Pattern Recognition Letters 7 (1988) 135–144
North-Holland

March 1988

Dynamic guard zone for self-supervised learning

S.K. PAL, A. PATHAK and C. BASU

Electronics and Communication Sciences Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700035, India

Received 19 February 1987

Abstract. The dimension of the guard zone along with its bounds for the Generalised Learning Algorithm (Pathak and Pal, 1986) is determined for optimum learning. The dimension is found to be dynamic depending on the input sequence and the current estimates of classification parameters. Incorporation of this higher-order knowledge in a supervisory program improves the system performance. The performance is again found to be affected if the guard zone is shrunk/expanded for 'very weak'/'not too weak' estimates when speech data is considered to be input.

Key words. Dynamic guard zone, learning, speech recognition.

1. Introduction

A Generalised Guard-zone Algorithm (GGA) was described recently by Pathak and Pal (1986) for learning class parameters using a restricted updating program along with investigation of its stochastic convergence for optimum learning. Basically, it aims to detect outliers and reject them from the parameter-updating procedure. The algorithm is a generalisation of some existing ones (Chien, 1970; Pal et al.,1980, Pal, 1982) which were found to be useful for practical data.

The present work is a continuation of the GGA. It attempts mainly (i) to determine the dimension and bounds of the guard zone for optimum learning, (ii) to study the adaptive efficiency of the system in recognising a pattern with deliberately chosen poor (non-appropriate) estimates of the parameters representing classes, (iii) to study the effect of dimension of the guard zone on the system performance, and (iv) to investigate the effect of 'dynamic behavior' (higher-order knowledge based on input sequence) of the guard zone in acting as a supervisor on the decision of the classifier.

Two types of poor estimates, namely, 'very weak' and 'not too weak' are considered in order to study the adaptive efficiency of the system when Bayes' maximum likelihood ratio (Tou and Gonzalez, 1974) is considered as a classification tool. The merit of the dynamic property (over the fixed value) of the guard zone in a self-supervisory program is demonstrated for different input sequences. The results are also compared with those of fully-supervised case, when speech data in CNC (Consonant – Vowel Nucleus – Consonant) context is considered as input.

2. The Generalized Guard-zone Algorithm (GGA) (Pathak and Pal, 1986)

Let

$$X = [x_1, x_2, \ldots, x_N]', \quad X \in \mathbb{R}^N,$$

be an N-dimensional feature vector defined over a pattern class C.

Let us make the following assumptions:

(A1) The distribution of X over C is continuous.

(A2) This distribution depends on a q-dimensional parameter vector θ, some or all of which need to be learned.

(A3) The distribution of X over C is such that $E(X)$ exists and is equal to μ.

(A4) the dispersion matrix of $\boldsymbol{X}$, namely,

$$\operatorname{Disp}(\boldsymbol{X}) = \boldsymbol{\Sigma} = [(\sigma_{ij})] \quad \text{exists.}$$

Before stating the algorithm itself, let us define a guard-zone formally as follows:

Definition. Let S be a metric space and δ a metric defined on it. Then for any point $\boldsymbol{a} \in S$, a guard-zone $G(\boldsymbol{a}, \lambda)$ having an 'extent' λ is the subset of S defined by

$$G(\boldsymbol{a},\lambda) = \{\boldsymbol{X} : \delta(\boldsymbol{a},\boldsymbol{X}) \leq \lambda\}, \quad \text{where } \lambda \geq 0. \qquad (1)$$

Clearly, $G(\boldsymbol{a}, \lambda)$ is nothing but a closed ball in S, with radius λ, centered at $\boldsymbol{a}$ with respect to the metric δ. Let $S = \mathbb{R}^N$ and a metric d be defined as

$$d^2(x,y) = (x-y)'A(x-y), \quad x,y \in \mathbb{R}^N,$$

$\boldsymbol{A}$ being a symmetric, positive definite matrix. Then we proceed to the algorithm as follows:

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots$ be the sequence of learning (or training) samples, randomly selected from C, that is, assumed to be independently and identically distributed. We restrict ourselves to the case where $\boldsymbol{\theta}$ includes $\boldsymbol{\mu}$ and/or elements of $\boldsymbol{\Sigma}$ only.

The generalized guard-zone algorithm (GGA) for estimating $\boldsymbol{\theta}$ recursively is as follows:

$$\hat{\boldsymbol{\theta}}_{(t)} = \begin{cases} f(\boldsymbol{X}_{(t)}) & \text{for } t = 1, \\ \hat{\boldsymbol{\theta}}_{(t-1)} - a_{(t)} \boldsymbol{Y}_{(t)} & \text{for } t > 1; \end{cases} \qquad (2)$$

$$\boldsymbol{Y}_{(t)} = \begin{cases} \hat{\boldsymbol{\theta}}_{(t-1)} - f(\boldsymbol{X}_{(t)}) \\ \quad \text{if } \boldsymbol{X}_{(t)} \in G(\hat{\boldsymbol{\mu}}_{(t-1)}, \lambda_{(t)}), \\ 0 \quad \text{otherwise;} \end{cases} \qquad (3)$$

$\hat{\boldsymbol{\theta}}_{(t)}$: the t-th stage estimate of $\boldsymbol{\theta}$;

$\{a_{(t)}\}$: a sequence of positive numbers, with $a_{(t)} \leq 1, \forall t$;

f : $\mathbb{R}^N \to \mathbb{R}^q$ is a continuous mapping, defining an unbiased statistic for $\boldsymbol{\theta}$;

$\hat{\boldsymbol{\mu}}_{(t-1)}$: the $(t-1)$-th stage GGA estimate of $\boldsymbol{\mu}$;

$G(\hat{\boldsymbol{\mu}}_{(t-1)}, \lambda_{(t)})$
$= \{\boldsymbol{X} : \boldsymbol{X} \in \mathbb{R}^N, d_{(t)}(\boldsymbol{X}, \hat{\boldsymbol{\mu}}_{(t-1)}) \leq \lambda_{(t)}\}$;
$d^2_{(t)}(x,y) = (x-y)'A_{(t)}(x-y)$;

$\boldsymbol{A}_{(t)}$: a symmetric, positive definite matrix, which may or may not be a function of $\boldsymbol{X}_i$ and/or $\hat{\boldsymbol{\theta}}_i$, $i = 1(1)t$;

$\lambda_{(t)}$: a positive number, prespecified.

In essence, what this algorithm does is use only those training samples for updating the estimate, which lie within the corresponding guard-zone centred at the preceding estimate of the mean. Training samples which lie outside it are ignored and the estimate is kept unchanged at the corresponding stages. Such a region (guard-zone) therefore forms the basis of a supervisory program.

The convergence of the GGA for estimating $\boldsymbol{\theta}$ for different cases is provided by Pathak and Pal (1986) using a stochastic approximation procedure. The earlier algorithms of Chien (1970) and Pal et al. (1980) in the same line have also been found to be formulated from the GGA.

3. Dynamic behavior of the guard zone

It is obvious from the previous discussion that the choice of λ (the dimension of the guard zone) plays a crucial role so far as the estimation of the parameters along with their convergence and classification efficiency are concerned. While it is not a very simple problem to obtain some sort of an optimal value without making additional assumptions, one can obtain certain bounds for λ from the viewpoint of convergence of the class parameters. The size of the guard zone may then be experimentally determined using some linear combination of those bounds.

As seen in our earlier work (Pathak and Pal, 1986), one of the conditions necessarry for having some form of stochastic convergence of the estimates to the true value was

$$p_{(t)} = P[d_{(t)}(\boldsymbol{X}_{(t)}, \hat{\boldsymbol{\mu}}_{(t-1)}) \leq \lambda_{(t)} \mid \hat{\boldsymbol{\mu}}_{(t-1)}] > \delta \qquad (4)$$
$$\text{for some } \delta > 0,$$

i.e., the probability of $d_{(t)}(\boldsymbol{X}_{(t)}, \hat{\boldsymbol{\mu}}_{(t-1)})$ being less than or equal to the dimension of guard zone is strictly greater than zero.

By virtue of the Lemma given below we have

$$d^2_{(t)}(\boldsymbol{X}_{(t)}, \hat{\boldsymbol{\mu}}_{(t-1)}) \geq \pi_{(t)\min} \| \boldsymbol{X}_{(t)} - \bar{\boldsymbol{\mu}}_{(t-1)} \|^2 = l^2_{(t)}, \quad \text{say}, \qquad (5a)$$

$$d^2_{(t)}(\boldsymbol{X}_{(t)}, \hat{\boldsymbol{\mu}}_{(t-1)}) \leq \pi_{(t)\max} \| \boldsymbol{X}_{(t)} - \bar{\boldsymbol{\mu}}_{(t-1)} \|^2 = L^2_{(t)}, \quad \text{say}, \qquad (5b)$$

where $\pi_{(t)\min}$ and $\pi_{(t)\max}$ are respectively the smallest and largest eigenvalues of $\boldsymbol{A}_{(t)}$. As the $\boldsymbol{A}_{(t)}$'s are assumed to be symmetric positive definite, we must have

$$\pi_{(t)\max} \geq \pi_{(t)\min} > 0 \quad \forall t, \tag{6a}$$

$$0 \leq l_{(t)} \leq L_{(t)}. \tag{6b}$$

Now, $\lambda_{(t)}$ can not be less than or equal to $l_{(t)}$ as that would mean that $p_{(t)} = 0$ which violates the condition (4). Also, $\lambda_{(t)}$ can not be greater than or equal to $L_{(t)}$, as that would mean that $p_{(t)} = 1$, which is not desirable because all the samples would then be accepted by the supervisor for updating $\boldsymbol{\theta}$.

Thus, one must necessarily have

$$\sqrt{\pi_{(t)\min}} \, \| X_{(t)} - \hat{\mu}_{(t-1)} \| = l_{(t)} < \lambda_{(t)} < L_{(t)} = \sqrt{\pi_{(t)\max}} \, \| X_{(t)} - \hat{\mu}_{(t-1)} \|. \tag{7}$$

The value of $\lambda_{(t)}$ is therefore found to be bounded between $l_{(t)}$ and $L_{(t)}$ in order to have convergence of the estimates of classification parameters to their true values. From equation (7) it is also interesting to note that the dimension of the guard zone is dynamic (varying) and its value at the t-th stage depends on the $(t-1)$th stage-estimate of mean vector and the value of $\pi_{(t)}$ i.e., the $(t-1)$th stage estimate of the matrix A. This adaptive (expanding–shrinking) behavior of the guard zone $G(\boldsymbol{a}, \lambda_{(t)})$ centred at $\boldsymbol{a}$ enables to accept sometimes a sample having a larger distance from $\boldsymbol{a}$ while discarding another one with smaller distance for parameter updating procedure. This was not the case with the algorithms of Chien (1970) and Pal et al. (1980) where such a parameter was considered to be fixed throughout the learning process. In other words, the supervisory program uses here a higher-order level of knowledge depending on the input sequence.

Having the lower and upper bounds $l_{(t)}$ and $L_{(t)}$ respectively for $\lambda_{(t)}$, we may take their weighted average, namely

$$\lambda_{(t)} = (1-\alpha) l_{(t)} + \alpha L_{(t)}, \quad 0 < \alpha < 1, \tag{8}$$

in order to describe the dynamic behavior of the extent of the guard zone at the t-th stage of learning.

It is to be noted here that condition (7) is violated in case the matrix $A_{(t)}$ is a scalar multiple of the identity matrix for any t. This type of situation is also explained in Section 5.1.

Lemma. *If A is a symmetric matrix of order p then*

$$\pi_p \leq \frac{\boldsymbol{x}' A \boldsymbol{x}}{\boldsymbol{x}' \boldsymbol{x}} \leq \pi_1, \quad \boldsymbol{x} \in \mathbb{R}^p,$$

where π_p and π_1 are respectively the smallest and largest of the p roots of the equation

$$|A - \pi \boldsymbol{I}| = 0,$$

that is, they are respectively the smallest and largest eigenvalues of A (both non-negative). $\boldsymbol{I}$ denotes the identity matrix of order p.

4. Maximum likelihood classifier

In order to demonstrate the effectiveness of the GGA in discarding doubtful (unreliable) samples from estimating parameters, we have considered Baye's maximum likelihood classifier (Tou and Gonzalez, 1974) for taking decision on an unknown pattern $\boldsymbol{X}$. For an m-class pattern recognition problem, let

μ_j	: the true mean vector for the class C_j, $j = 1(1)m$;
Σ_j	: the true dispersion matrix for C_j;
$P(C_j)$	: the apriori probability for C_j;
$p(X\|C_j)$	: the probability density of X in C_j.

Then the decision rule is: Classify $\boldsymbol{X}$ into C_k iff

$$P(C_k)p(X|C_k) > P(C_j)p(X|C_j) \quad \text{for } j \neq k;\ j,k = 1(1)m. \tag{9}$$

If $P(C_j) = 1/m$ for all j and the conditional density of $\boldsymbol{X}$ given C_j is assumed to be normal $\mathrm{N}(X;\ \mu_j, \Sigma_j)$ i.e., if

$$p(X|C_j) = (2\pi)^{-N/2} |\Sigma_j|^{-1/2} \exp[-\tfrac{1}{2}(X-\mu_j)'\Sigma_j^{-1}(X-\mu_j)] \tag{10}$$

then the decision rule is, decide $X \in C_k$ iff

$$D_k(X) = \max_j \{D_j(X)\}, \quad j, k = 1(1)m, \tag{11}$$

where

$$D_j(X) = \ln|\Sigma_j| + (X-\mu_j)'\Sigma_j^{-1}(X-\mu_j). \tag{12}$$

ln stands for natural logarithm.

5. Decision parameter of the supervisor

As explained before, after having an unknown sample $X_{(t)}$ classified by the maximum likelihood classifier, the next task is to check whether $X_{(t)}$ falls within the guard zone (acting as supervisor) centred at the estimated mean $\hat{\mu}_{j(t-1)}$ of the recognised class. This is determined by the decision parameter of the supervisor which is defined for the j-th class as

$$(\mathrm{DPS})_{j(t)} = \sum_{n} [(x_{nj(t)} - \hat{\mu}_{nj(t-1)})/\hat{\sigma}_{nj(t-1)}]^2 \quad (13)$$
$$n = 1(1)N; j = 1(1)m.$$

where $\hat{\sigma}_{nj(t-1)}$ denotes the $(t-1)$th estimate of the n-th component of standard deviation of j-th class. It is clear that this is a special case of GGA (Section 2) when

$$\boldsymbol{\theta}' = [\boldsymbol{\mu}' \vdots \boldsymbol{\varphi}'], \quad (14a)$$

$$\boldsymbol{\varphi}' = [\sigma_{11}\sigma_{12}\cdots\sigma_{1N}\sigma_{22}\sigma_{23}\cdots\sigma_{2N}\cdots \sigma_{(N-1)(N-1)}\sigma_{(N-1)N}\sigma_{NN}], \quad (14b)$$

$$q = N + N(N-1)/2 = N(N+1)/2, \quad (14c)$$

$$\boldsymbol{A} = [\mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2, \ldots, \sigma_N^2)]^{-1}, \quad (14d)$$

with $\sigma_n^2 = \sigma_{nn}$ (the variance along the n-th component) and

$$a_t = 1/t. \quad (14e)$$

The supervisor then accepts the decision made by the classifier that $X_{(t)}$ is from the k-th class only if

$$(\mathrm{DPS})_{k(t)} \leq \lambda^2_{k(t-1)}, \quad (15)$$

$\lambda_{k(t-1)}$ being the dimension (extent) of the guard zone for the k-th class at the $(t-1)$th stage. The parameter $\boldsymbol{\theta}$ (equation 14(a)) representing the mean and the co-variance matrix for the k-th class and the dimension of the guard zone λ for the k-th class are then correspondingly updated for that input $X_{(t)}$. Otherwise the decision is considered to be doubtful and no other alteration of the components of $\boldsymbol{\theta}_k$ and λ_k is made.

It is to be mentioned here that this decision parameter would lead to hyper-ellipsoidal shapes of the guard zones. Since the system does not need any additional information and uses the inherent properties of the distribution of only a subset of the parameters as used by the classifier itself, it may be called a *'self-supervisory system'*.

5.1. The bounds for $\lambda_{(t)}$

Under the conditions given in equations (14), the bounds for $\lambda_{(t)}$ (equation 8) for a class will become

$$l_{(t)} = \| \boldsymbol{X}_{(t)} - \hat{\boldsymbol{\mu}}_{(t-1)} \| / \hat{\sigma}_{(t-1)\max}, \quad (16a)$$

and

$$L_{(t)} = \| \boldsymbol{X}_{(t)} - \hat{\boldsymbol{\mu}}_{(t-1)} \| / \hat{\sigma}_{(t-1)\min}, \quad (16b)$$

where $\hat{\sigma}_{(t-1)\max}$ and $\hat{\sigma}_{(t-1)\min}$ denote respectively the largest and smallest values among the $(t-1)$th estimates of N standard deviation components in a class.

From the above equations it is seen that if for a particular class we have further

$$\hat{\sigma}_{11(t)} = \hat{\sigma}_{22(t)} = \cdots = \hat{\sigma}_{nn(t)} = \cdots = \hat{\sigma}_{NN(t)},$$

then $\hat{\sigma}_{(t)\max} = \hat{\sigma}_{(t)\min}$, $\boldsymbol{A}_{(t)}$ becomes a scalar multiple of the identity matrix and $l_{(t)} = L_{(t)} = \lambda_{(t)}$.

But this (as described in Section 3) violates the requirement $l_{(t)} < \lambda_{(t)} < L_{(t)}$. This also leads to $p_{(t)} = 1$ and condition (15) will therefore always be satisfied for updating the parameters of that class. In other words, the GGA becomes totally ineffective in acting as a supervisor.

5.2. Iterative algorithm for parameter estimation

In general, the input events which are to be classified are in a somewhat randomly mixed sequence. These samples after being classified and accepted by guard zone become members of certain classes and modify the centres, dispersions (variances and covariances) and guard zone dimensions of them.

Let $\hat{\mu}_{n(t)}$, $\hat{\sigma}_{ij(t)}$ and $\lambda_{(t)}$ represent the n-th component of the mean, the (i,j)th element of the covariance matrix, and the extent of the guard zone, respectively estimated by first t samples in a class. Then after the addition of another sample $X_{(t+1)}$ these parameters would be adjusted as follows:

$$\hat{\mu}_{n(t+1)} = \frac{t}{t+1}\hat{\mu}_{n(t)} + \frac{1}{t+1}x_{n(t+1)}, \tag{17a}$$

$$\hat{\sigma}_{ij(t+1)} = \frac{1}{t+1}C_{ij(t+1)} - \hat{\mu}_{i(t+1)}\hat{\mu}_{j(t+1)}, \tag{17b}$$

$$C_{ij(t+1)} = C_{ij(t)} + x_{i(t)}x_{j(t)}, \tag{17c}$$

$$C_{ij(t)} = \sum_q x_{iq}x_{jq}, \quad q = 1(1)t, \tag{17d}$$

$$\lambda_{(t+1)} = (1-\alpha)l_{(t+1)} + \alpha L_{(t+1)}, \\ 0 < \alpha < 1, \tag{17e}$$

$$l_{(t+1)} = \| X_{(t+1)} - \hat{\mu}_{(t)} \| / \hat{\sigma}_{(t)\max}, \tag{17f}$$

$$L_{(t+1)} = \| X_{(t+1)} - \hat{\mu}_{(t)} \| / \hat{\sigma}_{(t)\min}, \tag{17g}$$

$$\hat{\sigma}_{(t)\min} = \min_i \sqrt{\hat{\sigma}_{ii(t)}}, \tag{17h}$$

$$\hat{\sigma}_{(t)\max} = \max_i \sqrt{\hat{\sigma}_{ii(t)}}, \tag{17i}$$

$$i, j, n = 1(1)N.$$

6. Method of recognition

Figure 1 shows the block diagram of a self-supervised recognition system. The model uses a classifier based on Bayes' maximum likelihood ratio which measures the similarity between the different representative vectors and the input vector and then assigns the input to the class for which the representative vectors show maximum similarity. To study the adaptive capability of the system with *dynamic* guard zone in recognising a pattern, the initial values of these parameters are deliberately chosen to be different (estimated by a poor subset of training samples) from their true values. It is seen from section 5.2 that we need primarily only $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ to be estimated from some sets of training samples. The other parameters are being automatically derived from those estimates.

After the classification of $\boldsymbol{X}$ into the k-th class, the task of the supervisor is to compute $\lambda_{k(t)}$ and then to judge whether the sample $\boldsymbol{X}$ is within the specified guard zone as defined by $\lambda_{k(t)}$ around $\hat{\boldsymbol{\mu}}_{k(t)}$. If it does, the decision of the classifier that $\boldsymbol{X}$ is from the k-th class is accepted by the supervisor and the parameters of that class are updated by $\boldsymbol{X}$. Otherwise, there will be no alternation of the class parameters before the next input.

In fully supervised learning, the decision of the classifier is verified by an external supervisor and the class parameters are altered only if the classification is found to be correct.

7. Implementation to speech sounds

The previously mentioned algorithm was implemented on a set of 871 Telugu (an important Indian Language) vowel sounds in CNC (Consonant – Vo-

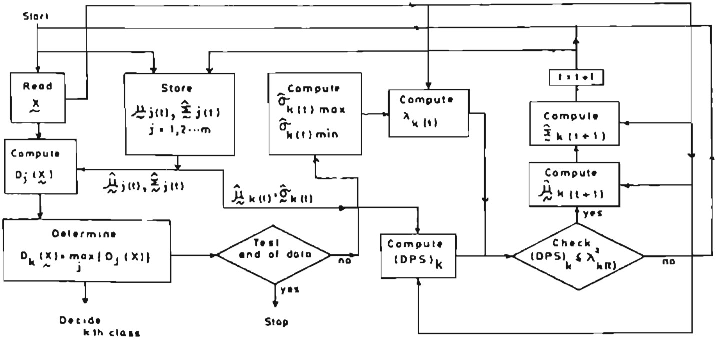
Start
Read X
Store $\hat{\mu}_{j(t)}, \hat{\Sigma}_{j(t)}$ j = 1,2...m
Compute $\hat{\sigma}_{k(t)\max}$ $\hat{\sigma}_{k(t)\min}$
Compute $\lambda_{k(t)}$
t := t+1
Compute $\hat{\Sigma}_{k(t+1)}$
Compute $\hat{\mu}_{k(t+1)}$
yes
Compute $D_j(X)$
$\hat{\mu}_{j(t)}, \hat{\Sigma}_{j(t)}$
$\hat{\mu}_{k(t)}, \hat{\sigma}_{k(t)}$
Determine $D_k(X) = \max_j \{D_j(X)\}$
Test end of data
no
yes
Stop
Decide kth class
Compute $(DPS)_k$
Check $(DPS)_k \le \lambda^2_{k(t)}$
no

Figure 1. Block diagram of dynamic self-supervised recognition.

wel Nucleus – Consonant) context uttered by three male speakers in the age group of 30 to 35 years. The first three vowel formant frequencies F_1, F_2 and F_3 were considered as recognition features to classify ten vowel classes (∂, a: , i, i: , e, e: , u, u: , o and o:) including long and short categories. Since the short and long categories of a vowel differ only in duration, these were pooled together resulting in six groups (∂, a: , I, E, U and O) which differ only in phonetic feature. Figure 2 shows the distribution of the Telugu vowels in the $F_1 - F_2$ plane. Although the shorter and longer types of vowels I, E, U and O are treated similarly, they were given individual class parameter values.

The set of data for each class has been found to follow the normal distribution (Pal, 1978). Therefore, the use of the Bayes' classifier for normally distributed patterns (Section 4) and the assumption (as made in the earlier work of Pathak and Pal (1986)) that the "probability of misclassification of the input patterns falling within the guard zone constructed around the central tendency of a class distribution is substantially low" are well justified here.

Now we are interested here mainly in studying:

(a) the adaptive efficiency of the system in recognising vowel sounds starting with the poor (non-appropriate) estimates of the parameters representing the classes;

(b) the effect of α (i.e., the weighting co-efficient for determining the dimension of guard zone) on the performance of the system with an attempt to determine experimentally its optimum value; and

(c) the effect of 'dynamic behavior of the guard zone' in acting as a supervisor on the decision of the classifier.

The first part of the investigation involved the computation of $\hat{\mu}$ and $\hat{\Sigma}$ values with only five samples selected randomly from the utterances of (i) single speaker and (ii) three speakers so that the initial estimates may be designated as 'very weak' and 'not too weak', say respectively. Recognition efficiency obtained with such weak representative parameters was compared with that of a fully supervised system for different input sequence.

The above experiment was then repeated for different values of α namely, 0.1, 0.2, 0.3, ..., 0.7, 0.8, 0.9 for demonstrating the second part of the investigation.

In order to exhibit the third part of our interest, the performance of the classifier for the aforesaid cases was compared with those obtained when the extent λ is taken to be fixed throughout the learning process. Two such fixed values considered here are

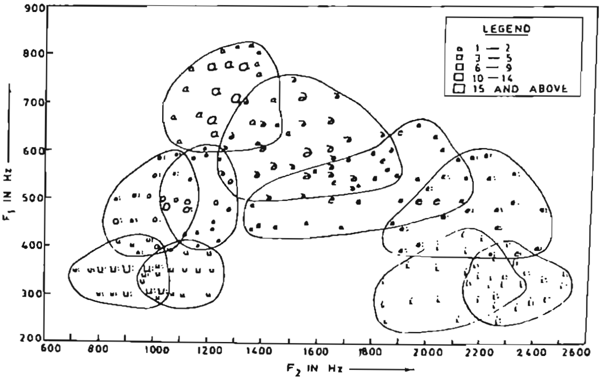

Figure 2. Vowel diagram in the $F_1 - F_2$ plane.

(i) $\lambda_{(1)}$ (i.e., the value of λ generated by the system at $t = 1$), and (ii) around 1/2 (an optimum value obtained by Pal et al. (1980) and Pal (1982) with fixed guard zone dimension for two different types of classifiers). In other words, this part also gives a comparison of the proposed algorithm (GGA) with the existing ones based on similar concept.

8. Experimental results

Since the performance of an adaptive system depends much on the sequence of incoming samples, the experiment was repeated several times for different orders of appearance of the events in the sample space. Figures 3 and 4 illustrate, for three such typical instances, the variation of cumulative recognition score after every 100 samples for different values of α. Figure 3 corresponds to cases when the training set of 5 samples are taken only from a single speaker whereas, the results corresponding to cases when all the speakers are considered for drawing those 5 samples are depicted in Figure 4. Results obtained with self-supervised learning are compared in each case with those for fully-supervised (FS) case.

As expected, Figure 4 (with 'not too weak' initial parameters) shows higher recognition score than Figure 3 where initial class parameters were selected to be 'very weak'. With such very weak representative parameters, the system could not improve significantly its performance even for the fully-supervised case (Figure 3). This is not the case with Figure 4, where fully-supervised learning is found to provide an overall increase (≈ 8%) in recognition score.

From Figures 3 and 4 it is seen that when the initial estimates are 'very weak', good system performance is observed for values of α ranging from 0.7 to 0.9 i.e., high λ-value whereas, the range is found to be 0.1 to 0.3 i.e., low λ-value for 'not too weak' initial estimates (Figure 4). This means, when the initial estimates are not so bad, a very lenient supervisor on lifting a strict check on the incoming samples may affect the system performance by shifting the mean and co-variance values away from their true ones. On the other hand, the guard zone needs to be flexed more, for the bad estimates, in order to strengthen the estimates by allowing higher proportion of correct to incorrect samples more available.

It is also to be noted from Figure 3(b) that the performance corresponding to higher λ-value (i.e., higher α-value) is better even than the case of FS learning, while the results corresponding to low α-value are worst among the three instances (Figures 3(a)–(c)). Under investigation it is revealed that the first few sets of input sequence provided here very good proportion of correct to incorrect samples. As a result, incorporating/discarding them by expanding/shrinking the guard zone improved/declined the estimates, and hence the recognition score, significantly.

Finally, the effect of dynamic property of the supervisory program is demonstrated through Table 1. Here we have considered, as a typical illustration, only three samples from class o: which were correctly identified by the classifier. It is seen that the higher order knowledge (obtained from the input sequence) of the supervisor enables the sample which has largest distance (Euclidean) to get selected for updating procedure while rejecting the remaining two even with smaller distances. Had the value of $\lambda_{(t)}$ been fixed at 1/2 and 2 throughout the learning process, the response would have been 'rejected' and 'accepted' respectively in all the three instances.

Table 1
Supervisor's response for updating procedure

Actual class	Recognised class (C_k)	Euclidean distance from C_k	$\lambda_{(t)}$	Response	$\lambda_{(t)}$	Response	$\lambda_{(t)}$	Response
	o:	9.35×10	0.67	reject	0.5	reject	2.0	accept
	o:	2.16×10^2	1.58	accept	0.5	reject	2.0	accept
	o:	6.9×10	0.492	reject	0.5	reject	2.0	accept

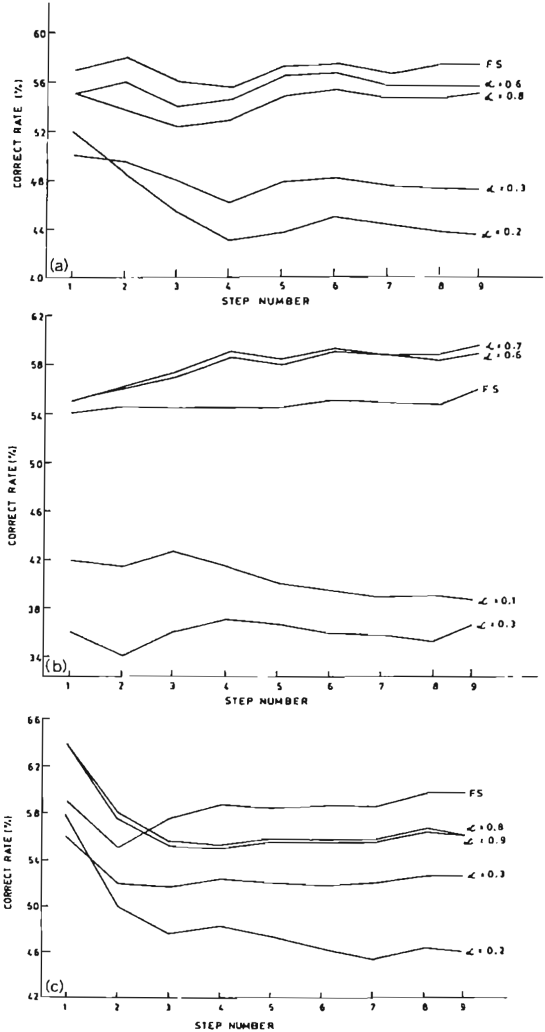

Figure 3. System performance curve when initial estimates are considered to be 'very weak'.

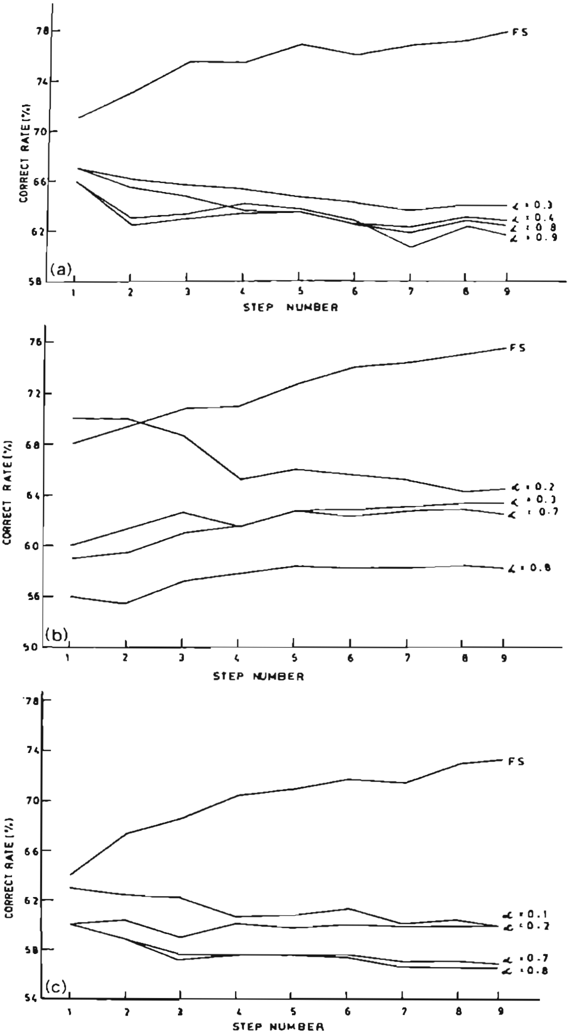

Figure 4. System performance curve when initial estimates are considered to be 'not too weak

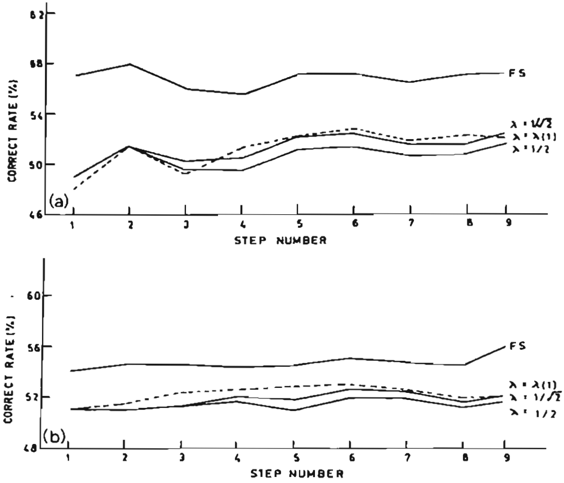

Figure 5. System performance curve for fixed λ value and 'very weak' initial estimates.

The superiority of the varying $\lambda_{(t)}$-value over the fixed λ-value in improving system's performance is illustrated in Figure 5 when λ is kept fixed at $\lambda_{(1)}$ (initial value generated by the system) and, $1/\sqrt{2}$ and 1/2. Here the input sequences for Figures 5(a) and 5(b) are the same as in Figures 3(a) and 3(b). The results corresponding to very weak initial estimates (estimated with five samples taken from a single speaker) are only shown here as an illustration.

9. Conclusion

The dimension $\lambda_{(t)}$ of guard zone along with its bounds for the GGA (Pathak and Pal, 1986) is determined in order to have its convergence for optimum learning. It is found that $\lambda_{(t)}$ is dynamic, depending on the input sequence and previous estimates of the parameter $\boldsymbol{\theta}$. The incorporation of this higher-order knowledge (based on the input sequence) in a self-supervisory program improves the classification efficiency compared to fixed-λ case. Again, the performance of the system may be affected if the guard zone is expanded/shrunk for 'not too weak' / 'very weak' estimates of $\boldsymbol{\theta}$.

References

Chien, Y.T. (1970). The threshold effect of a nonlinear learning algorithm for pattern recognition. *Inform. Sci.* 2, 351-358

Pal, S.K. (1978). *Studies on the Application of Fuzzy Set Theoretic Approaches in Some Problems of Pattern Recognition and Man Machine Communication by Voice.* Ph.D. Thesis, Calcutta University, India.

Pal, S.K., A.K. Dattta and D. Dutta Majumder (1980). A self supervised vowel recognition system. *Patt. Recog.*, 12, 27-34

Pal, S.K. (1982). Optimum guard zone for self-supervised learning. *IEE Proc.* 129, 9–14.

Pathak, A. and S.K. Pal (1986). A generalised learning algorithm based on guard zones. *Patt. Recog. Lett.* 4, 63 69.

Tou, J.T. and R.C. Gonzalez (1974). *Pattern Recognition Principles.* Addison – Wesley, London.