STANFORD ARTIFICIAL INTELLIGENCE
MEMO AIM-198

STAN-CS-73-364

# ESTIMATION OF PROBABILITY DENSITY USING SIGNATURE TABLES FOR APPLICATION TO PATTERN RECOGNITION

BY

R. B. THOSAR

SUPPORTED BY
ADVANCED RESEARCH PROJECTS AGENCY
ARPA ORDER NO. 457

MAY 1973

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

STANFORD ARTIFICIAL INTELLIGENCE LABORATORY
MEMO AIM-198

MAY 1973

COMPUTER SCIENCE DEPARTMENT
REPORT NO. STAN-CS-73-364

# ESTIMATION OF PROBABILITY DENSITY USING SIGNATURE TABLES FOR APPLICATION TO PATTERN RECOGNITION

by

R. B. Thosar

ABSTRACT: Signature table training method consists of cumulative evaluation of a function (such as a probability density) at pre-assigned co-ordinate values of input parameters to the table. The training is conditional: based on a binary valued "learning" input to a table which is compared to the label attached to each training sample. Interpretation of an unknown sample vector is then equivalent of a table look-up, i.e. extraction of the function value stored at the proper co-ordinates. Such a technique is very useful when a large number of samples must be interpreted as in the case of samples must be interpreted as in the case of speech recognition and the time required for the training as well as for the recognition is at a premium.
However this method is limited by prohibitibe storage requirements, even for a moderate number of parameters, when their relative independence cannot be assumed. This report investigates the conditions under which the higher dimensional probability density function can be decomposed so that the density estimate is obtained by a hierarchy of signature tables with consequent reduction in the storage requirement. Practical utility of the theoretical results obtained in the report is demonstrated by a vowel recognition experiment.

This research was supported in part by the Advanced Research Projects Agency of the Office of the Secretary of Defense under Contract No. SD-183.

The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects

## Acknowledgements

The author wishes to express his deepest gratitude to Dr. A.L.Samuel for introducing him to the concept of signature tables and for the suggestion of it's application to speech recognition. Prof. T.M.Cover provided encouragement with many a stimulating discussion. The author is thankful to Prof. John McCarthy for the opportunity to work at the Artificial Intelligence Project as a visiting research associate.

## 1. Introduction

Signature table training has been described by Samuel[1] where it is used in a program which plays the game of checkers. The input to the tables are parameters which evaluate specific aspectes of a board situation. The output from the table hierarchy is a number which represents in a sense, "figure of merit" for the board. This board evaluation is then used in the search for the best possible move.

The signature table scheme has been extended and modified to adapt it for use in speech recognition[2]. The tables are used to compute the postiriori probability of a specific sound feature such as voicing, a front vowel etc. or a sound class such as a phonemic category, being present. These probabilities are used for classification of the sound in Bayesian sense (the actual implementation makes a compound decision using local context). However this scheme makes several implicit, simplifying assumptions with regards to mutual independence between sets of input parameters. This report describes a method of probability density estimation using signature tables which does not require independent set of parameters and still requires storage of the same order.

The concept of signature tables is best illustrated by a simplified table arrangement(Fig.1) similar to the one used in the speech recognition program[2]. The two-level arrangement has six inputs. the Fi's represent the frequencies of amplitude maxima(formants) in the vowel spectrum and the Ai's are their corresponding amplitudes. The parameters are divided into two sets as inputs to the two first level tables. The outputs from the first level tables are used as inputs to the second level table. The input marked "V" is a 1-bit learning input and indicates whether the vowel "V" is the tag attached to the current sample vector or it is not.

The inputs are quantized into adequate range of values, 8 in this case. A signature table has an entry for every possible combination of inputs. Thus the size of first level tables is 512. Each entry (one computer word) is divided into three fields. The field [p] is incremented by 1 if "V" is indicated otherwise field [q] is incremented. The output of each entry is computed as:

$$P("V"|\ F1,F2,F3) = p/(p+q) \qquad [1.1]$$

which directly gives the postiriori probability of the class "V" for the specific entry shown in Fig.1. This value is also quantized to a prespecified accuracy, say 3-bits, and is stored in the output field [r]. The second level table processes its input in a like manner. Thus the output of the second level table is the probability

$$P("V"|\ P(F1,F2,F3),P(A1,A2,A3)) \qquad [1.2]$$

whereas the probability we require is

$$P("V"|\ F1,F2,F3,A1,A2,A3) \qquad [1.3]$$

Thus while any one table generates the required form of probability density(eq. 1.1), the two level arrangement generates 1.2, which is equivalent to 1.3 only if the parameters in the sets (F1,F2,F3) and (A1,A2,A3) are mutually independent.

The main objective of this report is to investigate the conditions under which the decomposition of a multidimensional function into two or more factors is valid so that a hierarchy of tables as shown in the above example still generate a true higher dimensional probability density as in the equation 1.3 while mostly retaining all the advantages of the signature table method.

The three advantages that emerge from this method of training as it has been used in the past are as follows:

1)Essentially arbitrary inter-relationships between the inputs are taken into account by any one table. The only loss of accuracy is in the quantization.

2)The training is a simple process of accumulating counts.The training samples are introduced sequentially, and hence simultaneous storage of all the samples is not required.

3)The process linearizes the storage requirements. The example shown requires 2*512+64, 1088 entries instead of $2^{(3*6)}$ , 256 K entries, were the entire space to be represented.

Before investigating the conditions under which the decomposition of a multidimensional space is valid it will simplify the explanation if we first consider a specific example in qualitative terms only.

Consider the simple table arrangement as shown in Fig.2 where we wish to take account of 5 input parameters, each requiring say, 3 bits for its specification. Were we to do all this in one table it would require $2^{15}$ or 32,768 entries, instead of the 1024 entries required for the two level arrangement shown, when the output from Table 1 is also quantized to 3 bits. What we require as the output from the first level table is some function which represents the contribution made by the inputs to the first table so that the output from the second table truely represents the conditional probability that CLASS has been represented by the specific values of the inputs. Thus, in order to utilize the Bayesian decision rule we want to determine

$$P(CLASS|A,B,C,D,E)$$

where A B C D and E represent the input parameters. It is easier to determine the inverse probability during the training phase in accordance with the rule

$$P(CLASS|A,B,C,D,E) = P(A,B,C,D,E|CLASS) * P(CLASS) / P(A,B,C,D,E).$$

The divisor on the right hand side appears as a common factor in the conditional probabilities for all the classes and hence need not be taken into account. P(CLASS), the apriori of a class may either be known for the recognition problem under consideration, or it can be estimated from the sample set used during the training. The remaining factor P(A,B,C,D,E|CLASS) is to be determined using the signature table arrangement shown in Fig.2.

Consider the expansion

$$P(A,B,C,D,E|CLASS) = P(D,E|A,B,C,CLASS) \; . \; P(A,B,C|CLASS)$$

the second factor on r.h.s., the marginal probability is independent of the other inputs and hence given directly by the counts accumulated in first table (field [p] Fig.2) with appropriate normalization by the the total counts in the table. We now focus our attention on the first factor, the conditional probability P(D,E|A,B,C,CLASS) computed by the second table. We observe that the input marked f(A,B,C) partitions this table into sections. In the expanded case where Table 2 has one such section for every entry in Table 1, there would have been enough of such sections to allow for every combination of values assignable to A,B and C. However we are not finally interested in which particular points in (A,B,C,D,E) space are associated with each region (in the situation where the variables are continuous, each region would be an equiprobable surface), but only the value of the probability. Therefore we need only allow enough space to represent these values to the desired accuracy(3 bits in this example). Thus the outputs stored in the first level table (field [r]) correspond to those entries which must be grouped together and have identical value. In effect what we are doing is to reduce the dimensionality of the space represented by the second table by 2 (i.e. from {A,B,C,D,E} to {D,E,f(A,B,C)}) by grouping together those points in (A,B,C) space which give the same overall probability in (A,B,C,D,E) space.

There is apparent circularity in this argument, that is we must know the overall probability - which we are ultimately trying to find - in order to accomplish the grouping in the lower dimensional space. However we shall prove in later section that there exists a mapping f(A,B,C|CLASS) of the function P(A,B,C,D,E|CLASS) which achieves the desired grouping.

In a manner similar to the one used to obtain marginal probabilities, the function f(A,B,C|CLASS) is accumulated iteratively in field [q] in each table entry. The values in the fields [q] are then quantized to a desired accuracy and stored in the output field [r] of each entry.

So far we have dealt exclusively with the process of training, how the data is entered into the tables. Interpretation of an unknown sample or evaluation of it's class conditional probability may be explained with reference to Fig.2. Suppose that the unknown input is (A',B',C',D',E'). The marginal density is given by

$$P(A',B',C'|CLASS) = [p_1] / SIGMA\ [p_1]$$

and the conditional probability is given by the second table as

$$P(D',E'|[r_1],CLASS) = [p_2] / SIGMA\ [p_2].$$

The required probability is then obtained as the product of these two factors.

The above highly simplified example also indicates an important property of the signature table method. During both the training and interpretive phases a table derives all the necessary information from the outputs of it's immediate predecessors. This fact can greatly simplify the programs required for the construction and the execution of complicated signature table networks. For more details of the programs and for more general speech specific aspects of the signature table usage, the reader is refered to Samuel[3].

In the next section we obtain the conditions under which the decomposition of a higher dimensional probability density function into a product of two lower dimensional functions is valid. In the following section we outline a method of estimation of marginal and conditional probability densities which occur in the decomposition. The signature table is shown to be an approximation of this method of density estimation. The last section describes results of experiments performed using a rather small set of training samples. The objective is to demonstrate the feasibility of this method, rather than obtain statistically valid error bounds, were this technique used for probability density estimation per se, rather than for a classification or a recognition task.

## 2.0 Decomposition of Probability Density Functions

The simplified example given in the previous section made an implicit assumption that the random variables were discrete, each with 8 distinct states representable by 3 bits. It is more illuminating however to obtain the conditions for decomposition in the continuous domain. The discrete situation may then be treated as a special case.

Let $x_1, x_2, .. , x_n$ be N continuous, non-independent random variables. Let $p(x_1, x_2, .., x_n | C)$ define a class conditional probability density function which is continuous everywhere in $R^N$ space. The objective is to factorize this function such that each factor represents a function whose dimensionality is less than N. Write

$$p(x_1, x_2, ..x_i, ..x_n | C) = p(x_{i+1}, ..x_n | x_1, x_2, ..x_i, C) . p(x_1, ..x_i, C) \quad [2.1]$$

also let

$$= p(x_{i+1}, ..x_n | f(x_1, ..x_i), C) . p(x_1, ..x_i, C) \quad [2.2]$$

Consider the factorization in equation 2.1. The second factor, the marginal density has a dimensionality I which is less than N. But the first factor, although a conditional, has implicit dimensionality of N. In equation 2.2 we have grouped together the conditioning variables $(x_1, ..x_i)$ by as yet undefined functional f, to give a mapping of I dimensions to 1. So the dimensionality of this factor is essentially (N-I+1).

Now define

$$f(x_1, x_2, \ldots x_i) = p(x_1, x_2, \ldots x_i, x_{i+1} = 0, \ldots x_n = 0 \mid C)$$

or for notational convenience,

$$= d(x_1, x_2, \ldots x_i \mid C) \qquad [2.3]$$

called the degenerate probability density function with the variables $x_{i+1}$ through $x_n$ set to zero, or any other convenient, arbitrary set of constants. With this definition of the function f we shall show that equations 2.1 and 2.2 are equivalent.

Consider a partition in the $R^N$ space generated by setting

$$p(x_1, x_2, \ldots x_n \mid C) = K \quad \text{where} \quad 0 \le K < 1 \qquad [2.4]$$

and denote the set of all solution vectors which satisfy 2.4 by $S_N(X_N)$ where $X_N$ is a N-vector. Also let

$$d(x_1, x_2, \ldots x_i \mid C) = K$$

and denote the solution set of this equation by $S_I(X_I)$. Clearly from the definition of the degenerate function 2.3,

$$\text{for any} \quad (X_I \in S_I) \supset ([X_I = X_N] \in S_N) \qquad [2.5]$$

with the proviso that only the first I terms in the vectors $X_I$ and $X_N$ need match.

Now consider within this partition (defined by 2.4), the factor which determines conditional density in equation 2.2, namely,

$$p(x_{i+1},..x_n \mid d(x_1,..x_i),C) = p(x_{i+1}\ ...x_n \mid (X_I \in S_I),C)$$

$$= p(x_{i+1},..x_n \mid (X_N \in S_N),C) \quad \text{from 2.5}$$

$$= p(x_{i+1},..x_n \mid x_1,x_2..x_i,C)$$

Since other terms in the equations 2.1 and 2.2 are identical and the condition 2.5 holds true for the complete range of values K (in 2.4) may take, the required equivalence holds in general.

To summarize the result obtained in this section, we have shown that it is possible to factorize a N dimensional probability density function into a I dimensional marginal and a (N-I+1) dimensional conditional probability density function. The explicit dimensionality of the product(eq. 2.2) is either I or (N-I+1), which ever is greater. The savings in the storage requirement can therefore be quite significant.

## 3.0 Estimation of Probability Density

In the previous section we have obtained the conditions for the decomposition which requires the estimation of three different functions:

1) the marginal density $p(x_1,x_2,...x_i)$,

2) the degenerate density $d(x_1,x_2,...,x_i)$,

and 3) the conditional density $p(x_{i+1},\ldots x_n \mid d(x_1, x_2, \ldots x_i))$.

Further we require that the training algorithm to be used for the estimation be iterative so as to admit one sample at a time. The non-parametric method of density estimation which uses superposed potential functions appears to be the best candidate which satisfies all these conditions. In its most general form, the density is estimated by the summation

$$p_M(X) = 1/M * \sum_{m=1}^{M} \mathrm{psi}(X, X_m) \qquad [3.1]$$

where X is a random vector variable, $X_m$ is the $m^{th}$ sample, M the total number of samples available and psi is any one of the admissible potential or kernel functions. Parzen[4] has obtained the conditions on psi under which 3.1 becomes a valid probability density in one dimension. Murthy[5] has generalized these conditions for N dimensions. A concise description of various admissible forms of psi and the related conditions may be found in Andrews[6], Sec. 4.3.

We shall use the Gaussian kernel

$$\mathrm{psi}(X, X_m) = (2\pi\alpha^2)^{-N/2} * \exp(-[X-X_m]^t * [X-X_m] / 2\alpha^2)$$

$$= K * \exp(.)$$

which was first proposed by Sebestyen[7] and also used later by Specht[8] to obtain trainable polynomial discriminant functions. The attractive property of this kernel is that the constant α may be chosen so as to produce required smoothness in the generated density. Small values of α cause each sample to stand out in the summation 3.1, whereas larger values of $\alpha$ give a smoother surface. The iterative form of the summation using the Gaussian kernel is simply,

$$p_M(X) = (M-1)/M * p_{M-1}(X) + K/M * \exp(.)$$

The marginal densities can use this form directly. The degenerate densities become

$$d_M(x_1,x_2,..x_i) = p_M(x_1,x_2,..x_i,x_{i+1}=0,...,x_n=0)$$

The estimation of the density which is conditional to the degenerate is clearly a second level process, in the sense that it presumes a stabilized, consistent degenerate estimate. It also implies that we need to estimate a density of the form $p(x_{i+1},..x_n)$ for every possible value the degenerate may take. It appears therefore that evaluation of the conditional factor in the continuous domain is infeasible. Thus we must quantize the range of values of the degenerate generated by the training process at the first level and then obtain the conditional densities for each of these values at the second level.

## 3.1 Pragmatic Considerations

The preceding discussion may give an impression that the fact that we have achieved a decomposition which leads to a simpler estimation problem (as a reduction in dimensionality) is largly illusiory. Reason being that the estimation of the conditionals hides an inherently higher dimensional problem. The argument is certainly valid in continuous domain. However every practical problem involves a discretization at some stage. At best it would be based on the signal to noise ratio in the measurement and at the worst a more crude quantization dictated by available resources. Now, our claim is that the method outlined first simplifies the problem by decomposition and then gives you a control over the error which will be propogated to the higher level: merely by quantization of the degenerate to a required accuracy. In the next section where the signature table method of estimating the density is described, we outline possible modes that can be used to quantize the degenerate. These modes of quantization appear reasonable for the problem at hand, namely multicategory pattern recognition.

## 4.0 Estimation Using Signature Tables

A signature table assumes explicit quantization (not necessarily uniform) of the inputs. There is a unique entry or a signature-type in a table for all the possible combinations generated by quantized inputs. Therefore any function which is to be evaluated by a table is known only at those points.

First consider the estimation of a marginal probability density for say, 3 variables. Using the iterative formulation 3.4,

$$p_M(x_1,x_2,x_3) = (M-1)/M * p_{M-1}(x_1,x_2,x_3) + K/M *$$

$$\exp\{-[(x_1-x_1^m)^2+(x_2-x_2^m)^2+(x_3-x_3^m)^2]/2\alpha^2\}$$

where $K=(2\pi\alpha^{2})^{-3/2}$ for $N=3$ and $(x_1^m, x_2^m, x_3^m)$ is the $m$th training sample. Now consider an arbitrary table entry $(x_1^i, x_2^j, x_3^k)$. The factor (M-1/M) which normalizes the accumulated density with M-1 samples can be neglected even for moderate values of M. Thus the new density after $M$th sample is obtained by adding the increment

$$K/M * \exp\{-[(x_1^i - x_1^m)^2 + (x_2^j - x_2^m)^2 + (x_3^k - x_3^m)^2]/2\alpha^2\} \qquad [4.1]$$

to the count stored for this entry. Apparently the above increment must be computed and added for each entry in the table. However since the Gaussian kernel decays exponentially the number of table entries for which the increment is significant can be small. The entry for which the increment is maximum is given by

$$\{\min|x_1^i - x_1^m|, \min|x_2^j - x_2^m|, \min|x_3^k - x_3^m|\}$$

where i, j and k are varied over the respective range of quantization. Other entries for which the increment might be significant are the neighboring entries. Thus the search for the entries which must be modified is straightforward.

Estimation of the degenerate density is done in analogous manner. The degenerate variables are set to zero or other more convenient constant and do not figure in location of an entry. Thus if $x_3$ were a degenerate in the previous example then the increment for entry $(x_1^i, x_2^j)$ would be

$$K/M * \exp\{-(x_3^m)^2/2\alpha^2\} * \exp\{-[(x_1^i - x_1^m)^2 + (x_2^j - x_2^m)^2]/2\alpha^2\} \qquad [4.2]$$

and the degenerate contribution factor would be same for all the entries.

In the foregoing analysis we have assumed that all the variables are continuous. However if some variables are inherently discrete or can be reasonably assumed to be so, then the computational requirements may be reduced considerably. The difference terms in the exponential factor become zero for the discrete variables. If all the variable are discrete then the maximum increment becomes K/M, and the increments for the immediate neighbors may be obtained by efficient table look up procedures.

The third type of density to be estimated, the conditional has a general form (dropping the indices used in the decomposition)

$$p(x_1, x_2 .. \mid d(y_1, y_2 .. )).$$

Assume that previous training gives consistent estimates of $d(y_1, y_2 ..)$ and we have "suitably" quantized the range of d into Q intervals. Thus the value of Q is also stable and consistent. Now the required conditional density is obtained by a set of Q second level tables. Each table in effect estimates a separate marginal density function, or

$$p_M^j(x_1, x_2, x_3, \ldots | j) \; .. \qquad j = 1 \text{ to } Q,$$

where the j merely acts as a switch to choose one of the Q second level tables.

## 4.1 Quantization of the Degenerate Density

Appropriate quantization of the degenerate density turns out to be the focal issue in this approach. The degree of quantization determines the trade-off between savings in storage(and computation) against the required accuracy. At the risk of being repetitious we may say that if the degenerate is not quantized at all then the storage required is same as for a large, single, one level table. Whereas fewer the intervals into which the degenerate is quantized, more is the reduction in the storage.

There are two possible approaches to the quantization problem. We may treat the estimation of a class conditional probability independently of other classes or we may cross reference between classes at lower levels in the signature table hierarchy. First consider quantization of a degenerate independently of other classes. Also assume that we are resource bound and the number of intervals into which it must be divided (Q) is prespecified.

Division of the degenerate range in Q equal intervals may be ruled out. This would mean that intervals reflecting lower density would have very few samples at the next level. Q intervals spaced equally over logarithm of the degenerate will give more equitable distribution of samples at the next level.

Though this would not be optimal if the underlying multivariate density does not have some exponential form, or if it is multimodal.

The quantization may be made dependent upon the data accumulated in the table itself. Each interval is chosen such that the integral of the associated marginal density over that interval is 1/Q. Computationally this involves 1) ordering the accumulated degenerate values and 2) summing up the corresponding marginal values until the sum equals 1/Q of the total and placing the interval boundary at this point. The process is time consuming as it requires sorting the degenerate values in a table. However this quantization need only be done after sufficiently large number of samples have been processed so as to give stable quantization boundaries. This method of degenerate quantization was used for the recognition experiment reported in the next section.

The quantization may be made error bound if Q is not prespecified. It would also involve a sort of the degenerate values. The interval boundaries may then be located such that every degenerate value is within the specified error from the nearest boundary.

The quantization methods discussed so far attempt to get the best possible multivariate density estimate for a single class. With simultaneous quantization of the appropriate degenerates of all the class categories to be recognized, it may be possible to minimize the misclassification rate and also minimize the storage requirement.

First consider a simple case with only two classes. The quantization boundaries should be placed at a value where the two degenerate functions are equal. The approximate boundry location may be found by comparing the ordered sets of degenerate values. Clearly, the misclassification rate is given by the sum of all the marginal densities for which the the degenerates are below this value. The exact boundry value may then be chosen so as to minimize these sums. The implication of following such a procedure in a two class situation is that we have located the optimum discriminant boundry, and 1 bit quantization (Q=2) of the degenerate is sufficient.

In a multi-category situation a similar procedure involves simultaneous location of the discriminant boundaries between all the possible pairs of class categories so as to minimize the overall misclassification rate. A sub-optimal approach is possible which avoides the minimization problem. For a given class locate all the cross-over points where the degenerates of all other classes are nearly equal the degenerate of this class and provide a quantization with maximum resolution in the cross-over range.

This approach to quantization with cross reference between the classes to be recognized, would certainly produce near optimal results. However it would tend to be highly sensitive to the stationarity of the underlying probability distributions. One bad category whose probability distribution changes with time could drastically alter the overall performance of the system. However if each class conditional probability density is obtained independently of the other classes, then a drifting category would affect only the nearby categories and could easily be identified.

## 5. Experiments

Experiments in vowel recognition illustrate the application of the signature table method of probability density estimation discussed above. The data used for the experiments is derived from 51 words spoken by one speaker. Vowel data extracted from 26 words is used for training the tables. Data derived from the remaining 25 words is used for recognition.

The speech is digitized with a sampling rate of 20 kHz and 12-bit quantization. Frequency domain representation of the speech is obtained by taking 256 sample FFT(12.8 msec.) with 128 sample overlap between successive FFTs. A set of 19 parameters, such as three major peaks in specified frequency ranges and their corresponding amplitudes, energies average) in certain frequency regions etc., are obtained by measurements on each FFT. However, in the following limited experiments we have used only the three vowel formants and their amplitudes, since these six parameters are known to be significant for vowel perception.

Each parameter measurement is scaled and quantized to a 6-bit value in the first instance. For more details of the parameterization, the reader is refered to [2]. Every parameter vector, one every 6.4 msec. of speech, is given an appropriate tag according to the vowel category to which it belongs. This labeling is done by visual inspection of the speech waveform. The mnemonics that have been used to identify the 12 vowel categories are given in Table 1.

Three types of analysis are performed to give a comparative evaluation of the signature table method.

1) A nearest-neighbor analysis [9] is done using the full 6-bit range of the six parameters. The results of this analysis serve as a guide for the

interpretation of other results and also give a feel for the inherent overlap in the data.

2) The six parameters are assumed to be independent of each other. The quantization is reduced from 6-bits to 3-bits and a single input signature table is used to find the class conditional probability for a parameter for each category. Thus, this method requires 72 tables each having 8 entries, a total storage of 576 words. The six dimensional joint probability for a category is obtained as the product of the six individual probabilities.

3) The probability density estimation method outlined in this report gives the third set of results. The implementation details and the approximations used therein, are given in the next section.

## 5.1 Implementation Details

The block schematic of the five level signature table arrangement used to generate the required 6 variable density is shown in Fig.3. All the inputs are quantized to 3 bits. The size of each table is thus 64 words, The total storage required is therefore 5*64*12, or 3840 words.

Since all the inputs are discrete, the exponential factor in the evaluation of a marginal density (eq. 4.1) becomes 1. The multiplicative factor K/M ensures convergence of the estimate for large values of M. But it also has the effect of weighing down the the samples which come later in the training sequence.Since the maximum number of training samples for a category in the present experiment was only 86, all the samples were given equal weight. Therefore the increment used to to generate the marginal density is 1.

The evaluation of a degenerate increment involves an exponential factor(eq. 4.2). For example, the first table in Fig.3 has four degenerate inputs. Therefore this factor would be

$$\exp(-(A1^2 + A2^2 + F1^2 + F2^2)/2*\alpha^2)$$

The quantization methods discussed in Sec. 4.1 show that the absolute value of the degenerate is unimportant so long as the relative ordering of the entries within a table is maintained. The exponential may therefore be approximated by using only the first term in the expansion for each input, i.e.

$$(1.-(A1/7)^2)*(1.-(A2/7)^2)*(1.-(F1/7)^2)*(1.-(F2/7)^2).$$

Also the contribution of a training sample to its neighbors is assumed to be negligible as the quantization itself is rather gross. The factor α, which determines the smoothness of the distribution is also neglected.

Now consider how the six term conditional probability is obtained using the various marginal densities computed by the lower level tables in Fig.3. In the factorization

$$p(F1,F2,F3,A1,A2,A3) = p(F2|\ d(F1,F3,A1,A2,A3))*p(F1,F3,A1,A2,A3)$$

the first factor on the r.h.s. is given by the marginal output of Table 5. The second factor is the five term marginal output of the Table 4, which in turn has been obtained using a similar factorization. The configuration given in Fig.3 is repeated for all the 12 categories.

As the total number of training samples, 738, is rather small and each level in the table hierarchy must be some what stable before meaningful training of the next higher level can be done, the same data is used in five passes over the set of tables. After each pass the next higher level tables are enabled, so that accumulation of counts in appropriate entries can be started.

## 5.2 Results

Exactly the same data as used for the training is used for the classification experiment. The objective is to determine the error introduced by the gross quantization and the hierarchic organization of the signature tables. Clearly, the nearest-neighbor procedure would give 100% correct classification. Therefore, scatter matrix shown in Table 2 is produced by finding the sample vector which is closest without giving an exact match. The overall correct figure of 65.9% indicates that on the average 34% of the training samples have a neighbor of a different kind. The classification result using the the present method (74.3%, Table 3) shows that even with 3-bit quantization, the probability estimate does improve the classification. As expected, the results obtained when independence is assumed are poorer(58.1%, Table 4). The vowels which are weakly articulated (AS,I,A,U) have more variability in the data and tend to get swamped by the stronger vowels.

The recognition results obtained with unknown samples which have been extracted from 25 words, are given in Tables 5,6 and 7. The reader is probably appalled by the low overall recognition rates. This "bad" example has been chosen with a purpose. None of the vowels used in the training set occur in exactly the same phonemic context as in the recognition set. Because of the high context sensitivity of some of the vowels, in particular AS,I,A,AR and U, most of the unknown samples tend to fall in the empty space between the "learned" categories. The result of the nearest-neighbor analysis is 34.6% (Table 5) compared with the signature table method result of 32.9% (Table 6). The signature tables thus have a comparable performance , even with 3-bit quantization. The apparently superior performance obtained when independence between the inputs is assumed (36.1%), is at the expense of the weaker categories as seen in Table 7.

However, to get back to the main purpose of choosing this particular example, the signature table method allows us to defer making a decision until a wider context has been analyzed. A compound decision can then be made using this contextual information. If one is allowed to consider the second choices in the above example, even when no context is taken into account, the recognition score increases to 48.8%, a rise of 16% as shown in Table 8.

It is also obvious that the context sensitivity in this data set is some what contrived, 1) by rather arbitrary division of the list of 51 words into two sets, and 2) by using only 738 samples for training against the 465 used in recognition. Clearly, in actual usage the training set would have an order of magnitude more samples and also those would be derived from a more representative context.

## 6. Conclusions

We have shown that it is possible to decompose a higher dimensional probability density function into two factors whose dimensionality is less than the original function. The decomposition is iterative and hence the dimensionality of the functions which are actually evaluated can reduced to any desired order. However the errors are propogated in each iteration and savings in storage accrued must be balanced against the desired accuracy.

The signature table method of training is shown to be effective for the estimation of the various probability density functions which arise as the result of the decomposition. The signature table method 1) does not require assumptions regarding the underlying probability distribution, 2) allows sequential introduction of the training samples, and 3) is very efficient: the estimation process reduces to simple counting when the input variables are discrete.

The disadvantages of the method are 1) the errors tend to propagate from one level to the next in the table hierarchy and, 2) the number of training samples required grows in proportion to the number of levels, so as to ensure overall convergence of the final estimate.

The first set of classification experiments show that even with possibly the worst decomposition, where a six dimensional density is decomposed into 5 levels with 2 inputs each, and with a reduced 3-bit quantization, the signature table method has a better performance than the corresponding nearest-neighbor and independence-assumption experiments.

The second set of recognition experiments is some what artificially contrived to highlight the application of this method to speech recognition and other similar applications where high degree of context sensitivity makes it imperative to have some confidence estimate of the purely local decision. A compound decision can then be made using these local estimates. Ofcourse, the basic assumption here is that the combinetorics of the problem rule out a compound decision which is based on the basic feature measurements over all of the desired context.

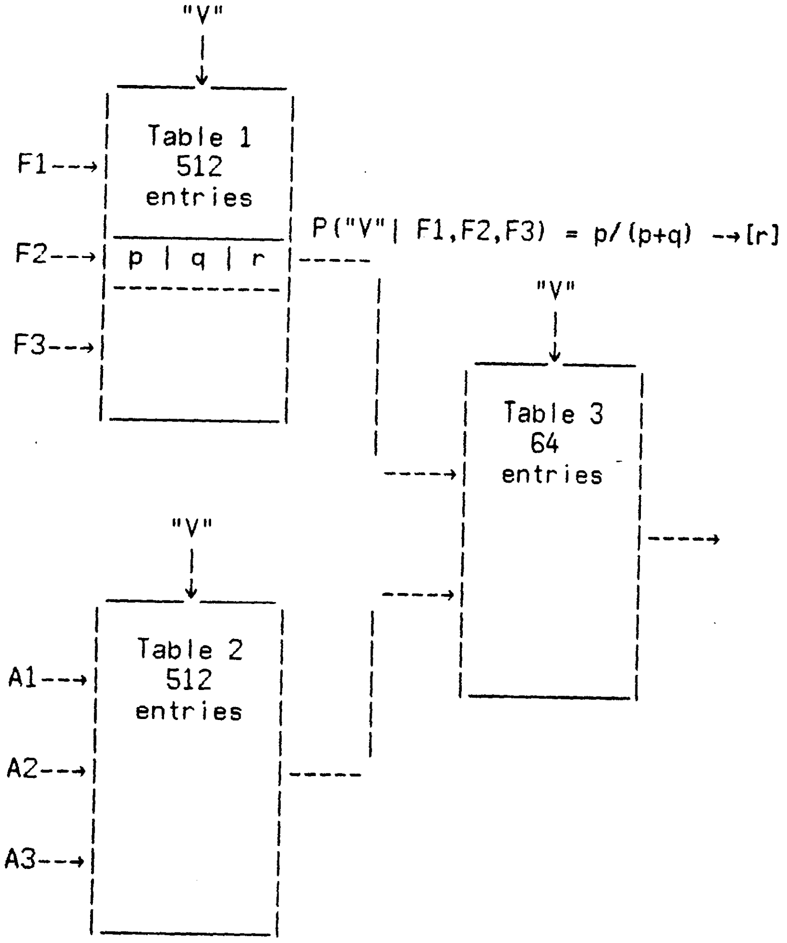

Fig.1 A Simplified Example of a Two-level Signature Table Arrangement Used in the Speech Recognition Program [2].

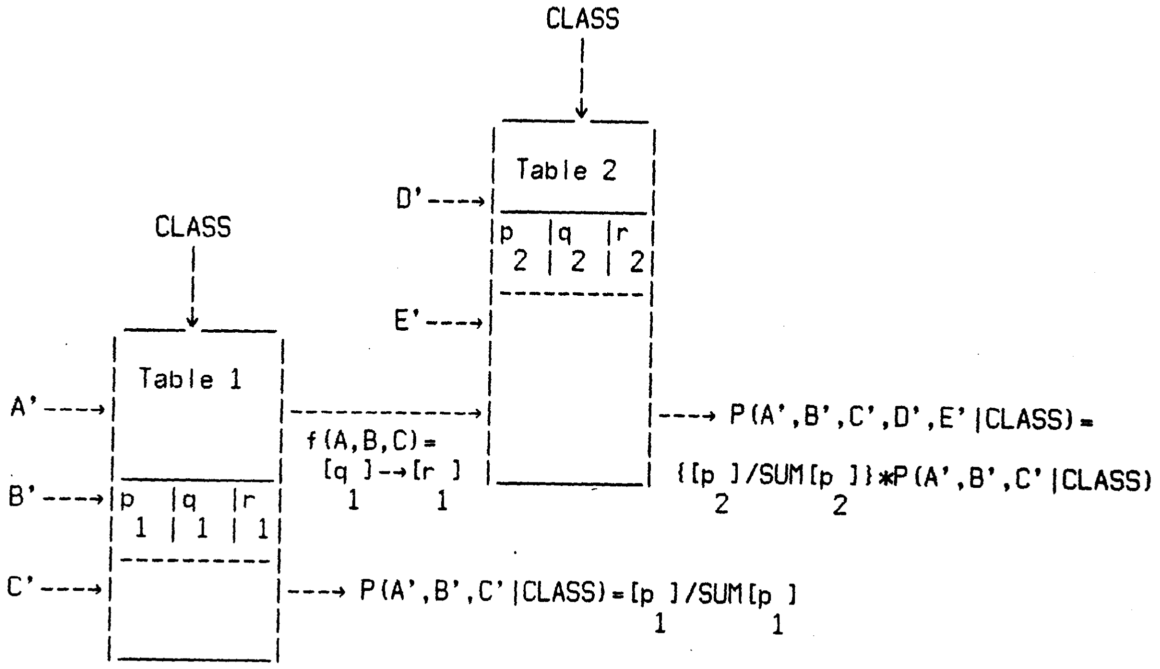
CLASS
Table 2
D'
$p_2$ $q_2$ $r_2$
E'
CLASS
Table 1
A'
B'
$p_1$ $q_1$ $r_1$
C'
$f(A,B,C) = [q_1] \rightarrow [r_1]$
$P(A',B',C',D',E'|CLASS) = \{[p_2]/SUM[p_2]\} * P(A',B',C'|CLASS)$
$P(A',B',C'|CLASS) = [p_1]/SUM[p_1]$

Fig.2. A Simplified Two-level Signature Table Arrangement

Specific values of of the inputs say, A',B' and C', when concatenated form an address which points to the entry shown. During training if CLASS is indicated for this input then the column $[p_1]$ is incremented by 1 and column $[q_1]$ is incremented by the function f. The table outputs in column $[r_1]$ are obtained by quantization of the values in $[q_1]$. The probability calculation performed during the recognition phase is as shown in the figure.

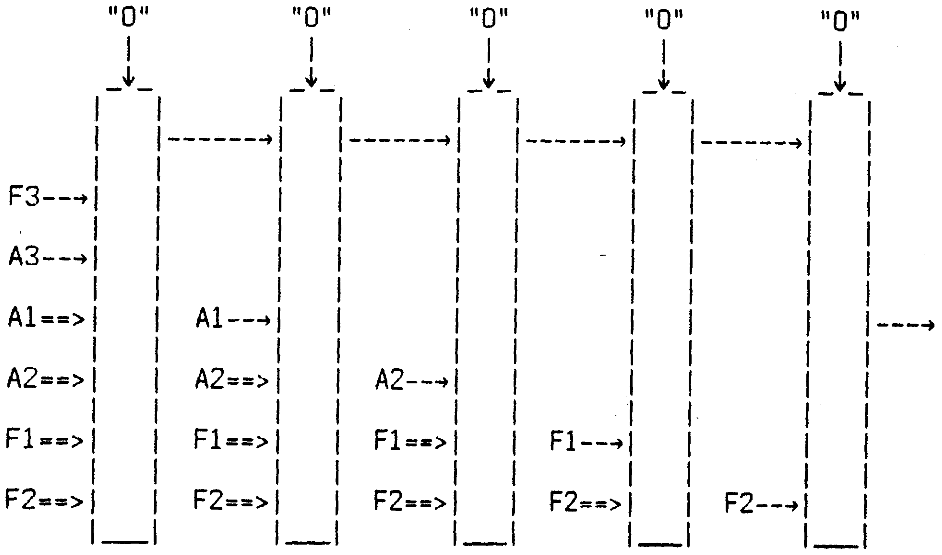

Fig.3 A Six-input, Five-level Signature Table Arrangement Used in the Experiments. A ==> Indicates a Degenerate Input.

|  |  |  |  |
|---|---|---|---|
| EE : beet | AE : bat | E : bait | I : bit |
| AS : a.but | AA : bar | AR : bird | A : but |
| AW : bought | OO : boot | O : boat | U : book |

Table 1. Vowel Mnemonics Used in the Experiments

Given

| | | EE | AE | E | I | AS | AA | AR | A | OO | U | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F | EE | 60 | 2 | 1 | 5 | 1 | | 10 | 5 | | 3 | |
| o | AE | 3 | 62 | 6 | 4 | | | | | | 1 | |
| u | E | 3 | 5 | 52 | 7 | 2 | | 1 | 1 | | 2 | |
| n | I | 5 | 2 | 2 | 31 | 2 | | 5 | | | 6 | |
| d | AS | 2 | 2 | | 3 | 21 | 1 | 1 | 3 | 4 | 4 | 5 |
| | AA | 4 | | | | 1 | 21 | 4 | | | 6 | |
| | AR | 3 | | 4 | 6 | 1 | 8 | 50 | 5 | | 4 | 5 |
| | A | 1 | | | | 1 | 1 | 5 | 49 | 4 | 3 | 3 |
| | OO | 1 | | | 3 | 1 | | | 2 | 56 | 2 | 2 |
| | U | 2 | | 1 | 8 | 2 | 3 | 3 | 6 | 2 | 28 | 6 |
| | O | | 2 | | | 6 | 1 | 7 | 3 | 2 | 8 | 57 |
| Total | | 84 | 75 | 66 | 67 | 38 | 35 | 86 | 74 | 68 | 67 | 78 |
| %Found | | 71 | 83 | 79 | 46 | 55 | 60 | 58 | 66 | 82 | 42 | 73 |

Overall correct 65.98 %

Table 2. Classification Result of Nearest-Neighbor Analysis

Given

| | | EE | AE | E | I | AS | AA | AR | A | OO | U | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F | EE | 51 | 1 | 1 | 1 | | | | 1 | | | |
| o | AE | 4 | 54 | 3 | 3 | | | 1 | | | 1 | |
| u | E | 6 | 1 | 53 | 2 | 1 | | 5 | 1 | | 2 | |
| n | I | 4 | 7 | 4 | 46 | | | | | | | |
| d | AS | 4 | 1 | 2 | 9 | 29 | | 7 | | 2 | 7 | 5 |
| | AA | 3 | 5 | 2 | | 4 | 33 | 6 | 6 | | 4 | |
| | AR | 1 | | | 1 | | | 45 | 1 | | 1 | 1 |
| | A | 2 | | | 1 | | | 5 | 59 | | 1 | 2 |
| | OO | 5 | | 1 | 4 | 1 | | 5 | | 66 | 2 | |
| | U | 2 | 2 | | | 1 | 2 | | 3 | | 44 | 2 |
| | O | 2 | 4 | | | 2 | | 12 | 3 | | 5 | 68 |
| Total | | 84 | 75 | 66 | 67 | 38 | 35 | 86 | 74 | 68 | 67 | 78 |
| %Found | | 61 | 72 | 80 | 69 | 76 | 94 | 52 | 80 | 97 | 66 | 87 |

Overall correct 74.26 %

Table 3. Classification Result Using Signature Tables

Given

|  |  | EE | AE | E | I | AS | AA | AR | A | OO | U | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F | EE | 50 | 4 | 4 | 12 | 4 | 8 | 1 | 2 | 4 | 3 |  |
| o | AE | 7 | 63 | 11 | 9 |  | 1 |  | 1 |  |  |  |
| u | E | 4 |  | 45 | 4 | 4 |  | 6 |  |  | 2 |  |
| n | I | 1 | 1 | 2 | 25 | 7 |  | 1 |  |  | 8 | 1 |
| d | AA |  |  |  |  |  | 2 |  |  |  |  |  |
|  | AR | 18 | 3 | 3 | 9 | 12 | 17 | 66 | 11 | 8 | 8 | 9 |
|  | A | 2 | 1 |  |  | 1 | 1 | 2 | 38 | 2 | 1 | 1 |
|  | OO | 2 |  |  | 2 |  |  | 1 | 10 | 52 | 3 |  |
|  | U |  |  | 1 | 6 | 1 | 2 |  | 1 | 2 | 24 | 3 |
|  | O |  | 3 |  |  | 9 | 4 | 9 | 11 |  | 18 | 64 |
| Total |  | 84 | 75 | 66 | 67 | 38 | 35 | 86 | 74 | 68 | 67 | 78 |
| %Found |  | 60 | 84 | 68 | 37 | 0 | 6 | 77 | 51 | 76 | 36 | 82 |

Overall correct 58.13 %

Table 4. Classification Result with Independence Assumption

Given

|  |  | EE | AE | E | I | AS | AA | AR | AW | O |
|---|---|---|---|---|---|---|---|---|---|---|
| F | EE | 42 |  | 3 | 7 |  | 16 | 10 | 10 | 1 |
| o | AE | 2 | 16 | 14 | 21 | 6 | 1 | 3 | 2 |  |
| u | E | 3 | 7 | 16 | 17 | 10 |  | 1 |  |  |
| n | I | 3 |  | 8 | 13 | 1 | 1 | 15 |  |  |
| d | AS | 2 |  | 3 | 3 | 6 | 7 | 10 |  | 8 |
|  | AA | 3 |  |  |  | 3 | 9 | 1 | 3 | 3 |
|  | AR | 4 |  | 2 | 9 | 4 | 4 | 27 | 1 | 8 |
|  | A |  |  | 2 | 1 | 2 | 3 |  |  | 8 |
|  | OO | 4 |  |  |  |  |  |  |  | 1 |
|  | U | 2 |  | 13 | 4 | 3 | 1 | 4 |  | 5 |
|  | AW |  |  |  |  |  |  |  |  |  |
|  | O |  |  |  |  | 8 | 1 |  | 2 | 32 |
| Total |  | 65 | 23 | 61 | 75 | 43 | 43 | 71 | 18 | 66 |
| %Found |  | 65 | 70 | 26 | 17 | 14 | 21 | 38 | 0 | 48 |

Overall correct 34.62 %

Table 5. Recognition Result of Nearest-Neighbor Analysis

Given

| | | EE | AE | E | I | AS | AA | AR | AW | O |
|---|---|---|---|---|---|---|---|---|---|---|
| F | EE | 31 | 1 | | 4 | | 3 | 1 | | |
| o | AE | 5 | 9 | 7 | 16 | 5 | 6 | | 3 | |
| u | E | 3 | 4 | 28 | 12 | 2 | | 9 | | |
| n | I | 8 | 2 | 12 | 8 | 1 | 1 | 8 | | 3 |
| d | AS | 2 | 4 | 11 | 24 | 14 | 5 | 29 | | 8 |
| | AA | 3 | 1 | 2 | 2 | 4 | 23 | 1 | 9 | 9 |
| | AR | | | | 1 | 1 | 2 | 7 | | 4 |
| | A | 2 | | | 1 | 2 | 2 | 2 | | 4 |
| | OO | 6 | | | 4 | 1 | | 9 | | |
| | U | 1 | 2 | | 1 | 5 | 1 | 4 | 1 | 5 |
| | O | 4 | | 1 | 2 | 8 | | 1 | 5 | 33 |
| Total | | 65 | 23 | 61 | 75 | 43 | 43 | 71 | 18 | 66 |
| %Found | | 48 | 39 | 46 | 11 | 33 | 53 | 10 | 0 | 50 |

Overall correct 32.9 %

Table 6. Recognition Result Using Signature Tables

Given

| | | EE | AE | E | I | AS | AA | AR | AW | O |
|---|---|---|---|---|---|---|---|---|---|---|
| F | EE | 39 | 1 | 7 | 14 | 3 | 11 | 1 | 6 | 2 |
| o | AE | 7 | 21 | 16 | 29 | 6 | 5 | 1 | 1 | |
| u | ,E | 1 | 1 | 23 | 15 | 7 | 3 | 5 | | |
| n | I | 3 | | 7 | 10 | 3 | 1 | 19 | | |
| d | AR | 14 | | 8 | 4 | 9 | 21 | 32 | 5 | 15 |
| | A | | | | | | 1 | 2 | | 6 |
| | OO | 1 | | | 1 | | | 4 | | |
| | U | | | | 2 | 8 | 1 | 7 | 1 | |
| | O | | | | | 7 | | | 5 | 43 |
| Total | | 65 | 23 | 61 | 75 | 43 | 43 | 71 | 18 | 66 |
| %Found | | 60 | 91 | 38 | 13 | 0 | 0 | 45 | 0 | 65 |

Overall correct 36.13 %

Table 7. Recognition Result with Independence Assumption

Given

|  |  | EE | AE | E | I | AS | AA | AR | AW | O |
|---|---|---|---|---|---|---|---|---|---|---|
| F | EE | 39 |  |  | 3 |  | 1 | 1 | 4 |  |
| o | AE | 4 | 17 | 7 | 7 |  | 2 | 3 | 5 | 5 |
| u | E |  | 1 | 35 | 8 | 2 | 2 | 4 |  | 3 |
| n | I | 4 | 1 | 6 | 21 | 1 | 2 | 4 |  | 1 |
| d | AS | 8 | 2 | 6 | 17 | 24 |  | 16 | 1 | 1 |
|  | AA | 1 | 1 |  | 3 | 2 | 29 | 3 | 6 | 2 |
|  | AR | 2 |  |  | 5 | 2 |  | 17 |  | 3 |
|  | A | 2 |  |  | 3 | 3 | 2 | 3 | 1 | 4 |
|  | OO | 2 |  |  | 4 | 1 | 2 | 9 |  |  |
|  | U | 2 |  | 6 | 2 | 3 | 2 | 3 |  | 2 |
|  | O | 1 | 1 | 1 | 2 | 5 | 1 | 8 | 1 | 45 |
| Total |  | 65 | 23 | 61 | 75 | 43 | 43 | 71 | 18 | 66 |
| %Found |  | 60 | 74 | 57 | 28 | 56 | 67 | 24 | 0 | 68 |

Overall correct 48.82 %

Table 8. Recognition Result with Second Choice Considered

## References

1] A.L.Samuel,"Some Studies in Machine Learning Using the game of Checkers.II-Recent Progress", IBM Jour. of Research and Development,11,601-617,November,1967.

2] R.B.Thosar and A.L.Samuel,"Some Preliminary Experiments in Speech Recognition Using Signature Table Learning," SUR Note 43, ARPA Network Information Center, NIC 11621, Stanford Research Institute, Menlo Park, California 94025.

3] A.L.Samuel and R.B.Thosar,"Recognition of Continuous Speech: Segmentation and Classification Using Signature Table Adaptation," Stanford AI Memo under preparation, Computer Science Dept., Stanford, California 94305.

4] E.Parzen,"On Estimation of a Probability Density Function and Mode", Ann.Math.Stat.,33,1065-1076,September,1962.

5] V.K.Murthy,"Estimation of Probability Density," Ann.Math.Stat.,36, 1027-1031,June,1965.

6] H.C.Andrews,Introduction to Mathematical Techniques in Pattern Recognition,Wiley-Interscience,1972.

7] G.S.Sebestyen,"Pattern Recognition by an Adaptive Process of Sample Set Construction," IRE Trans. on Information Theory,IT-8,S82-S91, September,1962.

8] D.F.Specht,"Generation of Polynomial Discriminant Functions for Pattern Recognition," IEEE Trans.on Electronic Computers,EC-16, 308-319,June,1967.

9] T.M.Cover and P.E.Hart,"Nearest Neighbor Pattern Classification," IEEE Trans. on Information Theory,IT-3, 21-27, Jan. 1967.

## Index Terms

Probability density estimation, decomposition of multivariate density functions, signature tables, speech recognition, pattern recognition.

Affiliation of author.

The author is with the Artificial Intelligence Project, Computer Science Department, Stanford University, Stanford, CA 94305, on leave of absence from the Tata Institute of Fundamental Research, Bombay, India.

This research was supported by the Advanced Research Projects Agency of the Office of the Secretary of Defense under Contract SD-183.

Figure and Table Captions.

1] Fig.1 A Simplified Example of a Two-level Signature Table Arrangement Used in the Speech Recognition Program [2].

2] Fig.2 A Simplified Two-level Signature Table Arrangement.

Specific values of the inputs say, A',B' and C', when concatenated form an address which points to the entry shown. During training if CLASS is indicated for this input then the column $[p_1]$ is incremented by 1 and column $[q_1]$ is incremented by the function f. The table outputs in column $[r_1]$ are obtained by quantization of the values in $[q_1]$. The probability calculation performed during the recognition phase is as shown in the figure.

3] Fig.3 A Six-input, Five-level Signature Table Arrangement Used in the Experiments. A ==> indicates a Degenerate Input.

4] Table 1. Vowel Mnemonics Used in the Experiments.

5] Table 2. Classification Result of Nearest-Neighbor Analysis.

6] Table 3. Classification Result Using Signature Tables.

7] Table 4. Classification Result with Independence Assumption.

8] Table 5. Recognition Result of Nearest-Neighbor Analysis.

9] Table 6. Recognition Result Using Signature Tables.

10] Table 7. Recognition Result with Independence Assumption.

11] Table 8. Recognition Result with Second Choice Considered.