Linguistic Recognition System Based on Approximate Reasoning

SANKAR K. PAL

and

DEBA PRASAD MANDAL

Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta 700 035 India

Communicated by Abraham Kandel

ABSTRACT

A linguistic recognition system based on approximate reasoning has been described which is capable of handling various imprecise input patterns and of providing a natural decision. The input feature is considered to be of either linguistic form or quantitative form or mixed form or set form. An input has been viewed as consisting of various combinations of the three primary properties SMALL, MEDIUM and HIGH possessed by its different features to some degree. The various uncertainty (ambiguity) in the input statement has been managed by providing/modifying membership values heuristically to a great extent. Unlike the conventional fuzzy set theoretic approach, the sets SMALL and HIGH have been represented here by π-functions. The weight matrices corresponding to various properties and classes have been taken into account in the composition rule of inference in order to make the analysis more effective. The natural output decision is associated with a confidence factor denoting the degree of certainty of the decision, thus providing a low rate of misclassification as compared to the conventional two-state system. The effectiveness of the algorithm has been demonstrated on the speech recognition problem.

1. INTRODUCTION

Zadeh has developed a theory of approximate reasoning [1] based on fuzzy set theory. This theory aims at modeling the human reasoning and thinking process with linguistic variable in order to handle both soft and hard data as well as various types of uncertainty. Many aspects of the underlying concept have been incorporated in designing decision-making systems [2–6] along with their applications.

655 Avenue of the Americas, New York, NY 10010 0020-0255/92/$5.00

The present work is an attempt to demonstrate the application of approximate reasoning in designing a general purpose linguistic recognition system. This is part of the investigation on the project "Approximate Reasoning and Knowledge Based Linguistic Recognition System" being carried out in the ECSU, Indian Statistical Institute, Calcutta.

In the conventional statistical [7] or syntactic [8] classifiers, the input patterns are quantitative (exact) in nature. The patterns having imprecise or incomplete information are usually ignored or discarded from their designing and testing processes. The impreciseness (or ambiguity) [9, 10] may arise for various reasons. For example, instrumental error or noise corruption in the experiment may lead to having partial/unreliable information available on a feature measurement F, viz., F is about 500, say (mixed form), or F is between 400 and 500, say (set form). Again, in some cases the expense incurred in extracting exact value of feature may be high, or it may be difficult to decide on the actual salient features to be extracted; on the other hand, it may become convenient to use the linguistic variables and hedges, e.g., small, medium, high, very, more or less, etc. in order to describe feature information (viz., F is very small). There has recently been an attempt [11] to provide the design concept of a classifier which needs enough *a priori* knowledge from experts, in linguistic form only, regarding the classification problem (viz. medical diagnosis).

The proposed classifier is designed to be capable of handling all the aforementioned impreciseness in pattern without consulting any expert. The classifier can be viewed as "general" because it takes feature input in both exact and inexact forms. As the linguistic representation contains summarized information. it is difficult to convert the linguistic information into a quantitative form. On the other hand, it is easier to convert any information into linguistic form. Keeping this in mind, the algorithm considers only three primary linguistic properties, namely, SMALL, MEDIUM, and HIGH so that any input information can be thought of as possessing various combination of these properties to some degree. Based on these properties, the various membership values of imprecise input are assigned and modified to a great extent heuristically. The compatibility functions for SMALL and HIGH have been represented by π functions. Note that this is a major deviation from the standard fuzzy set theoretic approach where these are represented by $(1 - S)$- and S-type functions, respectively. Since all the primary feature properties are not equally important in characterizing a class, a concept of weighting coefficient has also been introduced. The system uses Zadeh's compositional rule of inferences [1] and gives a natural output decision associated with its certainty (or validity). Finally, the effectiveness of the proposed linguistic system has been demonstrated on a speech recognition problem, where the classes have ill-defined boundaries and the input feature information have the aforementioned impreciseness.

In Section 2, an introduction to fuzzy set theory along with the concept of linguistic variable and approximate reasoning is provided. Section 3 gives a brief description of the recognition system. The description of different blocks are provided in Sections 4, 5, and 6. In Section 7, a discussion on the nature of output is given. Results on speech recognition problems are discussed in Section 8. Section 9 finds the conclusion.

2. FUZZY SET THEORY

DEFINITION 1 [12, 13]. Let X be a set, called the universe. The characteristic function μ_A of a classical subset A of X takes its value in the two-element set $\{0,1\}$ and is such that $\mu_A(x) = 1$ if $x \in A$ and 0 otherwise. A fuzzy set A has a characteristic function taking its value in the interval $[0,1]$. $\mu_A(x)$ is the grade of membership of $x \in X$ in A. A is symbolically denoted as

$$A = \{(x, \mu_A(x))\}, \qquad x \in X. \tag{1}$$

In a fuzzy set, the transition between membership is gradual rather than abrupt.

Further, the property p defined on an event x is a function $p(x)$, which can have values in the interval $[0, 1]$. A set of these functions $p(x)$, which assigns the degree of possession of some property p by an event x, constitutes what is called a property set. For example, p_n may denote the property that the outer boundary of a pattern is circular or straight line or that a person is blonde, tall, etc. or that a number is high, medium, small, etc. [2].

Assignment of membership function of a fuzzy set is subjective in nature and depends on the problem. It cannot be assigned arbitrarily. There are two standard membership functions [2, 14], called the S-function and the π-function. S-function is defined as

$$S(x;\alpha,\beta,\gamma) = \begin{cases} 0 & \text{for } x \leqslant \alpha, \\ 2[(x-\alpha)/(\gamma-\alpha)]^2 & \text{for } \alpha \leqslant x \leqslant \beta, \\ 1-2[(x-\gamma)/(\gamma-\alpha)^2] & \text{for } \beta \leqslant x \leqslant \gamma, \\ 1 & \text{for } x \geqslant \gamma, \end{cases} \tag{2}$$

where β is the crossover point $[S(\beta;\alpha,\beta,\gamma) = 0.5]$. In general, $\beta = (\alpha+\gamma)/2$.

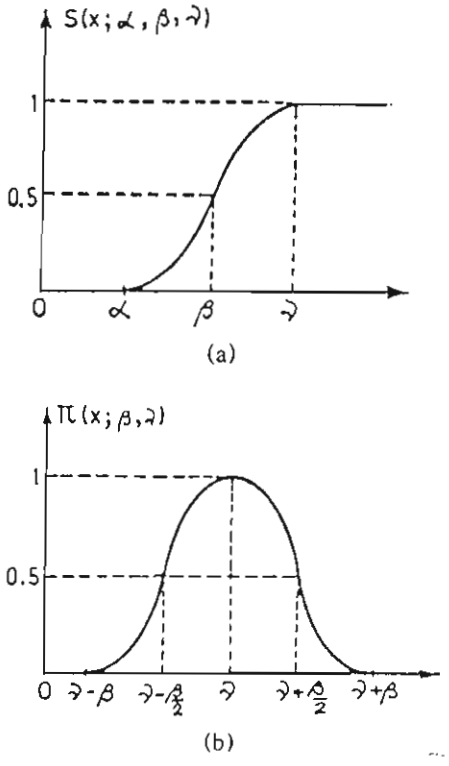

Fig. 1. a) S function and b) π function.

π-function is a combination of S and $(1-S)$ functions, defined as

$$\pi(x;\beta,\gamma) = \begin{cases} S(x;\gamma-\beta,\gamma-\beta/2\gamma,) & \text{for } x \leqslant \gamma, \\ 1-S(x;\gamma,\gamma+\beta/2,\gamma+\beta) & \text{for } x \geqslant \gamma. \end{cases} \tag{3}$$

Figure 1 shows the general structures of the standard S-function and π-function. The S-function denotes the compatibility function for the sets "x is large" whereas the π-function denotes the compatibility function for the set "x is γ." In other words, the sets SMALL, MEDIUM, and HIGH can be represented by $(1-S)$-, π-, and S-functions, respectively.

2.1. LINGUISTIC VARIABLES [15, 16]

By a linguistic variable, we mean a variable whose values are not numbers but words or sentences in a natural language.

DEFINITION 2 [15]. A linguistic variable is characterized by a quintuple $(X, T(X), U, G, M)$, in which X is the name of the variable; $T(X)$ is the term set of X; U is the universe of discourse; G is a syntactic rule which generates the terms in $T(X)$; and M is a semantic rule which associates with each linguistic value X its meaning where $M(X)$ denotes a fuzzy subset of U. For a particular X, the name generated by G is called a term.

EXAMPLE 1. Suppose X is a linguistic variable with the label "height" with $U = [0, 250]$. Terms of this linguistic variable, which are fuzzy sets, could be called "tall," "short," "very tall," and so on. The base variable U is the height in cm of persons. $M(X)$ is the rule that assigns a meaning, that is, a fuzzy set to the term

$$M(\text{tall}) = \{x, \mu_{\text{tall}}(x)\}, \qquad x \in [0, 250],$$

where

$$\mu_{\text{tall}}(x) = \begin{cases} 0 & \text{for } x \in [0, 150] \\ \left[1 + \{(x - 150)/10\}^{-2}\right]^{-1} & \text{for } x \in [150, 250]. \end{cases}$$

$T(X)$ will define the term set of the variable X. In this case

$$T(\text{Height}) = \{\text{tall, very tall, not very tall, quite tall, short, more or less short, ...}\}$$

where $G(X)$ is a rule which generates the terms in the term set $T(X)$.

DEFINITION 3 [15]. A linguistic hedge or a modifier is an operator which modifies the meaning of a term or more generally of a fuzzy set. If A is a fuzzy set, then the modifier m generates the composite term $b = m(A)$.

The mathematical models which are used very frequently for modifiers are

$$\text{Concentration: } \mu_{\text{CON}(A)} = (\mu_A(x))^2, \tag{4a}$$

$$\text{Dilation: } \mu_{\text{DIL}(A)} = (\mu_A(x))^{1/2}, \tag{4b}$$

Contrast intensification:

$$\mu_{\text{INT}(A)}(x) = \begin{cases} 2(\mu_A(x))^2 & \text{for } \mu_A(x) \in [0, 0.5], \\ 1 - 2(1 - \mu_A(x))^2 & \text{otherwise.} \end{cases} \tag{4c}$$

Generally, the following linguistic hedges are associated with the above-mentioned mathematical operators: If A is a term (a fuzzy set), then

$$\text{very } A \equiv \text{CON}(A), \tag{5a}$$

$$\text{more or less } A \equiv \text{DIL}(A) \tag{5b}$$

$$\text{plus } A \equiv A^{1.25}, \tag{5c}$$

$$\text{slightly } A \equiv \text{INT}\,[\text{plus } A \text{ and not (very } A)]. \tag{5d}$$

2.2. *APPROXIMATE REASONING*

By approximate reasoning, we mean a type of reasoning which is neither very exact nor very inexact. Consider a fuzzy proposition p of the form

$$p \equiv x \text{ is } Q, \tag{6a}$$

where x is a name of an object and Q is a label of a fuzzy subset of a universe X. p can be expressed by the relational assignment equation [14] as

$$R(A(x)) = Q, \tag{6b}$$

where A is an implied attribute of x, i.e., an attribute which is implied by x and Q; R denotes a fuzzy restriction on $A(X)$ to which the value Q is assigned by the relational assignment equation.

EXAMPLE 2. Let

$$p \equiv \text{This tomato is very red.}$$

So the corresponding relational assignment equation will be

$$R(\text{color (this tomato)}) \equiv \text{very red.}$$

For an illustration of approximate reasoning, let us consider the proposition p as a premise. The conclusion corresponding to an implication may be as follows:

Implication: If a tomato is red, then the tomato is ripe.
Conclusion: This tomato is very ripe.

In 1973, Zadeh [1] suggested the composition rule of inference for the above type of fuzzy conditional implication. Although, other authors [17, 18]

have suggested different methods, we have restricted ourselves to Zadeh's composition rule of inference while developing the system.

DEFINITION 4 [1]. Let A and B denote fuzzy sets in X and $X \times Y$, respectively. Then the composition rule of inference asserts that the solution of the relational assignment equations

$$R(x) = A \quad \text{and} \quad R(x, y) = B$$

is given by

$$R(y) = A \circ B = C, \tag{7}$$

where $A \circ B$ is the max–min composition of A and B.

EXAMPLE 3. Let the universe be $X = \{1, 2, 3, 4\}$.

A = little = $\{(1, 1.0), (2, 0.6), (3, 0.2), (4, 0.0)\}$

$B \equiv$ "approximately equal" be a fuzzy relation defined by

	1	2	3	4
1	1.0	0.5	0.0	0.0
2	0.5	1.0	0.5	0.0
3	0.0	0.5	1.0	0.5
4	0.0	0.0	0.5	1.0

Applying the max–min composition, $C(y) = A \circ B$ yields

$$C(y) = \max_x \min\{\mu_A(x), \mu_B(x, y)\}$$

$$= \{(1, 1.0), (2, 0.6), (3, 0.5), (4, 0.2)\}$$

$\equiv$ approximately little.

3. LINGUISTIC RECOGNITION SYSTEM

3.1. CONCEPT OF REPRESENTING A PATTERN

Based on the above-mentioned theories, let us now describe a recognition system which is capable of handling input pattern having feature information in a) linguistic form, b) quantitative form, c) mixed form, and d) set form. The primary linguistic terms or properties, under consideration, are SMALL, MEDIUM,

and HIGH so that each feature F in any aforementioned form can be converted to have these properties to some degree. Therefore, a pattern $X = [F_1, F_2, \ldots, F_N]$ can be represented as

$$X = \begin{vmatrix} \mu_{\text{SMALL}}(F_1) & \mu_{\text{MEDIUM}}(F_1) & \mu_{\text{HIGH}}(F_1) \\ \mu_{\text{SMALL}}(F_2) & \mu_{\text{MEDIUM}}(F_2) & \mu_{\text{HIGH}}(F_2) \\ \vdots & \vdots & \vdots \\ \mu_{\text{SMALL}}(F_N) & \mu_{\text{MEDIUM}}(F_N) & \mu_{\text{HIGH}}(F_N) \end{vmatrix}. \tag{8}$$

Representation of the imprecise input X through their primary properties SMALL, MEDIUM, and HIGH basically implies that the entire dynamic range of each feature has been divided into three overlapping subregions corresponding to these primary properties. So the whole feature space is divided into 3^N overlapping subspaces.

Let us now define a Property Combination Vector (henceforth PCV) consisting of 3^N components, which represent the various combinations of the basic properties SMALL, MEDIUM, and HIGH as possessed by the features of X. The components, viz. (F_1 is SMALL, F_2 is SMALL, ..., F_N is SMALL), (F_1 is SMALL, F_2 is SMALL, ..., F_N is MEDIUM), ..., (F_1 is HIGH, F_2 is HIGH, ..., F_N is HIGH), correspond to the above mentioned 3^N subregions of the feature space. In other words, we can write for $N = 2$

$$PCV = \begin{vmatrix} F_1 \text{ is SMALL}, & F_2 \text{ is SMALL} \\ F_1 \text{ is SMALL}, & F_2 \text{ is MEDIUM} \\ F_1 \text{ is SMALL}, & F_2 \text{ is HIGH} \\ \vdots & \vdots \\ F_1 \text{ is HIGH}, & F_2 \text{ is HIGH} \end{vmatrix}, \tag{9}$$

having $3^2 = 9$ components. These are explained in Figure 2, where the entire feature space has been decomposed into overlapping (fuzzy) subregions in order to represent the impreciseness in input feature information. Each component of PCV corresponds to one of the nine regions and represents a combination of the property sets "F_1 is p_i" and "F_2 is p_j," p_i and p_j taking values from {SMALL, MEDIUM, HIGH}. Value of a component of PCV therefore denotes the degree (joint possibility) to which F_1 and F_2 possess the properties p_i and p_j, respectively, i.e., the degree of belonging to one of the nine subregions of the feature space.

Now a pattern class can be viewed as consisting of all these property combinations to some extent. Depending on the significance (weight) of

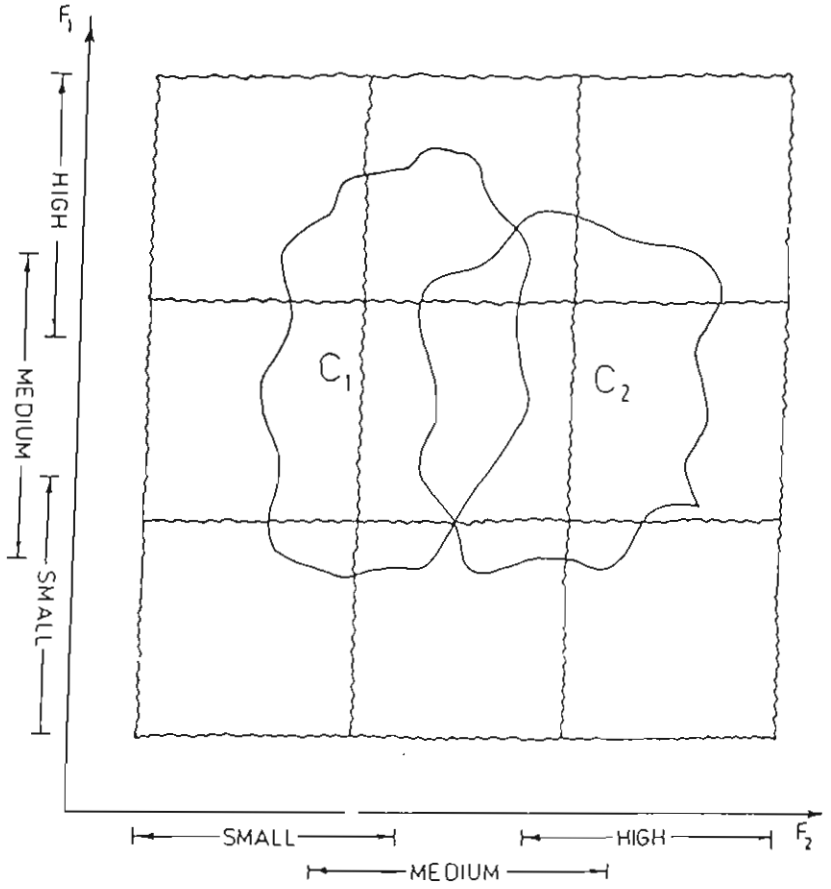

Fig. 2. Feature space showing nine overlapping regions in terms of the properties SMALL, MEDIUM, and HIGH. Curl lines denote fuzzy boundaries.

individual components of PCV in characterizing a class, a characterizing vector

$$CV_j(X) = \{cv_{ji}(X)\}_{i=1,2,\ldots,3^N}, \qquad j = 1,2,\ldots,M, \tag{10}$$

may be determined. Here M denotes the number of classes $C_1, C_2, \ldots, C_M$. The ith element of $CV_j(X)$ denotes the degree of possessing the ith property combination by X given that the X is from class C_j.

EXAMPLE 4. Consider the problem of identifying whether a student is from the science stream (C_1) or the humanities stream (C_2), from his marks in mathematics (F_1) and literature (F_2) in a combined examination. Assume that the marks are characterized by the linguistic properties SMALL (poor), MEDIUM, and HIGH (good). Let the marks obtained be 80 and 50, say, in mathematics and literature, respectively.

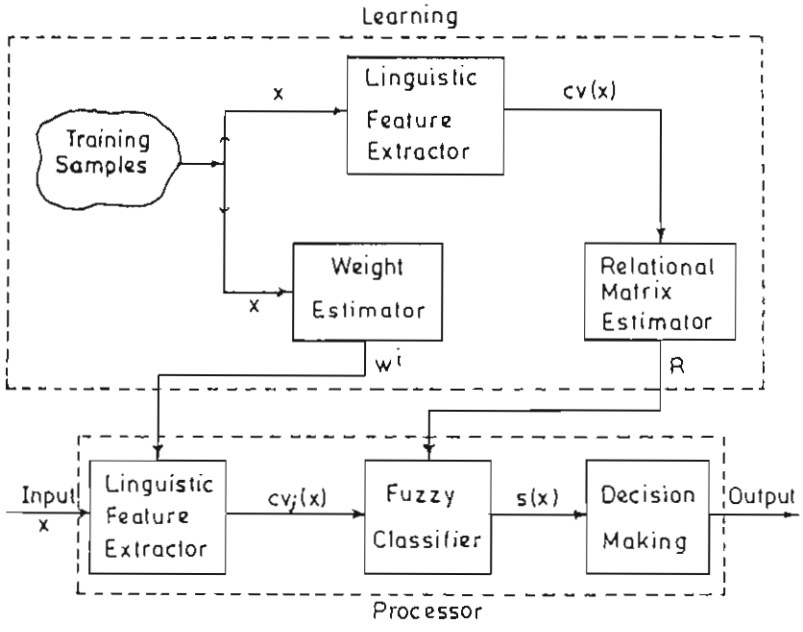

Fig. 3. Block diagram of the Linguistic Recognition System.

Consider a particular property combination (marks(math) is good, marks(lit) is good). Then, obviously, the degree of possessing this property combination is higher if the student is from science stream because the mark in mathematics is more important (i.e., has more weight) than that in literature, in characterizing a science student. The converse is true if the student is from the humanities stream. Therefore, CV(X) for the science stream will be different from that of the humanities.

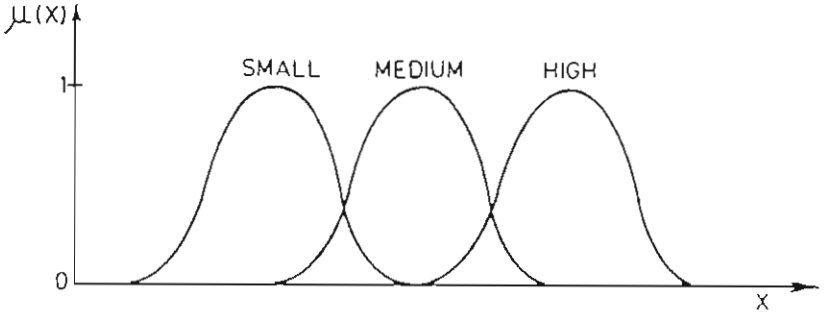

Fig. 4. Coexistance structure of the compatibility functions for SMALL, MEDIUM, and HIGH.

3.2. *BLOCK DIAGRAM*

The block diagram of such a recognition system is shown in Figure 3. It consists of two sections, namely Fuzzy Learning and Fuzzy Processor. The learning section uses the training samples and outputs a relational matrix and weight matrices to the processor. The fuzzy processor uses the relational matrix and the weight matrices to give a natural decision output regarding the classes from which the pattern X may come. Note that both the sections are having the block Linguistic Feature Extractor (LFE), which takes the input pattern X and outputs characteristic vectors. During learning, it outputs a characteristic vector $\mathrm{CV}(X)$ corresponding to the input X from the training samples, whereas, in the processor, it works as a weighted LFE, so that it outputs M characteristic vectors $\mathrm{CV}_j(X)$ with the help of weight matrices corresponding to an unknown pattern X.

The relational matrix R denotes the compatibility for the various pattern classes corresponding to the elements of PCV. The relational matrix is obtained from training samples. Each column of R corresponds to a class and each row of that column denotes the degree to which a class should be characterized (based on training samples) by the corresponding component of PCV.

Again all the features and even all the properties may not be of equal importance for characterizing a class. So it is reasonable to provide weights corresponding to feature properties in order to determine the characteristic vectors $\mathrm{CV}_j(X)$, $j = 1, 2, \ldots, M$, of X.

The $\mathrm{CV}_j(X)$'s are the input to the fuzzy classifier. It uses the relational matrix R to determine the degree of belonging of the unknown pattern to different pattern classes. This is done by the composition rule of inference taken between $\mathrm{CV}_j(X)$'s and R. The fuzzy classifier therefore gives as output a class similarity vector

$$S(X) = \{s_1(X), s_2(X), \ldots, s_j(X), \ldots, s_M(X)\}, \tag{11}$$

where $s_j(X)$ denotes the degree of similarity of a pattern X to the jth class.

Ambiguity (uncertainty) in the fuzzy decision, provided by $S(X)$, is then determined by computing CF (Confidence or Certainty Factor). The higher the value of CF, the greater is the contrast between the $\max_j\{s_j\}$ and the remaining, and hence the stronger is the decision. Depending on the value of CF, the final output of the recognition system is given in natural form.

The linguistic system described above may therefore be viewed as a generalized classifier providing natural (fuzzy and/or hard) output from both

fuzzy and deterministic input. Before describing the operations in different blocks of Figure 3, let us explain the compatibility function considered for assigning membership values corresponding to the sets SMALL, MEDIUM, and HIGH.

3.3. *COMPATIBILITY FUNCTION*

It is known that the data in linguistic form contains summarized information. So, for the recognition purpose, the primary properties SMALL, MEDIUM, and HIGH of a feature which reflect its linguistic information must be defined in such a way that they represent properly a set of data and provide a summarized information on them. Keeping this in mind, the sets SMALL, MEDIUM, and HIGH have all been represented by π-functions [Equation (3)].

Figure 4 shows the coexistence structure of the various compatibility functions under one particular feature. Depending on the problem, different values for β and γ will be fixed to represent the primary property sets with respect to each feature.

It is to be noted that consideration of π-function as the compatibility functions for SMALL and HIGH is a major deviation from the standard approach in the fuzzy set theory where these are usually represented by $(1 - S)$- and S-functions, respectively. The reason for this deviation is explained below.

In fuzzy set theory, the set "very small" is a subset of the set labeled "small," i.e., "very small" possesses a relatively high membership value in the class "small." But this is not always intuitively appealing. For example, consider the problem of scoring in a class. We usually use the term "very bad" to indicate marks between 10% and 30%, say, and "bad" to indicate marks between 20% and 40%, say. So by "bad" and "very bad" humans highlight two different data sets, although one overlaps other. Therefore, it is reasonable here to decrease the membership value for "very small" in the class "small." To incorporate this view, we have considered the membership function for SMALL as a π-function, instead of $(1 - S)$-function. Similar argument holds for the class HIGH. As an example, by "tall," it indicates the height roughly 5 ft 6 in. to 6 ft 6 in., say, but by "very tall," it indicates the height between 6 ft and 7 ft, say.

4. LINGUISTIC FEATURE EXTRACTOR (LFE)

The input pattern of the recognition system as mentioned earlier can be any of the four forms, namely, linguistic form, quantitative form, mixed form, and set form. First of all, each feature information is considered separately to determine its membership value corresponding to the properties SMALL,

MEDIUM, and HIGH. The way it has been done for various forms of input is furnished in the next section.

4.1. *LINGUISTIC FORM*

Here the input pattern information is provided in linguistic form, e.g., the information as "F_1 is small" or "F_1 is more or less medium" or "F_1 is very high," etc.

When the input statement contains only the primary terms, its membership values for the sets SMALL, MEDIUM, and HIGH are assigned as

$$\begin{aligned} \text{small} &\equiv \{0.8/\text{S}, 0.2/\text{M}, 0.0/\text{H}\} \\ \text{medium} &\equiv \{0.2/\text{S}, 0.8/\text{M}, 0.2/\text{H}\} \\ \text{high} &\equiv \{0.0/\text{S}, 0.2/\text{M}, 0.8/\text{H}\}. \end{aligned} \tag{12}$$

There may be statements with linguistic hedges, very, more or less, slightly, etc. When the statement contains "very small," its membership value for the property SMALL, as explained in the previous section, will be decreased and the membership value for the property MEDIUM will further be decreased. On the other hand, by "more or less small," the membership value for the set SMALL will be decreased, but the membership value for MEDIUM will be increased.

Similar is the case with HIGH also, but the converse is true for MEDIUM. Here we will increase the membership value for MEDIUM and decrease the membership value for both SMALL and HIGH, when the input statement contains "very medium." Similarly, by "more or less medium," the membership value for MEDIUM is decreased and the membership values for both SMALL and HIGH are increased.

The modifications of the membership values may be carried out in a similar manner for other possible linguistic hedges. To increase the membership value, the DIL [Equation (4b)] operation is used and to decrease the membership value, the CON [Equation (4a)] operation is used.

4.2. *QUANTITATIVE FORM*

The information in this form is in exact numerical terms, like "F_1 is 500," say.

In this case, find the membership value for different linguistic feature properties (i.e., SMALL, MEDIUM, HIGH) by their corresponding membership functions ($\pi(500; \beta, \gamma)$), whose parameter values are assigned beforehand according to the nature of various features.

4.3. MIXED FORM

The information is provided in the form of the mixture of linguistic hedges and quantitative terms such as "F_1 is about 400" or "F_1 is more or less 400."

Since the linguistic term increases the impreciseness in the information, the membership values of the statement as a whole, for different primary properties, should be lower than that of the quantitative term alone. The amount of decrease will be determined according to the linguistic hedges. As an example, for the information "about 400," the membership value will be decreased from $\mu(400)$ using

$$\mu(\text{about } 400) = \{\mu(400)\}^{1.25}, \tag{13}$$

where μ represents the membership value corresponding to a property set.

The above-mentioned modification of the membership value will be reflected in the confidence factor (CF) of the classifier output.

4.4. SET FORM

Like the mixed form the information here is also a mixture of linguistic hedges and quantitative terms. The only difference lies with the nature of linguistic hedges. The linguistic hedges which are used here in the set form are "less than," "more than," "between," etc., such that the data reflected are a set and at least one boundary of the data set is known. The examples of this form are "F_1 is less than 400, say" or "F_1 is more than 400, say" or "F_1 is between 400 and 500, say."

First the membership values for the various primary properties with respect to the quantitative terms (e.g., 400) are calculated. We know that the compatibility functions considered for the primary sets are all standard π-functions of the form $\pi(x;\beta,\gamma)$ where γ is the ideal point, i.e., the point where the membership value is 1.0. Modify the membership value $\mu(400)$ to obtain that of "less than 400" as

$$\mu_s(\text{less than } h) = \begin{cases} \{\mu_s(h)\}^{1/2} & \text{if } h \geqslant \gamma, \\ \{\mu_s(h)\}^{2} & \text{if } h \leqslant \gamma, \end{cases} \tag{14}$$

corresponding to a primary property.

As an example, consider the statement "Ram is less than 25 years old." This means Ram's age is more likely to be around 20 years. Therefore, if $25 > \gamma$ for a primary property set, the μ value of the statement "less than 25"

will be higher than that of $\mu(25)$, because "around 20 years," is more towards the value γ. The reverse is true for $25 < \gamma$. Equation (14) reflects these facts.

The modification of the membership values may be made similarly for the hedges "more than" or "greater than" and any other which are used to represent data in the set form.

Now there may be information with statements having connector "and" or "but" (e.g., F_1 is less than 500 and/but more than 400). In this case, first find the two membership values, as explained above, considering two statements separately. The resultant membership value can be found by their geometric mean as

$$\begin{aligned}&\mu_s(\text{less than } h_1 \text{ and more than } h_2)\\ &\quad = \{\mu_s(\text{less than } h_1) * \mu_s(\text{more than } h_2)\}^{1/2}.\end{aligned} \tag{15}$$

If there are statements like "F_1 is between 400 and 500," then it is equivalent to the statement "F_1 is greater than 400 and/but less than 500" and proceed as in the previous case.

It may happen that the information about some particular feature is fully missing. In this case it is reasonable to assign some low (say, 0.2) membership value to all the primary linguistic properties, i.e.,

$$\text{no information} \equiv (0.2/\text{S}, 0.2/\text{M}, 0.2/\text{H}). \tag{16}$$

The above-mentioned discussion shows the way how the impreciseness/uncertainty in the input feature information has been handled by providing/modifying membership value heuristically to a great extent. The logic behind the assignment of membership value is also intuitively appealing.

Determination of Characteristic Vectors (CV)

After obtaining the membership values of features for the properties SMALL, MEDIUM, and HIGH class membership of a pattern, $\text{CV}_j(X)$, corresponding to all property combinations (i.e., the elements of PCV) is then computed. Let us consider the ith component of PCV which represents the property combination

$$(p_1^i, p_2^i, \ldots, p_N^i). \tag{17}$$

where, depending on the value of i, p_m^i denotes one of the primary properties SMALL, MEDIUM, and HIGH and represents the set "F_m is p_m^i." So the ith

element of the characterized vector, $CV_j(X)$, i.e., the membership value corresponding to the ith element of the PCV, is defined as

$$cv_{ji}(X) = \sum_{m=1}^{N} \mu_{F_m p_m} * w^m(i,j), \tag{18}$$

where $\mu_{F_m p_m}$ is the membership value of the set "F_m is p_m" for an input pattern. $w^m(i,j)$ is the weight of the mth feature for the jth pattern class corresponding to the ith property combination (i.e., ith element of PCV). Note that $cv_{ji}(X)$ [Equation (18)] is the weighted arithmetic mean of the membership values of the individual feature properties. Output of LFE will have M vectors and each vector contains 3^N elements.

As mentioned in Section 3.2, the LFE in the processor works basically as a weighted LFE, and that is what has been described in this section to result in M characteristic vectors $CV_j(X)$. On the other hand, the LFE in the learning section does not take weight matrices into consideration and thus provides a single output vector irrespective of the classes from the training samples.

5. LEARNING

This section takes input from the training samples and estimates weight matrices and a relational matrix. It has three blocks, namely, Linguistic Feature Extractor (LFE), Weight Estimator, and Relational Matrix Estimator. The LFE determines a single characteristic vector $CV(X)$ corresponding to the sample X of the training patterns. The weight matrices and the relational matrix are similarly estimated from the training samples as follows.

5.1. WEIGHT MATRICES

It is the fact that all the features and hence the properties are not of equal importance to determine a pattern class. For example, for the first element of PCV, all the properties (i.e., "F_1 is SMALL," "F_2 is SMALL," ..., "F_N is SMALL) may not have equal importance in characterizing a class. So it leads us to define some weights corresponding to various feature properties to find the membership value of a pattern, corresponding to the elements in PCV, for a class. For N features and the M-class problem, we determine N weight matrices of order $3^N \times M$ where rows stands for elements of PCV, columns stand for different pattern classes. and each matrix corresponds to a particular feature. N such matrices can therefore be represented as $W^1, W^2, \ldots, W^N$, where the mth matrix W^m, corresponds to the mth feature, i.e., it represents

the weights (importance) of the mth feature information to determine the membership value of a pattern to the various classes corresponding to the entries in PCV.

In order to illustrate this, let us assume that the (i, j)th elements (i.e., the elements corresponding to the ith property combination $(p_1^i, p_2^i, \ldots, p_N^i)$ and the class C_j) of the weight matrices $W^1, W^2, \ldots, W^N$ are $\theta_1^{ij}, \theta_2^{ij}, \ldots, \theta_N^{ij}$, respectively. Then θ_1^{ij} represents the weight of the property "F_1 is p_1^i," θ_2^{ij} represents the weight of the property "F_2 is p_2^i,"..., θ_N^{ij} represents the weight of the property "F_N is p_N^i," to determine the ith element of the characteristic vector corresponding to the class C_j. Then N weight matrices may therefore be viewed as a single matrix of order $3^N \times M$, where the (i, j)th element is $(\theta_1^{ij}, \theta_2^{ij}, \ldots, \theta_N^{ij})$.

Determination of the Weight Matrices

Weight matrices $W^1, W^2, \ldots, W^N$ are determined from training samples. Consider the ith element of PCV [Equation (17)]. Let n_j be the number of training samples from jth pattern class. Find

$$\mathrm{SAd}_{F_m p_m} = \sum_{k=1}^{n_j} \mathrm{Ad}^k_{F_m p_m}, \qquad m = 1, 2, \ldots, N, \tag{19}$$

where

$$\mathrm{Ad}^k_{F_m p_m} = \left(1 - \mu^k_{F_m p_m}\right), \qquad k = 1, 2, \ldots, n_j.$$

$\mu^k_{F_m p_m}$ is the membership value of the kth training sample with respect to the set "F_m is p_m": $\mathrm{Ad}_{F_m p_m}$ denotes the absolute deviation of $\mu^k_{F_m p_m}$ from 1 (i.e., from ideal membership value) and $\mathrm{SAd}_{F_m p_m}$ is the sum of $\mathrm{Ad}^k_{F_m p_m}$ for all n_j samples.

So, $W^m(i, j)$, the entry corresponding to the ith row (i.e., ith element of PCV) and jth column (i.e., jth pattern class) of the mth weight matrix W^m (i.e., the weight matrix for mth feature), is

$$W^m(i, j) = \frac{(1/\mathrm{SAd}_{F_m p_m})}{\sum_{n=1}^{\Lambda} (1/\mathrm{SAd}_{F_m p_n})}. \tag{20}$$

Varying i, j, and m, all the entries of the weight matrices $W^1, W^2, \ldots, W^N$ are determined.

5.2. *RELATIONAL MATRIX*

The relational matrix R denotes the compatibility of the various pattern classes corresponding to the elements of PCV. The order of R is $3^N \times M$. Here columns correspond to the pattern classes and rows correspond to the property combinations, i.e., the components of PCV. In other words, we can write for $N = 2$ and $M = 3$

$F_1 \times F_2$	C_1	C_2	C_3
(S,S)	μ_{11}	μ_{12}	μ_{13}
(S,M)	μ_{21}	μ_{22}	μ_{23}
⋮	⋮	⋮	⋮
(H,H)	μ_{91}	μ_{92}	μ_{93}

If a pattern has the property "F_1 is SMALL and F_2 is SMALL," then μ_{11}, the first entry, will denote the membership value of the pattern, based on that property, to be in the class C_1. The case for all other entries of the relational matrix is similar. The concept of the relational matrix has already been provided in Section 2.

Determination of the Relational Matrix

The relational matrix R is determined from training samples. Consider here also the ith element of PCV which represents the property combination $(p_1^i, p_2^i, \ldots, p_N^i)$. Let $\mathrm{CV}(X^k) = \{\mathrm{cv}_i(X^k)\}$ be the characteristic vector for the kth training sample X^k, i.e., $\mathrm{cv}_i(X^k)$ denotes the membership value for ith element of PCV (i.e., for the property combination $(p_1^i, p_2^i, \ldots, p_N^i)$ for X^k. So

$$\mathrm{cv}_i(X^k) = \frac{1}{N} \sum_{m=1}^{N} \mu_{F_m p_m}(X^k), \qquad k = 1, 2, \ldots, n_j \tag{21}$$

where $\mu_{F_m p_m}(X^k)$ is the membership value of the kth training sample X^k for the jth pattern class corresponding to the property "F_m is p_m" of the ith property combination (i.e., ith element of PCV). Note that the weight matrices have not been considered here.

So $R(i, j)$, the (i, j)th element of the relational matrix R, is

$$R(i, j) = \frac{1}{n_j} \sum_{k=1}^{n_j} \mathrm{cv}_i(X^k), \qquad i = 1, 2, \ldots, 3^N, \quad j = 1, 2, \ldots, M, \tag{22}$$

where n_j is the number of training samples taken from jth pattern class.

Note that, in order to estimate the weight matrices and the relational matrix, we consider only those training samples which make nonzero contribution toward the components of PCV.

6. FUZZY PROCESSOR

It consists of three blocks, namely, Linguistic Feature Extractor (LFE), Fuzzy Classifier, and Decision Maker. The LFE gives characteristic vectors $CV_j(X)$, $j = 1, 2, \ldots, M$, as output for an input X. Its function has already been described in Section 4. The $CV_j(X)$'s are used to determine the class similarity vector $S(X)$ which denotes the degree of similarity of the input pattern X to the various pattern classes. The Decision Making block gives a natural output along with its degree of certainty based on the similarity vector $S(X)$.

6.1. FUZZY CLASSIFIER

The classifier incorporates composition rule of inference, which is described in Section 2.3. Here Zadeh's composition rule of inference has been modified by replacing the min operator (which is an connective operator) of max–min operation by the geometric mean (GM) (which gives collective information). The class similarity vector $S(X)$ $[= (s_j(X))]$ is determined as

$$s_j(X) = CV_j(X) \circ R = \max_{\substack{i=1,2,\ldots,3^N \\ m=1,2,\ldots,M}} \{cv_{ji}(X) * R[i,j]\}^{1/2}. \tag{23}$$

where $cv_{ij}(X)$ is the ith element of the characteristic vector $CV_j(X)$ for the unknown pattern X and $R[i,j]$ is the (i,j)th entry of the relational matrix R.

So a class similarity vector $S(X)$, of dimension M, in terms of fuzzy membership value for different pattern classes, comes from the classifier for an input pattern X.

EXAMPLE 5. Suppose we have only one feature, say F, and three pattern classes, say C_1, C_2, and C_3. Assume that the weights of all the feature properties are same for these classes, i.e., there is a unique characteristic vector. Let the characteristic vector be $CV(X) = [0.7, 0.3, 0.0]$ for an input pattern X and the relational matrix R be

F	C_1	C_2	C_3
SMALL	0.2	0.7	0.4
MEDIUM	0.8	0.1	0.1
HIGH	0.4	0.3	0.8

The similarity vector $S(X)$ will be

$$S(X) = \mathrm{CV}(X) \circ R$$

$$= [0.49 \quad 0.70 \quad 0.53].$$

From $S(X)$ it appears that pattern X is inclined to the class C_2.

6.2. *DECISION MAKING*

The class similarity vector $S(X)$ is analyzed here. In the way of analysis, it finds the CF value and, depending on this value, gives a linguistic output, which is indeed the output of the recognition system.

Confidence Factor (CF)

Now we need to find out some measure which will reflect the amount of difficulty in arriving at a single output by minimizing ambiguity in the similarity vector $S(X)$. It has to be mentioned here that the impreciseness in the input information has been reflected in characterized vectors $\mathrm{CV}_j(X)$. The concept of difficulties in arriving at a single output from the $S(X)$ will be clear from the following set of similarity vectors. Consider the following four output similarity vectors with three pattern classes:

$$S^1(X) = [0.9 \quad 0.2 \quad 0.0],$$

$$S^2(X) = [0.9 \quad 0.5 \quad 0.2],$$

$$S^3(X) = [0.9 \quad 0.8 \quad 0.2].$$

$$S^4(X) = [0.9 \quad 0.9 \quad 0.7].$$

It is clear from the above output vectors that the difficulty in deciding the class C_1 is increasing in the order $S^1(X)$, $S^2(X)$, $S^3(X)$, and $S^4(X)$. It can be said that the difficulties in assigning particular pattern class depend not only on the highest entry in the similarity vector $S(X)$ but also on its difference with other entries in $S(X)$. Based on this, a measurement of confidence factor (CF) is

defined as

$$\mathrm{CF} = \frac{1}{2}\left[\{s_{mxd}(X)\}^{f_{mxd}} + \frac{1}{(M - f_{mxd})}\sum_{i=1}^{M}\{s_{mxd}(X) - s_i(X)\}\right],$$

$$0 \leqslant \mathrm{CF} \leqslant 1, \quad (24)$$

where f_{mxd} is the frequency of the highest entry in $S(X)$, $s_i(X)$ is ith entry of the $S(X)$, $s_{mxd}(X)$ is the highest entry in $S(X)$, and M is the number of pattern classes. From Equation (24) it is seen that the higher the value of CF, the lower is the difficulty on deciding a class and hence the greater is the degree of certainty of the output decision. For the above-mentioned example, the CF values are 0.850, 0.725, 0.650, and 0.505 for $S^1(X)$, $S^2(X)$, $S^3(X)$, and $S^4(X)$, respectively.

7. OUTPUT

Based on the value of CF, the system makes the following decisions in order to give output in linguistic (natural) form.

I. The assigned pattern class is "definitely true" if CF is in [0.8, 1.0] and there is no second choice of pattern class.
II. The assigned pattern class is "true" and there is a second choice of pattern class if CF lies in [0.6, 0.8).
III. The assigned pattern class is "more or less true" and there is a second choice of pattern class if CF lies in [0.4, 0.6).
IV. The assigned pattern is "not false" for CF lying in [0.1, 0.4) and there is no second choice of pattern class.
V. The classifier is unable to recognize the pattern class from the given input information if CF lies in [0.0, 0.1).

To give a second choice of pattern class, we find the confidence factor (CF_2) for the second highest entry in the similarity vector $S(X)$ by the same formula as in (24). We will give the second choice of the pattern class, if $CF_2 \geqslant 0.2$.

Some typical output forms are:

I. This is very likely to be C_1 (CF = 0.89).
II. This is likely to be C_1 (CF = 0.72) but not very unlikely to be C_2 (CF = 0.32).
III. This is more or less likely to be C_1 (CF = 0.45) but not unlikely to be C_2 (CF = 0.35).

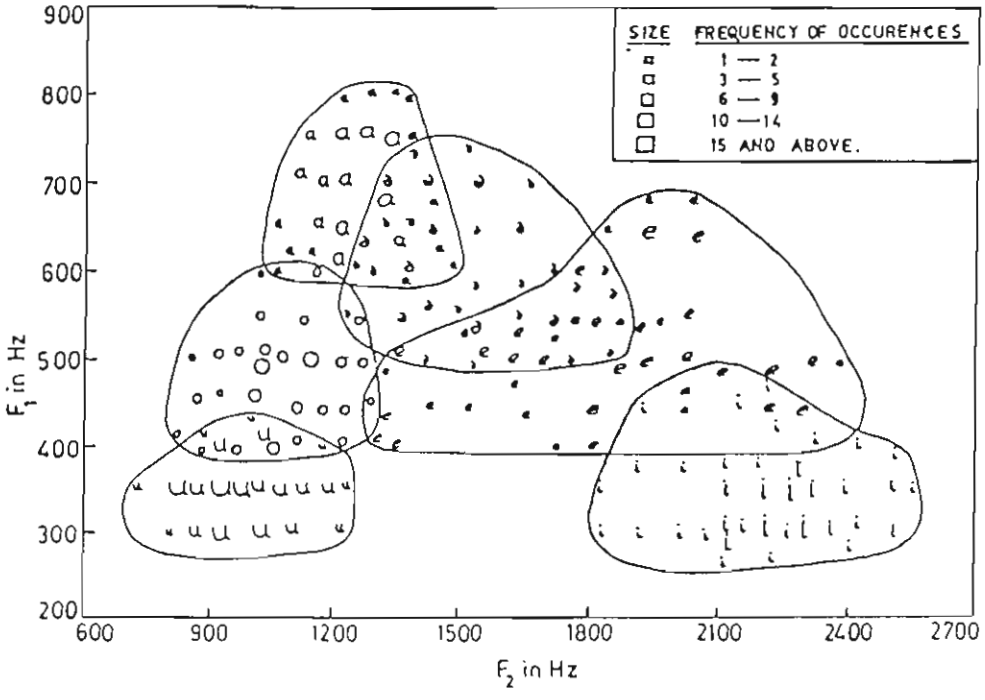

Fig. 5. Vowel classes in the $(F_1 - F_2)$ plane.

IV. This is not unlikely to be C_1 (CF = 0.28).
V. Sorry, it is difficult to recognize.

There may be some cases where there are multiple entry with highest value in the similarity vector $S(X)$. In that case, there will not be a second choice of pattern class. The form of the output here is

VI. This is likely C_1 or C_2 (CF = 0.52).

8. IMPLEMENTATION AND RESULTS

The above-mentioned algorithm was implemented on a set of Indian Telegu vowel sounds in a consonant–vowel–consonant context uttered by three speakers in the age group 30–35 years. Figure 5 shows the typical feature space of six vowels (δ, a:, i, u, e, o) considering 871 deterministic data corresponding to F_1 and F_2. F_1 and F_2 denote the first and second vowel formant frequencies which were obtained through spectrum analysis of the speech data. The boundaries of the classes are seen to be ill-defined (fuzzy).

The testing data set consists of 871 deterministic and 102 imprecise data. This imprecise (F_1, F_2) information was coded to various linguistic forms, viz.,

(small, more or less medium), (700, between 1800 and 2200), (about 600, high), (small, —), etc. by trained personnel. It is to be mentioned here that these imprecise samples were ignored in earlier work [19–22] which was incapable of handling them.

The compatibility functions assigned corresponding to F_1 and F_2 are as follows:

For F_1 feature:

$$\mu F_{1\text{SMALL}}(x) = \pi(x; 200, 325),$$

$$\mu F_{1\text{MEDIUM}}(x) = \pi(x; 200, 525),$$

$$\mu F_{1\text{HIGH}}(x) = \pi(x; 200, 725).$$

For F_2 feature:

$$\mu F_{2\text{SMALL}}(x) = \pi(x: 450, 1000),$$

$$\mu F_{2\text{MEDIUM}}(x) = \pi(x; 450, 1600),$$

$$\mu F_{2\text{HIGH}}(x) = \pi(x; 450, 2200).$$

The overall recognition score for various sizes of samples is shown in Figure 6 by divided-bar diagram. The recognition scores shown are obtained by averaging those corresponding to five different training sets of a specified size. The individual recognition scores are shown, as an illustration, in Figure 7 only for 10% training samples. The recognition scores are grouped in four categories, namely, first correct choice, combined correct choice, second correct choice, and fully wrong choice. Here the first correct choice set includes those samples for which the classifier's first choice agrees with their actual class. Combined correct choice includes those samples while one of the combined choice is correct. The second correct choice includes those samples, for which their second choice corresponds to the actual vowel class. Vowels not falling under the above-mentioned categories are termed as misclassified or fully wrong choice.

A list of some typical output is given below for illustration.

F_1 F_2

(300, 900): This is most likely to be /u/ (CF = 0.82).

(250, 1550): Sorry, it is difficult to recognize.

(450, 2400): This is likely to be /e/ (CF = 0.68), but not unlikely to be /i/ (CF = 0.51).

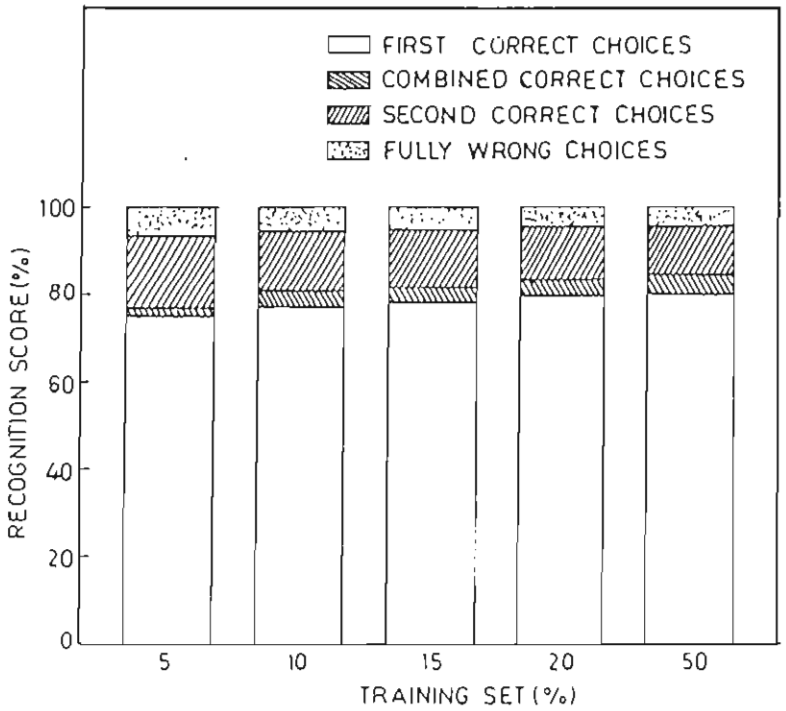

Fig. 6. Pie diagram showing the overall recognition score for different sizes of training samples.

(700, 1300): This is more or less likely to be /a:/ (CF = 0.49) but not unlikely to be /e/ (CF = 0.32).
(900, 1400): This is not very unlikely to be /a:/ (CF = 0.25).
(600, 1200): This is likely to be /o/ or /δ/ (CF = 0.42).
(small, very small): This is likely to be /u/ (CF = 0.74) but not very unlikely to be /o/ (CF = 0.25).
(bigh, more or less small): This is more or less likely to be /a:/ (CF = 0.52).
(between 500 and 600, 1600): This is likely to be /e/ (CF = 0.63), but not unlikely to be /δ/ (CF = 0.45).
(more than 650, high): This is more or less likely to be /e/ (CF = 0.42).
(about 350, ---): This is likely to be /i/ or /u/ (CF = 0.48).

These natural outputs confirm the vowel diagram in Figure 5. Note that, for the input information (250, 1550), the system is unable to recognize the vowel, as this information has a very much insignificant similarity with the vowel classes. This has been regarded as misclassified while computing the recognition score of the system. Further, for the input information (about 350, ---) (here "---" indicates that there is no information about F_2 feature), the

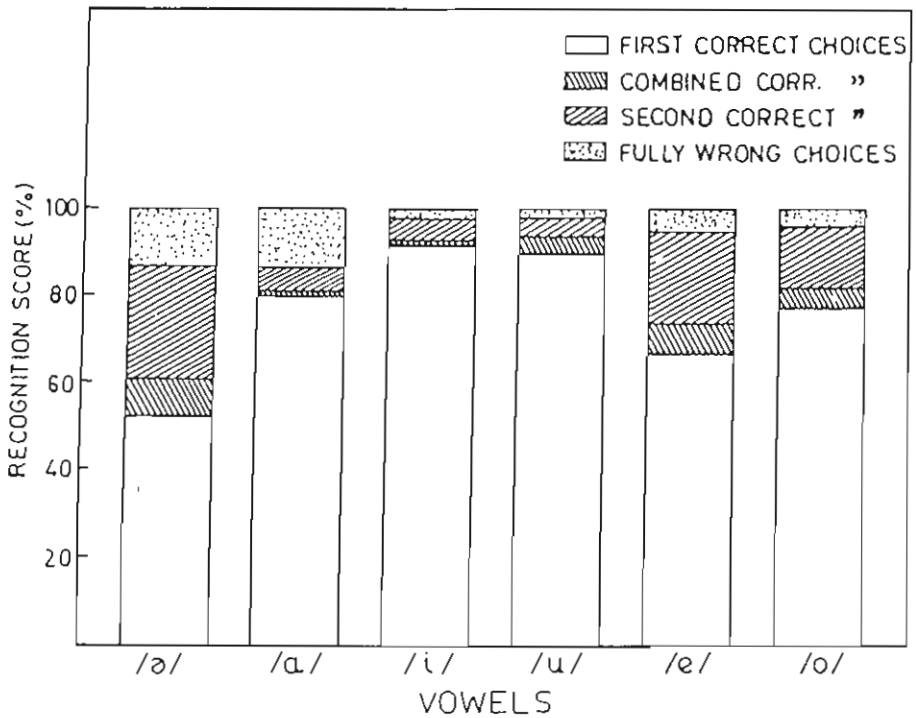

Fig. 7. Pie diagram showing the recognition score of the individual vowels for 10% training sample.

system finds some similarity of this information with the vowel classes /i/ and /u/, on the basis of the F_1 feature information only.

9. CONCLUSIONS AND DISCUSSION

A recognition system having the flexibility of accepting linguistic input and providing output decision in natural form along with its degree of certainty has been formulated. The problem of recognizing vowel sound in the consonant–vowel–consonant context has been considered, as an illustration, to demonstrate the effectiveness of the system.

It is observed that the confusion in recognizing a sample considering the first choice lies, in general, only with the neighboring classes constituting a vowel triangle. Similar findings were also obtained with previous investigations [19–22] considering deterministic input/output. The overall recognition score corresponding to first choice is quite satisfactory considering the fact that it accepts approximate feature information and the information relates only to

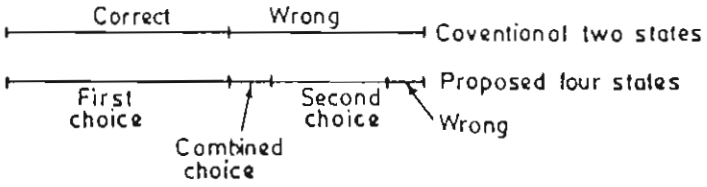

Fig. 8. Four state vs. conventional two state output.

F_1 and F_2. Feature F_3, which was incorporated in [19–22], has not been considered here.

Furthermore, since it provides natural output in four states, it has very low ($\approx 5\%$) misclassification rate as compared to those ($\approx 20\%$) in [19–22], which give two state hard decision like "correct" or "wrong." This is explained in Figure 8, where the category "wrong," in two state conventional system, has been decomposed into three categories in the proposed system. Because of the flexibility the proposed system has therefore a provision of improving its efficiency significantly by incorporating combined and second choices under the control of a supervisory scheme.

It has further to be mentioned here that the system has been programmed considering only three primary properties SMALL, MEDIUM, and HIGH. Incorporation of additional subset property (e.g., very small, more or less small, etc.) will definitely improve the system performance because it will lead to reduction of the impreciseness in input linguistic information by generating more subregions in Figure 2. For example, if we include another property "very small," say, it will result in 4^N subspaces. Again, the primary properties considered need not necessarily be the same for all the features.

The authors gratefully acknowledge Professor D. Dutta Majumder for his interest in this work and Mr. S. Chakraborty for drawing the diagrams.

REFERENCES

1. L. A. Zadeh, Fuzzy logic and approximate reasoning, *Synthese* 30:407–428 (1977).
2. S. K. Pal and D. Dutta Majumder, *Fuzzy Mathematical Approach to Pattern Recognition*, Wiley (Halsted Press), New York, 1986.
3. E. Sanchez, Medical diagnosis and composite fuzzy relations, in *Advances in Fuzzy Set Theory and Applications* (M. M. Gupta, R. K. Ragade and R. R. Yager, Eds.), North-Holland, Amsterdam, 1979, pp. 437–444.
4. R. R. Yager, Multiple objective decision using fuzzy subsets, *Int. J. Man-Mach. Stud.* 9:375–382 (1977).
5. M. M. Gupta, A. Kandel, W. Bandler, and J. B. Kiszka, *Approximate Reasoning in Expert Systems*, North-Holland, New York, 1985.

6. A. Kandel, *Fuzzy Mathematical Techniques with Applications*, Addison-Wesley, Reading, Mass., 1986.
7. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
8. K. S. Fu, *Syntactic Pattern Recognition and Applications*, Academic, London, 1982.
9. L. A. Zadeh, The role of fuzzy logic in the management of uncertainty in expert systems, *Fuzzy Sets Syst.* 11:199–223 (1983).
10. R. Martin-Clouaire and H. Prade, On the problems of representation and propagation of uncertainty in expert systems, *Int. J. Man-Mach. Stud.* 22:251–264 (1985).
11. A. K. Nath and T. T. Lee, On the design of a classifier with linguistic variable as inputs, *Fuzzy Sets Sys.* 11:265–286 (1983).
12. L. A. Zadeh, Fuzzy sets, *Inform. Control* 8:338–353 (1965).
13. A. Kaufmann, *Introduction to the Theory of Fuzzy Subsets—Fundamental Theoretical Elements*, Academic, New York, 1975.
14. L. A. Zadeh, K. S. Fu, K. Tanaka, and M. Shimura, *Fuzzy Sets and Their Application to Cognitive and Decision Process*, Academic, London, 1975.
15. L. A. Zadeh, The concept of linguistic variable and its application to approximate reasoning—II, *Inform. Sci.* 8:301–357 (1975).
16. R. R. Yager, Validation of fuzzy linguistic models, *J. Cybernet.* 8:17–30 (1978).
17. J. F. Baldwin, A new approach to approximate reasoning using a fuzzy logic, *Fuzzy Sets Sys.* 2:309–325 (1979).
18. R. R. Yager, Approximate reasoning and possibility model in classification, *Int. J. Comp. Inform. Sci.* 10:141–175 (1981).
19. A. K. Datta, N. R. Ganguli, and S. Ray, Maximum likelihood methods in vowel recognition: a comparative study, *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-4:683–689 (1982).
20. S. K. Pal and D. D. Majumder, Fuzzy sets and decisionmaking approaches in vowel and speaker recognition, *IEEE Trans. Sys. Man Cybernet.* 7:625–629 (1977).
21. S. K. Pal, Optimum Guard zone for self-supervised learning, *IEE Proc. Part E* 129:9–14 (1982).
22. S. K. Pal, A. Pathak, and C. Basu, Dynamic guard zone for self-supervised learning, *Patt. Recog. Lett.* 7:135–144 (1988).

Received 14 June 1989; revised 9 November 1989