Phoneme-based Word Recognition by Neural Network

- A Step Toward Large Vocabulary Recognition -

Akihiro Hirai[†] Alexander Waibel*

[†]Systems Development Lab., Hitachi, Ltd.
1099 Ohzenji Asao Kawasaki, 215 JAPAN
*Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.

Abstract In this paper we present a neural network based word recognition system extendable to large vocabulary isolated word recognition. The system consists of (1) time-delay neural networks (TDNNs) for phoneme spotting, and (2) a higher level network and a DP-time alignment procedure for word recognition. TDNN-based phoneme spotting networks are used whose role is to fire when a particular phoneme is input. A higher level network then improves these phoneme firing patterns in view of an idealized phoneme sequence. For training of the higher level network, DP-matching is used to determine idealized phoneme firing patterns which are nearest to the actual phoneme firings. During recognition, the system selects the most probable word by applying DP-matching to the outputs of the higher level network. Speaker-dependent and isolated word recognition experiments show that word recognition rates of around 92% can be achieved for medium-size vocabularies.

1 Introduction

A number of recent studies[1, 2, 3, 4] (also, see [5]) suggest that neural networks designed specifically for speech recognition can yield excellent phoneme recognition performance. Encouraged by these results, we have begun to apply these networks to large vocabulary word recognition.

At the word level, it has been shown that neural networks[6, 7, 8, 9] achieve outstanding performance results for various small vocabulary spoken word recognition tasks. However, most of them rely on the availability of sufficient speech patterns of each word for training. The resulting need for considerable amounts of data renders such an approach impractical for large and changing vocabularies. It is very hard to collect a lot of training data for each word in a practical application. It is also time and resource consuming to train a neural network over a large amount of data.

In order to overcome this problem in large vocabulary systems and to exploit the phoneme recognition ability of neural networks, we adopt phonemes as underlying subword units that make up the words to be recognized. In this approach, phonemes are recognized first, and words are determined by the sequence of phonemes found in the input string. Training data for phonemes could be large enough for training as long as the vocabulary is large and sufficient training patterns are available for each phoneme.

Adopting such a phoneme-based approach, we build a word recognition system using both neural networks and a DP-matching procedure[10]. We perform speaker-dependent, isolated word recognition experiments for evaluation.

In the following, we first present the architecture of the word recognition system. We then describe the phoneme spotting networks and the higher level network. We finally report on experimental evaluations of the system.

2 Word Recognition System

In our system, phonemes are recognized first and then words are recognized based on this phoneme recognition. Figure 2.1 shows the whole architecture. It consists of the following components.

(1) phoneme spotting network

Phoneme spotting networks scan the time frequency patterns of an input word. Each of them are trained to fire only when a particular phoneme is found in the input, resulting in a sequence of phoneme firings over the frames of the input word.

(2) higher level network

Phoneme firings are not necessarily perfect. They sometimes have misfirings or deletions. The higher level network acts as a postprocessor to smooth and correct these patterns in order to raise the recognition accuracy.We apply DP-matching for recognizing words, taking phoneme sequences in the dictionary as reference data, and the outputs of the higher level network as input data.

(3) dictionary

Correct phoneme sequences for the words to be recognized are taken from a dictionary. It can have several phoneme sequences for a word.

3 Phoneme Spotting Network

Phoneme spotting networks recognize phonemes of an input word. Each of them fires only for a particular phoneme. A Time-Delay Neural Network(TDNN) architecture[1, 2] was used. In a TDNN, a unit in a layer is connected to a unit in the upper layer directly and with delays[1]. Each of these connections has different weights. A TDNN unit can have the ability to relate and compare current input with the past history of events by this structure.

The architecture of a phoneme spotting network is shown in Figure 3.1. Sixteen melscale spectral coefficients serve as input to the input layer of the network. Input speech, sampled at 12 kHz, was hamming windowed and a 256-point FFT computed every 5 msec. Melscale coefficients were computed from the power spectrum[1] and adjacent coefficients in time collapsed resulting in an overall 10 msec frame rate. The coefficients over a 150 msec time interval were then normalized to lie between -1.0 and $+1.0$ with the average at 0.0. In Figure 3.1, X output unit means arbitrary phoneme unit and non-X output unit is the complement to X, i.e., anything else. Each phoneme spotting network is trained using the Back-propagation Learning Procedure[11].

4 Higher Level Network

4.1 Structure

The higher level network reshapes the phoneme firings that the phoneme spotting networks produce. The structure of the network

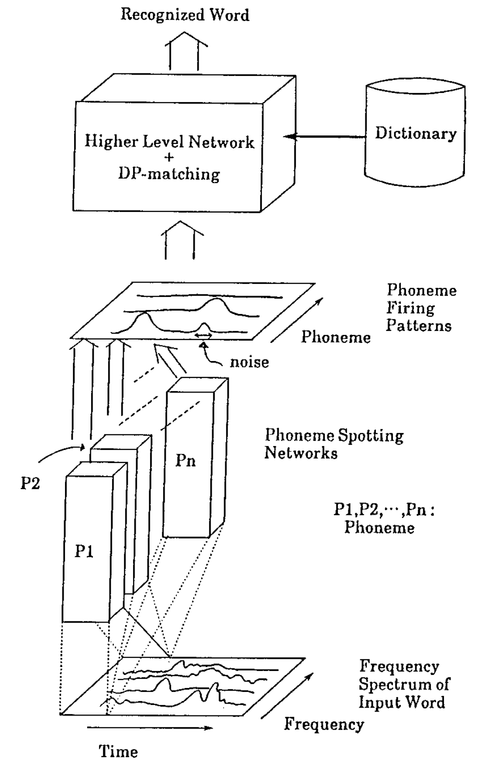

Figure 2.1 Word Recognition System by Neural Network

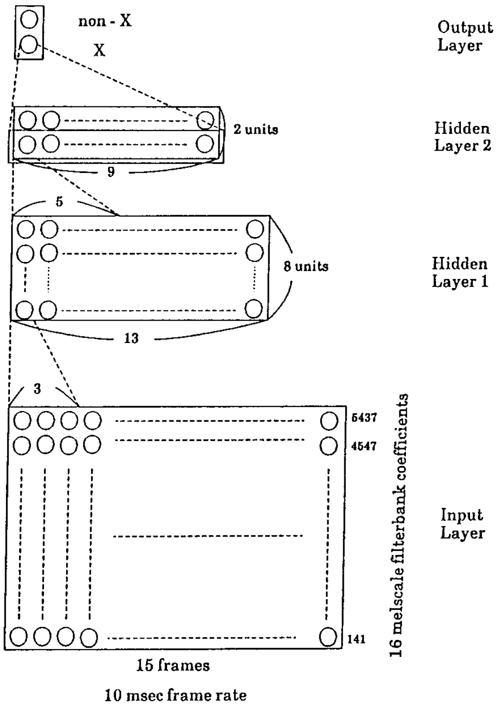

Figure 3.1 Phoneme Spotting Network Architecture

is shown in Figure 4.1. This network has 2 layers of units, input and output layer. The input layer is fully interconnected to the output layer. Basically, this network takes one frame of phoneme firings as input, and outputs one frame of improved phoneme firings. It scans all frames of an input word and produces reshaped phoneme firings. Figure 4.2 shows a spatially expanded figure of this network. In this figure, the weights of the corresponding connections in the time shifted copies are the same.

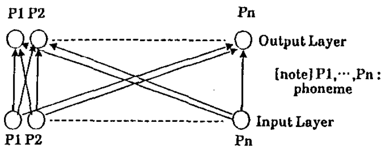

Figure 4.1 Higher Level Network Structure

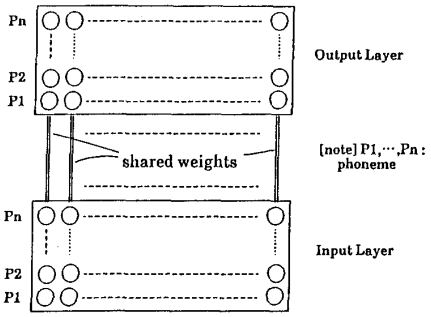

Figure 4.2 Higher Level Network (spatially expanded view)

Use of TDNN for Higher Level Network :The higher level network mentioned above is composed of normal units, not TDNN units. However, we can also adopt TDNN architecture for the network so that it could use contextual information. This is only one variation of the network.

4.2 Training

In order to determine the ideal phoneme firings for each frame, a DP-matching procedure was used. Taking phoneme firings for an input word as input data, and the lexical phoneme sequence from the dictionary as reference data, DP-matching determines the phoneme firings that are nearest to the input firings. We adopted a slope constraint for DP-matching as shown in Figure 4.3. The DP-equations are as follows.

$$g(1,1) = d(1,1)$$

$$g(i,j) = \min \begin{bmatrix} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 1.5 \times d(i,j) \end{bmatrix}$$

where $d(i,j)$ is the Euclidean distance between the i-th frame of phoneme firings and the vector representing the j-th phoneme of a word's phoneme sequence. In practice, we double each phoneme in the lexical phoneme sequence for the reference data. These conditions were introduced to avoid phoneme deletions among similar words (e.g., "kikaku" → "kiku". Both are Japanese words. "kikaku" means "plan". "kiku" means "chrysanthemum".).

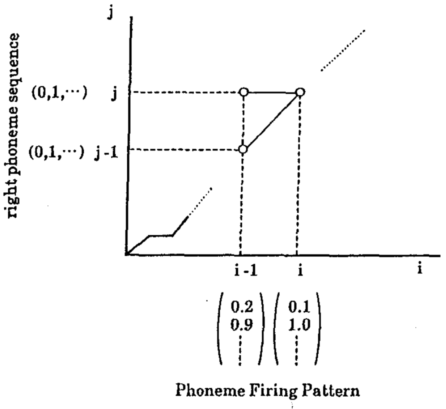

Figure 4.3 Slope Constraint for DP-matching

First, we apply the regular back-propagation to all time shifted copies corresponding to one time aligned phoneme based on the ideal phoneme sequence, and derive weight changes for those connections. Then, the weight changes are averaged over the phoneme interval in order to avoid the effect of the differences in phoneme durations. Weights are then changed accordingly. The weights are changed once for each phoneme interval at each iteration.

Dynamic Application of DP-matching : In the previous section, the alignment of ideal firings is determined before training and is never changed during training. Therefore, if noisy input phoneme firings lead to an inappropriate DP-alignment, suboptional network training will result. In order to avoid this problem, DP-matching can be applied to the outputs of the higher level network during training. In this case, output targets are changed after each iteration according to the optimal time alignment. If noise that disturbed DP-matching were decreased during training, the desired outputs would lead to improved results. We will call this dynamic application of DP-matching "dynamic alignment" and the previous method of using DP-matching "static alignment". They are similar in spirit to the work proposed by Sakoe et al[8].

4.3 Recognition

During recognition, the network takes phoneme firings of an input word as input, and outputs improved phoneme firings. We apply DP-matching to this improved firing taking all phoneme sequences in the dictionary as reference data. The word whose phoneme sequence has the least DP error value is chosen as the recognized word.

5 Recognition Experiments

5.1 Experiment with small vocabulary

In order to check the ability of our word recognition system, a preliminary small experiment was carried out first. The task is speaker-dependent, isolated word recognition.

5.1.1 Data

(1) Phoneme data(for phoneme spotting networks)

For efficiency we limited the number of phonemes. The 6 most frequently occurring phonemes "a", "i", "u", "t", "k", "g" were chosen. They also reflects typical recognition difficulties. Phoneme spotting networks for these 6 phonemes were made using the training data as follows.

The source of the training data : ATR (Advanced Telecommunications Research Institute International) speech database 5240 common Japanese words[1] spoken by one male native Japanese speaker(MAU) was used. All utterances were recorded in a sound proof booth and digitized at a 12 KHz sampling rate.

Training token : The actual training tokens are extracted from the training set based on the hand labels in the database. The range of each tokens is 150 msec. We extracted training tokens for silent parts from the training set, too.

(2) Word data (for higher level network)

We collected utterances of words whose pronunciations consist of only the above mentioned 6 phonemes from the same ATR database. The silent parts were eliminated from the utterances by hand. We collected 79 utterances(79 different words). These utterances were separated into training set and testing set by taking advantage of a large number of homophones in the database.

Table 5.1 Recognition Rates (Small Vocabulary)

kinds of network / kinds of data	without higher level network	higher level network (1 frame window)	TDNN higher level network (3 frame window)	TDNN higher level network (5 frame window)
testing data (20)	92.5 %	95.0 %	95.0 %	97.5 %
training data (59)	91.5 %	93.2 %	95.8 %	96.6 %
whole data (testing data + training data) (79)	91.8 %	93.7 %	95.6 %	96.8 %

training data = 59 utterances
(59 different pronunciations)
testing data = 20 utterances
(15 different pronunciations)

5.1.2 Experiments

We ran experiments as shown below. In all cases, 1000 iterations were run.

(1) Word recognition by applying DP-matching directly to phoneme firings

(2) Higher level network with normal units (1 frame window)

(3) Time-delayed higher level network
  (a) 3 frame window
  (b) 5 frame window

(4) Higher level network with normal units (1 frame window) and dynamic alignment
  (a) initial weights are random
  (b) initial weights are the ones obtained by 1000 iterations of training with static alignment.

5.1.3 Results and Discussions

The recognition rates obtained in this experiments are shown in Table 5.1. When the higher level network with normal units is used, the recognition rate for each case is better than without a higher level network. The time-delayed higher level network works better than the network with single frame units. The highest recognition rate is obtained by a time-delayed higher level network using a 5 frame window, i.e., a network incorporating a maximum

Table 5.2 Effects of Dynamic Alignment

recognition rate

kinds of training / kinds of data	dynamic(1000)	static(1000) +dynamic(1000)	static(1000) static+(1000)
testing data (20)	----	95.0 %	95.0 %
training data (59)	----	96.6 %	93.2 %
whole data (testing data + training data) (79)	----	95.6 %	93.7 %

[note]
static (n) : n iteration training by static alignment
dynamic (n) : n iteration training by dynamic alignment
---- : training failure

amount of contextual information.

The effects of using dynamic alignment are shown in Table 5.2. Dynamic alignment caused learning failure when the initial weights were random. Dynamic alignment training whose initial weights are the ones obtained by 1000 iteration training with static alignment improved recognition rate for training data.

5.2 Experiment with medium-size vocabulary

An experiment with a medium-size vocabulary was carried out in order to judge the recognition ability of the higher level network. In this experiment, we used another TDNN-based phoneme spotting network provided by Sawai et al[12]. We can know the invariant features of the higher level network by using different phoneme spotting networks.

5.2.1 Data

We limited the number of phonemes in order to limit the use of computer resources. We chose 10 frequent phonemes "a", "i", "u", "e", "o", "t", "k", "h", "r", "s". We then collected utterances of words whose pronunciations consist of only the above mentioned 10 phonemes from the same ATR database, and for whom homophones exist in the database. As a result, we obtained 225 utterances (225 different words, 96 different pronunciations). We separated these utterances into training set and testing set as follows.

training data = 96 utterances (96 different pronunciations)
testing data = 129 utterances (96 different pronunciations)

Each word was represented by a string of phonemes for its most likely pronunciation. A small number of alternate pronunciations was also introduced in the phonemic dictionary.

5.2.2 Experiments

The following experiments were carried out. In all cases, 1000 iterations were run.

(1) Word recognition by applying DP-matching directly to phoneme firing
(2) Higher level network with normal units (1 frame window)
(3) Time-delayed higher level network
 · 5 frame window
(4) Time-delayed higher level network with dynamic alignment
 · 5 frame window. Initial weights were obtained after 1000 iteration of training with static alignment

5.2.3 Results and Discussions

The recognition rates obtained in this experiment are shown in Table 5.3. When the higher level network with single frame window is used, the recognition rate is higher than the case without the higher level network. It is because the network raises insufficient firings near phoneme boundaries so that silent parts between phonemes could be shorter, or decreases noise.

The time-delayed higher level network works better than the network with single frame window. It has more power to raise insufficient firings by looking at the neighboring frames and knowing what phoneme is dominant in the neighboring frames. Thus, the problem with large distances between two phonemes is reduced. Dynamic alignment also contributes to increasing recognition rates because of its fine tuning ability as observed in 5.1.

The higher level network cannot compensate for complete omissions of phonemes and isolated strong incorrect firings. This is the main reason of misrecognition.

Table 5.3 Recognition Rates (Medium-size Vocabulary)

kinds of network & training kinds of data	without higher level network	higher level network with normal units (1 frame window)	TDNN higher level network (5 frame window)	TDNN higher level network (5 frame window & dynamic alignment)
testing data (129)	81.4 %	87.5 %	89.1 %	91.9 %
training data (96)	74.0 %	80.2 %	89.6 %	92.7 %
whole data (testing data + training data) (225)	78.2 %	84.4 %	89.3 %	92.9 %

6 Conclusions

We have presented a phoneme-based word recognition system based on neural networks. We have found recognition rates in excess of 90% in speaker-dependent, medium-size vocabulary experiments. We believe that it may serve as a good first step toward the development of large vocabulary word recognition systems based on neural networks.

We have also presented the techniques which are effective in rasing recognition rate. One is the use of time-delayed higher level network and the other is given by joint optimization of DP-matching and back-propagation that we have called "dynamic alignment". These techniques might be generally applied to a variety of systems.

Three avenues of research might be particularly appropriate. First, our system should be extended to include all 24 phonemes in the Japanese database. Networks achieving excellent performance (up to 98%) in spotting *all* phonemes are nearing completion[13] and should be incorporated in our system. The second is to optimize the phoneme spotting networks for use in word recognition by allowing error back-propagation to proceed from the output layer of the higher level network all the way through the phoneme spotting networks down to the underlying speech signal. The third is better duration control. At present, no information for the expected duration of each phoneme is used in our models. A third extension from which we expect additional performance improvements is therefore the introduction of such duration control.

Acknowledgements

We thank ATR Interpreting Telephoney Research Laboratories for allowing us to use its speech database. We also thank Dr. Takeshi Kawabata, Mr. Nobuo Hataoka, Mr. John Hampshire and Mr. Joseph Tabelskis for valuable comments and Ms. Cindy Wood for setting up the data environment for us.We'd like to thank Prof. Jaime G. Carbonell of Carnegie Mellon University and Mr. Singi Domen, General Manager of Systems Development Laboratory, Hitachi Ltd., for giving us the oppertunity of this research, too.

References

[1] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.ASSP-37, March 1989.

[2] A. Waibel, H. Sawai, K. Shikano. Consonant and Phoneme Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks. *Proc. of the ICASS*, Glasgow, May 1989.

[3] R. L. Watrous. *Speech Recognition Using Connectionist Networks.* PhD thesis, University of Pennsylvania, November 1988.

[4] E. McDermott and S. Katagiri. Phoneme Recognition Using Kohonen's Learning Vector Quantization. ATR Workshop on Neural Networks and Parallel Distributed Processing, Osaka Japan, July 1988.

[5] R. P. Lippmann. Review of Neural Networks for Speech Recognition in *Neural Computation*, MIT press, March 1989.

[6] D. J. Burr. Speech Recognition Experiments with Perceptron. in *Neural Information Processing Systems*(E. Anderson Ed.), pp.144-153, American Institute of Physics, 1988.

[7] R. P. Lippmann and B. Gold. Neural-Net Classifiers Useful for Speech Recognition. *Proc. of International Conference on Neural Networks*, IEEE, San Diego, June 1987.

[8] H. Sakoe, R. Isotani and K. Iso. Speaker-Independent Word Recognition Using Dynamic Programming Neural Networks. *Proc. of the ICASS*, Glasgow, May 1989.

[9] L-Y. Bottou. Reconnaissance de la Parole par Reseaux multi-couches. *Proc. of Neuro-Nimes 88*, November 1988.

[10] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE, Transactions on Acoustics, Speech and Signal Processing*, vol.ASSP-26, No.1, pp.43-49, February 1978.

[11] D. E. Rumelhart and G. E. Hinton and R.J. Williams. Learning Representations by Back-Propagating Errors. *Nature*, vol.323, pp.533-536, October 1986.

[12] H. Sawai, A. Waibel, P. Haffner, M. Miyatake and K. Shikano. Parallelism, Hierarchy, Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes/CV-Syllables. *Proc. of IJCNN*, vol.2, pp.81-88, Washington D.C., June 1989.

[13] M. Miyatake, H. Sawai, K. Shikano. Improvement on Phoneme Spotting Experiment by a Large Phonemic Time Delay-Neural Network. *Proc. of ICASS*, 1990 [in press].