(Reprinted from *JASI* Volume VI Number 4, October 1978, pp. 117–123)

COMPUTER RECOGNITION OF VOWEL SOUNDS USING A SELF-SUPERVISED LEARNING ALGORITHM

S.K. PAL, A.K. DATTA & D. DUTTA MAJUMDER

Electronics & Communication Sciences Unit, Indian Statistical Institute, Calcutta 700 035

The paper describes a model for computer recognition of vowel sounds uttered by a number of speakers in CNC context using first three formants (F_1, F_2 and F_3) as acoustic features. The formants are extracted through spectrum analysis carried over a large number of Telugu (one of the major Indian Languages) words. The distribution of the samples are studied and the boundaries among vowel classes are seen to be fuzzy. The prototype vectors for each of the categories are chosen from this formant space.

The method uses a single pattern training procedure for self-supervised learning and maximum value of fuzzy membership function is the basis of recognition. The fuzzy membership value is considered to be a function of weighted Euclidean distance where the reciprocal of standard deviation of the features is used as a weighting coefficient. The algorithm with selected representative points with a number of guard zones which are ellipsoidal in the three-dimensional feature space around the representative vectors of the classes is taken as a supervisor. The dimensions of the different guard zones selected for investigations are 1st, 1/2th, 1/4th, 1/6th and 1/8th part of the variances of the vowel categories. An optimum zone for self-supervised learning is found to correspond to 1/2th part of the class variances beyond which the machine loses its efficiency. A comparison with a nonadaptive recognition system has also been included.

1. INTRODUCTION

The problem of speech recognition involves multi-level decision processes[1–10] ranging from recognition of vowel, consonant, isolated word by a single speaker and limited vocabulary system for a few trained speakers to connected speech recognition system with unlimited vocabulary for large number of speakers and speech understanding systems [2,11]. Since speech is a pattern of biological origin and carries information regarding the message, the speaker, his health and mood, it is found to a considerable extent to be fuzzy in nature. There exists no precise boundary due to inherent vagueness (fuzziness) rather than randomness in the patterns. Again, since the conditional densities of classes are not known and only a small number of design samples are available, a classifier based on similarity or dissimilarity measures within the framework of fuzzy language theory [12] appears to be suitable to their recognition [13–15]. Of course, both the stochastic and fuzzy techniques of classification can be derived from the two probabilistic concepts, namely, statistical independence (stochastic) and logical implication (fuzzy) [16,17]. Zadeh[18] proposed a probabilistic measure over the fuzzy events where the probability of a fuzzy event is equal to the expected value of its membership function.

Learning is a process which improves the system's performance by acquiring necessary information for decision during the system's operation. The decision in the process is based on the information learned (estimated) and obtained from the observed patterns and if the information learned gradually approaches the true information, then the decision will eventually approach the optimal decision as if all the desired information of each pattern class is known. Therefore the performance of the system in classifying a pattern during the system's operation is gradually improved.

The present paper confines itself to demonstrate the adaptive efficiency of a system in self-supervised recognition method with vowel sounds in CNC (Consonant – Vowel Nucleus – Consonant) context starting with arbitrary representative vectors different from true representatives of the population. The test set contains about 900 utterances constituting a three dimensional vector space. For self-supervised learning, different guard zones were selected whose semiaxes are represented by the $(1/\lambda)$ th (λ = the zone controlling parameter = 1, $2^{1/2}$, 2, $6^{1/2}$, $8^{1/2}$) part of the respective standard deviations initially selected for each of the vowel classes. Similarity measure in the classifier is based on the fuzzy membership value. The results obtained by optimum guard zone are compared with those of the fully-supervised and nonadaptive recognition methods. General purpose digital computer, Honeywell 400 was used for analysis.

2. THE MODEL OF VOWEL RECOGNITION SCHEME

Pattern recognition is considered as a process of decision making in which a new input is decided to be a member of a particular group by the comparison of its attributes with those of previously known members of that class. The present study of vowel recognition covers the following aspects :

a) The recognition process is based on adaptive classification techniques.

b) The behaviour of the model when the representative vectors for each of the categories are not the mean vectors.

c) There is no higher level of supervisor based on different sources of knowledge like linguistic, semantic, syntactic constraints and

d) The supervision programme is based on the inherent properties of the features which may lead to an improvement over the nonsupervised and non-adaptive recognition schemes.

The Telugu vowel patterns (/ ∂ / , / a/, /i/ , /i:/ , /u/, /u:/ , /e/ , /e:/ , /o/ and /e:/) are selected in a CNC combination, as the vowel qualities are considerably influenced in connected speech by the adjoining consonants. All these speech samples are recorded by five adult male speakers on an AKAI tape recorder. From these five informants three denoted X, Y, Z are chosen on the basis of listening experiments based on the opinion of ten listeners. The frequency analysis is done on a Kay Sonagraph model 7029 A. The first three formants are taken at the steady state of the first nuclear positions. Out of 871 samples, the third formant for 384 samples is not clearly located. For these vowels, the average value of the third formant for that vowel corresponding to the particular speaker is taken. The measured value of the three features F_1, F_2 and F_3 is thought to constitute a three dimensional feature space Ω_x. The significant information available about the event thus could be expressed as a three-dimensional feature vector, $X = (F_1, F_2, F_3)$, $X \epsilon \Omega_x$. The coordinates of X would have numerical values indicating the amount of each property of the event.

All the features of a particular sample point represented by the different coordinate directions are not equally significant in defining the characteristics of a class to which like events belong, it is reasonable to assume that the features with larger variation is less characteristic in nature. Therefore in measuring closeness or similarity, lower weight is to be given to features having large variations. In the present experiment, the features with increasing variance have been weighted with decreasing values of a "feature weighting" coefficient, W_n, n = 1, 2,3 and standard deviation of the formants as weighting coefficients were studied.

Of the ten Telugu vowels, long and short categories, viz., /i, i:/, /u, u:/, /e, e:/ and /o, o:/ were found to differ from one another mainly in duration but phonetically they are not distinctively different. With this background, the number of pattern classes to be recognised is reduced to six, namely, / ∂ /, /a:/, /I/, /U/, /E/ and /0/ which are phonetically different from one another. Therefore, the above designed feature space could be viewed to be constituted by various pattern classes C_1, C_2, C_{3s} and C_{3l}, and C_{4s} and C_{4l} and C_{5s} and C_{5l} and C_{6s} and C_{6l} where subscripts s and l stand for shorter and longer categories. Through classification, this three-dimensional feature space is to be divided into such regions which contain vowels differing only in phonetic features. Some of these regions would contain two subregions, one for the short vowels another for the longer ones.

The next task before classification is, obviously, the selection of reference vectors or "prototypes" denoting the representative points of each class. Now we are interested here in studying the recognition of vowel sounds on adaptive classification basis with the nonappropriate prototype vectors representing the classes. A part of the investigations in the above line has already been reported[19] using non-adaptive, supervised and nonsupervised procedures where the prototype points and corresponding weighting coefficients of a specified class were obtained from five utterances of a single speaker (Z) selected randomly from each of the categories. In the present experiment the representative vector of a vowel class was chosen just outside the boundary of an ellipsoid having the three axes equal to the respective standard deviations of the features and mean of the class as the centre. This is explained in figure 1 where $\langle \cdot \rangle$ denotes estimated value. The standard deviations for defining weighting coefficients corresponding to these representative points were obtained from a specified training set of samples selected randomly from the classes.

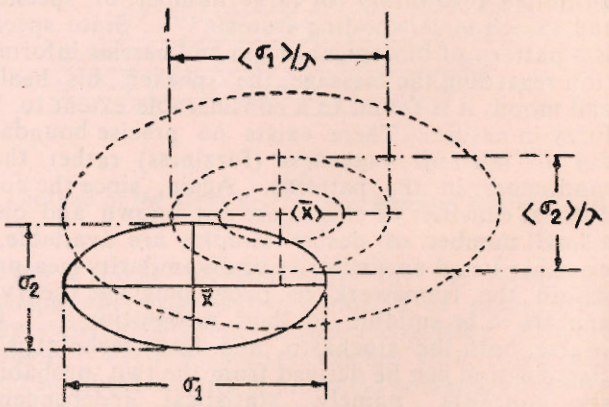

Fig. 1
Selection of Representative Vectors

A supervisory scheme is proposed for self-supervised learning of the recognition model where different zonal boundaries as shown in fig. 1 by dotted

ellipses are described around the selected representative points. The three semiaxes of the various guard zones are respectively chosen to be $(1/\lambda)$th ($\lambda = 1, 2^{1/2}, 2, 6^{1/2}$ and $8^{1/2}$) part of the corresponding standard deviations obtained from the training set of samples. The object of the supervisory programme is to check the decision of the classifier and inhibit the updating procedure if and only if the sample representing the utterance falls outside the guard zones. Otherwise the decision of the classifier is accepted to be correct and the representative and the weight vector of the recognised class are modified with the addition of the new input pattern before the new input sample is allowed to enter into the system.

The last and final task of pattern recognition system is classification by a suitable classifier, whose function is to examine a maximum similarity between the reference vectors of a class and a new input vector. For the vowel classes /I/, /U/, /E/ and /0/, the closeness is measured separately for both the shorter and longer subgroups, and input pattern is assigned to a class which is associated with maximum similarity for either of the subgroups. A suitable classifier based on "maximum fuzzy membership function" along with the learning algorithm is described in the subsequent sections.

3. DECISION RULE AND LEARNING ALGORITHM

Consider an N-dimensional feature vector space Ω_x containing m ill-defined pattern classes to be recognised with defined set of N-dimensional prototypes $R_1, R_2, R_3 \cdots R_j, \ldots R_m$ such that

$$R_j^{(l)} \in R_j$$

$l = 1, 2, \ldots h_j$, h_j is the number of reference vector in set R_j.

3.1. Fuzzy Membership Function:

A class of events $x_1, x_2, \ldots x_n$ in the universe of discourse U is defined as a Fuzzy set (A), if their transition from membership to nonmembership is continuous rather than abrupt. $x_1, x_2, \ldots x_n$ represent the supports of A at which the value of membership function $\mu_A(x)$ characterising the grade of membership of x in A is positive ranging between zero and one. A fuzzy set A with its finite number of supports could therefore be viewed as

$$A = \{\mu_A(x_i), x_i\}$$

where $\mu_A(x_i)$ being the grade of membership of x_i in A denotes the degree to which an event x_i may be a member of or belong to A and as it approaches to unity, the grade of membership of x_i in A becomes higher. For example $\mu_A(x_i) = 1$ indicates strictly the containment of the event x_i in A and if x_i on the other hand does not belong to A, $\mu_A(x_i) = 0$. Any intermediate value would represent the degree upto which x_i could be a member of A.

The decision of the classifier for the purpose of recognition of an unknown pattern $X = [x_1, x_2, \ldots x_n, \ldots x_N]$ is based on the magnitude of the fuzzy membership function corresponding to jth ($j = 1, 2, \ldots m$) class:

$$\mu_j(X) = [1 + \{ d(X, R_j) / F_d \}^{F_e}]^{-1} \qquad \ldots(1)$$

where F_e is the exponential fuzzifier,

F_d is the denominational fuzzifier,

and $d(X, R_j) = m_l in \|X - R_j^{(l)}\|$

$$\|X - R_j^{(l)}\| = \left[\frac{\Sigma}{n} \left(\frac{x_n - \bar{x}_{jn}^{(l)}}{\sigma_{jn}^{(l)}} \right)^2 \right]^{0.5}$$

denotes the weighted Euclidean distance [20,21] between unknown pattern X and l th reference vector $R_j^{(l)}$ in j th class in which $\bar{x}_{jn}^{(l)}$ and $\delta_{jn}^{(l)}$ are the mean and standard deviation of the features along nth coordinate in jth class and correspond to l th prototype.

The fuzzifiers have the effect of altering the ambiguity in a set and hence overall recognition score.[13-15] The membership function is defined in such a way that it maps N-dimensional feature space into an m-dimensional membership space which is a unit hypercube and should satisfy the following conditions:

(i) $\mu_j(X) \to 0$ as $d(X, R_j) \to \infty$

(ii) $\mu_j(X) \to 1$ as $d(X, R_j) \to 0$

and (iii) $\mu_j(X)$ increases as $d(X, R_j)$ decreases

Therefore the membership function $\mu_j(X)$ having positive value in the interval [0, 1] denotes the degree to which an event X may be a member of or belong to jth class and classificatory decision rule would be as follows:

decide: $X \in C_k$ if $\mu_k(X) > \mu_j(X)$

$j = k = 1, 2, \ldots m, j \neq k.$

3.2. Iterative Algorithm for Parameter Estimation:

The components of reference vector and weight vector for a class used in the decisional algorithm are respectively the mean and reciprocal of standard deviation of the components of the feature vectors. The reciprocal of standard deviation is found to provide appropriate phase weights to patterns for their proper classification.[13-15, 22]

The basic idea of the recognition system is to draw unknown samples randomly one after another and build up their classes. The samples those are inserted to a given class modify the centres and relative weights on the axes of the classes. The iterative procedure adopted here is therefore the "centre - variance adjustment algorithm" in which the weighted Euclidean distance used in membership function reflects an ellipsoidal shape of each cluster.

If $x_{n(t)}$ and $^3\sigma_{n(t)}$ represent the mean and variance of a class along nth co-ordinate axis, estimated by first t samples, we note

$$\bar{x}_{n(t)} = \frac{1}{t}\sum_i x_i \qquad \ldots (2a)$$

and $$\sigma^2_{n(t)} = \frac{1}{t}\sum_i (x_i - \bar{x}_{n(t)})^2$$

$$= \frac{1}{t}\sum_i x_i^2 - \bar{x}^2_{n(t)}, \quad i = 1, 2, \ldots t$$

$$= \frac{1}{t} C_{n(t)} - \bar{x}^2_{n(t)} \qquad \ldots (2b)$$

where $C_{n(t)} = \sum_i x_i^2$, $i = 1, 2, \ldots t$

Let another sample $x_{(t+1)}$ fall in this class. Then mean and variance are adjusted as follows:

$$\bar{x}_{n(t+1)} = \frac{t}{t+1}\bar{x}_{n(t)} + \frac{1}{t+1} x_{(t+1)} \qquad \ldots (3a)$$

$$C_{n(t+1)} = C_{n(t)} + x^2_{(t+1)} \qquad \ldots (3b)$$

$$\sigma^2_{n(t+1)} = \frac{1}{t+1} C_{n(t+1)} - \bar{x}^2_{n(t+1)} \qquad \ldots (3c)$$

All these equations (3) provide us with an iterative algorithm for estimation of the mean and variance vectors, given successive samples.

3.3. Algorithm for Self-supervised Learning Scheme

In the self-supervised adaptive learning scheme a "guard zone" is used to serve the role of a supervisor in providing correct labels to the samples and also to postpone classification of doubtful samples (lying outside the guard zones) so as to ensure minimisation of the effect of wrong classification on the initial parameters of the algorithm. The "decision parameter of the supervisor" (DPS) on which the guard zone is based on is defined with respect to jth class as

$$(DPS)_j = \sum_n \left(\frac{x_n - \bar{x}_{jn}}{\sigma'_{jn}} \right)^2 \qquad \ldots (4a)$$

where $\sigma'_{jn} = \sigma_{jn}/\lambda$

$n = 1, 2, \ldots N$

and λ (a positive constant) is termed as "zone-controlling parameter" which controls the dimension of the guard zone in N-dimensional vector space Ω_x.

After having the knowledge of DPS – values for all the classes, the supervisor accepts the decision made by the classifier that $X \in C_k$ only if,

$$(DPS)_k \leqslant 1 \qquad \ldots (5)$$

Otherwise the decision is declared to be wrong and no other alteration of the mean and variance vectors of the kth class is made for that input sample. It is to be noted here that, these decision parameters would lead to ellipsoidal shape of the guard zones.

4. EXPERIMENTAL RESULTS

The distribution of the samples uttered by three speakers in $F_1 - F_2$ plane of the vector space is sketched in Fig. 2, where the boundaries among the vowel classes are seen to be ill-defined. The position of the mean vectors alongwith the standard deviation of the vowel classes are shown in Fig. 3 in three-dimentional feature space. The length of the arrows represents the magnitude of the respective standard deviations in the same scale. Table I shows the components of the prototype vectors and the corresponding standard deviations selected as initial parameters of the recognition system.

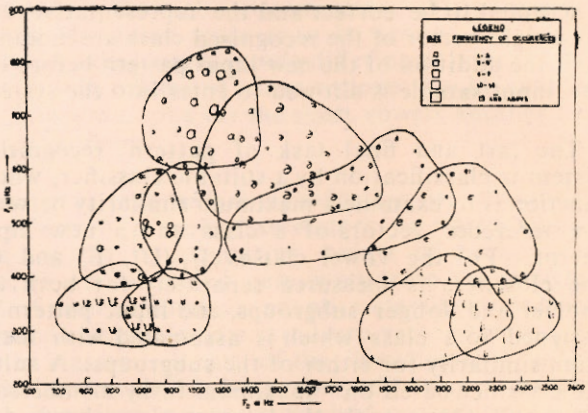

Fig. 2
Distribution of Telugu Vowels in $F_1 - F_2$ Plane

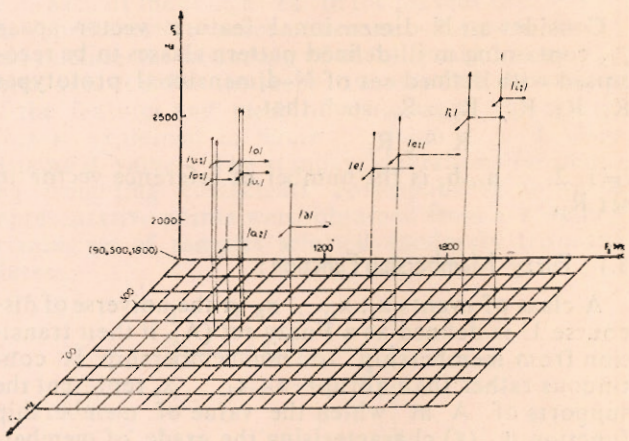

Fig. 3
Location of the Mean Vectors with Standard Deviations of Vowels

Fig. 4 shows the flow chart for the recognition scheme where the method consisted of nonadaptive, supervised and nonsupervised recognition of vowel sounds. The nonadaptive scheme with fixed mean and weighted vectors is presented to demonstrate the efficiency of system adaptation to the new input patterns. Block diagram of the supervisor using self-supervised learning schemes based on guard zones is explained in fig. 5.

With the defined set of representative vectors as shown in Table I, the system started to recognise

TABLE I

Initial Values of Mean and Standard Deviation Vectors

Vowel	Mean			Standard Deviation		
	F_1 Hz	F_2 Hz	F_3 Hz	F_1 Hz	F_2 Hz	F_3 Hz
ə	650	1535	2396	86	214	154
a:	715	1180	2362	79	74	192
i	280	2100	2795	50	83	130
i:	300	2220	2732	40	105	187
u	340	1050	2437	47	108	203
u:	320	865	2532	33	72	112
e	530	2150	2591	65	130	160
e:	550	1965	2725	73	397	175
o	500	1065	2561	73	101	218
o:	525	965	2570	40	51	227

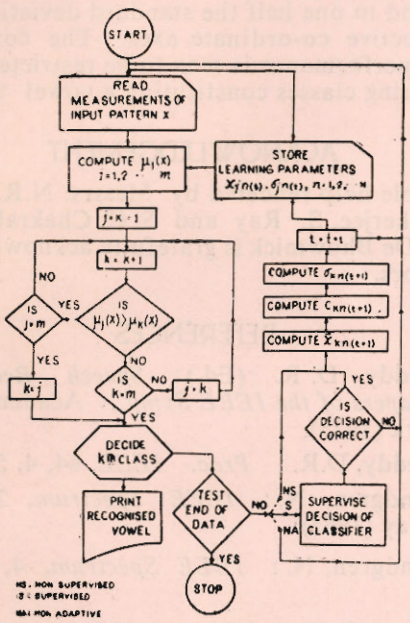

Fig. 4

A Flow Chart for Recognition System

unknown utterances of all the three speakers taken in a random sequence from the sample space Ω_X. Though for vowels /I/, /U/, /E/and/0/the longer and shorter varieties are pooled together, they were given individual reference vectors and weight vectors computed over the respective set of training samples. Thus in the present experiment, m=6 N=3, h=1 for/ə/and/a:/and h=2 for/I/,/U/,/E/and/0/. Computing membership values with respect to all the classes an input utterance is assigned to kth (k=1, 2, ... 6) class associated with maximum μ - value.

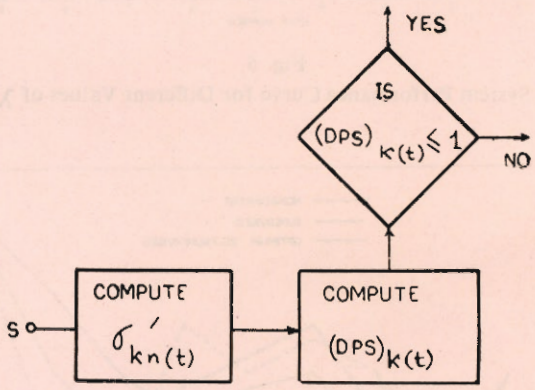

Fig. 5

Block Diagram of Supervisor for Self-supervised Learning Algorithm

Fig. 6 illustrates the variation of recognition score with successive input patterns for different values of λ, the zone controlling parameter, where the rate of correct classification after every 100 input samples was noted and plotted. As λ increases, the dimension of the guard zones decreases and the corresponding DPS - values increase. Therefore, the chances of correct samples correcting the representative vectors decrease. As a result of this, the system performance approaches the nonadaptive recognition case. With the decrease in the value of λ, the zone boundaries on the other hand increase making reduction in the DPS - values. The system then behaves more akin to nonsupervised recognition algorithm where the chances of wrong samples vitiating the representative vectors increase.

The results obtained using self-supervised learning algorithms are compared in fig. 7 with those obtained

with nonadaptive and fully-supervised learning algorithms. In fully-supervised learning, the decision of the classifier is verified by an external supervisor and class parameters are altered only if the classification is found to be correct. Otherwise no alterations of the representative and weight vectors are made.

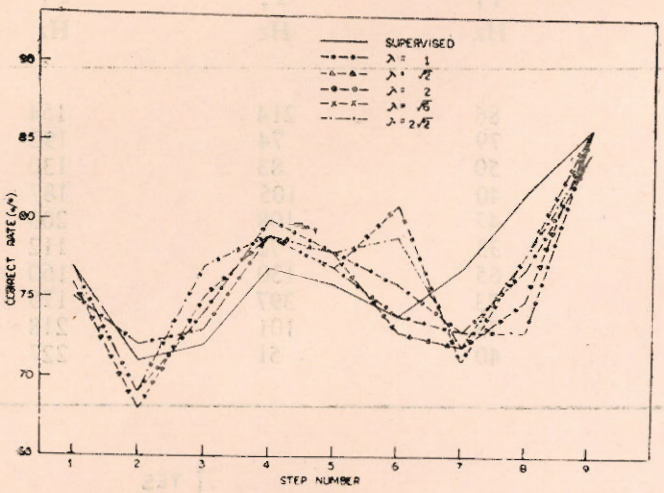

Fig. 6
System Performance Curve for Different Values of λ

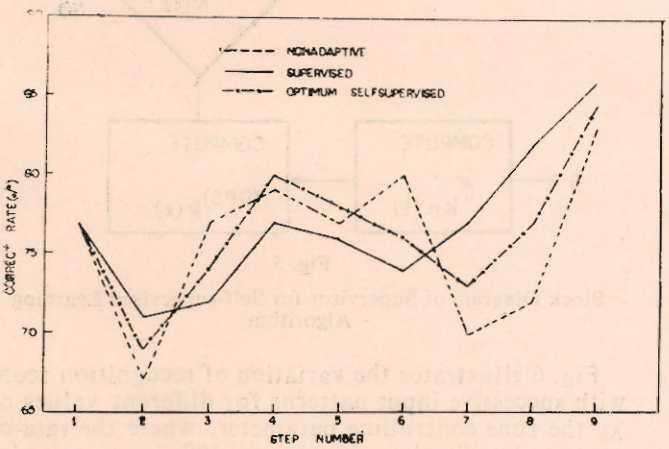

Fig. 7
System Performance Curve

Deviations of the self-supervised performance curves based on mean square distance from that obtained with fully-supervised procedure at every instance are tabulated in Table 2. The curve corresponding to $\lambda = 2$ shows the best match with that of fully-supervised performance. The classifier with guard zones corresponding to $\lambda = 2$ is expected to avail highest proportion of correct to incorrect samples so that after the several utterances being dealt with by the classifier, the class representative and weight vectors are likely to approach their respective true values.

TABLE II

Mean Square Distance from the Fully-supervised Performance Curve for different Values of λ

λ^2	1	2	4	6	8
Mean Square Distance	12.0	11.33	7.58	11.69	15.0

5. CONCLUSIONS

A self-supervised learning algorithm from the standpoint of fuzzy set theoretic concepts is presented and implemented to Telugu vowel sound recognition using first three formant frequencies and single pattern training procedure. System performance for different guard zones selected around the initial representative vectors is studied. With the shrinkage of the zone boundaries, the system behaves like a nonadaptive recognition system whereas the nonsupervised performance is approached by relaxing the boundaries. Optimum results are obtaind when the semiaxes of the guard zones defined for the classes correspond to one half the standard deviations along the respective co-ordinate axes. The confusion in machine performance is seen to be restricted only to neighbouring classes constituting a vowel traingle.

ACKNOWLEDGEMENT

Valuable help rendered by Messrs. N.R. Ganguli, B. Mookherjee, S. Ray and S. C. Chakraborty and Smt. S. De Bhowmick is gratefully acknowledged by the authors.

REFERENCES

1. Reddy, D. R. (Ed.): *Speech Recognition : Invited papers of the IEEE Symp.*, Academic Press, New York (1975).

2. Reddy, D.R.: *Proc. IEEE*, 64, 4, 501 (1976).

3. Lindgren, N.: *IEEE Spectrum*, 2, March, April, May (1965).

4. Lindgren, N.: *IEEE Spectrum*, 4, June, 75 (1967).

5. Klatt, D.H. and Stevens, K.N.: *IEEE Trans. Audio Electroacoust.* AU-21, 210 (1973).

6. Weinstein, C.J., McCandless, S.S., Mondshein, L. F. and Zue, V.W.: *IEEE Trans. Acoust., Speech, Sig. Process.*, ASSP – 23, 54 (1975).

7. Dutta Majumder, D. and Datta, A.K.: *JIETE*, 15, 233 (1969).

8. Sharma, V.V.S. and Yegnanarayana, B.: *Proc. All India Interdisc. Symp. on Dig. Tech. and Pat. Recog.*, Calcutta, I. S. I. Feb. 15–17 (1977), To appear.

9. *Trans. IEEE. on Acoust., Speech and Sig. Process.*, Special Issue on IEEE Symp. on Speech Recognition, ASSP – 23, 1, (1975).

10. *Proc. IEEE*, Special Issue on Man-Machine Communication by Voice, 64, 401 (1976).

11. Woods, W. A. and Makhoul. J. : *Artificial Intelligence*, 5, 73 (1974).

12. Zadeh, L. A., Fu, K. S., Tanaka, K. and Shimura, M. (Eds.) : *Fuzzy Sets and their Applications to Cognitive and Decision Processes*, Academic Press, New York (1975).

13. Dutta Majumder, D. and Pal, S. K. : *Proc. IEEE Int. Conf. on Cybernatics and Society*, Washington, DC, September 19–21, (1977), To appear.

14. Pal, S.K. and Dutta Majumder, D.: *IEEE Trans. Syst. Man and Cyberns.*, SMC – 7, 625 (1977).

15. Pal, S.K. and Dutta Majumder, D.: *IEEE Trans. Syst. Man and Cyberns.*, SMC – 8, 302 (1978).

16. Gaines, B.R.: *Electronics Letters*, 11, 188 (1975).

17. Stallings, W. : *IEEE Trans. Syst. Man and Cyberns.* SMC – 7, 216 (1977).

18. Zadeh, L.A. : *J. Math. Analysis and Applications*, 23, 2, 421 (1968).

19. Pal, S.K., Datta, A.K. and Dutta Majumder, D. *Int. J Syst., Sciences*, **9**, 887 (1978).

20. Sebestyen, G.S. : *Decision Making Processes in Pattern Recognition*, The Macmillan Co., New York (1962).

21. Meisel, W.S. : *Computer-Oriented Approaches to Pattern Recognition*, Academic Press, New York (1972).

22. Dutta Majumder, D., Datta, A.K. and Pal S.K. : *JCSI*, 7, 14, (1976).