SPOKEN DIGIT RECOGNITION USING NEURAL NETWORKS TRAINED BY INCREMENTAL LEARNING

Tetsuya Hoya, Hiroyuki Kamata and Yoshihisa Ishida

Department of Electronics and Communication
Meiji University

ABSTRACT - In this paper, a new method to develop a practical spoken word recognition system is proposed. In order to improve the recognition rate, the incremental training technique is used to re-configure the original network. The experimental results show that the proposed method is worthy of introduction into the system for the improvement on the recognition performance.

INTRODUCTION

In these days, variety use of artificial neural networks (ANNs) in the field of signal processing has been reported. ANNs have also been attracted special interests and applied in speech recognition system. However, the introduction of ANNs in such practical applications as pattern cognition systems still requires a large amount of memory and the iterative training procedure. Recently, the concept of incremental learning has been studied as one of the possible analogous learning manners of actual brain cells. In this paper, we propose a new approach to recognize correctly the wrongly classified patterns by employing the technique of incremental learning.

THE PROCEDURE OF INCREMENTAL LEARNING

Incremental learning is a technique that makes it capable of reducing the training effort without spoiling the old information or retraining the networks over the whole training set whenever a new category is added. According to the paper of Ramani[1], in a three-layered network the hidden layer is the one that characterizes the invariant relationship between the input and output patterns; this relationship should be held and is of essential even if a new category is added to the classification problem.

We introduced the idea of which a wrongly classified pattern by the original network is regarded as a new category to the speech recognition system, and when this occurs a new output unit is added in the output layer. This approach can make sense provided that the problem is small enough not to collapse the existing weight vector spaces, namely, enough to preserve the current recognition performance of the system without any deterioration.

At the stage of re-configuration of the original network, the weight vector between the input and hidden layers should thus be remained the same and only the weight vector between the hidden layer and the new output unit needs to be trained. In Fig. 1, the re-configured network which correctly classifies the two wrongly classfied patterns by the original network is shown.

Fig. 2 shows the speech processing steps. The speech signal is sampled at 10KHz, and sub-divided into 16 frames. Both the beginning and end points of the speech signal are determined by the fixed threshold level. The respective frame data have the time length of 25.6ms, and are pre-processed through an adaptive inverse filter composed of seven-order sub-filters. Nakajima *et al.*[2] have developed such an adaptive filter in order to eliminate the influence of the sound source from the individual utterance and this technique is superior to the pre-emphasis processed through the conventional fixed first-order digital filter. Fig. 3 shows the block diagram of this adaptive filter. In this figure, each sub-filter from Af_1 to Af_4, and Af_6 compensates the slope of speech spectra envelope due to the glottis wave. This filtering operation is expressed by the transfer function:

$$Af_i(z) = 1-\varepsilon_i z^{-r}. \qquad \text{... (1)}$$

The coefficients ε_i of these filters are obtained as follows:

$$\varepsilon_i = \frac{R(r)}{R(0)}, \qquad \text{... (2)}$$

where

r=1 for i=1~4, 6
r=2 for i=5
r=3 for i=7.

$R(\cdot)$ is the autocorrelation coefficient for the input signal to each filter.

In the paper of authors[3], a fast algorithm for the estimation of the adaptive inverse filter, mentioned above, from the autocorrelation coefficients has been developed. Provided that the power spectra of the input sequence of the sub-filter Af_i is written as $P_i(m)$, the corresponding autocorrelation coefficient $R_i(k)$ and $P_i(m)$ constitute the Fourier transform pair:

$$R_i(k) = \frac{1}{N}\sum_{m=0}^{N-1} P_i(m)\exp(j\frac{2\pi mk}{N}), \qquad \text{... (3)}$$

$$P_i(m) = \sum_{k=0}^{N-1} R_i(k)\exp(-j\frac{2\pi mk}{N}). \qquad \text{... (4)}$$

Similarly, provided that the power spectra of the sub-filter Af_i is written as $F_i(m)$, the corresponding autocorrelation coefficients $Q_i(k)$ can be expressed by the following equation:

$$Q_i(k) = \frac{1}{N}\sum_{m=0}^{N-1} F_i(m)\exp(j\frac{2\pi mk}{N}). \qquad \text{... (5)}$$

The autocorrelation coefficient for the input signal to i+1-th order sub-filter is given as:

$$R_{i+1}(k) = \frac{1}{N}\sum_{m=0}^{N-1} P_i(m)F_i(m)\exp(j\frac{2\pi mk}{N}). \qquad \text{... (6)}$$

Substituting Eq.(4) and (5) for Eq. (6) yields

$$R_{i+1}(k) = \sum_{h=0}^{N-1} R_i(h)Q_i(k-h). \qquad \text{... (7)}$$

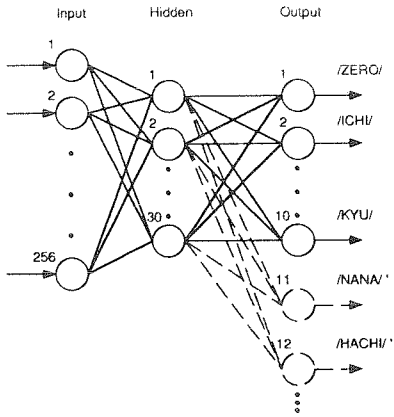

Fig. 1. A three-layered neural network with two additional output units

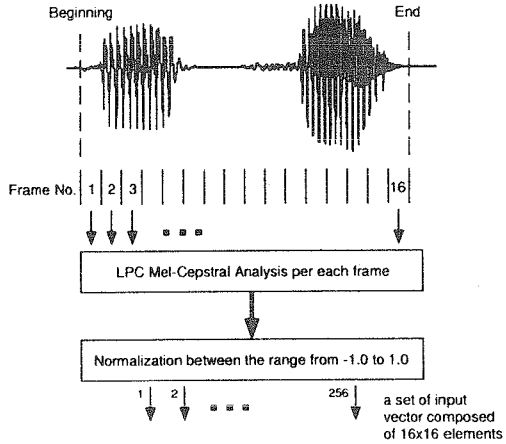

Fig. 2. Speech processing steps from the speech signal to a set of input vector for ANNs

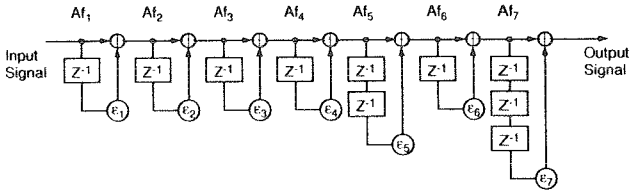

Fig. 3. Adaptive inverse filter composed of seven-order sub-filters

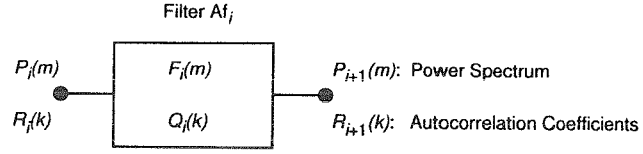

Fig. 4. The sub-filter circuit Af_i

Using the convolution sum with the operator *, we can rewrite Eq. (7) as follows:

$$R_{i+1}(k) = R_i(k) * Q_i(k). \qquad \text{... (8)}$$

Fig. 4 illustrates the circuit of the sub-filter Af_i.

The other way, the power spectrum $F_i(m)$ of the sub-filter is written as:

$$F_i(m) = |Af_i(z)|^2$$

$$= 1 + \varepsilon_i^2 - 2\varepsilon_i \cos\left(\frac{2\pi mr}{N}\right). \qquad \text{... (9)}$$

Secondly, Eq. (5) can be rearranged as follows:

$$\begin{aligned} Q_i(k) &= 1 + \varepsilon_i^2 && \text{for } k = 0 \\ &= -e_i && \text{for } k = \pm r \\ &= 0 && \text{for } k \neq 0, \pm r. \end{aligned} \qquad \text{... (10)}$$

Replacing $Q_i(k)$ in Eq. (8) by its value as given by Eq. (10), the relation between the autocorrelation coefficients , $R_i(k)$ and $R_{i+1}(k)$, of the input and output sequences in the filter Af_i can finally be written in the form of a simple recursive equation as below:

$$R_{i+1}(k) = (1 + \varepsilon_i^2) R_i(k) - \varepsilon_i R_i(|k - r|) - \varepsilon_i R_i(k + r). \qquad \text{... (11)}$$

Fig. 5 shows the architecture for calculating the autocorrelation coefficients by Eq. (11).

After performing the above mentioned pre-emphasis, 16 parameters per frame are obtained by LPC-Mel-Cepstrum analysis, and are normalized between the range from -1.0 to 1.0. Therefore, 16x16 parameters per speech signal are given as a set of input vector of ANNs.

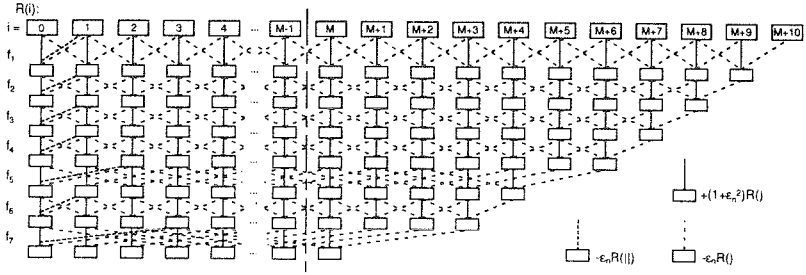

Fig. 5. Architecture for calculation of autocorrelation coefficients by Eq. (11)

EXPERIMENT

In the experiment, we performed spoken digit word recognition from 0 to 9 (i.e. /ZERO/, /ICHI/, /NI/, / SAN/, /YON/, /GO/, /ROKU/, /NANA/, /HACHI/, and /KYU/ corresponding to English digits of /ZERO/, / ONE/, /TWO/, /THREE/, /FOUR/, /FIVE/, /SIX/, /SEVEN/, /EIGHT/, and /NINE/). The corresponding digit data used for training were gathered from 5 male speakers, and the data used for recognition were collected from other 5 male speakers. Both data sets were recorded in an ordinary computer room of our laboratory.

A three-layered network in the original form has 256 units in the input, 30 units in the hidden, and 10 units in the output layer, respectively as shown in Fig. 1(by solid line). Two additional output units were required in this experiment corresponding to the same number of the wrongly classfied patterns by the original network. The elements of the additional weight vectors between the hidden layer and the new-created output unit have initially random values, and is trained by the well-known Back-Propagation

Table 1. Confusion matrix for speaker-independent word recognition by the conventional method

In \ Out	0	1	2	3	4	5	6	7	8	9	Average
0	100										
1		100									
2			100								
3				100							
4					100						98.7%
5						100					
6							100				
7								93.3		6.7	
8		6.7							93.3		
9										100	

Table 2. Confusion matrix for spoken word recognition by the re-configured network

In \ Out	0	1	2	3	4	5	6	7	8	9	Average
0	100										
1		100									
2			100								
3				100							
4					100						100.0%
5						100					
6							100				
7								100			
8									100		
9										100	

Algorithm[4] with the momentum method[5][6]. In Table 1 and 2, the experiments result are summarized.

CONCLUSION

This paper presents a novel method to improve the recognition rate using incrementally trained neural networks. However, future work includes the consideration on the case that the problem is drastically changed, and therefore the structure of the network comes to be more complicated. We hope that the proposed method gives a good chance to take a further step forward the development of new spoken word recognition system.

REFERENCES

[1] N. Ranami, "Incremental Learning Using Hidden Layer Activations - Test On Fish Identification Data", Proceedings of IJCNN'92, Vol. 1, p640-645, June 1992.

[2] T. Nakajima, T. Suzuki, H. Ohmura, S. Ishizaki and K. Tanaka, "Estimation of vocal tract area function by adaptive deconvolution and adaptive speech analysis system", The Jounal of the Accoustical Society of Japan, Vol. 34, No. 3, P. 157, 1978 (in Japanese).

[3] H. Kamata, Y. Ishida and Y. Ogawa, "High speed estimation of vocal tract area function using digital signal processor", Trans. IEE Japan, Vol. 110-D, No. 7, P. 773, 1990.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation", Parallel Distributed Processing, Vol. 1, MIT Press, 1986.

[5] J. M. Zurada, "Introduction To Artificial Neural Systems", West Publishing Co. Inc., 1992.

[6] K. Nakajima, K. Inuma, and Neuron net group, "Neuro Computer", Gijutsu-Hyoron Co. Inc., 1989.