TR-I-0122

Constructive Neural Network for Speech Recognition

構成的ニューラルネットによる音声認識

Takeshi KAWABATA

川端　豪

Nov. 7, 1989

ABSTRACT

A new method to combine phone TDNNs for constructing word or continuous speech networks is proposed. The method is called Constructive Neural Network (CNN). The CNN method has the following important features:

(1) Any word network can be composed of neural phone verifiers according to its phonetic representation.

(2) Non-linear time warping is realized using only a network structure including FIR filters which control the duration between each phone pairs independently.

(3) Word-level backpropagation learning can be used for training the TDNN phone recognizers.

ATR自動翻訳電話研究所

CONSTRUCTIVE NEURAL NETWORK FOR SPEECH RECOGNITION

Takeshi KAWABATA

ATR Interpreting Telephony Research Laboratories Seika-cho, Souraku-gun, Kyoto 619-02 Japan

1. INTRODUCTION

In recent years, neural approaches for pattern recognition have again advanced by utilizing the back propagation learning algorithm [Rumelhart 86]. In the field of speech recognition, a neural network which has time-delay elements can accurately discriminate subtle differences in Japanese phones such as /b/, /d/ and /g/ [Waibel 89]. The Time Delay Neural Network (TDNN) has the capability to recognize time-shifted patterns such as speech.

This paper describes a new speech recognition method, which is called the Constructive Neural Network (CNN), to combine TDNN phone recognizers for constructing word or continuous speech recognition networks. The CNN method has the following important features:

(1) Any word network can be composed of neural phone verifiers according to its phonetic representation.

(2) Non-linear time warping is realized using only a network structure including FIR filters which control the duration between each phone pairs independently.

(3) Word-level backpropagation learning can be used for training the TDNN phone recognizers.

2. COMPOSITION OF WORD NETWORKS

A phone recognition unit is formalized as a phone verifier (Fig.1), which has an input and output for intermediate firing patterns. The neural phone verifier receives an intermediate firing pattern and updates it according to the phone verification result.

Figure 2 shows an example of a composed network for the Japanese word /asa/. Three neural phone verifiers, /a/, /s/ and /a/ , are concatenated with

intermediate firing patterns. Only when all phone recognition processes are executed successfully will the /asa/ network output a word firing pattern with high intensity. Any word network can be composed of the neural phone verifiers according to its phonetic representation.

3. INTERNAL STRUCTURE OF NEURAL PHONE VERIFIER

Figure 3 shows an internal structure of a neural phone verifier. An input firing pattern is delayed by Finite Impulse Response filter 1 and summed with phone feature intensity which is extracted by a TDNN (Time Delay Neural Network). The summed intensity pattern is reformed by a sigmoid function and output to a next phone verifier after being delayed by FIR filter 2.

FIR filter 1 controls the duration between the left phone boundary and a typical phone feature. FIR filter 2 controls the duration between a typical phone feature and the right phone boundary. Because the FIR filters control the duration between each phone pairs independently, non-linear time warping is realized using only the network structure.

In actual implementation of the system, these impulse responses are approximated using the Gaussian distribution as shown in Fig.4.

A neural cell, which contains an accumulator and a sigmoid operator, is located between the FIR filters. This neural cell works as a soft AND circuit. As shown in Fig.5 and Fig.6, basic logical operators (AND/OR) can be realized using neural cells. Thus, a neural network can simulate any deterministic mechanisms.

In the current implementation (Fig.3), a word network is composed of TDNN phone outputs and their AND connections. Figure 7 shows the network for the Japanese word /ikioi/ and an example of its intermediate firing patterns.

4. WORD-LEVEL BACKPROPAGATION LEARNING

The network composed by the CNN method is a simple feed forward network, therefore, word-level backpropagation learning can be used for training the TDNN phone recognizers.

When a composed word network does not fire with high intensity at a

correct word period, the system puts a word-level supervisory signal and backpropagates it through the whole word network. The dE/dw value is calculated and accumulated for each TDNN connection. After processing all training words, the weight of each connection is updated using averaged dE/dw. Figure 8 shows an example of miss detection for a word network (/kaigi/). The reason for this miss detection is because its /i/-TDNN was not trained sufficiently. Figure 9 shows word-level backpropagation of δ (dE/do). The ill-trained /i/-TDNN is automatically found and trained correctly.

5. RECOGNITION EXPERIMENT

The proposed method was tested by word recognition experiments using 165 phonetically balanced words uttered by a male speaker. TDNN phone recognizers are trained by 2 methods:

(1) By using a phone database extracted from odd words of 5,240 isolated word database. The word database was uttered by the same speaker.

(2) By using the above database and word-level learning for 165 words extracted from the 5,240 words. The training and test words use the same vocabulary, however, were uttered separately. In the prototype system, only a simplified TDNN structure is implemented. Thus, the system does not yet discriminate b/d/g, p/t/k and m/n/N . The TDNN structure is shown in Fig.10.

Table 1 shows the word recognition score. Using word-level learning, The word recognition rate is improved from 81.8% to 90.9% for 165 phonetically balanced words. Clearly, word-level learning is effective for tuning TDNN phone recognizers.

6. CONCLUSION

A new method to combine phone TDNNs for constructing word or continuous speech networks is proposed. Using the method called CNN, any word network can be composed of the neural phone verifiers according to its phonetic representation. In this method, non-linear time warping is realized using only a network structure, and word-level backpropagation learning can be used for tuning the TDNN phone recognizers.

ACKNOWLEDGEMENT

The CNN method and its first implementation were developed by the author while at Carnegie Mellon University. The author would like to thank Dr. Akira Kurematsu, President of ATR Interpreting Telephony Research Laboratories, for his guidance and Dr. Alexander Waibel, Research Computer Scientist of CMU Computer Science Department, for supporting the research environment at CMU. The author also wishes to thank Mr. John Hampshire, Mr. Nobuo Hataoka, Mr. Ajai Jain and other members of the CMU neural speech group for discussing the CNN architecture.

REFERENCES

[Rumelhart 86] Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: "Learning Internal Representations by Error Propagation", PDP, Vol.1, Chap.8, The MIT Press (1986)

[Waibel 89] Waibel, A., Hanazawa, T., Hinton, G. E., Shikano, K., and Lang, K. J.: "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans on ASSP, Vol.37, No.3, pp.328-339 (Mar. 1989)

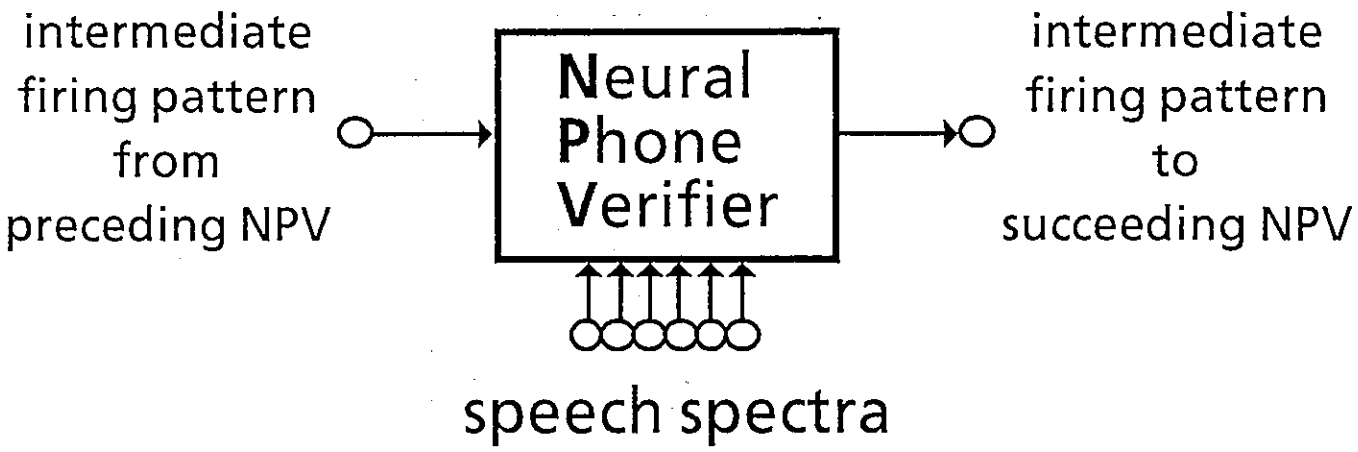


Fig.1 Neural Phone Verifier

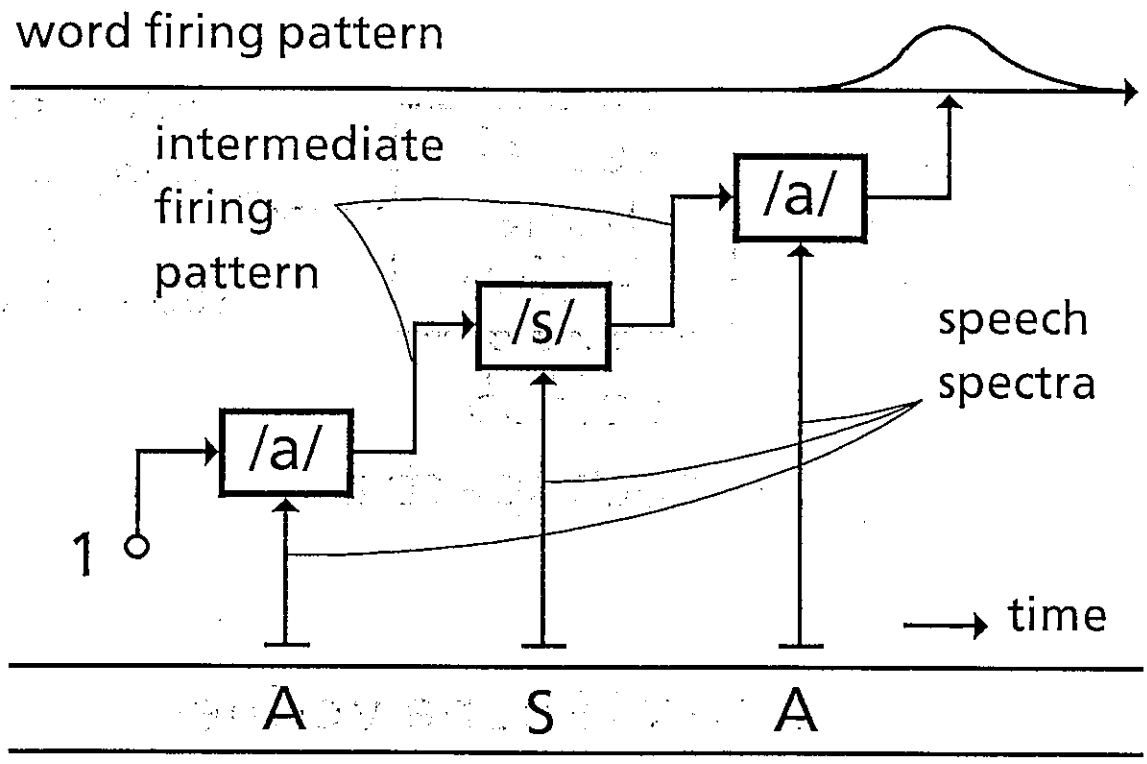


Fig.2 Composition of the neural network for the Japanese word /asa/

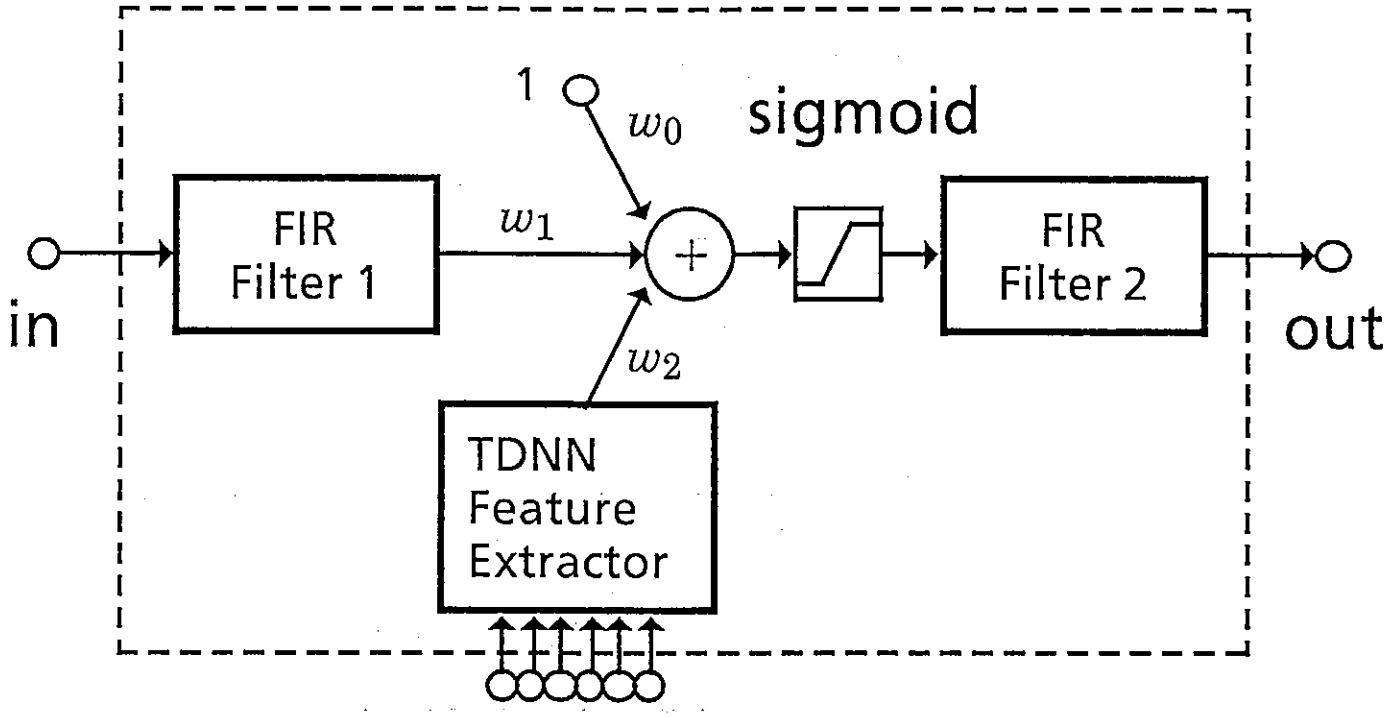


Fig.3 Simplest internal structure of
Neural **P**hone **V**erifier

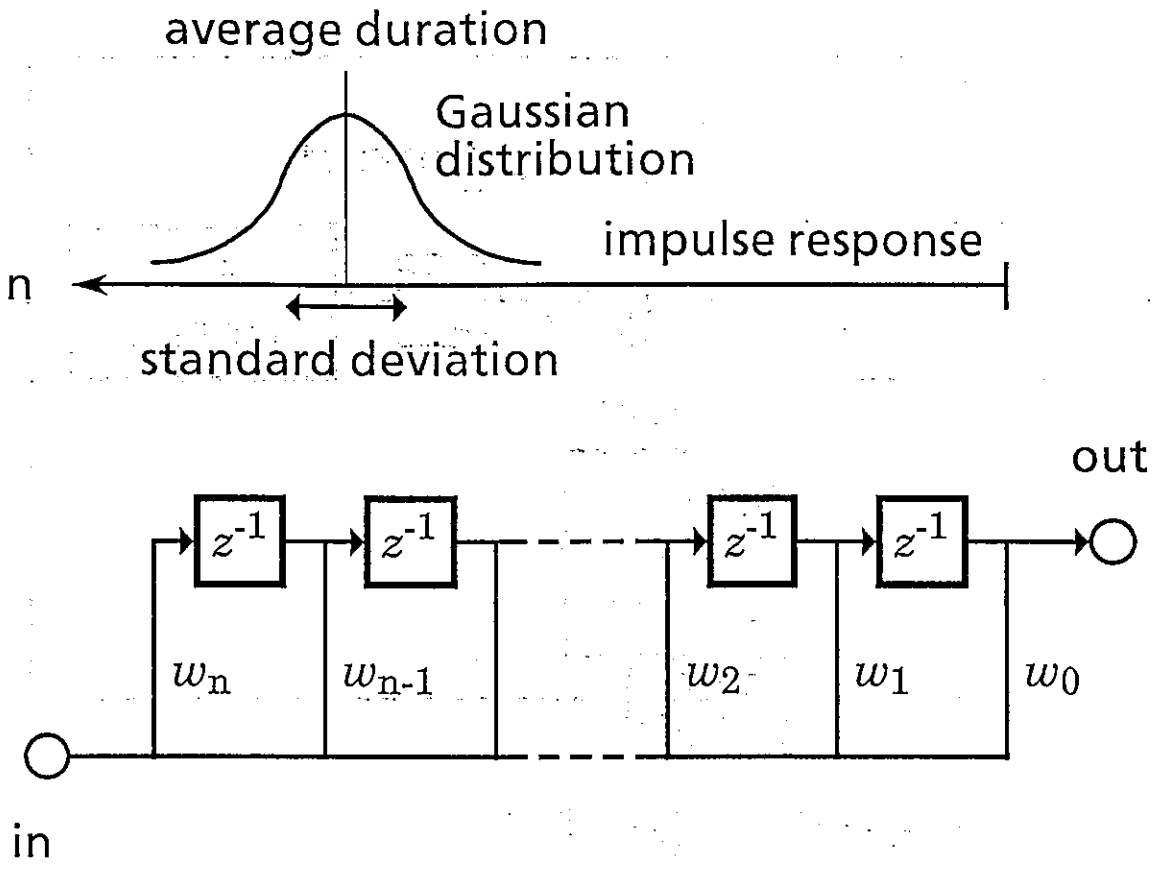
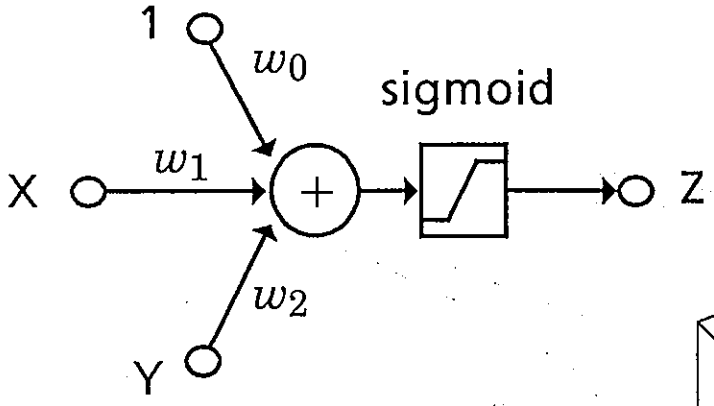


Fig.4 FIR filter for duration control



$$Z = X \otimes Y$$

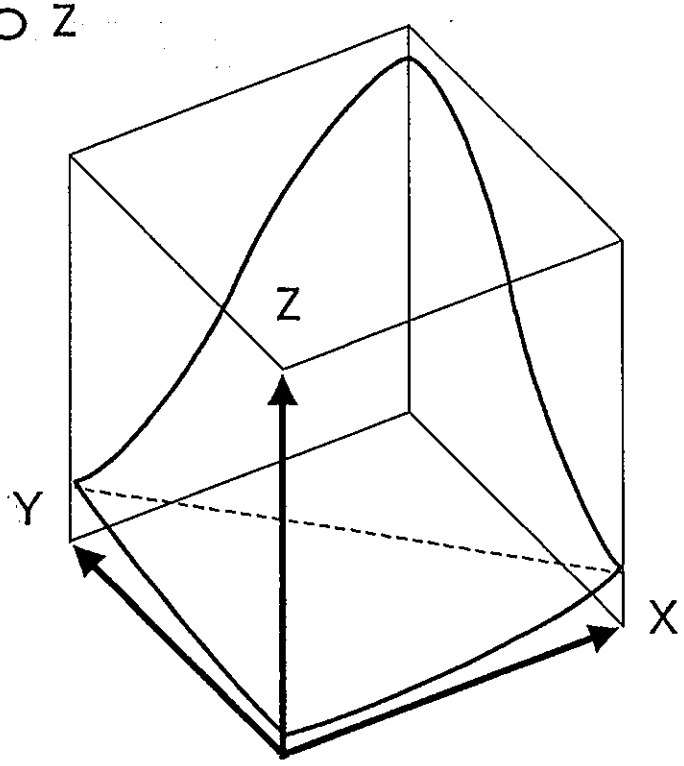
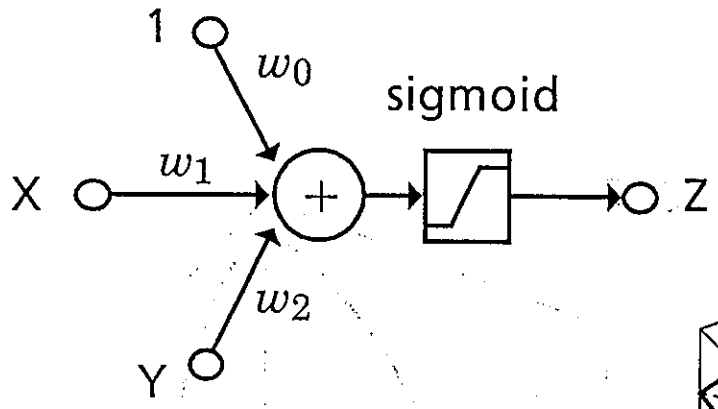


Fig.5 Neural AND circuit



$Z = X \oplus Y$

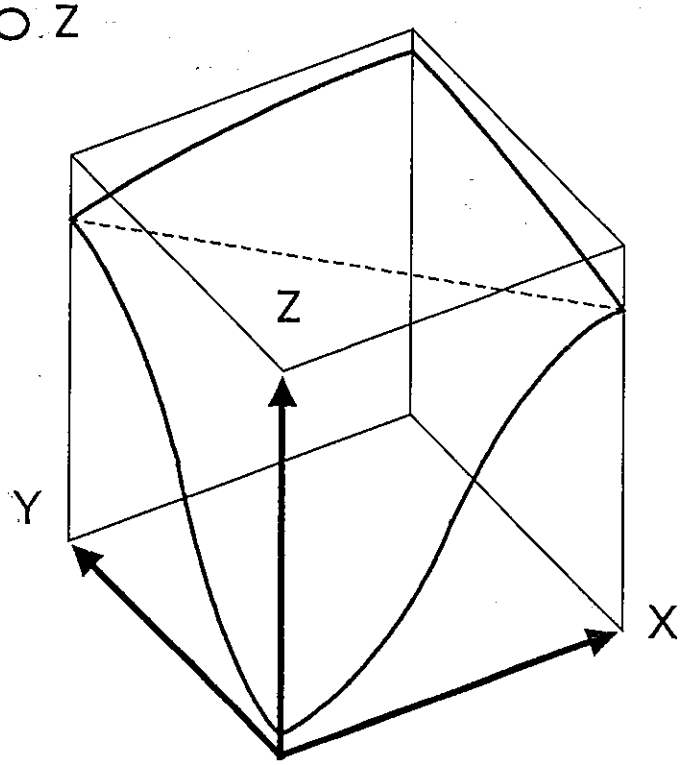


Fig.6 Neural OR circuit

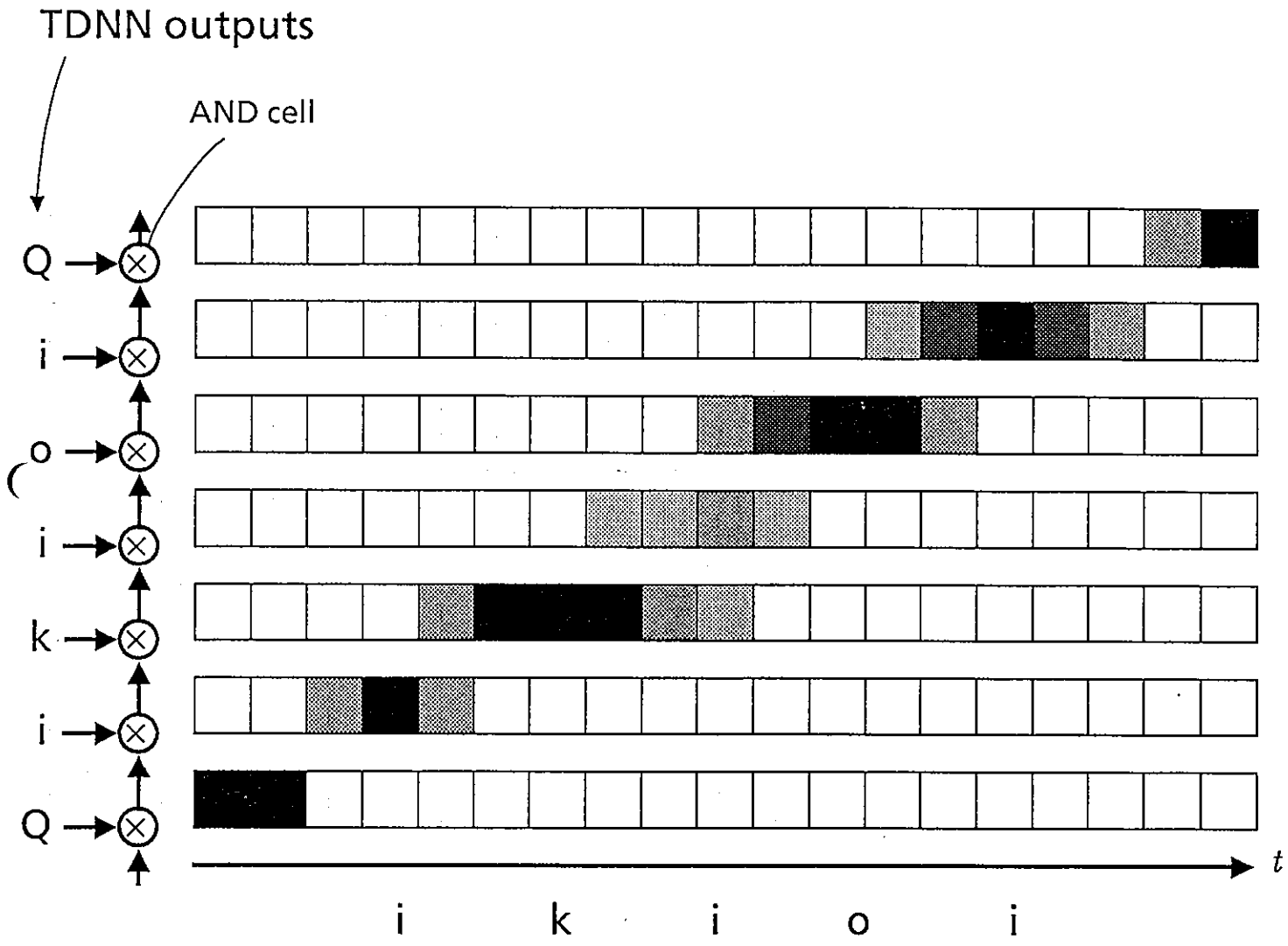


Fig.7 Word network (/ikioi/) and an example of intermediate firing patterns

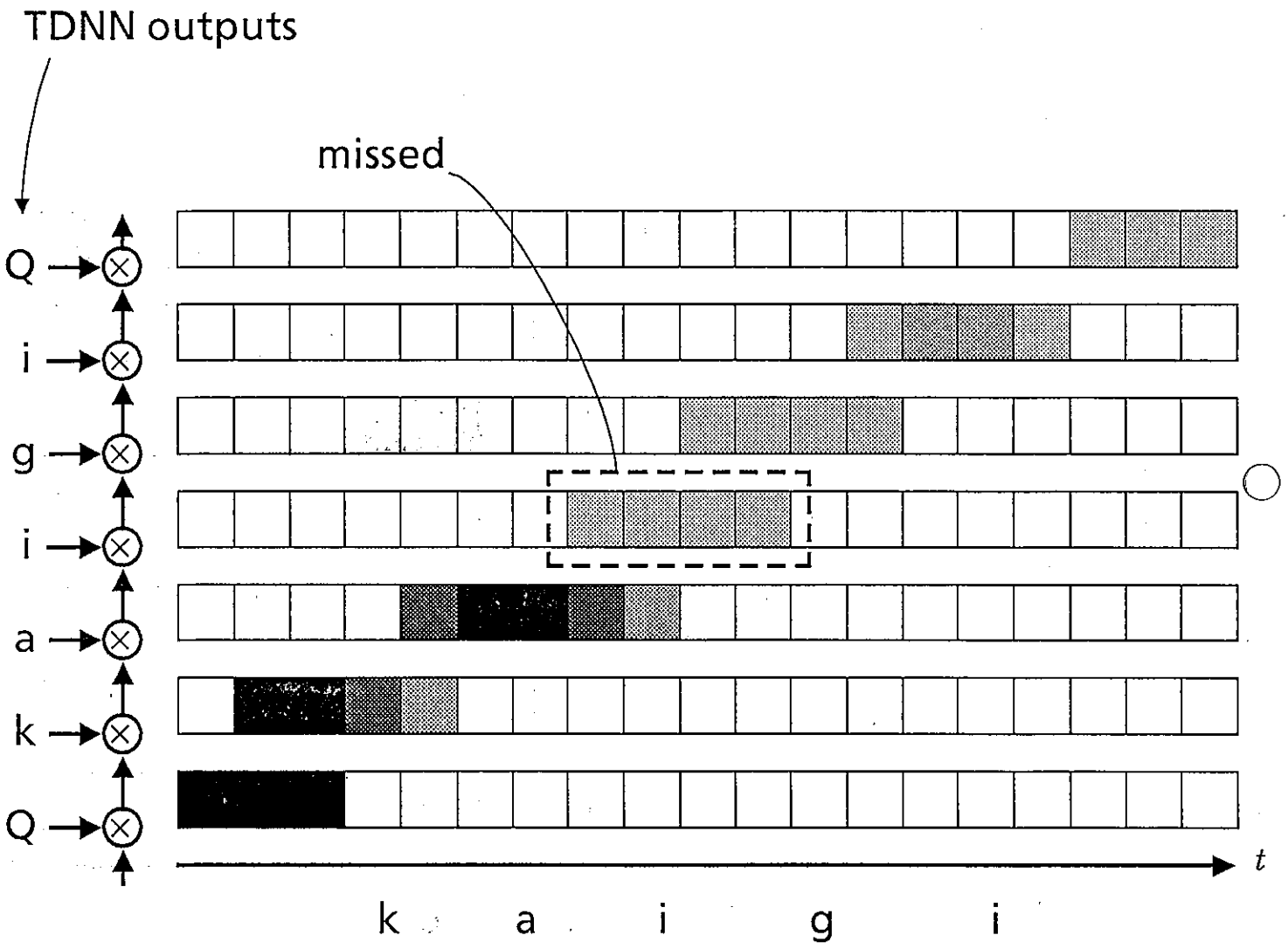


Fig.8 Example of miss detection

Word network (/kaigi/) does not fire with high intensity at the correct word period because its /i/ TDNN was not trained sufficiently.

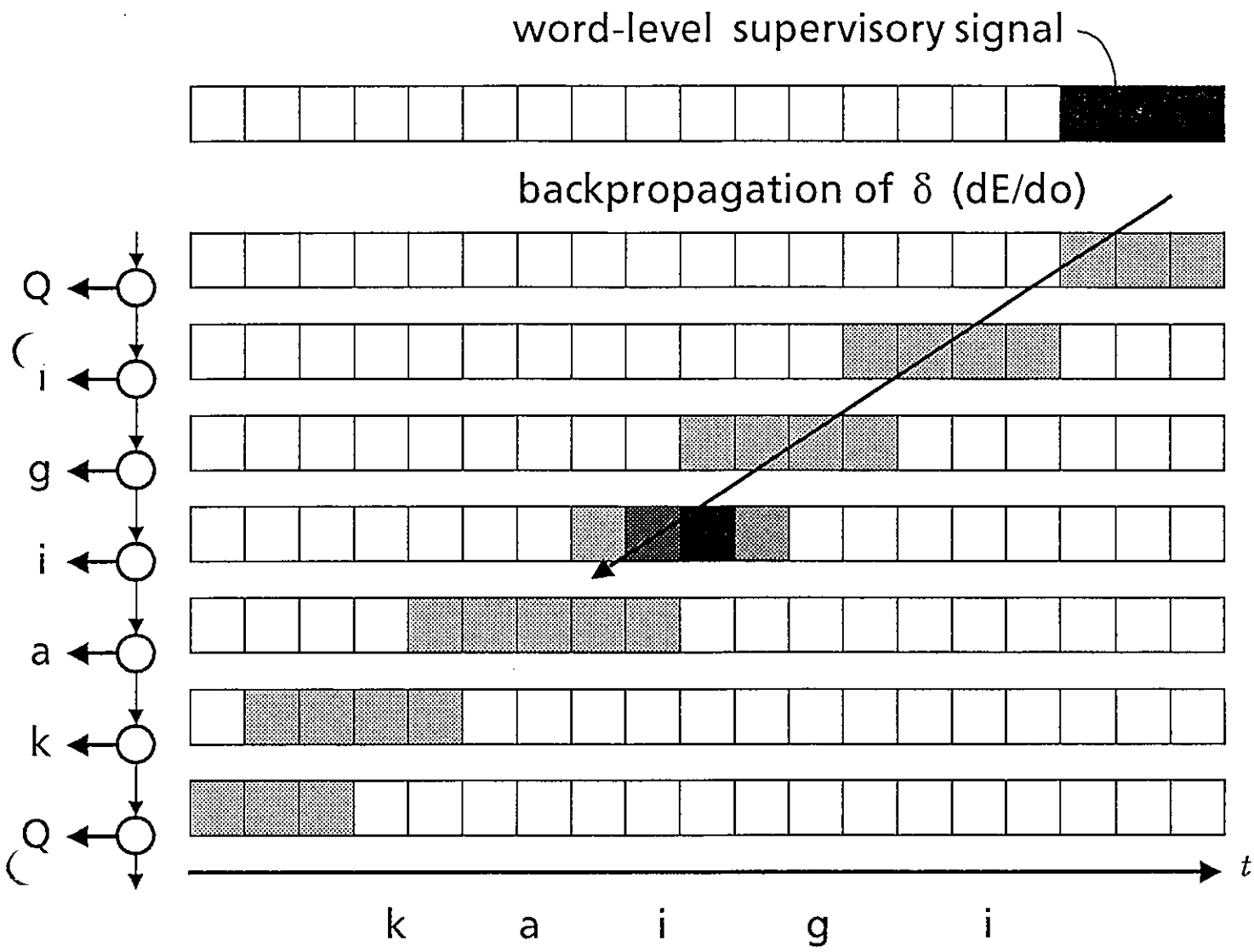


Fig.9 Word-level backpropagation of δ (dE/do)

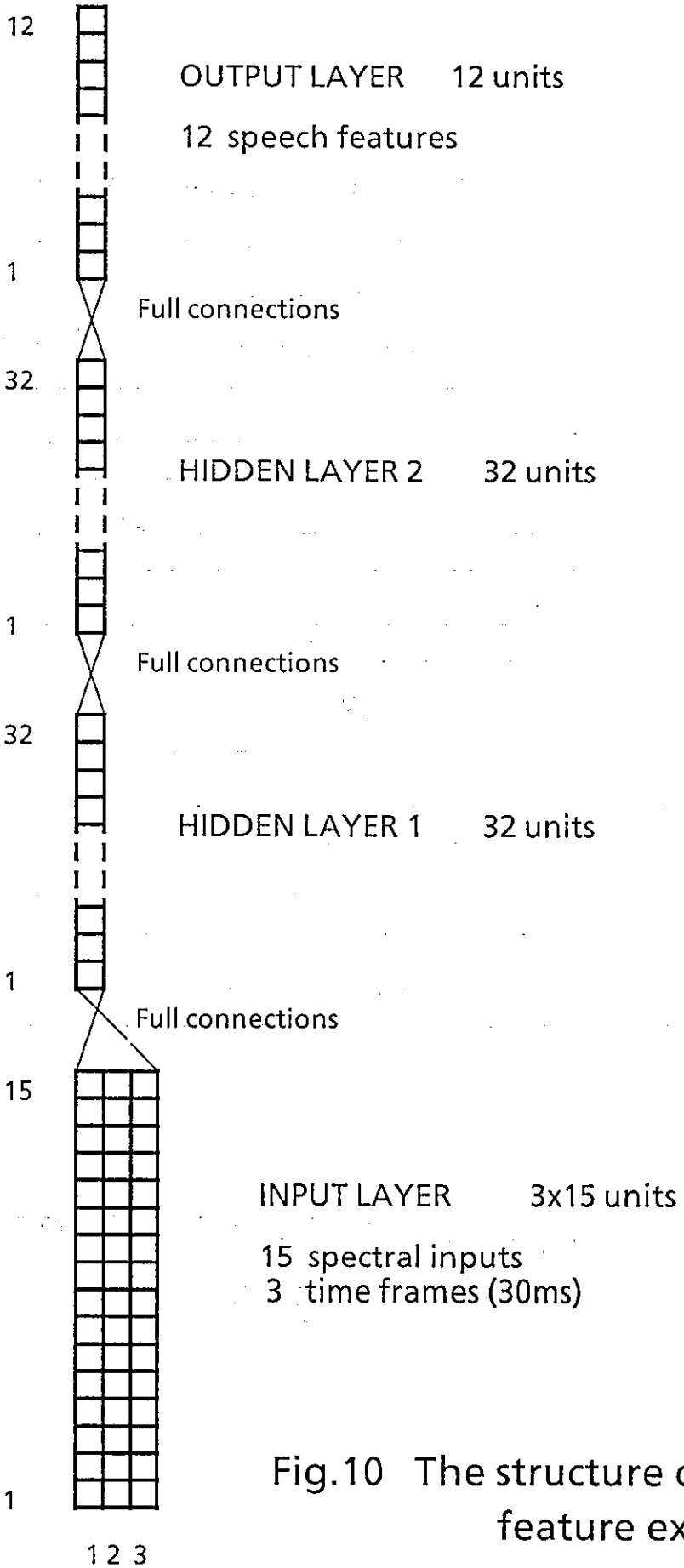


Fig.10 The structure of the TDNN feature extractor

Table 1 Word Recognition Rates (%)

rank	phone training	word training
1	81.8	90.9
≦2	87.3	97.0
≦3	90.9	97.6
≦4	91.5	98.2
≦5	91.5	98.2