

RECURRENT NEURAL NETWORKS FOR PHONEME RECOGNITION

Takuya Koizumi, Mikio Mori, Shuji Taniguchi, and Mitsutoshi Maruya

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

ABSTRACT

This paper deals with recurrent neural networks of multi-layer perceptron type which are well-suited for speech recognition, specially for phoneme recognition. The ability of these networks has been investigated by phoneme recognition experiments using a number of Japanese words uttered by a native male speaker in a quiet environment. Results of the experiments show that recognition rates achieved with these networks are higher than those obtained with conventional non-recurrent neural networks.

1 INTRODUCTION

The conventional neural networks of multi-layer perceptron (MLP) type have been increasingly in use for speech recognition and also for other speech processing applications. Those networks work very well as an effective classifier for vowel sounds with stationary spectra, while their phoneme discriminating power deteriorates considerably for consonants which are characterized by variations of their short-time spectra. This may be attributable to a fact that feedforward multi-layer neural networks are inherently unable to deal with time-varying information like time-varying spectra of speech sounds. One way to cope with this problem is to incorporate feedback structure in the networks to provide them with an ability to memorize incoming time-varying informations. Incorporating feedback structure in feedforward networks results in so-called recurrent neural networks (RNNs) which have feedback connections between units of different layers or connections of self-loop type.

Recently a simple recurrent network [1], which has feedback connections of self-loop type around hidden layer units, has been proposed as an attractive tool for recognizing speech sounds including voiced plosive sounds. This network will be called Type 2 RNN.

A different recurrent network [2] which has feedback connections around output layer units has been found to surpass this Type 2 RNN and conventional feedforward networks in

phoneme discriminating power. This network will be designated as Type 1 RNN. The ability of the network has been investigated by phoneme recognition experiments using a number of Japanese words uttered by a native male speaker in a quiet environment. Results of the experiments show that recognition rates achieved with a single or a set of Type 1 RNNs are higher than those obtained with a single or a set of Type 2 RNNs or ordinary feedforward networks.

In what follows, both recurrent networks will be described in detail, then the results of the experiments will be discussed, comparing the performance of the Type 1 RNN with that of the Type 2 RNN and feedforward network.

2 THE TWO RECURRENT NEURAL NETWORKS

2.1 The structure of the Type 1 and 2 RNNs

The Type 1 recurrent neural network has three layers, input, hidden, and output layers. Each of the output layer units has a feedback connection with itself, i.e., a self-loop, as shown in Fig.1. The output of each input layer unit at time $t-1$ is fed, through connections between the input and hidden layers, to all the hidden layer units at time t and in the same manner the output of each hidden layer unit at time $t-1$ is supplied, through connections between the hidden and output layers, to all the output layer units at time t . The output at time $t-1$ of each output layer unit is fed back to itself at time t . The Type 2 recurrent neural network has a similar structure as that of the Type 1 network except for the fact that it has a connection of self-loop type around each of the hidden layer units, as depicted in Fig.2.

2.2 The training of the networks

In training the Type 1 RNN, weights at t of all the connections between the input and hidden layers as well as connections between the hidden and output layers are affected by all the input vectors to the input layer before time t , while in the Type 2 RNN input vectors prior to time t have no effect upon the weights of connections between the hidden

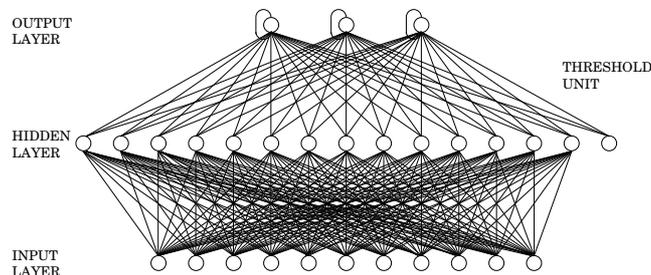


Figure 1: The structure of the Type 1 RNN.

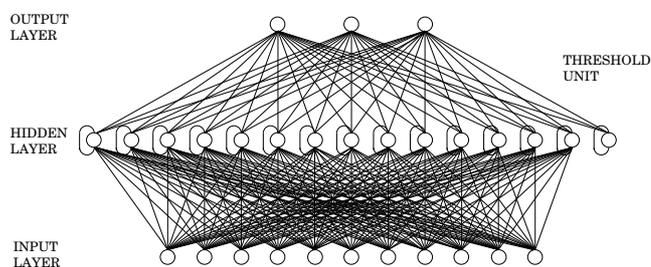


Figure 2: The structure of the Type 2 RNN.

and output layers before time t . This difference in the effect of training between both RNNs will certainly lead to a difference in phoneme discriminating power. We will see this shortly in the results of the experiments.

3 PHONEME RECOGNITION EXPERIMENTS

To compare the phoneme discriminating power of Type 1 RNN with that of the Type 2 RNN and ordinary feed-forward network (MLNN), the phoneme recognition experiments were performed on these networks. The method and results of the experiments 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. Those words were sampled at 20kHz and digitized.

By way of representing each phoneme token as 11 frames of waveform data, a 25.6ms time window is shifted over each phoneme token, 10ms at a time, producing 1, 7, and 3 frames of waveform data for transient, consonant, and vowel portions of token, respectively, for consonants, and 11 frames for a stationary portion of token for vowels.

Those 11 frame phoneme data derived from phoneme tokens will be called phoneme samples. Those phoneme samples were equally divided into samples for training the RNNs and samples for evaluating the performance of the RNNs with an equal frequency of having each class of succeeding vowels in the training and test samples.

In order to take account of spectral variations of phoneme samples, running spectra were calculated for phoneme samples. Figure 3 illustrates how running spectral vectors, which will be called phoneme vectors, are generated from speech waveform. First, a short-time power spectrum is calculated using the FFT for each frame of phoneme samples. Here the short-time spectrum is defined as the average output power of each of 16 bandpass filters with an equal bandwidth of 1.1 Bark in the range between 200Hz and 6kHz, each of which is fed with the phoneme samples as input. A seven-frame (7x25.6ms) time window is shifted over those 11 frames of short-time power spectra, one frame at a time. This produces a 112-dimensional running spectral vector (phoneme vector) at each position and five such phoneme vectors altogether for each phoneme sample, which will be used as input vectors for the RNNs.

3.2 The number of hidden layer units

The number of hidden layer units is an important factor of multi-layer neural networks, since it determines not only recognition accuracy but also computation time in training those networks. A preliminary phoneme recognition experiment was performed for 33 different phonemes of Japanese, varying the number of hidden layer units, to find out that an appropriate number of hidden layer units is 40. The numbers of input and output layer units are set equal to the dimensionality of input vectors and the number of phoneme categories to be classified, respectively. The back-propagation algorithm was utilized to train the RNNs.

3.3 Phoneme recognition using a single RNN

A phoneme recognition experiment was carried out to evaluate the phoneme discriminating power of each of the RNNs and MLNN. Table 1 shows a result of the experiment and

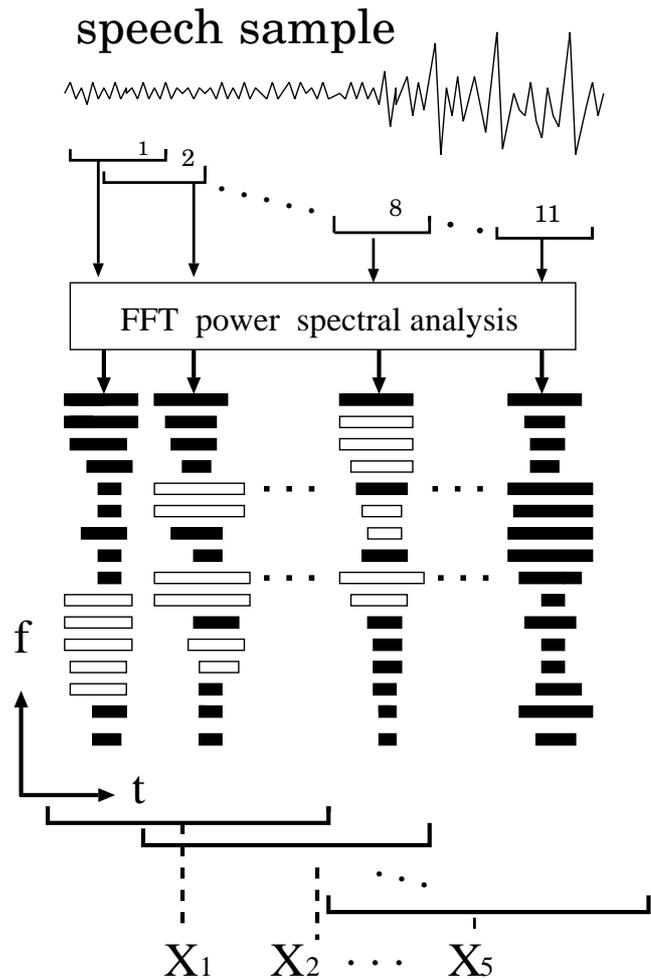


Figure 3: The generation of phoneme vectors.

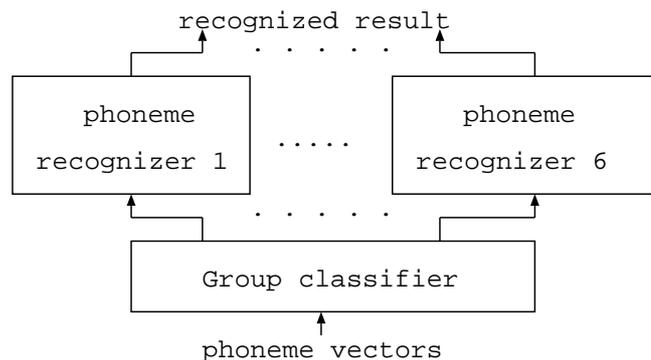


Figure 4: The group classification scheme.

Table 1: Recognition rates(%) of the recognizer using a single RNN.

| phonemes | Type1 | Type2 | MLNN | phonemes | Type1 | Type2 | MLNN | phonemes | Type1 | Type2 | MLNN |
|----------|-------|-------|------|----------|-------|-------|------|----------|-------|-------|------|
| /k/ | 78.3 | 60.5 | 48.2 | /h/ | 90.6 | 70.8 | 67.5 | /N/ | 87.1 | 72.5 | 60.6 |
| /ky/ | 94.0 | 46.0 | 34.0 | /hy/ | 44.4 | 66.7 | 11.1 | /ng/ | 86.7 | 72.3 | 70.7 |
| /t/ | 87.0 | 69.4 | 65.3 | /f/ | 94.4 | 94.4 | 88.9 | /y/ | 75.3 | 66.3 | 60.2 |
| /p/ | 66.7 | 54.2 | 33.3 | /z/ | 97.4 | 95.7 | 83.5 | /w/ | 97.3 | 96.0 | 93.3 |
| /g/ | 86.6 | 91.0 | 74.6 | /j/ | 87.8 | 63.0 | 59.8 | /r/ | 75.1 | 67.8 | 50.5 |
| /gy/ | 0.0 | 0.0 | 33.3 | /ch/ | 64.1 | 52.3 | 35.9 | /ry/ | 84.6 | 76.9 | 53.8 |
| /d/ | 89.5 | 80.5 | 86.8 | /ts/ | 94.8 | 98.5 | 74.7 | /a/ | 99.5 | 99.0 | 97.5 |
| /b/ | 71.2 | 54.1 | 49.1 | /m/ | 81.3 | 67.2 | 58.8 | /i/ | 86.5 | 81.7 | 78.8 |
| /by/ | 25.0 | 100. | 50.0 | /my/ | 25.0 | 75.0 | 100. | /u/ | 81.8 | 80.0 | 74.5 |
| /s/ | 78.3 | 69.2 | 76.1 | /n/ | 77.1 | 68.7 | 63.0 | /e/ | 98.2 | 94.5 | 95.3 |
| /sh/ | 78.3 | 55.3 | 47.3 | /ny/ | 83.3 | 66.7 | 16.7 | /o/ | 96.7 | 94.3 | 90.5 |
| | | | | | | | | average | 84.9 | 75.1 | 68.5 |

Table 2: Phoneme groups.

| groups | phonemes |
|----------------------------|---------------------------------------|
| unvoiced plosives | /k/, /ky/, /t/, /p/ |
| voiced plosives | /g/, /gy/, /d/, /b/, /by/ |
| unvoiced fricatives | /s/, /sh/, /h/, /hy/, /f/, /ch/, /ts/ |
| voiced fricatives + glides | /z/, /j/, /y/, /w/, /r/, /ry/ |
| nasals | /m/, /my/, /n/, /ny/, /N/ |
| | /ng/ |
| vowels | /a/, /i/, /u/, /e/, /o/ |

compares recognition rates (%) of the Type 1 and 2 RNNs with those of non-recurrent MLNN. Table 1 clearly indicates that the performance of the Type 1 RNN surpasses that of the other networks. In particular, the average recognition rate achieved with the Type 1 RNN is 16.4% higher than that of the MLNN. This result may suggest that the self-loops within the network function as memories to memorize time-varying spectra of speech sounds and improve appreciably the phoneme discriminating power of the network.

3.4 Phoneme recognition using group classification scheme

When only a single RNN is used to recognize all the phonemes, computation time increases and recognition rate decreases with the number of phoneme categories. One way to cope with this problem is to divide the entire phoneme categories into several, say 6, phoneme groups by a group classifier RNN and then to recognize phonemes in each group by a phoneme recognizer RNN. We call this way of recognizing phonemes in two steps a group classification scheme. A flow diagram of this scheme is shown in Fig.4. Input phoneme vectors are classified into 6 phoneme groups by the group classifier RNN. This group classification is similar to the phoneme classification in phonetics, but is slightly modified, as one sees in Table 2. Phonemes are classified into the following 6 groups: unvoiced plosives, voiced plosives, unvoiced fricatives, voiced fricatives+glides, nasals, and vowels. Suppose an unknown phoneme vector was classified as one of unvoiced fricatives. The vector would be sent to a specific phoneme recognizer for this particular group and recognized result would be put out. One advantage of this scheme is that we can use the right network to obtain the best recognition result. The group classifier RNN has 112, 40, and 6 units in the input, hidden, and output layers, respectively, and the phoneme recognizer RNNs 112, 30, and 4 ~ 7 units

in the input, hidden, and output layers, respectively. Tables 3 and 4 show group classification accuracies and intra-group phoneme recognition rates, respectively, of Type 1 and 2 RNNs and MLNN. Obviously higher average accuracies are attained in the order of the MLNN, Type 2 RNN, and Type 1 RNN, both in the group classification and in the intra-group phoneme recognition. Table 5 compares overall recognition rates obtained through the use of the group classification scheme with recognition rates obtained by a single RNN. This result clearly indicates that the group classification scheme brings about an improvement of 3.2% in average recognition rate over the recognizer using a single Type 1 RNN.

4 CONCLUSIONS

Two new phoneme recognizers using a single or several recurrent neural networks have been described above. For the purpose of evaluating the ability of these recognizers phoneme recognition experiments were performed using a number of phoneme tokens derived from a Japanese word database. Findings from the results of the experiments can be summarized as follows:

1. The Type 1 and 2 RNNs surpass the MLNN in phoneme recognition accuracies. This may suggest that the feedback loops within the RNNs function as memories to memorize time-varying spectra of speech sounds and improve appreciably the phoneme discriminating power of the networks. The Type 1 RNN surpasses the Type 2 RNN in phoneme discriminating power. This may be due to the fact that training affects weights of all the connections in the Type 1 RNN, while it affects only part of connections in the Type 2 RNN.
2. The average recognition rate of the group classification scheme is 3.2% higher than that of the recognizer using a single Type 1 RNN.

REFERENCES

- [1] Watrous, R.L., Ladendorf, B., and Kuhn, G. "Complete gradient optimization of a recurrent network applied to /b/, /d/, /g/ discrimination," *J. Acoust. Soc. Am.*, 87: 1301-1309, 1990.
- [2] Koizumi, T., Taniguchi, S., Ishida, H., and Mori, M. "Recurrent Neural Networks for Phoneme Recognition," *Tech. Report of the Institute of Electronics, Information, and Communication Engineers*, SP94-1, 1-8, 1994.

Table 3: Group classification accuracies(%).

| phoneme | Type1 RNN | | Type2 RNN | | MLNN | | phoneme | Type1 RNN | | Type2 RNN | | MLNN | |
|---------|-------------|-----------|-------------|-----------|-------------|-----------|---------|-------------|-----------|-------------|-----------|-------------|-----------|
| | recog. rate | ave. rate | recog. rate | ave. rate | recog. rate | ave. rate | | recog. rate | ave. rate | recog. rate | ave. rate | recog. rate | ave. rate |
| /k/ | 86.9 | | 77.8 | | 66.6 | | /z/ | 96.5 | | 95.7 | | 93.0 | |
| /ky/ | 92.0 | 89.6 | 84.0 | 82.1 | 36.0 | 69.3 | /j/ | 92.6 | | 84.7 | | 66.1 | |
| /t/ | 96.3 | | 93.3 | | 81.5 | | /y/ | 92.2 | 85.8 | 84.3 | 78.9 | 77.7 | 70.3 |
| /p/ | 95.8 | | 83.3 | | 54.2 | | /w/ | 96.0 | | 94.7 | | 94.7 | |
| | | | | | | | /r/ | 79.4 | | 71.8 | | 63.8 | |
| | | | | | | | /ry/ | 94.9 | | 82.1 | | 66.7 | |
| /g/ | 88.1 | | 82.1 | | 76.1 | | /m/ | 98.5 | | 94.5 | | 88.9 | |
| /gy/ | 16.7 | | 33.3 | | 33.3 | | /my/ | 0.0 | | 0.0 | | 25.0 | |
| /d/ | 95.3 | 88.3 | 92.1 | 88.3 | 87.9 | 78.9 | /n/ | 98.9 | 96.2 | 95.4 | 92.2 | 92.4 | 83.2 |
| /b/ | 86.0 | | 90.1 | | 74.8 | | /ny/ | 83.3 | | 83.3 | | 100. | |
| /by/ | 0.0 | | 0.0 | | 0.0 | | /N/ | 95.2 | | 89.9 | | 71.1 | |
| | | | | | | | /ng/ | 92.0 | | 90.4 | | 88.3 | |
| /s/ | 99.2 | | 96.8 | | 92.9 | | /a/ | 99.2 | | 97.5 | | 97.2 | |
| /sh/ | 95.8 | | 92.7 | | 82.1 | | /i/ | 85.0 | | 83.3 | | 81.2 | |
| /h/ | 92.0 | | 84.9 | | 74.5 | | /u/ | 84.0 | 92.4 | 83.7 | 91.2 | 85.2 | 90.2 |
| /hy/ | 33.3 | 96.3 | 33.3 | 92.6 | 55.6 | 85.5 | /e/ | 98.8 | | 97.5 | | 97.5 | |
| /f/ | 92.6 | | 83.3 | | 72.2 | | /o/ | 95.2 | | 94.2 | | 89.8 | |
| /ch/ | 93.8 | | 88.3 | | 76.6 | | average | | 91.9 | | 88.1 | | 81.3 |
| /ts/ | 100. | | 97.9 | | 94.8 | | | | | | | | |

Table 4: Intra-group phoneme recognition rates(%).

| phoneme | Type1 RNN | | Type2 RNN | | MLNN | | phoneme | Type1 RNN | | Type2 RNN | | MLNN | |
|---------|-------------|-----------|-------------|-----------|-------------|-----------|---------|-------------|-----------|-------------|-----------|-------------|-----------|
| | recog. rate | ave. rate | recog. rate | ave. rate | recog. rate | ave. rate | | recog. rate | ave. rate | recog. rate | ave. rate | recog. rate | ave. rate |
| /k/ | 98.9 | | 96.9 | | 95.0 | | /z/ | 100. | | 100. | | 99.1 | |
| /ky/ | 98.0 | 98.3 | 100. | 94.5 | 96.0 | 94.6 | /j/ | 99.5 | | 96.3 | | 83.1 | |
| /t/ | 97.2 | | 88.7 | | 95.1 | | /y/ | 92.8 | 98.3 | 82.5 | 93.7 | 84.9 | 91.5 |
| /p/ | 87.5 | | 75.0 | | 66.7 | | /w/ | 100. | | 100. | | 94.7 | |
| | | | | | | | /r/ | 99.5 | | 94.3 | | 94.9 | |
| | | | | | | | /ry/ | 87.2 | | 87.2 | | 66.7 | |
| /g/ | 100. | | 94.0 | | 95.5 | | /m/ | 91.2 | | 86.8 | | 81.1 | |
| /gy/ | 33.3 | | 66.7 | | 33.3 | | /my/ | 100. | | 100. | | 100. | |
| /d/ | 98.9 | 96.5 | 95.3 | 94.3 | 95.3 | 91.4 | /n/ | 91.6 | 91.3 | 92.4 | 87.9 | 82.8 | 82.3 |
| /b/ | 95.0 | | 94.1 | | 88.3 | | /ny/ | 83.3 | | 50.0 | | 66.7 | |
| /by/ | 100. | | 100. | | 100. | | /N/ | 92.3 | | 86.5 | | 81.6 | |
| | | | | | | | /ng/ | 88.8 | | 89.4 | | 86.2 | |
| /s/ | 85.2 | | 87.7 | | 87.7 | | /a/ | 100. | | 100. | | 99.7 | |
| /sh/ | 83.1 | | 68.4 | | 67.1 | | /i/ | 99.7 | | 99.2 | | 98.2 | |
| /h/ | 99.1 | | 99.1 | | 97.6 | | /u/ | 97.7 | 99.3 | 96.7 | 98.9 | 96.0 | 97.7 |
| /hy/ | 77.8 | 87.6 | 77.8 | 84.0 | 33.3 | 81.1 | /e/ | 99.7 | | 99.2 | | 97.2 | |
| /f/ | 98.1 | | 100. | | 90.7 | | /o/ | 99.3 | | 99.7 | | 97.3 | |
| /ch/ | 74.2 | | 71.1 | | 69.5 | | | | | | | | |
| /ts/ | 95.4 | | 87.1 | | 75.8 | | | | | | | | |

Table 5: Recognition rates(%) of both schemes.

| phoneme | group classification scheme | a single RNN | phoneme | group classification scheme | a single RNN | phoneme | group classification scheme | a single RNN |
|---------|-----------------------------|--------------|---------|-----------------------------|--------------|---------|-----------------------------|--------------|
| /k/ | 85.8 | 78.3 | /h/ | 91.5 | 90.6 | /N/ | 89.9 | 87.1 |
| /ky/ | 90.0 | 94.0 | /hy/ | 33.3 | 44.4 | /ng/ | 80.9 | 86.7 |
| /t/ | 93.8 | 87.0 | /f/ | 92.6 | 94.4 | /y/ | 84.9 | 75.3 |
| /p/ | 87.5 | 66.7 | /z/ | 96.5 | 97.4 | /w/ | 96.0 | 97.3 |
| /g/ | 88.1 | 86.6 | /j/ | 92.1 | 87.8 | /r/ | 78.3 | 75.1 |
| /gy/ | 0.0 | 0.0 | /ch/ | 71.9 | 64.1 | /ry/ | 76.9 | 84.6 |
| /d/ | 94.2 | 89.5 | /ts/ | 95.4 | 94.8 | /a/ | 99.2 | 99.5 |
| /b/ | 82.0 | 71.2 | /m/ | 86.6 | 81.3 | /i/ | 85.0 | 86.5 |
| /by/ | 0.0 | 25.0 | /my/ | 0.0 | 25.0 | /u/ | 83.0 | 81.8 |
| /s/ | 84.4 | 78.3 | /n/ | 93.9 | 77.1 | /e/ | 98.5 | 98.2 |
| /sh/ | 79.2 | 78.3 | /ny/ | 66.7 | 83.3 | /o/ | 94.8 | 96.7 |
| | | | | | | average | 88.1 | 84.9 |