53 %, by Hidden Markov Model. A certain technical improvement is applied to our neuron model. by dividing data of a phoneme into two part. one for the former frames. the other for the latter frames. A remarkable improvement is obtained with average recognition rate of 79.79 %.

1. INTRODUCTION

Recently many studies have been focused on large vocabulary continuous speech recognition. The main issues are divided into two technological problems. One is how to improve good acoustic models such as triphone models with mixed Gaussian distribution and/or with tree based clustering. The other is how to reduce perplexity by using language models such as n-gram or context free grammer. It was reported recently that the former improvement was much more effective than the latter[1]. An increase of phoneme recognition rate by 1-2 % amounts to decreasing perplexity by 10~20 % in langnage model.

In the present paper, we try to apply the equation of stereo vision neural nets known to process a depth perception to speech recognition. In the stereo vision. a 3-dimensioned scene is imaged from two different points. left and right eyes. The local similarity(or disparity) between two 2-dimensional images are processed through the equation of neural nets with competition and coorperation. resulting in a clear depth perception[2.3.4.5]. In the speech recognition. we consider that characteristic features of each phoneme are stored in our memory through daily life training and that the input speech data are compared with the memorized data of each phoneme and their similarity (or disparity) measure is estimated for every phonemes.

We use the recent modified Coupled Pattern Recognition equation (CPR) for stereo vision[3.4.5] to process similarities among phonemes. When the equation is applied to phoneme recognition. it develops competition among activities of different phonemes and cooperation among those at neighbouring frames and the process known as winner-take-all selects a specific phoneme as a recognized one.beating others down to zero. We use a Gaussian probability density function(pdf) (represented by a mean vector and a covariance matrix of MFCC coefficients) to represent memorized data of each phoneme in our brain and the similarities of an input phoneme to the memorised ones are caculated. The simulation is performed for Japanese phoneme database supplied by ATR and ASJ[6.7]. It was shown that the neural net equation gave a clear recognition to each phoneme eventually.

It was shown that the recognition rates was much raised when each phoneme was devided into two parts. before and after the mid flame position of a phoneme data and its similalities were calculated separately. The CPR is applied to the similarities with best 5 hypotheses among 25 kinds of phonemes. The average rates were 79.79 % for speaker independent recognition and 81.00 % for speaker dependent recognition wheih were compared to 71.53 % and 82.84 % by HMMs. respectively.

2. SIMILARITY MEASURE AND NEURAL NET EQUATION

The speech(phoneme) recognition systems by neural net equation is divided into three main processes:

(1) A number of training speech data are stored and classfied for each phoneme. The data are supposed to be memorized in the brain with a standard form such as Gaussian pdf of cepstrums for each phoneme.

(2) An input phonemes is 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. The neural net equation processes the activity ξ_u^a to move toward a stable point after

the equation receives the similarity measure as an input and a recognition is achieved when it reaches to a stable state.

The memorized data are expressed in terms of Gaussian pdf for input o.

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

where μ_a is a mean value of training cepstrum data of a phoneme /a/.
Σ_a is given by

$$\Sigma_a = \frac{1}{N}\sum_{n=1}^{N}(o_n-\mu_a)(o_n-\mu_a)^t \quad (2)$$

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

$$\lambda_u^a = \frac{N(o_u;\mu_a,\Sigma_a) - < N >}{< N >} \quad (3)$$

where $< N >$ means an average over phonemes.

As a neural net equation we employ coupled pattern recognition(CPR) equation which was successful for recognizing stereovision. The CPR equation processes input similarity measures for two 2-dimensional data from left and right eyes and arrive at a definite depth perception through competition and cooperation processes. We use this equation for a speech recognition by using similarity measures λ_u^a between memorized pdf's and input phoneme data. CPR equation is given as

$$\dot{\xi}_u^a(t) = -\frac{dU}{d\xi_u^a(t)} \quad (4)$$

$$U(\xi_u^a(t)) = \frac{\alpha}{2}\xi_u^a(t)^2 - \frac{E}{3}\xi_u^a(t)^3 + \frac{C}{4}\xi_u^a(t)^4 \quad (5)$$

$$\alpha_u^a = -\lambda_u^a + (B+C)\sum_{a'\neq a}\xi_u^{a'}(t)^2 - D\sum_{u'=u-l}^{u+l}\xi_{u'}^a(t)^2 \quad (6)$$

where B.C.D.E are positive definite constants. α_u^a receives not only similarity λ_u^a but also influence of neighboring activities $\xi_{u'}^{a'}$. The second term in α_u^a represents a competition with activities ξ_u^a of other phonemes and the third term does cooperation among the same phonemes at the neighboring frames $u-l <= u' <= u+l$.

The solution of the CPR equation are determined by the input λ and initail values of ξ's. As stated later, however, they converge to the same values which are independent of the initial condition if they start from positive values.

3. CPR EQUATION WITH COOPERATION AND CONPETITION

Similarity measures λ are input to CPR equations (4)(5)(6) only through α and which plays an important role in the equations. Figure 1 shows typical potential form for positive and negative values of α. In the potential for $\alpha > 0$ given in figure 1(a). ξ tends to go to zero through CPR equations. In the potential for $\alpha < \frac{-E^2}{1.5C}$ given in figure 1(b). ξ goes to a certain positive value corresponding to the absolute minimum of the potential if initial ξ is set to a positive value. We call a winner neuron when ξ gets a positive value finally and a loser neuron when it becomes 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 2 shows an example

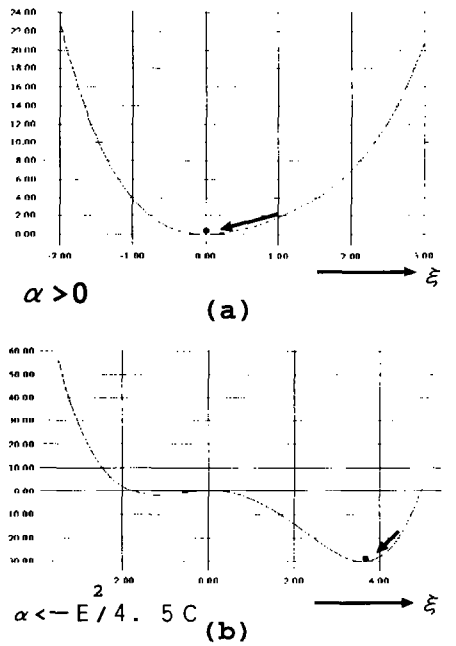

Figure 1: Potential function with B=0.25. C=1.25. D=0.60. E=3.00. l=4

of a similarity map where the input phoneme is actually pronounced as /n/ and the phonemes /n/./m/./o/./g/./w/ with best five similarities are selected. Figure 3 shows results of recognition after CPR equations are applied. /n/is a winner for the frames 1-11, while /m/ is a winner for the frames 12-15 in this example.

n	m	o	g	w
0,172569	0,007747	-0,179798	0,068170	-0,517374
0,047733	0,021844	0,012022	0,106935	-0,377080
-0,055958	-0,254189	0,174484	0,140137	-0,321096
-0,020677	-0,345811	0,166542	0,152411	-0,270617
0,071875	-0,109546	0,026478	0,047362	-0,181884
0,164128	-0,065376	-0,075502	0,000766	-0,187911
0,074848	0,021229	0,011177	-0,173780	-0,040727
0,075048	-0,128097	0,029788	-0,158120	0,028273
0,151001	-0,058196	-0,134349	-0,094952	-0,014505
0,181342	-0,005437	-0,214245	-0,072309	-0,070694
0,132347	0,084662	-0,163194	-0,224362	-0,046461
0,052027	0,157427	0,039553	-0,173396	-0,524618
0,112184	0,316814	0,044812	-0,315088	-0,632532
0,088750	0,277316	0,008593	-0,229108	-0,520211
0,064446	0,061100	0,028512	-0,384859	0,038372

Figure 2: An example of a similarity map

We conclude /n/ is correctly recognized in average. To get an understanding for the processes dynamically, we notice the potential form of figure 1 which is written for the

n	m	o	g	w
6.282958	-0.000000	-0.000000	0.000000	0.000000
6.044124	0.000000	0.000000	0.000000	0.000000
5.854819	0.000000	0.000376	0.000001	0.000000
5.706582	-0.000000	0.000000	0.000000	0.000000
5.592782	-0.000000	0.000000	0.000000	0.000000
5.479568	-0.000000	0.000000	0.000000	0.000000
5.218249	0.000000	-0.000000	0.000000	0.000000
4.969046	-0.000000	0.000000	0.000000	0.000000
4.728297	0.000000	-0.000000	-0.000000	-0.000000
4.477895	0.000000	0.000000	0.000000	0.000000
4.362889	0.000001	-0.000000	-0.000000	-0.000000
0.000006	3.567760	-0.000000	-0.000000	-0.000000
0.000000	3.731247	0.000000	-0.000000	-0.000000
-0.000000	3.905251	-0.000000	-0.000000	0.000000
-0.000000	4.125674	0.000000	-0.000000	-0.000000

Figure 3: Recognition results using CPR equation

typical values of α. The CPR equation let ξ's move to the local minimum.

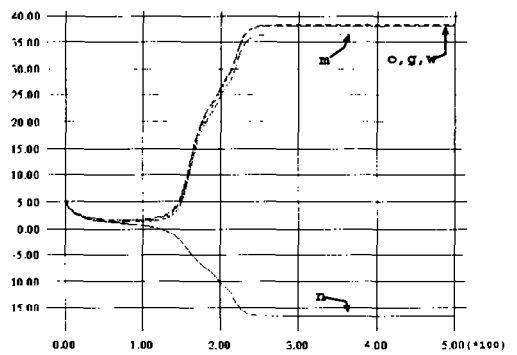

Figure 4: Time dependent behaviors for α

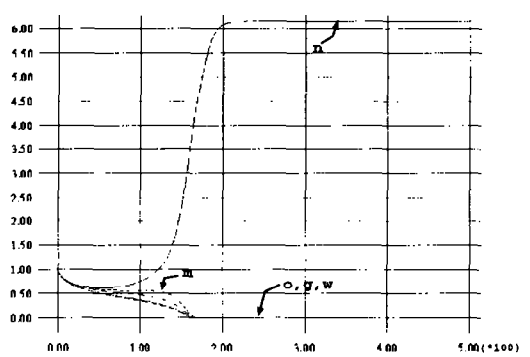

Figure 5: Time dependent behaviors for ξ

Since the stable solution for the equation is decided by the minima of the potential, the simulation shows that the solutions are uniquely determined as the minima of either figure 1(a) or figure 1(b), independent of the initial values of ξ's as far as we set the initial values of ξ's as positive definite in order to avoid to go to the local minimum for $\xi < 0$ in figure 1(b). We set all ξ's=1 in the following discussions.

Figure 4,5 show 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 2 is input. Initially, only the difference among phonemes comes from α's in CPR equation, where input λ's are different. In our parameter settings, whole α's start from the positive values as shown in figure 4 and whole ξ's begin to decrease from 1 as shown in figure 5 due to the potential form in figure 1(a) for $\alpha > 0$. Then, α for /n/ with the biggest λ begins to take negative values as the competition term decreases. When α^n becomes negative, the activity ξ^n turns to increase according to the potential form figure 1(b). Figure 5 shows ξ^n turns to increase. At this stage it is noticed that the cooperation term in α^n helps to reduce α^n and accelerate ξ^n in increasing. α's for other phonemes, on the other hand, begin to increase(Figure 4) because of the increase of their competition terms due to increasing ξ^n. Therefore, ξ's of other phonemes continue to decrease and finally goes to zero(Figure 5). Thus, the phoneme /n/ is recognized through CPR equations. In conclusion, cooperation and competition in CPR equation play a good role to make a definite recognition as shown in Figure 3.

4. TECHNICAL IMPROVEMENT AND EXPERIMENTAL RESULTS

To make Gaussian pdf for each phoneme from training data, we extracted labeled phonemes from ATR data[6] composed of 4000 words spoken by 10 male speakers repeatedly and from ASJ data of 500 sentences[7] by 6 male speakers. Input data for recognition experiment were composed of two kinds, one from speaker dependent(SD) data used for training and the other from new data of 216 words spoken by 3 male speaker independent(SI). The speech data were analyzed as follows:

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

To compare our neuron model with the conventional model, the phoneme recognition experiment was performed for Hidden Markov Model(HMM) with single mixture and three states, by using exactly the same data as was used for the neural net equation. We took 25 simple labeled phonemes from ATR data for recognition. The recognition results are shown in table 3 where our neuron model(N Model1) gives recognition rate 7 % less for SD data and 3 % higher for SI data than the HMM model.

We consider to improve our neuron model. If we look at the phoneme data, it is noticed that there are considerable differences in the characteristic features of cepstrum data between those of former half part and latter half part of frame data. Therefore, we decide to divide cepstrum data into two parts, the former half and the latter half of phoneme data. We make two Gaussian pdf's for each phoneme sep-

arately. Input data which are similarly divided into two parts are compared to corresponding part of the Gaussian pdf's separately and a similarity map is obtained. After CPR equation processed these similarity maps, the resulting recognition is shown in table 3(N Model2), where we see a remarkable improvement is obtained. Our neuron model gives recognition rate 8 % higher than the HMM does in average.

5. CONCLUSION AND DISCUSSION

The equation of neural nets for stereo vision was applied to speech recognion. We used CPR equation which has been shown to organize depth perception very well through competition and cooperation. We constructed pdf for each phoneme from a number of training data. The input data to be recognized were compared to the pdf's and the similarity measures were obtained for each phoneme. The CPR equation developed neuron activities by receiving these similarity measures as input. The recognition rate for 25 Japanese phonemes were 74.75 % in average which was compared to 71.53 %, by HMMs. A certain technical improvement was applied to our neuron model, by dividing data of a phoneme into two part, one for the former frames, the other for the latter frames. A remarkable improvement was obtained with average recognition rate of 79.79 %.

Competition and cooperation mechanism had an important role in raising the recognition rates. For the next task of the neuron model, we have to proceed to develop a system for continuous speech recognition. Although conventional techniques used for HMM's will be employed for the neuron model, we have a new situation characteristic of CPR equations. The most distinguished part of the model is the cooperation among neighboring frames. In the present paper, we treated frames as static coodinates. In the real time recognition, however, frame should be treated as time variable. Thus, we have two kinds of time variables. One is frame number and the other is the one from CPR equation. Therefore, we may have interesting situations where two time variables compete each other. It seems to resemble to a belt conveyorized system for automobile building, where there are two time variables competing each other. One is belt conveying time and the other is automobile building time. It seems that our goal is how to get a correct recognition at the end of belt by processing continuously coming speech data with cooperation and competition.

Kinds of data	Kinds of Model	Recognition rates
SD data	HMM	82.84
〃	N Model1	76.08
〃	N Model2	81.00
SI data	HMM	71.53
〃	N Model1	74.75
〃	N Model2	79.79

Table 2: The rates of speaker dependent(SD) and speaker independent(SI) recognition

Phoneme	HMM	N Model1	N Model2
NG	53.46	80.38	85.44
a	92.55	95.03	95.24
b	76.62	79.75	81.01
ch	84.62	86.15	87.69
d	69.84	73.44	84.38
e	64.77	85.98	89.77
f	64.29	46.67	46.77
g	57.14	49.35	63.64
h	63.46	51.92	59.62
i	69.16	85.71	92.21
j	97.01	94.03	97.01
k	55.25	70.13	69.70
m	61.90	47.17	54.72
n	44.30	42.50	46.25
o	70.58	92.18	95.27
p	64.00	45.24	42.86
r	62.34	44.26	55.19
s	89.01	85.32	85.71
sh	96.05	80.26	82.89
t	4.35	35.48	36.56
ts	65.22	86.96	86.96
u	94.78	48.82	67.06
w	84.38	63.64	75.76
y	61.36	70.45	65.91
z	87.76	87.76	85.71
ALL	71.53	74.75	79.79

Table 3: Recognition rates of test data

REFERNCES

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.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
5. Y.Yoshitomi, T.kanda, T.kitazoe "Neural Nets Pattern Recognition Equation for Stereo Vision" Trans.IPS.Japan:29-38, 1998
6. ATR Japanese Speech Database and Technical Report, Japan, 1988
7. ASJ Continuous Speech Corpus for Research, Japan, 1991