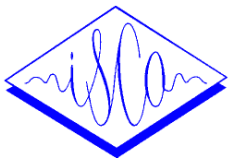

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

# SPEAKER INDEPENDENT SPEECH RECOGNITION BASED ON NEURAL NETWORKS OF EACH CATEGORY WITH EMBEDDED EIGENVECTORS

Yasuyuki Masai, Hiroshi Matsu'ura, Tsuneo Nitta

Information & Communication Systems Laboratory, TOSHIBA Corporation
70,Yanagi-Cho,Saiwai-Ku,Kawasaki,210,Japan

## ABSTRACT

This paper describes a speaker independent word recognition algorithm which is based on four layered neural networks with embedded eigenvectors. Eigenvectors in the Subspace Method (SM) are used as weights. In the SM, the accumulation of projection component values from an input pattern is used as a measure of similarity. In contrast to this, our proposed method utilizes each projection component value to achieve performance better than that of the SM. We propose the Subspace Training (SST) algorithm with the SM and the Decision Controlled Back Propagation Training (DCBPT) algorithm to reduce training times.

Training and recognition experiments were performed using a 26 word vocabulary consisting of train station names. The error rate of the SM was 1.3%. The error rate was reduced to 0.7% using the neural networks combined with the SM.

## 1 INTRODUCTION

Authors have been researching into speaker independent word recognition methods using the Subspace Method (SM) [1].

The SM effectively incorporates pattern fluctuations into an eigenvector set $\phi m$ through the use of the K-L transform. The similarity $S^{(k)}$ between the eigenvector set $\phi m^{(k)}$ of category K and a normalized input pattern $x$ is as follows:

$$S^{(k)} = \sum_{m=1}^{M} (x \cdot \phi m^{(k)})^2 \qquad (1)$$

where $(\cdot)$ denotes inner product and M is the number of eigenvectors. It is difficult to reject words outside the vocabulary for the SM because there is very little difference in similarity between category k and others. Dealing with this issue, the authors proposed a method which used a different rejection level for each category [2]. This issue arises because each projection component value is not directly used for recognition in the SM.

Recently, several interesting results based on neural networks applied to pattern recognition tasks have been published [3,4]. However, two major problems arise from applying adopting neural networks to speech recognition tasks. The problems are the reduction of training times and the decrease of recognition accuracy caused by an increase of vocavulary. The former can be examined in various ways such as from a network structural approach or an algorithmic approach [5]. For the latter , a neural network constructed in a modular fashion [6] and spotting CV-syllables using neural networks that discriminate between single CV-syllables [7] were proposed.

Dealing with these issues, authors propose 1) the use of four layered neural networks of each category with embedded eigenvectors to evaluate each projection component value, and 2) combining the neural networks with the SM that evaluates accumulation of projection component values.

## 2 NEURAL NETWORKS WITH EIGENVECTORS AND SYMMETRICAL NONLINEAR FUNCTION

Eigenvectors derived from training patters uttered by a number of speakers characterize the patterns. Fig.1 shows an example of an eigenvector set of the monosyllable /cjo/. The symbol " ● " indicates positive value of eigenvectors, and the symbol "⊛" indicates negative value. These values are biased for the expression. The first axis $\phi 1$ shows average of training patterns uttered by a large number of speakers. We can not distinguish the monosyllable /cjo/ from the monosyllable /sjo/ using only the first axis $\phi 1$ in which features of affricate are out of focus. The second axis $\phi 2$ shows fluctuation in the time direction. The third axis $\phi 3$ shows the affricate from which we can distinguish /cjo/ from /sjo/. The forth axis $\phi 4$ shows fluctuation in the frequency direction. Characteristics of other axes not shown in Fig.1 are not as clear as that of first four axes. From this example, we can see that weighting the axes differently with the similarity values rather than equally will allow /cjo/ to be better distin-

guished from /sjo/. We propose the neural networks to calculate the weights.

The neural networks are constructed for each category. The structure of the networks is shown in Fig.2. There are two units in the output layer. A unit $S_S^{(k)}$ is the accumulation of projection component values in the SM. Another unit $S_N^{(k)}$ is a measure of similarity that is derived from each projection component value, i.e., the output of a three layer neural network that uses each projection component value as input to the first hidden layer. Input signals to the neural networks are parameters of input speech and the number of input units is 256. We use eigenvectors obtained by the Subspace Training (SST) algorithm [1,8] as weights of the first hidden layer ($N_2$ = 15 units). Weights of the second hidden layer ($N_3$ = 17 units) and the output layer are calculated using the Decision Controlled Back Propagation Training (DCBPT) algorithm. Symmetrical nonlinear function is used as an activation function of the first hidden layer, and sigmoid functions are used for the second layer and the output layer.

Usually, sigmoid functions are used as activated functions of neural networks. However, sigmoid functions are not suitable when eigenvectors are used as weights. The neural networks represent only a one directional pattern fluctuation, as the range of sigmoid functions is not activated for negative values. For example, we consider an inner product between an input pattern and the fourth axis $\phi 4$ in Fig.1. The inner product takes a positive value when the input pattern is shifted to the left, and takes a negative value when the input pattern is shifted to the left. Therefore, it is difficult to recognize input patterns which fluctuate to the left of the time direction in neural networks with sigmoid functions. We should use a symmetrical function as the activation function of the first hidden layer. An example of the symmetrical function is as follows:

$$yj = | xj - (sin 2\pi xj) / 2\pi | \quad (2)$$

where x ($-1 \leqq x \leqq 1$) is the inner product between input pattern and eigenvector, and $|\ \ |$ denotes absolute value.

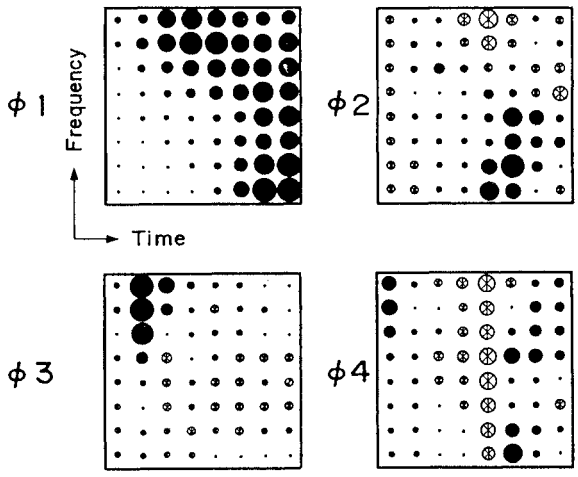


Fig.1 Example of eigenvector set (Monosyllable /cjo/)

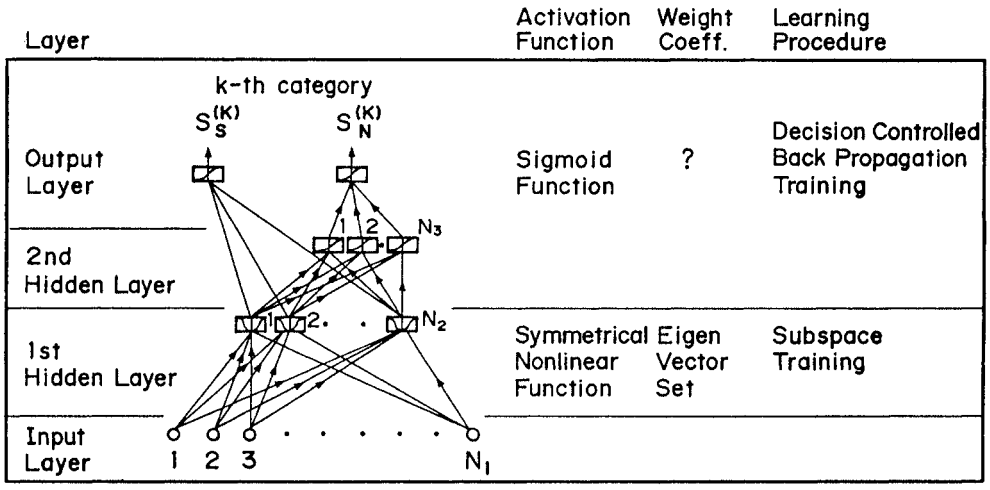


Fig. 2 Neural network structure

### 2.1 SST ALGORITHM

The SST algorithm is used to derive an eigenvector set from a correlation matrix. The algorithm is as follows.

**STEP 1** A correlation matrix R is calculated from training pattern set {fi} for each category k. An eigenvector set is derived from these correlation matrices through the use of the K-L transform.

**STEP 2** Training pattern set is recognized using eigenvector set.

**STEP 3** The correlation matrix R is updated using the training patterns which generate an error signal in STEP 2 as follows:

$$R'^{(k)} = R^{(k)} + \sum_{i=1}^{I} \mu i \; fi \; fi^{T} \qquad (3)$$

where $\mu i$ is a coefficient, I is the number of training patterns which generate the error signal for category k, and T denotes transposition.

**STEP 4** The eigenvector set is derived from updated correlation matrix R through the use of the K-L transform.

**STEP 5** STEP 2, 3 and 4 are iterated until the recognition score no longer increases.

### 2.2 DCBPT ALGORITHM

We proposed the DCBPT algorithm to reduce training times of Back Propagation algorithm [9]. In this algorithm, each unit in the input layer is initially connected to units in the first hidden layer by links having fixed weights corresponding to each eigenvector that is obtained using the SST algorithm. Adjusting weights on other links consists of feeding a part of the training patterns to the networks and then allowing the network to generate error values which are back-propagated to lower layers. This process is iterated until the recognition score no longer increases. The weights in the networks at this point are then used as initial values throughout the networks for further training on complete sets of training patterns. Only training patterns which generate error signals are learned by the networks. This process is shown as follows.

**STEP 1** The weights of the first hidden layer are fixed and other weights are trained on a part of training patterns with the Back Propagation algorithm.

**STEP 2** STEP 1 is iterated until the recognition score no longer increases for test patterns.

**STEP 3** The weights obtained in STEP 1 and 2 are used for recognition of complete set of training patterns.

**STEP 4** Only training patterns which generate error signals are used to adjust weight with Back Propagation algorithm.

**STEP 5** STEP 3 and 4 are iterated until the recognition score no longer increases for test patterns.

## 3 EXPERIMENTAL RESULTS

We performed speaker independent isolated word recognition experiments with the SM and the neural networks that we proposed.

### 3.1 EXPERIMENTAL CONDITIONS AND DATABASE

Training and recognition experiments were performed using a 26 word vocabulary consisting of train station names. The words were spoken once by 300 male and 300 female speakers. From this group, utterances from 250 male and 250 female speakers were used for training and the rest were used for recognition test.

The input speech signals are sampled at 12 kHz. Spectrum parameter are calculated as outputs of a 16 channel filter bank every 8 msec using a frame of 16 msec. Test data was spoken under noisy environment (65～75 dBA). Utterances that caused word boundary detection error were 1.5% and were excepted from the test patterns.

### 3.2 SUBSPACE METHOD

The recognition test was performed using an eigenvector set that was calculated using the SST algorithm. Similarity value between input pattern and the eigenvector set of each category is calculated using equation (1). The result is shown using the symbol "×" in Fig.3. The SST algorithm was iterated five times, and the lowest error rate was 1.3%.

### 3.3 NEURAL NETWORKS

First, the eigenvector set obtained in section 3.2 is fixed as weights of the first hidden layer. Next, the weights of other layers are modified through training on 100 patterns for each category out of training patterns with the Back Propagation algorithm (STEP 1 and 2 of the DCBPT algorithm). The recognition results of the neural networks as described above are shown using the symbol "△" in Fig.3. The error rate does not decrease to less than 5.2% at about 300 itera-

tions. Next, the weights were adjusted with the Back Propagation algorithm using training patterns which generated error signals (STEP 3,4 and 5 of the DCBPT algorithm). The results are shown using the symbol "▲" in Fig.3. The error rate dropped to 2.8% with additional training. The error rate of a neural network using only similarity value $S_N^{(k)}$ in Fig.1 is twice the error rate of the SM.

### 3.4 NEURAL NETWORKS COMBINED WITH SUBSPACE METHOD

We describe a method using two similarity values, $S_S^{(k)}$ and $S_N^{(k)}$ for recognition. In this method, evaluation is performed in two steps. Best score $S_{S1}$ and second score $S_{S2}$ of the similarity value $S_S^{(k)}$ are used for the first evaluation. The category of best score $S_{S1}$ is selected as recognition result, if this score is satisfied with one or both of the conditions that follows:

$$S_{S1} > S_0$$
$$S_{S1} - S_{S2} \geq \Delta S \qquad (4)$$

where $S_0$ and $\Delta S$ are threshold values. In another case, one of the two categories is selected according to the similarity value $S_N^{(k)}$. The result is shown using the symbol "○" and "●" in Fig.3 ($S_0$ = 0.88, $\Delta S$ = 0.33). The neural networks that were trained with the Back Propagation algorithm performed with an error rate of 0.9% , and the error rate was dropped to 0.7% with additional training with the DCBPT algorithm. The method we proposed above achieved half of the error rate of the SM. The results arise from the use of each projection component value in addition to the accumulation of these projection component values.

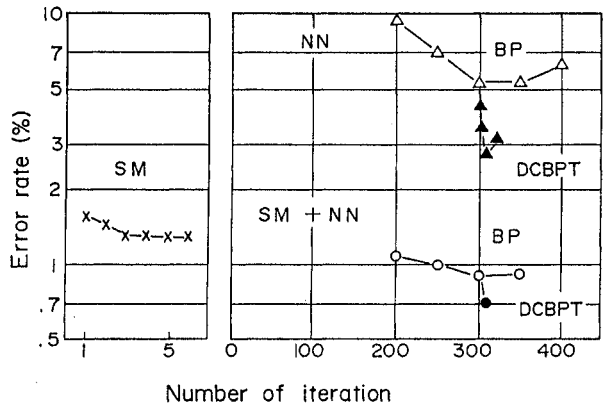


Fig.3 Error rate vs. the number of iteration

## 4 CONCLUSION

A speaker independent word recognition method which is based on four layered neural networks with embedded eigenvectors has been proposed. This method directly utilizes each projection component value from an input pattern in addition to the accumulation of these projection component values to achieve a performance better than that of the SM. The SST algorithm and the DCBPT algorithm were proposed to reduce training times of neural networks. Good performance was achieved on a speaker independent word recognition task.

We plan to apply these neural networks with embedded eigenvectors to rejection of the words outside the vocabulary.

## REFERENCES


[1] E.Oja : " Subspace Method of Pattern Recognition ", Research Studies Press, 1983.

[2] N.Sugi, T.Nitta, S.Kido, A.Nakayama, T.Shimada and T.Nishimura : " Development of Speaker Independent Word Recognition System for Ticket Vending Machine", IEICE Technical Report, SP89-24, pp.1-8, June 1989.

[3] D.E.Rumelhart and J.L.McClelland , " Parallel Distributed Processing; Explorations in the Microstructure of Cognition ", Volume 1 and 2 , MIT Press, Cambridge, MA, 1986.

[4] T.Kohonen : " Statistical Pattern Recognition with Neural Networks : Benchmarking Studies ", IEEE Proc. of ICNN, Vol.I pp.61-68, July 1988.

[5] D.S.Broomhead and D.Lowe : "Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks, RSRE Memorandum No.4148, March 1988.

[6] A.Waibel, H.Sawai and K.Shikano : " Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks" , IEEE Proc. of ICASSP-89, pp.112-115, May 1989.

[7] H.Sawai, A.Waibel, M.Miyatake and K.Shikano: "Spotting Japanese CV-Syllables and Phonemes Using Time-delay Neural Networks", IEEE Proc. of ICASSP-89, pp.25-28, May 1989.

[8] T.Nitta, K.Uehara and S.Watanabe : "Connected Word Recognition Based on Word Transition Network and Selective Scoring of Phonetic segments ", TEICE Trans. , Vol.J71-D No.9, pp. 1640-1649 Sept. 1988.

[9] T.Nitta : " Pattern Recognition using Neural Nets Based on Subspace Training and Back Propagation Training Algorithm ", Proceedings of the Spring Meeting of the Acoustical Society of Japan 1-6-15, March 1989.