Phoneme-based Word Recognition by Neural Network

- A Step Toward Large Vocabulary Recognition -

Akihiro Hirai[†] Alexander Waibel

August 1989

[†] He is a visiting researcher from Hitachi, Ltd. Hitachi is an affiliate member of CMT. His permanent address is
Systems Development Lab., Hitachi, Ltd., 1099 Ohzenji, Asao Kawasaki, 215
Japan.

Contents

discuss extensions currently under investigation and offer suggestions for possible future improvements.

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

3.1 Structure

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 D_1 through D_N as shown in Figure 3.1. Each of these connections

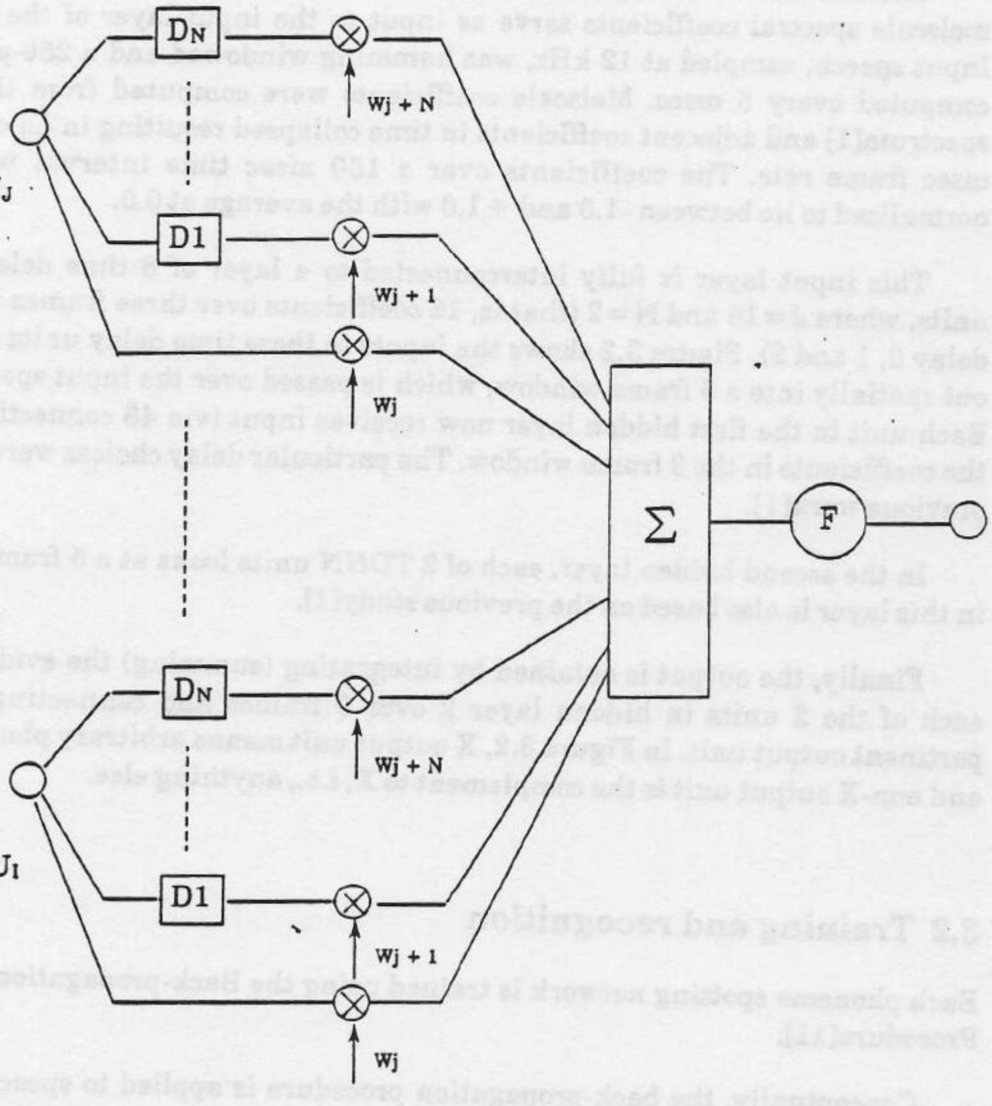
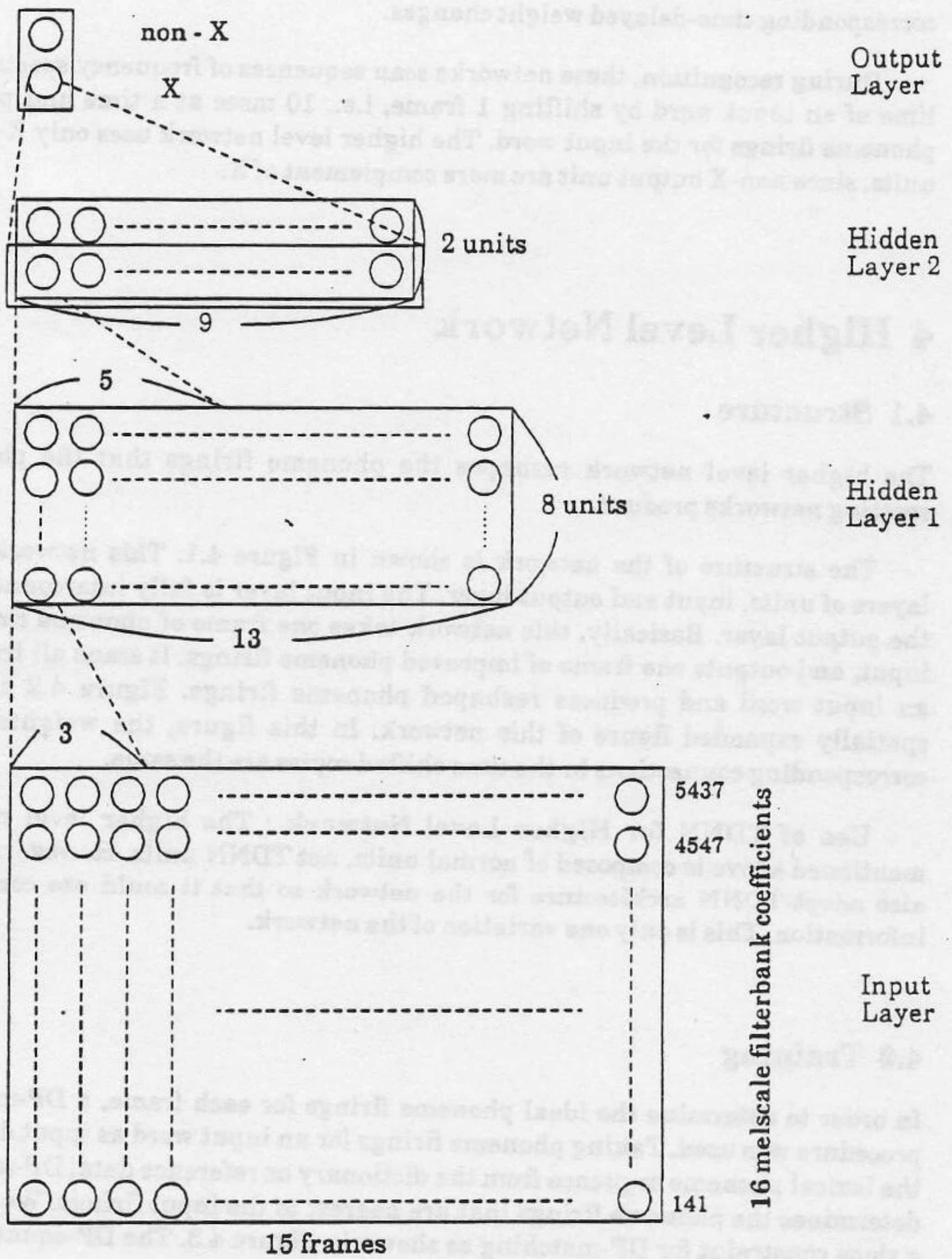


Figure 3.1 A Time Delay Neural Network (TDNN) unit



10 msec frame rate

Figure 3.2 Phoneme Spotting Network Architecture

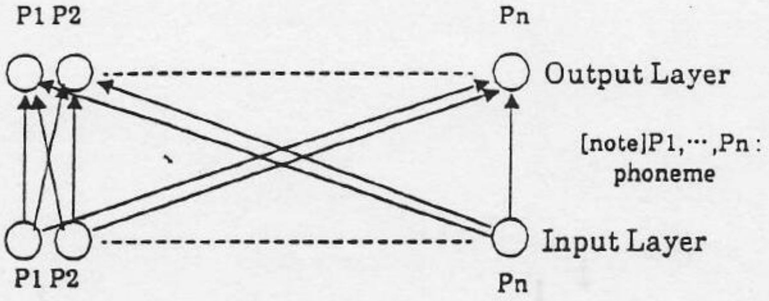


Figure 4.1 Higher Level Network Structure

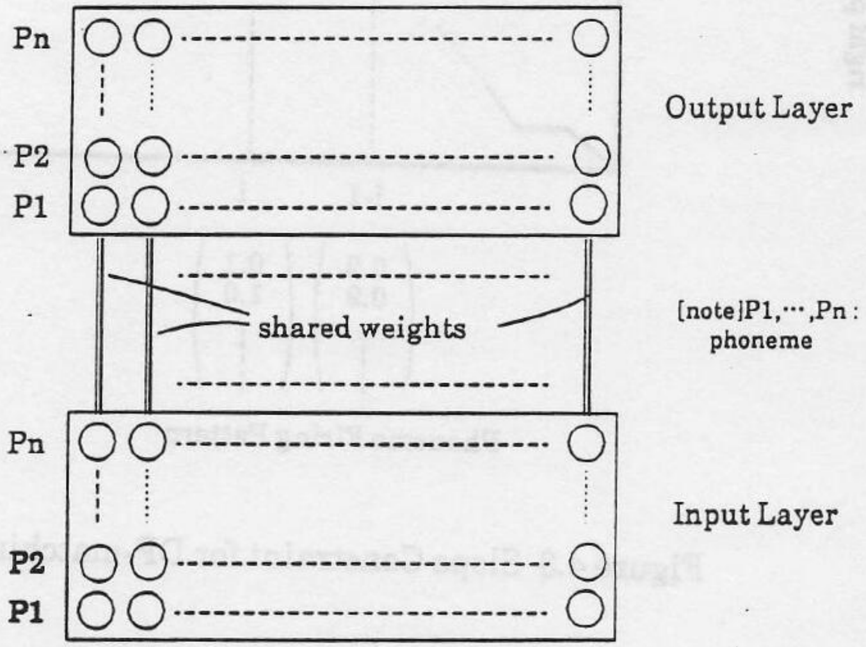


Figure 4.2 Higher Level Network (spatially expanded view)

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

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

reference data. These conditions were introduced to avoid phoneme deletions among similar words(e.g., "kikaku" → "kiku").

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 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.

(a) N = 2 (3 frame window)
(b) N = 4 (5 frame window)

(4) Higher level network with normal units 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

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 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. In Table 5.2, the recognition rate of a network whose weights were initiated by 1000 iteration training with static alignment, and which is then trained for 1000 more iterations using static alignment is shown for a fair comparison between static alignment and dynamic alignment.

5.1.4 Discussions

(1) Properties of phoneme firings

Phoneme firings that phoneme spotting networks produce have the following properties.

(a) Some neighboring phonemes overlap significantly
Figure 5.1 shows a phoneme firing pattern for the word "kaigi". "i" and "g" overlap. This behavior disturbs word recognition when two phonemes overlap completely.

(b) Durations of firings for some phonemes are short
An example is shown in Figure 5.2. It is a phoneme firing pattern for the word "au". There is a silent part between "a" and "u", since firing for "a" is too short. The DP-matching conditions were set adjusted to avoid the undesirable effects of this firing property.

Table 5.2 Effects of Dynamic Alignment

recognition rate

kinds of data \ kinds of training	dynamic(1000)	static(1000)
		+dynamic(1000)
			static(1000) +
			~~dynamic~~(1000)
			Static

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

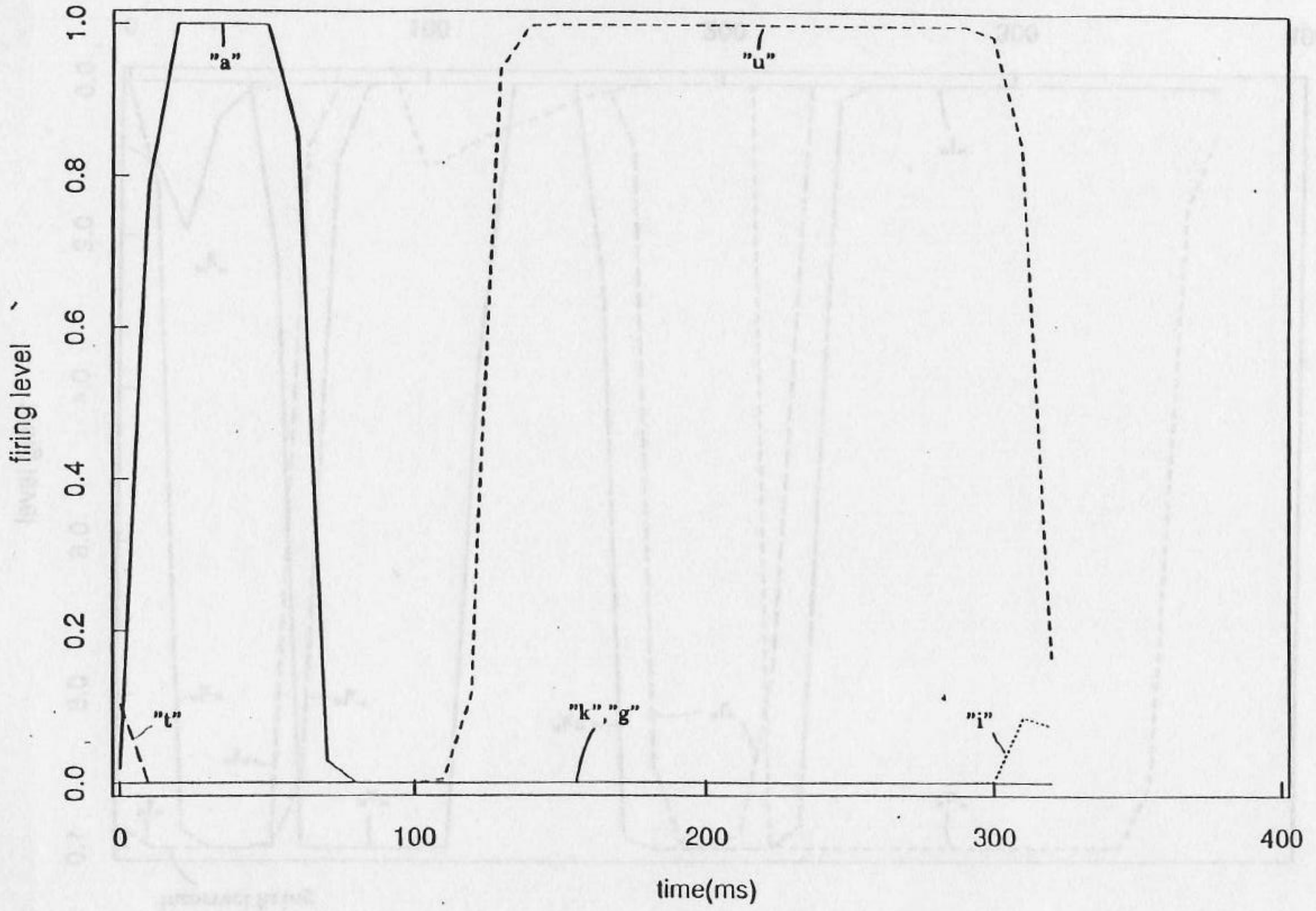


Figure 5.2 Phoneme Firing Pattern for the word "au"

(c) Particular phonemes fire whenever another particular phoneme fires

An example is shown in Figure 5.3. It is a phoneme firing pattern for the word "kaku". "t" fires for the leading "k". "t" always fires for a leading "k" of the other words, too. This sometimes causes misrecognition. In the case of "kaku", it is misrecognized as "taku".

(d) There are a some irregular misfirings

Sometimes phoneme spotting networks fire incorrectly and irregularly. This can lead to word recognition error.

(2) Abilities of a higher level network with normal units

In theory, it can distinguish linearly separable patterns in one frame, since it is a 2 layer network. It can perform a linear transformation on the input pattern. As a result of the experiment above, we found the following properties.

(a) increasing insufficient firing level

It can increase insufficient firing levels based on examples in the training data.

(b) decreasing low level noise

It can eliminate small local firing noise.

(c) eliminating overlapping firings

It can eliminate overlap. It is determined by the frequency in the training data, which phoneme firing is to be eliminated. For example, it can eliminate "i" for the overlapping of "i" and "g" in Figure 5.1. Then, the word can be recognized correctly by using the higher level network.

(d) eliminating wrong firings

It can eliminate wrong firings in some patterns. For example, it can eliminate the "t" firing at the leading "k" of Figure 5.3. Then, this word can be recognized correctly when a higher level network is used.

(3) misrecognition

We will describe some examples of misrecognition.

(a)misrecognition caused by incorrect firings

For the utterance of the word "igi", "i" continues to fire over "g" duration and "k" fires wrongly at the beginning of the word. The higher level network tried to eliminate the overlap but couldn't compensate these misfirings sufficiently.Then, when DP-matching is applied to the outputs of the higher level

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]. Because the network is trained using training data not only from restricted parts of phonemes but also other parts of word utterances and it rarely produces firings that disturb DP-matching completely, 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.

testing 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 patterns.

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

(3) Time-delayed higher level network
· N = 4 (5 frame window)

(4) Time-delayed higher level network with dynamic alignment
· N = 4 (5 frame window). Initial weights were obtained after 1000 iteration of training with static alignment

5.2.3 Results and Discussions

(1) Properties of phoneme firings

6.1 3 layer higher level network

6.1.1 Structure

The network structure(for one phoneme) is depicted in Figure 6.1. This is a spatially expanded structure and its time width corresponds to one phoneme duration. The structure of input layer and the 2nd layer is the same as the 2 layer network explored above. The weights of the corresponding connections between input and the 2nd layer in the time shifted copies are the same. Each output unit of this 2 layer network represents a phoneme.

In this network, output units are connected to a window of units of the 2nd layer like in the TDNN. In other words, activations of 2nd layer units corresponding to the same phoneme are connected to one output unit. Since phoneme durations are variable, connections between the 2nd and the output layer must be variable. So we call this architecture "Variable Time-Delay Neural Network".

The weights of the connections between the 2nd and output layer are normalized by phoneme duration length and shared by all time shifted copies.

6.1.2 Training and recognition

During training, ideal phoneme durations are determined by DP-matching using phoneme firings of an input word as input data and the correct phoneme sequence of the input word as reference data. We then apply regular back-propagation to all time shifted copies corresponding to one phoneme, and compute weight changes for the connections. The weight changes are averaged over the phoneme duration. In this manner, weights are changed once for one phoneme duration.

In recognition, mean square values are obtained for all the dictionary words and the word with the least error value is selected as the recognized word.

6.2 4 layer higher level network

6.2.1 Structure

The concept of the 4 layer network structure is shown in Figure 6.2. This network has 4 layers of units. It has one output unit for each word category. Input, the 2nd and the 3rd layer of units are the same as the 3 layer network mentioned above with the exception of the following modification. Each word has its own set of third layer units, but weights for units representing the same phoneme are

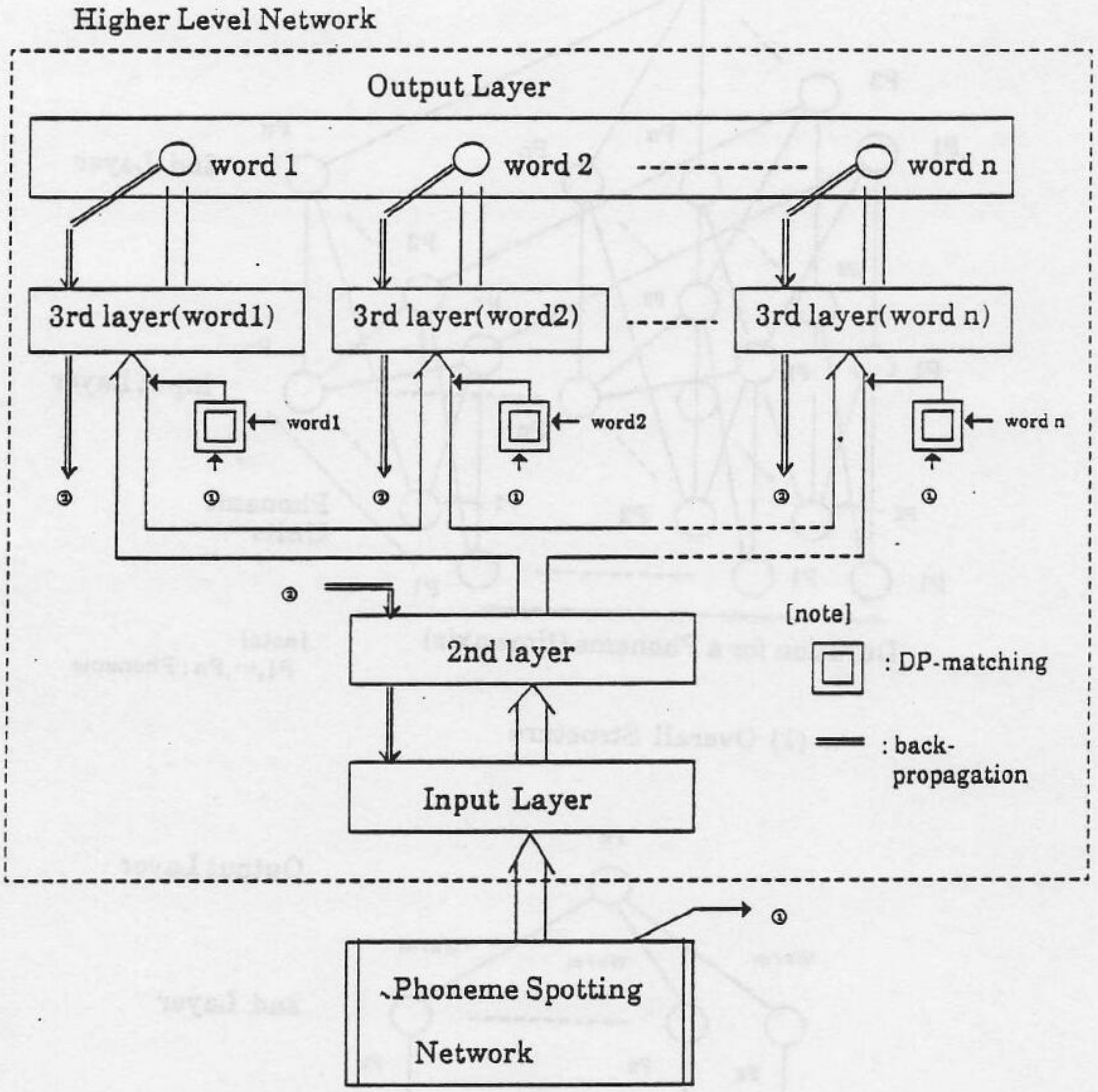


Figure 6.2 4 layer Higher Level Network Structure

Table 6.1 Recognition Rates of 3 and 4 Layer Higher Level Network

kinds of higher level network	small vocabulary			medium-size vocabulary		
	testing data (20)	training data (59)	whole data (79)	testing data (129)	training data (96)	whole data (225)
3 layer higher level network	65.0 %	71.2 %	70.9 %	41.4 %	41.7 %	41.3 %
4 layer higher level network	80.0 %	100 %	94.9 %	67.4 %	100 %	81.3 %

Hampshire and Joseph Tabelskis for valuable comments and Cindy Wood for setting up the data environment for us.

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].