3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SIMPLIFIED SUB-NEURAL-NETWORKS FOR ACCURATE PHONEME RECOGNITION

Takuya Koizumi, Shuji Taniguchi, Ken-ichi Hattori,
and Mikio Mori

Dept. of Information Science, Fukui University
3-9-1 Bunkyo, Fukui 910, Japan

ABSTRACT

This paper deals with a new phoneme recognition system comprising a set of a simplified version of the sub-neural network and the hidden Markov models (HMMs). The ability of this system has been investigated by a phoneme recognition experiment using a number of Japanese words uttered by a native male speaker in a quiet environment. The result of the experiment shows that recognition rates achieved with this system is higher than those attained with the LVQ-HMM system which is known to be one of the most reliable phoneme recognizers.

I. INTRODUCTION

There seem to be two conflicting objectives in the pursuit of highly reliable phoneme recognizer to be used as a front end of continuous speech recognition system. One is speaker independence and the other is, as a matter of fact, recognition accuracy. One of the most reliable phoneme recognizers which are capable of achieving to a degree both objectives is known to be the learning vector quantizer-hidden Markov model (LVQ-HMM) system [1].

Recently, a simple feed-forward neural network called sub-neural-network (SNN) has been proposed as an attractive tool for speech recognition by Liu et al. [2]. This network can considerably reduce computational time and expense in training compared with other types of network. A simplified version of this network, which will be designated as a simplified sub-neural-network (SSN), has been found to be more reliable than the original SNN. It has also been found that a set of the SSNs combined with the discrete HMMs to form a phoneme recognizer surpasses the LVQ-HMM or self-organizing feature map (SFM)-HMM system [3] in phoneme discriminating power. The ability of this SSN-HMM system has been investigated by a phoneme recognition experiment using a number of Japanese words uttered by a native male speaker in a quiet environment. The result of the experiment shows that recognition rates achieved with this system is higher than those attained with the LVQ-HMM and SFM-HMM systems.

In what follows, first the SSN and SSN-HMM system will be described in detail, then the result of the phoneme recognition experiment will be discussed, comparing the performance of the system with that of other systems including the SNN-HMM, SFM-HMM, and LVQ-HMM systems.

II. THE SSN-HMM SYSTEM

2.1 The structure of the SSN

Suppose we want to have a phoneme recognizer for discriminating voiced plosive sounds, /g/, /d/, and /b/. It can be realized using three simplified sub-neural-networks (SSNs) to discriminate those sounds and a decision layer, as illustrated in Fig.1.

Each of those networks has three layers, an input layer having as many units as the dimensionality of input vectors, a hidden layer with a few units, and an output layer having only a single unit. One of those networks can be used to discriminate a particular phoneme, say, /g/, from other phonemes, /d/ and /b/, belonging to the same group of voiced plosive sounds. Two similar networks can be formed also for discriminating in favor of phonemes of different categories in the same group. The output of each network is fed into the decision layer where a network which produces the largest output is chosen and the phoneme category represented by that network is put out as a recognized result. A set of networks and a decision layer as described above are necessary for each of phoneme groups to form a complete phoneme recognizer.

The SSN is characterized by its simple structure, reduced computational time in training, and high phoneme discriminating power compared with other types of network such as the SNN, conventional multi-layer perceptron (MLP), and Kohonen map. The high phoneme discriminating power of the SSN is primarily due to the fact that it has a very simple structure having only a single output unit. It has another advantage that adding a new phoneme category to a group of phoneme categories to be classified requires neither restructuring nor retraining of the entire phoneme recognizer but only the training of an added network for the new category and the alteration of the decision layer unlike the MLP which would need a complete restructuring and retraining.

2.2 The description of the SSN-HMM system

The SSN itself has a higher phoneme discriminating power than the conventional MLP, SNN, and Kohonen map. The output of a set of SSNs for a sequence of input vectors representing a class of phoneme can be transformed into a sequence of codes in the decision layer. If the discrete hidden Markov models (HMMs) are used to decode such sequences of codes into a string of recognized phonemes, then an SSN-HMM system as depicted in Fig.2 will result.

As mentioned earlier, this system is an excellent phoneme recognizer whose performance surpasses that

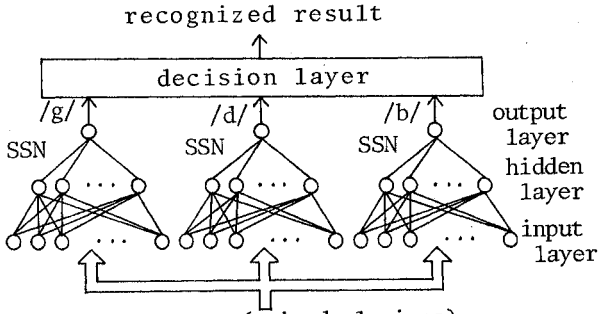


Fig.1. The structure of a phoneme recognizer comprising three SSNs for discriminating voiced plosives, /g/, /d/, and /b/ and a decision layer.

of other phoneme recognition systems like the LVQ-HMM system. In addition, it allows us to considerably save computational time and expense in training. To form a complete phoneme recognizer, however, one would need a phoneme group classifier which does classify a speech input into some prescribed different phoneme groups, e.g., unvoiced plosives, voiced plosives, fricatives, affricates, nasals, semivowels and glides, and vowels. This classifier can be realized using, for example, recurrent neural networks [4]. It would be possible to realize a complete phoneme recognizer with as many SSNs and HMMs as the number of phoneme categories, which classifies speech input into different phoneme categories in one step. Its performance, however, would be poorer than that of the phoneme recognizer which is made up of a cascade of the phoneme group classifier and SSN-HMM systems for discriminating phonemes in each phoneme group. Therefore, we will be concerned with the latter phoneme recognizer and bring into focus mainly the SSN-HMM systems in the following. The phoneme group classifier will be discussed somewhere else. Before going into the detail of phoneme recognition experiment, the discrete HMMs will be described briefly.

2.3 The discrete HMMs

In this work left-to-right three state models as depicted in Fig.3 were used. a_{ij} and b_{jk} represent the probability of transiting to state S_j given current state S_i and the probability of observing code k given current state S_j, respectively.

Parameters of the HMMs were estimated using the Baum-Welch algorithm. The parameters were readjusted lest the value of each parameter should become less than a very small number. This prevented the probability of each model from becoming zero, when a sufficiently large number of training vectors were not available.

III. PHONEME RECOGNITION EXPERIMENT

To compare the phoneme discriminating power of the SSN and SSN-HMM systems with that of other systems, a phoneme recognition experiment was performed on the proposed systems and some other phoneme recognition systems. The method and result of the experiment will be briefly described below.

3.1 Phoneme data and the method of analysis

A set of phoneme tokens was derived from a Japanese word database provided by ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. This database contains 5,240 Japanese words uttered by a single male speaker in a quiet environment, which were sampled at 20kHz and digitized.

Figure 4 illustrates how phoneme vectors are generated from speech waveform. By way of taking account of the spectral variations of phoneme tokens with time, each of which is represented as 15 frames of power spectra, a seven-frame (7×25.6ms) time window is shifted over those 15 frames, one frame at a time. This window produces a 112-dimensional phoneme vector at each position and 9 such phoneme vectors altogether for each phoneme token, which will be used as input vectors for the systems. Thus the number of sequences of codes which will be applied to the HMMs is also 9 for each phoneme token. Here the power spectrum of a phoneme token is defined as the average power of each of outputs of 16 bandpass filters with an equal bandwidth of 1.1 Bark in the range between 200Hz and 6,000Hz, each of which is fed with the phoneme token as input.

Those 9 phoneme vectors derived from a phoneme token will be called a phoneme sample. Among those phoneme samples odd-numbered samples were used as training data for the SSNs and HMMs and even-numbered ones as test data for evaluating the performance of the systems in order to have those samples equally divided between the training and test data for each class of succeeding vowels.

3.2 The number of hidden layer units

The number of hidden layer units is an important factor of neural networks like the SSN and MLP, since it determines not only recognition accuracy but also computation time in training of those networks. In order to determine an appropriate number of hidden layer units a recognition experiment was performed for

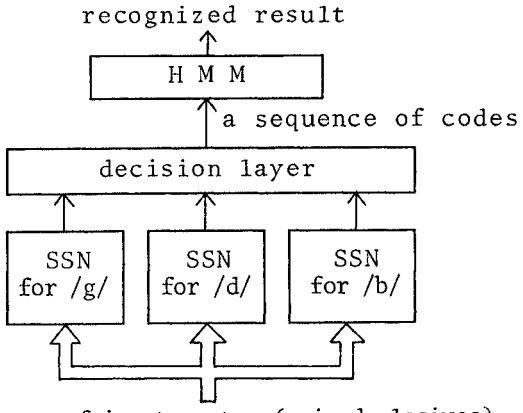
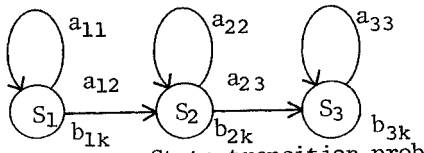


Fig.2. An SSN-HMM system for discriminating voiced plosive sounds.



S_i : State a_{ij} : State transition prob.
b_{jk} : Code output prob.

Fig.3. The left-to-right three state Markov model.

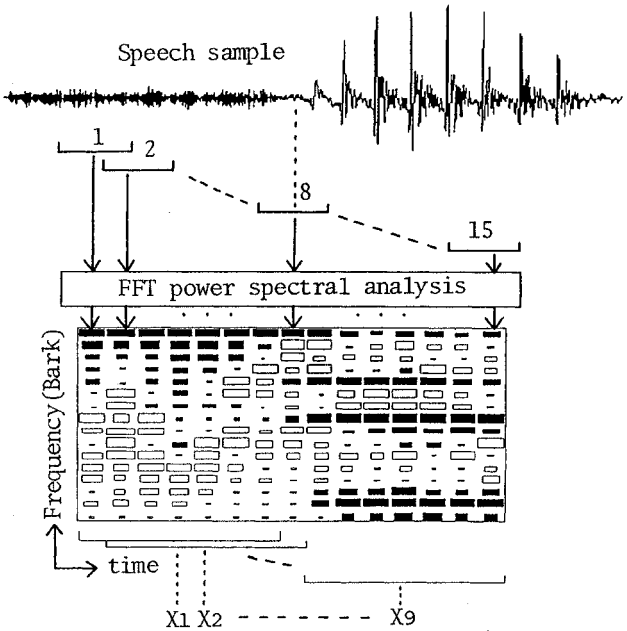


Fig.4. The generation of phoneme vectors.

6 groups of different phonemes, i.e., unvoiced plosives, voiced plosives, fricatives, affricates, nasals, and semivowels and glides, by varying the number of hidden layer units of the SSN, the result of which is shown in Table I.

This result clearly indicates that the SSN whose hidden layer has 3 or 4 units achieves recognition accuracies as high as those of the MLP that has a hidden layer with 13 units. Figure 5 shows computation time in training averaged over different phonemes of the SSN and SNN relative to that of the MLP which has a hidden layer consisting of 13 units as a function of the number of hidden layer units. This indicates that the SSN enables us to considerably save computational time in training compared with the MLP.

3.3 Corrective recognition

Corrective recognition is a way of improving the reliability of phoneme recognizers by introducing the principle of decision by majority into recognition strategy. Suppose we have a sequence of 9 recognized phoneme vectors for an unknown phoneme token as follows:

/k/, /k/, /t/, /k/, ?, /k/, /k/, /p/, /k/.

Here ? denotes a phoneme vector that was not recognizable. Five recognized phoneme vectors out of nine are /k/, thus the application of the "decision by majority" to this sequence yields /k/ as a correct category of the unknown phoneme token. This corrective recognition is applicable to the present recognition systems, since each of phoneme tokens is represented as 9 phoneme vectors as mentioned earlier.

3.4 The SSN and SNN systems

Recognition rates of the SSN and SNN systems and MLP with and without corrective recognition are compared in Table II for all phoneme categories.

3.5 Comparison of several systems

Since HMMs have a memory in the sense that they respond to a particular sequence of codes in the form of time series, they are capable of improving the discriminating power of neural networks when they are combined with neural networks. For example, combining the HMMs with the Kohonen map will result in the SFM-HMM and LVQ-HMM systems which are known to have a high phoneme discriminating power.

Table III compares recognition rates (%) of several systems including the SSN-HMM, SNN-HMM, SFM-HMM, and LVQ-HMM systems, each of which consists of particular neural networks and HMMs.

Table III clearly indicates that the performance of the SSN-HMM system surpasses that of other systems and the average phoneme recognition rate achieved with this system is higher than 97 per cent in case where the number of phoneme categories in one group does not exceed five.

3.6 The effect of adding a new category

Adding a new phoneme category to the SSN system requires neither restructuring nor retraining of all the networks in the system but only the training of an added network for the new category and the alteration of the decision layer. Thus adding a new phoneme category seems to have little effect on the performance of the system.

For example, in a vowel recognition experiment carried out with the SSN system average recognition rate turned out to be 95.8 per cent for four Japanese vowels, /a/, /e/, /i/, and /u/. This rate remained nearly unchanged (95.4 per cent) when a new vowel /o/ was added to the set of vowels, i.e., when a new network tuned to /o/ was added to the system and a necessary alteration of the decision layer was made.

IV. CONCLUSIONS

Two new phoneme recognition systems, the SSN and SSN-HMM systems have been described above. For the purpose of evaluating the ability of these systems a phoneme recognition experiment was performed using a number of phoneme tokens derived from a Japanese word database. The result of the experiment can be summarized as follows:

(1) The SSN enables us to considerably save computation time and expense in training since it has a very simple structure with fewer hidden and output layer units than the conventional MLP and SNN.

(2) Adding a new phoneme category requires neither restructuring nor retraining of all the networks in the SSN system unlike the MLP. In addition, adding a new category has very little effect on the phoneme discriminating power of the system.

(3) The average recognition rate for consonants of the SNN system is 3.4% higher than that of the MLP. With the SSN system it is possible to attain a 1.5% higher average recognition rate than the SNN system.

(4) The phoneme discriminating power of the SSN-HMM system surpasses that of other phoneme recognizers such as the SNN, SSN, SNN-HMM, SFM-HMM, and LVQ-HMM systems. The average recognition rate for consonants attained with the SSN-HMM system is as high as 97.6% and 1.7% higher than that of the LVQ-HMM system in case where the number of phoneme categories in one group does not exceed five.

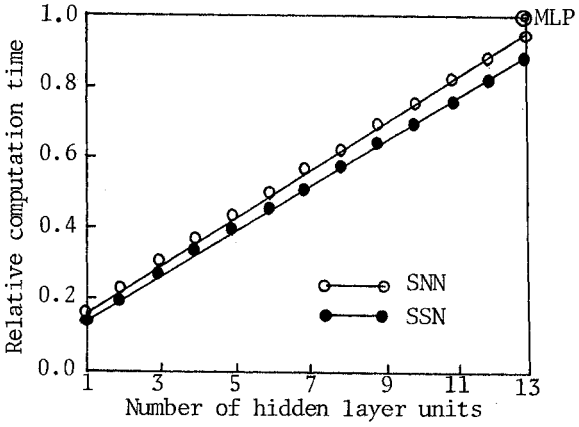


Fig.5. Computation time in training averaged over different phonemes of the SSN and SNN relative to that of the MLP which has a hidden layer consisting of 13 units versus the number of hidden layer units.

Table I. Recognition rate (%) versus the number of hidden layer units.

no. of units	SSN							MLP
phoneme	1	2	3	4	5	6	13	13
/k, t, p/	78.2	83.0	84.7	86.0	86.5	87.6	90.9	79.7
/g, d, b/	82.9	83.8	85.8	88.0	86.9	87.5	87.8	87.5
/s, sh, h, dz/	97.7	98.6	98.9	99.0	99.1	98.8	99.0	99.0
/ch, ts/	98.0	98.8	99.1	98.9	98.9	99.1	99.3	99.3
/m, n, N/	83.2	82.3	81.4	82.7	85.9	82.0	89.2	81.1
/y, r, w/	91.7	94.0	95.2	94.5	94.9	96.0	96.2	94.6

Table II. Recognition rates (%) of the SSN and SNN systems with and without corrective recognition. (SSNc and SNNc denote the SSN and SNN systems with corrective recognition.)

phoneme		SSNc		SSN		SNNc		SNN	
category	number	recognition rate	ave.	recognition rate	ave.	recognition rate	ave.	recognition rate	ave.
k	1150	98.2		92.5		99.0		94.4	
t	430	97.2	97.1	89.5	90.9	87.4	95.6	76.5	89.1
p	24	45.8		36.6		75.0		60.6	
g	250	94.4		89.8		94.4		88.9	
d	190	96.6	93.5	92.0	87.8	97.4	92.6	90.9	86.4
b	220	89.5		82.0		86.4		79.8	
s	500	100.0		99.3		99.8		98.9	
sh	310	100.0	99.8	99.2	99.0	100.0	99.8	97.9	98.6
h	210	100.0		99.2		100.0		98.9	
z	300	99.0		97.9		99.7		98.6	
ch	125	100.0	100.0	99.5	99.3	100.0	100.0	99.6	99.3
ts	190	100.0		99.2		100.0		99.1	
n	260	86.9		79.3		92.7		85.3	
m	470	96.8	95.1	89.8	89.2	84.3	91.2	77.5	86.1
N	490	97.8		93.9		97.1		94.7	
y	165	99.4		96.0		96.4		87.8	
r	730	98.5	98.5	96.3	96.2	98.9	98.1	97.7	95.3
w	75	96.0		95.4		94.7		88.9	
a	600	100.0		97.7		100.0		96.8	
e	600	98.5		95.9		99.2		94.6	
i	600	99.3	98.4	95.9	95.9	98.2	98.4	95.6	95.1
o	600	98.3		97.1		98.3		94.4	
u	600	95.7		93.1		96.5		93.9	

Table III. Recognition rates (%) of several systems including the SSN-HMM, SNN-HMM, SFM-HMM, and LVQ-HMM systems.

phoneme	SNN-HMM		SSN-HMM		SFM-HMM		LVQ-HMM	
	recognition rate	ave.	recognition rate	ave.	recognition rate	ave.	recognition rate	ave.
k	98.3		98.6		95.1		98.4	
t	97.2	97.0	98.6	97.3	93.8	94.0	96.5	97.4
p	29.2		8.3		50.0		66.7	
g	95.2		95.6		81.6		87.5	
d	97.4	93.9	97.4	94.4	93.7	86.7	92.6	89.1
b	89.6		90.5		85.6		87.8	
s	100.0		100.0		98.6		99.0	
sh	100.0	99.9	100.0	99.9	98.7	97.6	99.7	98.4
h	100.0		100.0		92.5		94.8	
z	99.7		99.3		98.3		98.4	
ch	100.0	100.0	100.0	100.0	91.4	93.5	93.8	96.0
ts	100.0		100.0		94.9		97.4	
n	93.5		93.9		87.4		93.5	
m	90.4	92.5	94.5	95.7	92.0	91.5	94.3	94.3
N	94.1		97.8		93.0		94.6	
y	99.4		98.8		98.8		100.0	
r	97.8	98.1	98.5	98.7	93.4	94.7	95.4	96.5
w	98.7		100.0		98.7		100.0	
average	96.7		97.6		93.6		95.9	

REFERENCES

[1] T.Koizumi, J.Urata, and S.Taniguchi. "The Effect of Information Feedback on the Performance of a Phoneme Recognizer using Kohonen Map," Proc. ICSLP-92, pp.1363-1366, 1992.

[2] F.Liu, J.Jiang, J.Cheng, and K.Yi. " A Neural Network Based on Subnets－SNN," Proc. ICSLP-92, pp.1459-1462, 1992.

[3] T.Koizumi, J.Urata, and S.Taniguchi. "A Phoneme Recognition Using Self-Organizing Feature Map and Hidden Markov Models," Proc. ICANN-91, pp.777-782, 1991.

[4] T.Koizumi, S.Taniguchi, H.Ishida, and M.Mori. "Recurrent Neural Networks for Phoneme Recognition," Tech. Report of the Institute of Electronics, Information, and Communication Engineers, SP94-1, pp.1-8, 1994.