A Study on the Syllable Recognition Using Neural Network Predictive HMM

*Soo-Hoon Kim, **Sang-Berm Kim, ***Si-Young Koh, and *Kang-In Hur

※This Paper is made Possible by the Dong-A University Research Grant in 1996.

Abstract

In this paper, we compose neural network predictive HMM(NNPHMM) to provide the dynamic feature of the speech pattern for the HMM. The NNPHMM is the hybrid network of neural network and the HMM. The NNPHMM trained to predict the future vector, varies each time. It is used instead of the mean vector in the HMM. In the experiment, we compared the recognition abilities of the one hundred Korean syllables according to the variation of hidden layer, state number and prediction orders of the NNPHMM. The hidden layer of NNPHMM increased from 10 dimensions to 30 dimensions, the state number increased from 4 to 6 and the prediction orders increased from the second oder to the fourth order. The NNPHMM in the experiment is composed of multi-layer perceptron with one hidden layer and CMHMM. As a result of the experiment, the case of prediction order is the second, the average recognition rate increased 3.5% when the state number is changed from 4 to 5. The case of prediction order is the third, the recognition rate increased 4.0%, and the case of prediction order is fourth, the recognition rate increased 3.2%. But the recognition rate decreased when the state number is changed from 5 to 6.

I. Introduction

The man-machine interface through speech has benefits in that it is fast and can be performed without special training. The establishment of the speech recognition technology is becoming an important research subject because the computer and telecommunication technology are developing rapidly[3, 4].

One of the previous speech recognition methods, Dynamic Programming method allows nonlinear elasticity in the time horizon of the time series pattern. It can be used to deal with the variations of the time series pattern, but can't deal with the variation of the spectrum itself, from the difference in individual speakers[3, 7].

The neural network can express the variations of spectrum as the sum of weight among units, and cover the defects of HMM by dealing with numbers of frame data in one time. But it is hard to deal with nonlinear elasticity of time series pattern like speech because the network does not include the time elements in its structure and weight[7]. On the other hand, the HMM method recognizes speech by treating the variations of the speech pattern statistically, and providing the statistic to the probability model. This method has the advantage that it is easy to deal with the variations of speech patterns, due to the individual differences, co-articulation, etc. However this method has several disadvantages. First, it is hard to determine the model structure, second, it needs large amounts of data and calculation power[3]. To overcome these disadvantages, the current HMM study adds the regression coefficient and the duration time probability to the parameter in order to include dynamic feature[3, 4, 5, 6].

In this paper, we composed the NNPHMM to provide the dynamic feature of the speech pattern for HMM. The network is trained to predict the future vector based on several last feature vectors, and defined every state of the HMM. In the experiment, we used the 100 Korean syllables, and compared the recognition rate of the NNPHMM as we increased the hidden layer and the state number according to prediction order from the second to the fourth order. We also compared the results for CHMM.

II. HMM MODEL

2.1 Continuous Distribution HMM

Left-to-right type HMM can be defined as finite-state-automata is showed in figure 1.

In the case of speech recognition using HMM, we trained the models as many standard patterns as we could. Then we should take the standard pattern which has the largest output probability as a recognition result for input

* Department of Electronic Engineering, Dong-A University.
** Department of Information Technology, Textile Polytechnic College.
*** Department of Electronic Engineering, Kyungil University

Manuscript Received : April 13, 1998.

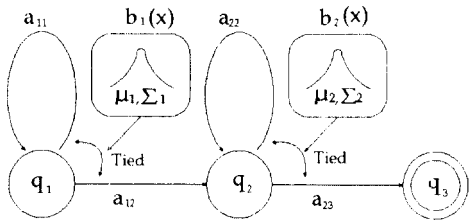


Figure 1. Continuous Distribution HMM.

pattern. In the continuous distribution HMM, the below Baum-Welch algorithm can be used to find, from the training data, the transition probability a_{ij}, from state i to j, and the output probability $b_{ij}(k)$ for symbol k in the transition process. If the number of states is N and the length of symbol series is T, the forward probability is $\alpha(i,t)\,(i=1,2,\cdots,N;t=1,2,\cdots,T)$, the back-ward probability is $\beta(j,t)\,(j=1,2,\cdots,N;t=1,2,\cdots,T)$ and $P(O|M)$ is the probability that model M generates symbol series $O=O_1O_2\cdots O_T$. The probability, which the transition from state i to state j happens at time t, can be defined as (1):

$$\gamma_t(i,j)=\frac{\alpha(i,t-1)\,a_{ij}b_{ij}(O_t,\mu_{ij},\sum_{ij})\beta(j,t)}{P(O|M)} \tag{1}$$

Thus, the estimation formula for the transition probability is the following (2) and (3):

$$a_{ij}=\sum_t \gamma_t(i,j)\,/\sum_t\sum_j \gamma_t(i,j) \tag{2}$$

$$b_{ij}(k)=\sum_{t,\,O_t=k} \gamma_t(i,j)\,/\sum_t \gamma_t(i,j) \tag{3}$$

If the output vector O_t follows the n-dimensional normal distribution, the output density function is as follows:

$$b_{ij}(O_t,\mu_{ij},\sum_{ij})=\frac{\exp\{-(O_t-\mu_{ij})'\sum_{ij}^{-1}(O_t-\mu_{ij})/2\}}{(2\pi)^{n/2}|\sum_{ij}|^{1/2}} \tag{4}$$

where μ_{ij} is the average of the output vector, $\sum_{ij}$ is the covariance matrix, and t is the transpose matrix, and -1 is the inverse matrix. The μ_{ij} and $\sum_{ij}$ can be estimated with the following (5) and (6):

$$\mu_{ij}=\sum_t \gamma_t(i,j)\,O_t\,/\sum_t \gamma_t(i,j) \tag{5}$$

$$\sum_{ij}=\frac{\sum_t \gamma_t(i,j)(O_t-\mu_{ij})(O_t-\mu_{ij})'}{\sum_t \gamma_t(i,j)} \tag{6}$$

2.2 Continuous Mixture HMM

In section 2.1, we explained how to estimate the parameters, given that the output probability density is continuously distributed. In general, the speech may not be approximated with only one normal distribution. In the transition process from i to j, output probability $b_{ij}(O_t)$ for vector O_t can be expressed as the sum of the weights of the M continuous distributions, as (7):

$$b_{ij}(O_t)=\sum_{m=1}^{M} \lambda_{ijm}b_{ijm}(O_t) \tag{7}$$

here

$$\sum_{m=1}^{M} \lambda_{ijm}=1,\qquad \int b_{ijm}(O)\,dO=1 \tag{8}$$

λ_{ijm} is the branch probability that the appearance probability of the m-th output probability density distribution, $b_{ijm}(O)$ is the m-th output probability density distribution. If the density distribution is normally distributed, $b_{ijm}(O)$ can be estimated with (9):

$$b_{ijm}(O)=b_{ij}(O_t,\mu_{ijm},\sum_{ijm}) \tag{9}$$

The transition probability can be estimated as in the formula (2). If $\gamma_t(i,j,m)$ is defined as (10)

$$\gamma_t(i,j,m)=\frac{\alpha(i,t-1)\,a_{ij}b_{ijm}(O_t)\beta(j,t)}{P(O|M)} \tag{10}$$

λ_{ijm}, μ_{ijm}, and $\sum_{ijm}$ are as follows

$$\lambda_{ijm}=\sum_t \gamma_t(i,j,m)\,/\sum_t\sum_m \gamma_t(i,j,m) \tag{11}$$

$$\mu_{ijm}=\sum_t \gamma_t(i,j,m)\,O_t\,/\sum_t \gamma_t(i,j,m) \tag{12}$$

$$\sum_{ijm}=\frac{\sum_t \gamma_t(i,j,m)(O_t-\mu_{ijm})(O_t-\mu_{ijm})'}{\sum_t \gamma_t(i,j,m)} \tag{13}$$

III. The Composition of NNPHMM

3.1 The NNPHMM

In the NNPHMM, the neural network is used to predict the future vector based on several last vectors, and defined each state of the HMM. Then the HMM trains the difference between the input vector y_t at time t and the output vector y_{it}' corresponding to each state i. This method uses the prediction value from the neural network, which is dynamically changing due to the effects of the prev-

ious feature vectors instead of the stable average vectors. That is, by using $b_{ij}(y_t, y_{it}', \Sigma_{ij})$ instead of $b_{ij}(y_t, \mu_{ij}, \Sigma_{ij})$, we can use (14) and (15) instead of (4) and (5), respectively, and re-estimate the variance with (16):

$$b_{ij}(y_t, y_{it}', \Sigma_{ij}) = \frac{\exp\{-(y_t - y_{it}')' \Sigma_{ij}^{-1}(y_t - y_{it}')/2\}}{(2\pi)^{n/2}|\Sigma_{ij}|^{1/2}} \quad (14)$$

$$\gamma_t(i,j) = \frac{\alpha(i,t-1)\, a_{ij} b_{ij}(y_t, y_{it}', \Sigma_{ij})\, \beta(j,t)}{p(y|M)} \quad (15)$$

$$\Sigma_{ij} = \frac{\sum_t \gamma_t(i,j)(y_t - y_t)(y_t - y_t)^t}{\sum_t \gamma_t(i,j)} \quad (16)$$

Because y_{it}' is estimated with the neural networks, the training is the combined process of both Baum-Welch algorithm and BP algorithm.

3.2 The Structure of NNPHMM

The NNPHMM is composed of the multi-layer perceptron and the continuous mixture distribution HMM. Figure 2 is one example of the NNPHMM, where $\alpha(i,t)$ is the forward probability at state i and time t, and $y(t)' = y(t) - \hat{y}_i(t)$

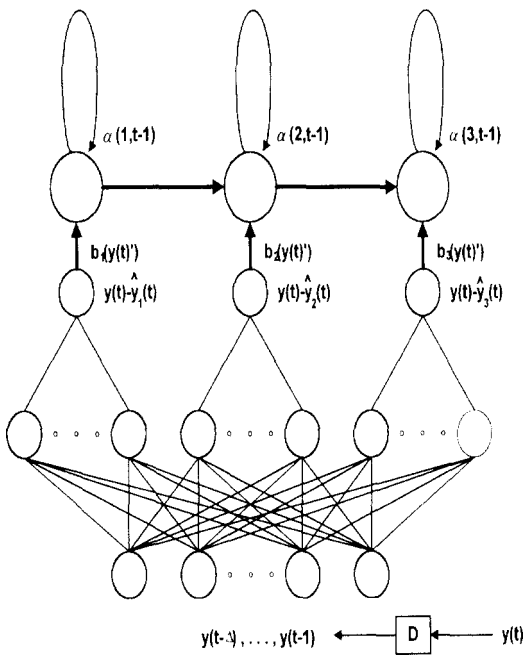


Figure 2. The NNPHMM.

The training process of the NNPHMM follows the sequence below.

① Finds the boundary of each state using previously trained HMM

② Train each neural network for the boundary derived from ①

③ Calculates the initial parameter of HMM from the prediction errors between the output of the neural network and the actual input vectors.

④ Repeats the process from ⑤ to ⑦ until the neural network have sufficient training.

⑤ Train the HMM parameter for the prediction errors.

⑥ Use the previously HMM parameter from the procedure ⑤ to calculate the boundary between each state.

⑦ Train the neural network for each state according to the boundary, generated from procedure ⑥

IV. The Recognition Experiment

4.1 The Speech DB and the Analysis Conditions

The speech analysis condition is summarized in table 1. Table 2 is the speech data of the 100 Korean syllables. The speeches are recorded 5 times by the 5 male speakers, age 20.

Table 1. The Condition of Speech Analysis.

A/D data	10 kHz, 12 Bit
High Frequency Emphasis	First Order Difference
Frame Period	5 ms
Analysis Window	Hamming Window
Window Length	20 ms
Feature Parameter	LPC Cepstrum(14th order) → LPC Melcepstrum (10th order)

Table 2. The Speech DB.

ga gan gal gam geo ge go gu gi
na nan nal nam ne no nu ni
da dan dal dam de do du di
la lan lam lo long lu li
ma man mal me mo mu mi
ba han bal bo bong bu bi
sa san sal sam se so su si
a an al am e ok u i
za zan zal zam ze zo zong zu zi
cha chan chal cham che cho chu chi
ka kil
ta tan tal tam to tu ti
pa pan pal po pi
ha han hal ham ho hu hi

4.2 The Result of the Experiments

In this experiment, we divided the 5 recorded speeches, spoken 5 times by the 5 male speakers into two groups; three were used for the training data and the remaining two were used for the test data. We used the three layer perceptron and did the training process 600 times using the BP algorithm. The HMM is continuous mixture HMM

with 3 mixture number. Maximum iteration number for reestimation was limited to 10 times.

Table 3 is the recognition rate according to the change of the prediction orders when the hidden layer is 10, 20, and 30 dimensions, respectively. Table 4 is the recognition rate according to the change of the prediction orders when the state number is 4, 5, and 6.

In the case of the prediction orders is the second order, the third order, and the fourth order, the input layer of the network is 20 units, 30 units, 40 units, respectively. The output layer is 10 dimensions.

Table 5 is the recognition rate of CHMM with 5 state number.

Table 3. The Recognition Rate According to The Variation of Hidden Layer.

Hidden Layer	Prediction Order	Recognition Rate(%)		
		Training	Test	Average
10 (Dimensions)	2nd	87.6	78.7	85.2
	3rd	90.0	79.1	85.6
	4th	90.1	79.3	85.8
20 (Dimensions)	2nd	89.7	78.5	85.2
	3rd	89.8	80.3	86.0
	4th	89.9	78.3	85.3
30 (Dimensions)	2nd	89.5	78.6	85.2
	3rd	89.7	79.0	85.4
	4th	90.0	79.6	85.9

Table 4. The Recognition Rate According to The Variation of State Number.

State Number	Prediction Order	Recognition Rate(%)		
		Training	Test	Average
4	2nd	87.1	73.5	81.7
	3rd	86.7	74.1	81.6
	4th	88.0	74.5	82.6
5	2nd	87.6	78.7	85.2
	3rd	90.0	79.1	85.6
	4th	90.1	79.3	85.8
6	2nd	87.3	74.0	82.0
	3rd	86.5	73.1	81.2
	4th	87.3	73.7	81.8

Table 5. The Recognition of CHMM.

Method	Recognition Rate(%)		
	Training	Test	Average
CHMM	98.0	85.6	91.8

In the result of experiment, initial sound of the vowels and /h/ caused the most frequent mis-recognition. These mis-recognized syllables occurred when final consonants were /n/, /l/, and /m/, similar. In the case that the initial sound is a vowel, when the prediction order increased from the second order to the fourth order, the recognition rates show the biggest improvement. On the other hand, the initial consonant /h/ shows no significant improvement in the recognition rate according to the changes of prediction orders.

V. Conclusion

In this paper, we composed the NNPHMM to provide the dynamic feature of the speech pattern for HMM. The network is trained to predict the future vector based on several last feature vectors, and defined every state of the HMM. In the experiment, we compared the recognition rate of the NNPHMM, by increasing the prediction orders from the second to the fourth order when the hidden layer is 10 dimensions. We also compared the results when the hidden layer increased from 10 to 30 dimensions. And we compared the recognition rate of NNPHMM by increasing the state number from 4 to 6 when the hidden layer is 10 dimensions. As a result of the experiment, the case of prediction order is the second, the average recognition rate increased 3.5% when the state number is changed from 4 to 5. The case of prediction order is the third, the recognition rate increased 4.0% and the case of prediction order is the fourth, the recognition rate increased 3.2%. But the recognition rate decreased when the state number is changed from 5 to 6.

When the state number is 4, the state number is too small to give the dynamic feature of speech pattern for the HMM.

When the state number is 6, it seems to be reduced the classification ability of neural network for speech patterns. Because, the output units of neural network increased according to the increase of state number.

As the prediction order increases, the variation of the recognition rate is less then 1% because the syllable data is too short to predict the future vector by NNPHMM sufficiently. So we will investigate application for word data or continuous speech data.

The recognition rate of CHMM is 91.8%. The result is better than the result of NNPHMM but we expect much more favorite results by finding the optimal model through a trial and error method and considering the effects according to the data. We will continue to experiment with other neural network structures and HMM models. The NNPHMM, combined with the HMM and the neural network, is becoming one of the most effective methods.

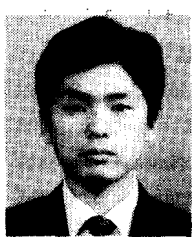
Reference

1. Nile L.T. and Silverman H.F.: "Combining Hidden Markov Model and Neural Network Classifiers", *Proc. ICASSP*, pp. 417-420 (1990)
2. E. Tsuboka, Y. Takada and H. Wakita: "Neural Predictive Hidden Markov Model", Proc. ICSLP-90, pp. 1341-1344 (1990)
3. S.H.Kim, J.J.Lee, and K.I.Hur: "Korean Continuous Speech Recognition Using Discrete Duration Control Continuous HMM", *The Journal of the Acoustical Society of Korea*, Vol.14 No.1 pp. 81-89 (1995)
4. S.H.Kim, J.Y.Sim, Y.J.Lee, S.Y.Ko, and K.I.Hur: "Syllable Recognition Using Neural Predictive HMM", Proc. *Korean Signal Processing Conf.* Part1. pp.239-242 (1996)
5. S.H.Kim, S.B.Kim, and K.I.Hur: "The Recognition of Korean Syllables Using Neural Predictive HMM", *Proc. ICSP-97*, pp.427-431 (1997)
6. S.H.Kim, C.K.Kim, S.B.Kim, and K.I.Hur: "A Comparative Study of NNPHMM According to The Variation of State Number", The Journal of Institute of Data Communication Dong-A University Vol.5 No.1 pp.129-135 (1997)
7. S.H.Kim, J.S.Kim, K.I.Hur: "A Study on The Phoneme Recognition Using Radial Basis Function Network" The Journal of the Korean Institute of Communication Sciences, Vol 22, No5. pp.1026-1035 (1997)

▲Soo Hoon Kim

Soo Hoon Kim was born in Pusan, Korea, on February 25, 1968. He received the B.S. and M.S. degrees from Dong-A University, Pusan, Korea, in 1991 and 1993, respectively. Since 1993 he is a Ph.D. Candidate in electronic engineering at Dong-A university. His current research interests are digital signal processing, speech recognition and synthesis, neural network.

▲Sang Berm Kim



Sang Berm Kim was born in 1969 in Pusan, Korea. He received the B.S. and M.S. electronic engineering from University of Dong-A in 1992, and 1994 respectively. He has been in the course of Ph.D. in Dong-A University. Since Mar, 1997, he has been with the department of information technology, textile polytechnic college, Daegu, as a instructor. His current interests are speech recognition, neural network.

▲Si-Young Koh

Si-Young Koh was born in Taegu, Korea, on August 16, 1952. He received the B.C. and M.S. degree in the department of electronic engineering from Yeungnam university, in 1977 and 1982, respectively, and the Ph.D. of engineering degree from Dong-A university, in 1922. Since 1986 he has been with Kyungil university, where he is a professor in the department of electronic engineering. His research interests are speech analysis, speech recognition and digital signal processing.

▲Kang-In Hur

Kang-In Hur was born in Pusan, Korea. He received the B.S. and M.S. degrees from Dong-A University, Pusan, Korea, in 1980 and 1982, respectively, and the Ph.D. degrees in Electronic Engineering from the Kyunghee University, Seoul, Korea, in 1990. Since 1984 he is a professor Dept. of Electronic Engineering at Dong-A University. During 1988 to 1989, he joined Dept. of Information and Computer Sciences, Tsukuba University, Japan as a visiting scientist. From 1992 to 1993, he joined Dept. of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, Japan as Post-Doctral researcher. His current research interests are digital signal processing, speech recognition and synthesis, neural network.