APPLICATION OP QUESTIONNAIRE THEORY TO PATTERN RECOGNITION

J. C. Simon and C. Roche

Institut de Programmation - 9, quai Saint Bernard PARTS 5e

ABSTRACT

Analogies between Pattern Recogn ition and Statistacal Communication Theory are pointed out. With the relative entropy between the "ideal attribute" and "heuristic attribute" sets as an index of performance, optimum selection of possible heuristic operators resuJts in a recognition decision in probabilitty , with the smallest possi ble average number *of* tests during a "sequential analysis"'.

On the other hand, a computer gencrati on of quasi-optimal operators may be done based on numerical attributes using the same entropy index of performance.

Experimental results are presented :

Optimal operator choice is demonstrated for recognition of handwri tten capital letter words and for vocoder speech recogniti on.

Computer buildi ng of quasi-optimal operators Is shown for digitized images. They are limited to first level linear numerical operators.

PRKLTMrNARIES

Information or Communi cation theory deals with the following problem :

A number of n mutually exclusive "events" or "messages" x_i are "realized" or "sent". As a consequence an "observation" yj may be made among m possible observati ons. Let X and Y be the corresponding sets, $p(x_i)$ be the a priori probability of xi and p(yj | xi) the conditional probability of observing yi if the message xi is sent. These last two setsof quantities usually are easy to obtain by experiment. From them, p(xi | yj) is computed through the Bayes formula :

$$p(x_i \mid y_j) = \frac{p(x_i)\, p(y_j \mid x_i)}{\sum_{i=1}^{n} p(x_i)\, p(y_j \mid x_i)}$$

Let us call H_1 and H_2 two extreme hypothesis.

H_1 - If $p(x_i \mid y_j) = 1$ for $j = k$, $p(x_i \mid y_j) = 0$ for $j \neq k$, the observation of y_k informs us with a certainty of the event x_i occurence. In addition if this is true for n y_j among m, i being different for each, any of the observations "informs" us completely on the event which took place.

H_2 - Let $\forall y_j,\ p(x_i \mid y_j) = p(x_i)$

The probability distribution of the events is not modifi ed by any observation knowledge , which then gives no information on X.

Information theory provides a measure of this information obtained by the observation knowledge

Among many possible presentations let us recall briefly the "Questionnaire theory" approach, cf. Pi card (13) , Terrenoi re (15) ;

The average minimum number of binary questions to be asked to obtain the prohahility one answer is given by

$$H[X] = \sum_{i=1}^{m} p(x_i)\, \llcorner\, p(x_i) \qquad [1]$$

$\llcorner$ being the base two logarithm. When y_j is tained, this quantity is given by

$$H[X \mid y_j] = -\sum_{i} p[x_i \mid y_j]\, \llcorner\, p[x_i \mid y_j] \qquad [2]$$

We are assured, cf. Fano (7), chap. 2 that

$$0 \leq H[X \mid y_j] \leq H[X] \qquad [3]$$

If H = 0, p [.] is equal to zero except for one value for which p [.] is equal to 1, we are then in the hypothesis H_1. H [X] is called the a priori entropy of X, H [X | y_j] the conditional entropy. Let H [X | Y] be the average on Y.

$$H[X \mid Y] = \sum_{j} p[y_j]\, H[X \mid y_j] \qquad [4]$$

Again $$0 \leq H[X \mid Y] \leq H[X] \qquad [5]$$

If H [X | Y] = 0 hypothesis H_1 is realized. Thus the decrease towards zero of H [X | Y] evaluates the information obtained by an observation. It is given by :

$$0 \leq I[X ; Y] = H[X] - H[X \mid Y] \qquad [6]$$

I [X ; Y] is called the "average mutual information".

$$I[X ; Y] = I[Y ; X] = \sum_{i,j} p[x_i, y_j]\, \llcorner \frac{p[x_i, y_j]}{p[x_i] \cdot p[y_j]} \qquad [7]$$

From [6] , it is clear that one may compute ei ther H [X | Y] or 1 [X ; Y] . This last quantity usually evaluates the information capacity of a communication channel. If hypothesis H2 is realized I [X ; Y] = 0, if hypothesis H_1 is, I [X ; Y] = H [Xj . Indeed 1 [X ; Y] is an evaluation of the average information brought by an observation.

§1 INTRODUCTION

1.1 DEFINITIONS

Most of the time, Pattern Recognition may be described by using the following elements :

a. A set Ix} of "primitive patterns" X. These patterns X may be themselves sets of "elementary patterns". The usual representation of an elementary pattern is a name and fixed or variable qualifiers$_t$ usually numerical values. For instance, a retina made of cells delivers an X, which is the set of elementary measures given by the cells. One of these measures is representated by a name, the variable numerical amplitude value and the fixed qualifiers of cell coordinates in the retina.

b A number of "operators" £j acting on X. Usually these operators are implemented by computer programs. Their final result may be called an attribute, which in concrete form is some computer register state. This attribute may be interpreted as a numerical value or a word on the primitive pattern names alphabet. We shall call the corresponding operator "numerical" or "linguistic" operators.

c [Sometimes an "ideal" operator £$_{id}$ delivers an attribute such that, with it alone, a recognition decision can be made with quasi-certainty. Most of the time, different attributes arc-used to make a decision. For this purpose another operator acting on the attribute is used, the result of which is the expected classification. This special operator is called by various names : "classifier", "detector", "rategorizer", "discriminant", "decision machine" etc... Its final result or attribute is the recognition of a specific pattern, represented by a name and eventually some qualifiers. The obtained set uf recognized patterns may be used as a new "primitive pattern", which in turn may be analysed by some other operators giving birth to new recognized patterns.

I.j. RELATIONS BETWEEN PATTERN RECOGNITION AND COMMUNICATION THEORY

From the above definitions, it is clear that an "operator" may be considered as a communication channel between the primitive pattern set

Let us admit that an "ideal operator" is available. A pattern X belonging to class P_i (1 < i < n) will always give the same (i) when analysed by &id. Most of the time, it will be a human observer. Then comparisons have to be established between that ideal operator (or channel) and thcjpractical operators (or channels) implemented into the computer. Let w(i) €Ω be the ideal operator attributes, qj(k) € Aj be the practical operator attributes.

Some operators being implemented, we will show in paragraph II, that their sequential use, to approach the ideal operator result, may be optimized by the comparisons between these operators and the ideal operator. Lewis (9) has proposed I[Ω ; Aj] as an evaluation function for "feature selection and ordering", cf. also Fu (8) chap, *2* §1. Benzecri (3), Bongard (5) chap. 7. However the analogy with Communication Theory has not been clearly set up. As a consequence the authors derive directly the entropy concept from linear factor analysis, cf. Tou (15). In fact, the main interest of our paragraph II applications is to show that this concept works for any type of operators, whether it is numerical or linguistic, as long as the requested statistical information is available.

In fact, it would be quite profitable to implement an automatic generation of operators by computers instead of using man-made programs.

In paragraph 111, we show how operators may be automatically generated and evaluated with "the same information performance index. For the envisioned example the best operators found are linear digital filters of a type familiar to image processing specialists. However the program application(II. 5)is not restricted to this type of linear operator.

The automatic generation of numerical operators (character!zers) has been initiated by Pr. Uhr cf. Uhr and Vossler (17), Uhr and Jordan (18) In a sense the learning studies in the field of Perceptrons, Factor analysis, Karhunen-Loeve expansion may be considered as automatic generation of linear numerical operators.

T. G. Evans (9) has made a review of similar efforts in the field of automatic generation of pattern grammars, called by us linguistic operators. Little has been achieved up to now and our statistical approach seems promising even in this domain, to which it may be applied as well.

To our knowledge, the use of the statistical information methods has not been proposed for the generation *of* any type of pattern recognition ope rators.

Remark

w(i)€ Ω being an attribute *of* the ideal operator &id a pattern X analysis may deliverw(i). This attribute classify X in the class Pj_.

An operator &j will deliver an attribute aj(k) € A: for the same pattern X.

by presenting a succession of pattern X, an array of the frequency of the simultaneous occurence of w(i) and aj(k) may be obtained. We will assume that this array converges towards the probability density matrix.

$$\| p [\omega(i), a_j(k)] \| \quad \begin{matrix} i = 1 \ldots n \\ k = 1 \ldots n_j \end{matrix}$$

This requires some stationnarity or at least ergodicity during the "learning" or "training" experiment, cf. Bush (6). For instance, if no pattern of one of the n envisioned class P_j is tested, no information is obtained on the corresponding frequency of occurence.

In the same way, the probability vectors $(p[\omega(i)])$ and $(p[a_j(k)])$ can be obtained. $p[\omega(i)]$ depends upon the way the learning experiment has been performed, it may vary for a new training set.

As Benzecri (4) remarked, the interesting classifying information is found in the relative probability distribution

$$p[\omega(i) \mid a_j(k)] = \frac{p[\omega(i), a_j(k)]}{p[\omega(i)]}$$

This probability distribution is different from the a priori $p[\omega(i)]$ distribution, the knowledge of the attribute $a_j(k)$ "informs" us regarding the probability that X belongs to one of the class P_i.

The relative entropy $H[\Omega \mid a_j(k)]$ estimates the information given by the arrival of $a_j(k)$. The average on A_j, $H[\Omega \mid A_j]$ is an index of performance of the operator $\mathcal{E}_j$ compared to $\mathcal{E}_{id}$, or if one prefers, of the two information channels, see [4].

If $H[\Omega \mid A_j] = 0$, then of course $\mathcal{E}_j$ is an ideal operator ; if $H[\Omega \mid A_j] = H[\Omega]$, no information has been obtained. The mutual information $I[\Omega ; A_j]$ measures the average decrease of entropy, see [6].

§II OPTIMAL SEQUENTIAL SELECTION OF OPERATORS

II.1. SELECTION OF OPERATORS THROUGH RELATIVE ENTROPY

Let $\mathcal{E}_1, \mathcal{E}_2 \ldots \mathcal{E}_j \ldots \mathcal{E}_p$ be p heuristic operators implemented by computer programs. We wish to use some of these operators sequentially to obtain a probability distribution such that a classification decision may be made.

Starting from a $p[\omega(i)]$ distribution, the use of $\mathcal{E}_j$ will give an attribute $a_j(k)$ and thus a new $p[\omega(i) \mid a_j(k)]$ distribution. Starting from $H[\Omega]$, the entropy is $H[\Omega \mid a_j(k)] \leqslant H[\Omega]$. The idea is to use the best average entropy operator. Thus to choose $\mathcal{E}_j$ such that $H[\Omega \mid A_j]$ is smallest, or $I[\Omega ; A_j]$ is biggest, see [7].

The next operator $\mathcal{E}_k$ will be selected among the remaining operators. The mutual entropy of the next step is $H[\Omega \mid A_j, A_k]$.

Let $H[\Omega \mid A^q] = H[\Omega \mid A_{j_1}, A_{j_2}, \ldots A_{j_q}]$

$$0 \leqslant H[\Omega \mid A^q] \leqslant H[\Omega \mid A^{q-1}] \leqslant \ldots$$

$$\ldots \leqslant H[\Omega \mid A] \leqslant H[\Omega] \qquad [8]$$

We are sure that at each step $H[\Omega / A^q]$ is a non increasing quantity ; a lower bound of $H[A \mid A]$ thus exists, cf. Fano *(7)*. Our process of choosing the highest decreasing available r.tep insures that this iower bound is approached as fast as possible. A new probability distribution will result from the process after the use of q attributes.

$$p[\omega(i) \mid a_{j_1}, a_{j_2} \ldots a_{j_q}] = p[\omega(i) \mid a^q]$$

This probability distribution approaches an ideal probability distribution if $H[\Omega \mid A^q]$ is very small.

In practice, an upper bound on the probability distribution is chosen such that if $p[\omega(i) \mid a^q]$ reaches that bound for $i = k$; then the process is stopped and the classification decision that $X \in P_k$ is made.

To obtain the probability distributions $(p[\omega(i) \mid a^q])$, Bayes' relation is used, with the restriction that it implies the statistical independence of the operators.This independence may be evaluated by the mutual information $I(A_j ; A_k)$ between the two operators $\mathcal{E}_j$ and $\mathcal{E}_k$, j and k taking the values considered in A^q.

When a large number of classes is searched, it is customary that the initial set of heuristic operators is not efficient in the sense that the lower limit $H[\Omega \mid A^q]$ is not small enough. Then the less informative operators are discarded, new ones are implemented.

II.2 INDEPENDENCE OF THE CHOSEN OPERATORS

Let us examine if the value of using $\mathcal{E}_1$ can be determined after the $\mathcal{E}_k$ choice.

Like the mutual informations $I[\Omega ; A_1]$ and $I[\Omega ; A_k]$, the mutual information $I[A_1 ; A_k]$ can be defined. It permits a determination of the efficiency of successive use of $\mathcal{E}_k$ and $\mathcal{E}_1$. The following relation holds :

$$I[\Omega ; A_1] + I[\Omega ; A_k] - I[A_k ; A_1]$$

$$\leqslant I[\Omega ; A_k, A_1] \leqslant I[\Omega ; A_1] + I[\Omega , A_k] \qquad [9]$$

$I[\Omega ; A_k, A_1]$ is the information given by the attributes belonging to $\mathcal{E}_k$ and $\mathcal{E}_1$. If the attribute from $\mathcal{E}_k$ is obtained, the "information step" gained with $\mathcal{E}_1$ is :

$$I[\Omega \mid A_k ; A_1] = I[\Omega ; A_k, A_1] - I[\Omega ; A_k] \qquad [10]$$

Thus from [9]

$$I[\Omega ; A_1] - I[A_k ; A_1] \leqslant I[\Omega \mid A_k ; A_1] \leqslant I[\Omega ; A_1] \qquad [11]$$

[11] shows that to have an efficient step, both conditions must be realized :

$$I [\Omega ; A_1] \text{ large,}$$

$$I [A_k ; A_1] \text{ small.}$$

The conditions can be generalized to estimate the (q + 1) operators interest, having had the informations provided by the q others.

N. B. The condition I [A_k ; A_1] small is essential for a practical application of Bayes formula, which permits computing p [w(i) | a^q] . Rigourously, Ak and A_1 should be independent to apply Bayes' formula.

II.3 CAPITAL HANDWRITTEN WORDS RECOGNITION

An optical system delivers for a written word of five letters a quantized pattern X, made of five subpatterns Xj, each representing a handwritten capital letter. Questionnaire methods as described above are applied to both stages : word and letters.

1. Wod recognition

The five operators of this stage are : {vhich letter is the j^{th} | j = 1, Z\. 5).

The program asks sequentially for the most informative letter detection, until the word probability detection reaches 80 %. Fifty words of five letters are to be read ; they were selected at random. On the average only two letters out of five had to be determined to obtain an BO % probability.

2. Letter recognition

When the $j^{\wedge '1}$ letter has to be "read", preprocessing of X; is made. From the primitive pattern Xj of 0 and i, "characteristic features" are extracted by a program. These features are straight segments. A new primitive pattern Y is obtained. Let us represent Y.

The first line contains the names of object or relations considered in the columns. Other lines are the analysis results or attributes for relations of features. Let us take the example of the letter A.

Y =

PRIM 1	PRIM 2	POS 1	POS 2	AGL
1	2	0	0	0
1	3	2	0	3
2	3	?	1	1

The second line contains features 1 and 1: (column 1 and 2") ; 0 in the third and fourth columns means that they are linked by their tails ; 0 in the fifth column codes the angle in octal representation. This representation is quite similar to Shaw's P. *b.* L. (Ih)

Nine operators are applied to a pattern Y.

&1 number of primitive features.

fij number of relations.

&3 number of primitive features found only in one line, i.e. linked to one other feature only.

fill number of graph cycles of a letter.

£5 number of features linked by the middle.

&6 POS 1 + 3 x POS 2 of second line ; this line is the first obtained through the preprocessing performed in a standard way.

£7 angle of second line (column five).

&8 angle of third line.

&9 apgle of fourth line.

Experience shows that only &1 , &2 and &8 are called consistently by the program. Alone they are sufficient to reach a decision on a letter with high probability. The other operators thus are rejected.

On the average four questions were sufficient to select a word among fifty with an 80 % probabi lity. For instance ZEBRE was tested that way :

Question 1 : $\mathfrak{E}_2$ on X_2 (second letter).

Question 2 : $\mathfrak{E}_2$ on X_1 (first letter).

Question 3 : $\mathfrak{E}_1$ on X_1.

Question 4 : $\mathfrak{E}_8$ on X_1.

Four questions were asked instead of the forty five possible questions.

II.4 SPEECH RECOGNITION

Description of the experiment

The goal of this program is to detect the highest probability "phonemes" in experimental patterns delivered by a "vocoder". Fig. 1 displays the vocoder output ; a "line" is obtained every *?"i* milliseconds, it gives the digitized values of the selective filters and the pitch value

a. Seventeen heuristic operators $\&_1, \ldots \&_{17}$ were implemented, such as :

   - existence of a pitch different from zero or not.
   - number of relative maxima (formants).
   - relative difference between first maximum and minimum.
   - difference between biggest maximum and smal lest minimum.
   - variation of average intensity between two successive lines.

   The operator choice was based on the need for being invariant to translations of the frequency and of the intensity axes (logarith mic scales).

b. "Training set". A human operator had to decide what is the phoneme to be recognized (setΩ) versus the pattern obtained with the vocoder. Fig. 1 shows some of the VTSU program results used by the human operator.

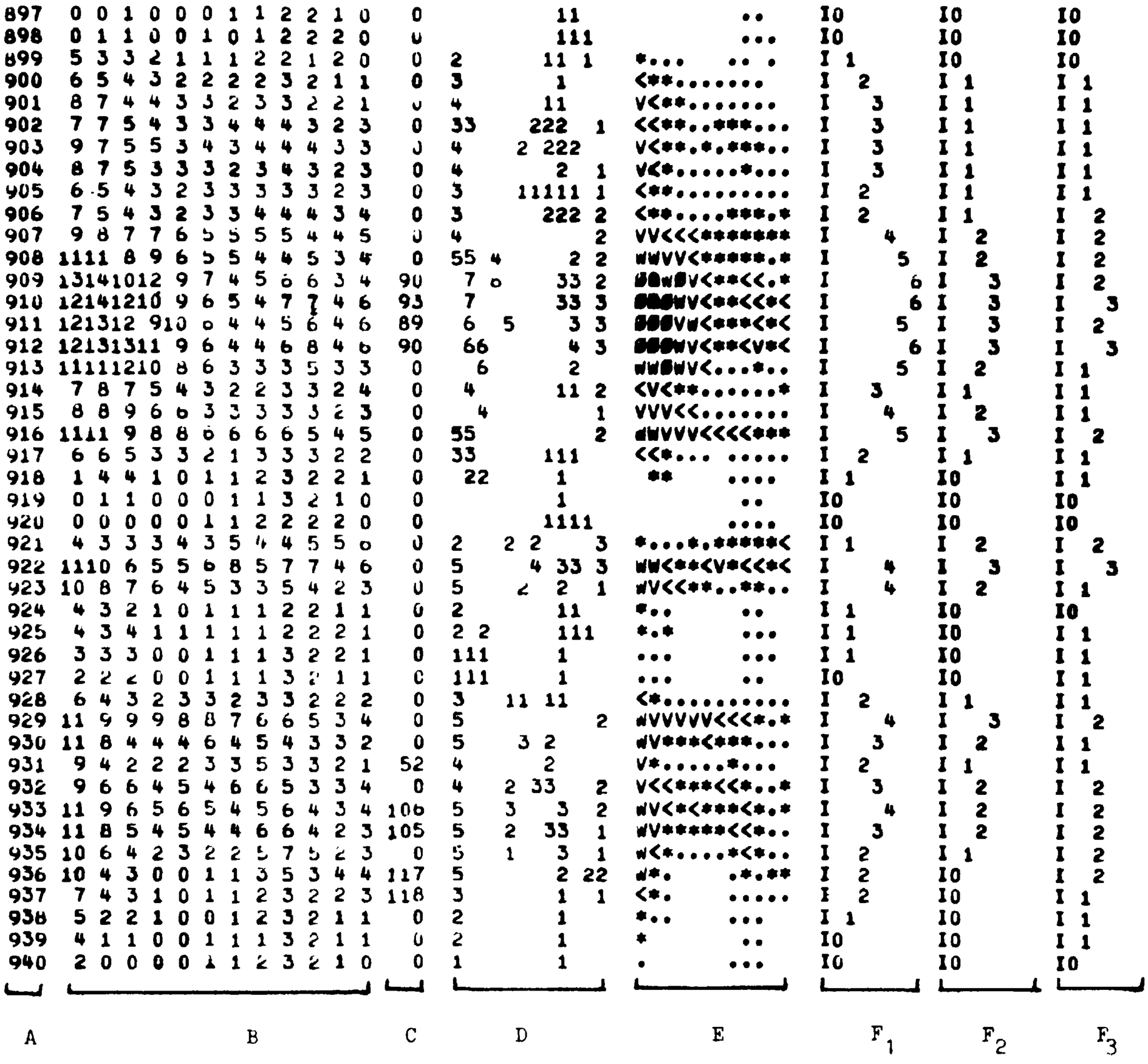


Fig. 1 VISU program giving the vocoder information when the french words "forte pluie" are pronounced

A Line number, B Filter digitized outputs.
C Pitch values, D "Formant" amplitude and frequency.
E Another display of filter outputs.
F_i Four filter output sum, i = 1 first to four, i = 2 five to eight, i = 3 nine to twelve.

c. The program PROBA builts up the matrix $\| p[\omega(i), a_j(k)] \|$ cf. Fig. 2.

d. The program CORREL gives all the $I[A_j ; A_k]$, $j \neq k$, $j = \in \{1, \ldots, 17\}$.

e. The program SYNTHE works according to the algorithm described in §II.1; II.2, on a new set of data, different from the training set. Fig. 3 shows the modifications of $(p(p[\omega(i) \mid a^q]))$ for q = 1, 2,..., 10.

Result

Results obtained

The results seem satisfactory at first view. For instance, Fig. 3 shows that STRATUS is recognized as :

?,S,S,S,S,R,T,T,R,A,A,A,R,R,R,R,R,T,?,U,U,U,R,S,S

The poor ability of the vocoder to detect S, R and L has to be remembered.

OPERATEUR NO : 1

REP NO :

	PROBCO																				PRE
	A	E	I	O	U	EE	AN	IN	OU	UI	JJ	B	D	F	J	K	L	M	N	P	
1	.075	.004	.030	.077	.080	.075	.035	.070	.011	.014	.007	.020	.000	.004	.024	.000	.077	.017	.033	.004	.671
2	.004	.004	.007	.004	.007	.004	.008	.004	.004	.007	.004	.011	.004	.028	.004	.014	.009	.004	.007	.017	.329

OPERATEUR NO : 2

REP NO :

	PROBCO																				PRE
	A	E	I	O	U	EE	AN	IN	OU	UI	JJ	B	D	F	J	K	L	M	N	P	
1	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.062
2	.006	.003	.003	.003	.006	.003	.003	.003	.003	.003	.003	.003	.003	.009	.003	.012	.004	.003	.003	.004	.166
3	.007	.003	.003	.003	.003	.006	.004	.003	.003	.003	.003	.004	.006	.007	.004	.003	.004	.003	.003	.003	.111
4	.028	.003	.003	.003	.004	.012	.003	.004	.003	.003	.003	.007	.003	.003	.009	.003	.003	.003	.006	.004	.142
5	.011	.006	.006	.003	.011	.016	.011	.007	.004	.004	.003	.004	.003	.006	.006	.003	.003	.003	.007	.006	.145
6	.006	.003	.006	.009	.019	.014	.009	.004	.004	.006	.003	.007	.003	.003	.003	.009	.011	.004	.009	.004	.141
7	.003	.003	.007	.011	.016	.006	.004	.003	.004	.006	.004	.004	.003	.003	.003	.003	.006	.003	.004	.003	.110
8	.003	.003	.009	.014	.009	.006	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.009	.004	.003	.122

OPERATEUR NO : 3

REP NO :

	PROBCO																				PRE
	A	E	I	O	U	EE	AN	IN	OU	UI	JJ	B	D	F	J	K	L	M	N	P	
1	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.062
2	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.062
3	.003	.006	.021	.021	.051	.011	.004	.007	.004	.011	.004	.012	.006	.015	.006	.011	.015	.011	.019	.007	.322
4	.003	.003	.003	.004	.004	.007	.006	.007	.003	.003	.003	.007	.003	.004	.007	.003	.004	.003	.006	.003	.112
5	.018	.003	.003	.029	.003	.034	.018	.003	.004	.003	.003	.003	.003	.003	.006	.004	.003	.003	.003	.006	.190
6	.009	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.076
7	.022	.003	.003	.003	.003	.003	.003	.003	.004	.003	.003	.003	.003	.003	.004	.003	.003	.003	.003	.003	.096
8	.006	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.071

OPERATEUR NO : 4

REP NO :

	PROBCO																				PRE
	A	E	I	O	U	EE	AN	IN	OU	UI	JJ	B	D	F	J	K	L	M	N	P	
1	.007	.002	.002	.005	.002	.002	.006	.002	.002	.002	.002	.004	.002	.004	.002	.004	.002	.002	.002	.002	.076
2	.018	.002	.002	.009	.002	.004	.011	.002	.004	.002	.002	.004	.002	.002	.004	.002	.002	.002	.006	.002	.109
3	.012	.004	.002	.004	.005	.004	.004	.002	.002	.005	.002	.005	.002	.002	.002	.002	.005	.002	.004	.004	.100
4	.006	.002	.004	.002	.004	.005	.002	.005	.004	.002	.002	.002	.002	.005	.004	.002	.002	.004	.002	.004	.096
5	.004	.004	.002	.002	.002	.026	.002	.005	.002	.002	.002	.002	.004	.002	.007	.002	.005	.004	.007	.004	.110
6	.004	.002	.002	.013	.026	.005	.004	.005	.004	.002	.004	.007	.004	.009	.004	.006	.007	.006	.006	.002	.146
7	.002	.002	.015	.015	.015	.007	.002	.002	.002	.005	.002	.004	.002	.002	.002	.005	.004	.002	.004	.002	.114
8	.002	.002	.004	.005	.004	.002	.004	.002	.002	.004	.002	.002	.002	.002	.004	.002	.002	.002	.002	.002	.067
9	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.050
10	.002	.002	.002	.004	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.002	.002	.002	.002	.002	.004	.061
11	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.059

OPERATEUR NO : 5

REP NO :

	PROBCO																				PRE
	A	E	I	O	U	EE	AN	IN	OU	UI	JJ	B	D	F	J	K	L	M	N	P	
1	.007	.003	.010	.026	.006	.030	.005	.010	.007	.007	.003	.006	.003	.010	.007	.012	.010	.007	.007	.009	.234
2	.020	.007	.008	.012	.033	.017	.008	.003	.009	.005	.003	.003	.005	.007	.008	.003	.007	.007	.014	.005	.209
3	.009	.003	.003	.005	.020	.008	.007	.005	.003	.005	.005	.003	.003	.005	.008	.003	.007	.005	.007	.003	.173
4	.014	.003	.003	.012	.003	.005	.005	.005	.003	.005	.003	.005	.003	.003	.003	.005	.005	.003	.005	.003	.125
5	.006	.003	.003	.007	.003	.003	.010	.003	.003	.003	.003	.006	.003	.005	.003	.003	.003	.003	.003	.005	.102
6	.003	.003	.003	.006	.007	.010	.005	.003	.003	.003	.003	.007	.005	.005	.003	.003	.003	.003	.005	.007	.135

OPERATEUR NO : 6

REP NO :

	PROBCO																				PRE
	A	E	I	O	U	EE	AN	IN	OU	UI	JJ	B	D	F	J	K	L	M	N	P	
1	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.006	.003	.003	.003	.003	.003	.075
2	.020	.003	.003	.003	.008	.003	.012	.004	.003	.003	.003	.003	.003	.003	.006	.009	.004	.003	.003	.003	.173
3	.029	.003	.004	.008	.019	.029	.009	.009	.004	.004	.003	.008	.003	.015	.012	.003	.004	.003	.013	.006	.240
4	.006	.004	.017	.012	.025	.022	.004	.004	.003	.006	.004	.009	.006	.004	.004	.003	.015	.006	.004	.004	.213
5	.003	.004	.004	.009	.006	.004	.004	.003	.004	.006	.003	.006	.003	.004	.003	.003	.004	.009	.006	.003	.122
6	.003	.003	.004	.031	.008	.003	.003	.003	.004	.003	.003	.003	.003	.003	.003	.003	.003	.006	.004	.003	.106
7	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.070

OPERATEUR NO : 7

REP NO :

	PROBCO																				PRE
	A	E	I	O	U	EE	AN	IN	OU	UI	JJ	B	D	F	J	K	L	M	N	P	
1	.003	.003	.005	.008	.005	.005	.003	.003	.003	.005	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.067
2	.008	.005	.022	.024	.020	.014	.005	.003	.007	.007	.005	.008	.007	.016	.014	.005	.007	.007	.012	.007	.202
3	.010	.005	.003	.016	.024	.027	.007	.010	.003	.005	.003	.005	.003	.003	.007	.012	.014	.008	.008	.003	.241
4	.022	.003	.003	.016	.020	.008	.018	.007	.005	.005	.003	.008	.003	.005	.003	.003	.005	.003	.008	.005	.173
5	.003	.003	.003	.003	.003	.005	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.071
6	.003	.003	.003	.006	.003	.010	.005	.003	.003	.003	.003	.007	.003	.005	.003	.003	.003	.003	.005	.007	.129

OPERATEUR NO : 8

REP NO :

	PROBCO																				PRE
	A	E	I	O	U	EE	AN	IN	OU	UI	JJ	B	D	F	J	K	L	M	N	P	
1	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.044
2	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.002	.002	.002	.002	.058
3	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.002	.002	.002	.002	.002	.065
4	.015	.002	.002	.003	.003	.002	.007	.002	.002	.002	.002	.004	.002	.004	.002	.003	.001	.002	.002	.002	.080
5	.013	.002	.002	.004	.003	.002	.006	.002	.003	.002	.002	.002	.002	.005	.003	.003	.002	.002	.002	.004	.075
6	.005	.002	.002	.003	.002	.002	.003	.002	.002	.002	.002	.003	.002	.004	.002	.002	.003	.002	.002	.003	.067
7	.004	.002	.002	.004	.003	.008	.002	.003	.002	.003	.002	.006	.004	.003	.003	.002	.002	.002	.002	.002	.080
8	.003	.003	.002	.003	.003	.009	.002	.003	.002	.002	.002	.002	.002	.004	.002	.002	.002	.002	.007	.002	.071
9	.002	.003	.004	.003	.007	.008	.003	.004	.003	.001	.002	.002	.002	.003	.004	.002	.003	.001	.006	.004	.062
10	.002	.002	.003	.004	.018	.008	.003	.002	.003	.001	.002	.004	.002	.002	.002	.002	.006	.003	.004	.002	.083
11	.002	.002	.005	.006	.007	.004	.002	.002	.002	.004	.003	.002	.002	.002	.002	.002	.004	.003	.003	.002	.066
12	.002	.002	.005	.014	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.004	.002	.002	.002	.060
13	.002	.002	.002	.004	.002	.002	.002	.002	.002	.004	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.044
14	.002	.002	.002	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.041
15	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.042
16	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.042

Fig. 2

$\| p [\omega(i), a_j(k)] \|$

For different ε_j used for operator evaluation

Fig. 3

Successive probability distribution of phonemes after operators applications.

Decision is made after ten applications.

II.5 GENERAL RECOGNITION PROGRAMMATION

The programs PRORA, CORREL, SYNTHE are general and may be utilized as an aid for "building a pattern recognition system :

(i) Take a "training set" and a "test set" of data Both of them are presented to the "teacher" or Ideal operator $\&_{id}$; the results are introduced in the computer.
(ii) Implement some "heuristic" operators $\{\&_j \mid J \in 1, \ldots p\}$.
(iii) See with CORREL if they are correlated by pair. Modify the operators until Max $1 \mid \&j ; \&_k] < l_s$, either by grouping or suppressing some of them.
(iv) Use SYNTHE and see if the cho sen number of operators is sufficient ; if not, add other operators to the set and go to (ii) ; if it is sufficient go to (iv).
(v) What are the operators used by SYNTHE? Use PROBA to see if a probability clustering is apparent. If possible, group some operators according to this clustering. Go to (iii).

If (v) is successful, only one operator remains, very close to the "ideal operator". Otherwise a pattern recognition process as described in §1.1. has to be started anew on the new attribute sets, giving birth to a new recognition level.

§I GENERATION OF QUASI OPTIMAL NUMERICAL OPERATORS

III.I PRINCIPLE

In paragraphe II, the operators $\&_j$ are supposed to be implemented by the human experimenter. The computer production of new operators certainly would be of interest. For this purpose, we tried to introduce a program that builds new operators by combining the attributes of some former operators.

Let us consider a "retina" made of elementary cells. The result registered by one of these cells may be considered as the attribute of a "hidden" operator. Let a-, a;... be the variable names of these cell attributes ; each of these is an element of the sets A_j , A; .. . Again the ideal operator $\&_{id}$ will result in classifying the X presented to the retina. Here this $£_{id}$ will be implemented by the experimenter himself, who will decide to which class X belongs.

In paragraph II, the process of using the string aj_1 , aj_2 . . . aj_q could be considered the result of the action of a "union operator", the attribute of which is an element of A_q (cf. II.1 of this paper). We found that the information value of a string was $H[\Omega \mid A^q]$.

Let us consider an operator $\&_{op}$ working on p attributes and giving a new attribute $a_{op} \in A_{op}$. The value of an operator $£_{op}$ is to reduce the dimensionality of the attribute space. It follows from the same source as the idea of defining "regions" by decision surfaces. But we can show, cf. Ash (1) p. 85, that the performance index $H[\Omega \mid A_{op}]$ satisfies :

$$H[\Omega \mid A_{op}] \geq H[\Omega \mid A^p] \qquad [12]$$

By reducing the dimensionality we must try not to loose much information.

III.2 EXPERIMENT

The pattern X is the result of an image digitizer of four levels (0, 1, 2, 3). The objects are simple shapes of nearly uniform level, cf. Fig. 5-a.

Let us call &T a parametered operator, such that it gives a 7 x 7 square subpattern. The parameters of &T, are the coordinates xT, yT of the square center. Inside this 7 x 7 square, *U9* attributes may be obtained for a certain x_T, yT. They may be considered to be the results of 49 suboperators &T(v), with 1 < v < 49.

The goal is to decide which class the subpattern selected by &T is from among the set $\Omega = \{.., ,N\}$. The elements of are determined by Sid, here a human experimenter working with a "training set", cf. Fig. 4.

Ω elements	(french)	(english)
.	BORD	BOUNDARY
	BLANC	WHITE
N	NOIR	BLACK

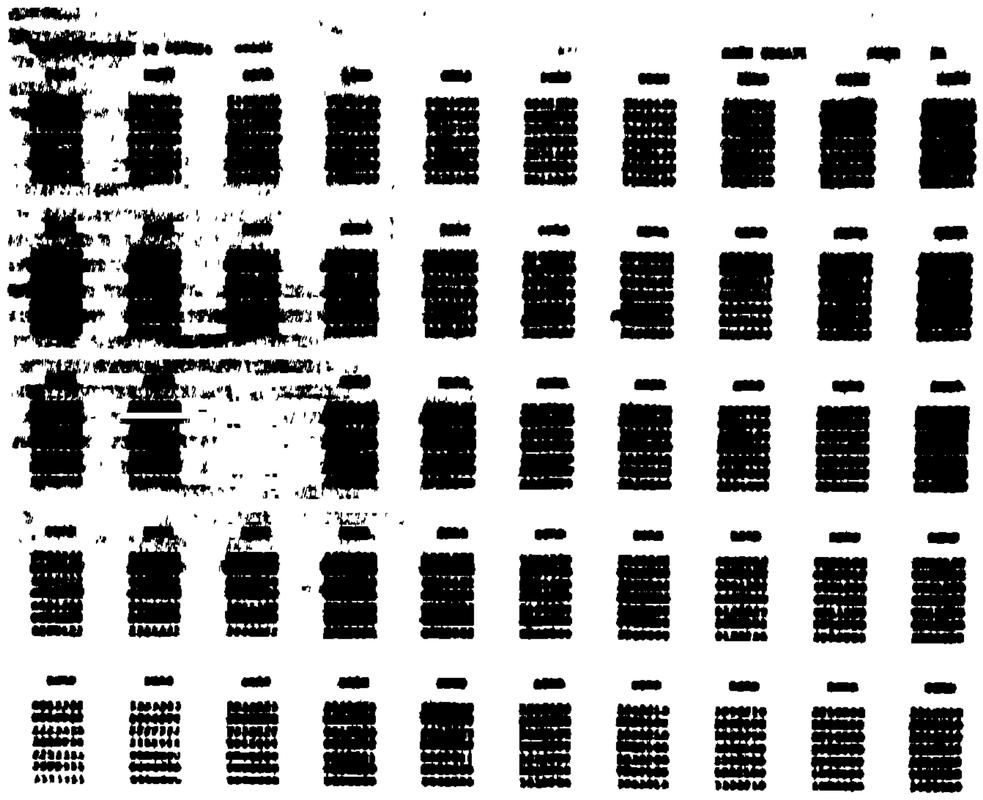
The training set is made of one hundred elements of each class, parts of which are shown on Fig. 4. The membership decision is made by the human ideal operator $\&_{id}$. In arbitrary units the entropy of Ω is $H[\Omega] = 143$.

Using the matrix $\|p[\omega(i) \mid a_j(k)]\|$ obtained with the training set, the operators $\&_T(v)$ performances are evaluated. The mutual informations between $\&_T(v)$ and $\&_{id}$ are given by Table I.

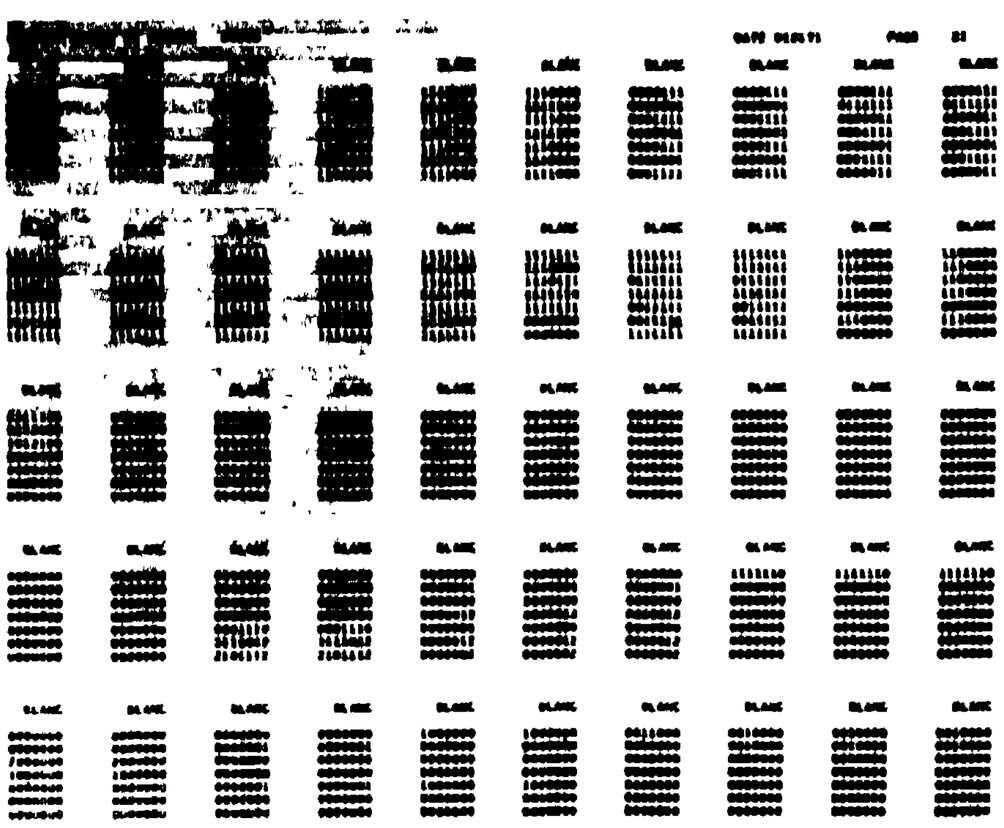
50	**52**	**57**	**57**	**52**	**51**	**45**
52	59	58	58	58	56	52
52	57	60	60	64	61	**51**
53	57	63	65	63	60	**51**
51	58	60	62	58	55	**52**
52	55	56	60	61	56	**47**
48	55	57	58	59	56	50

Table I

The quantities have to be compared to H(Ω) = 143. The center cell maximum mutual information value shows that the knowledge of the corresponding attribute alone is far from enabling us to make a correct decision.



Sample of the "Training set" for BOUNDARY



Sample of the "Training set" for WHITE

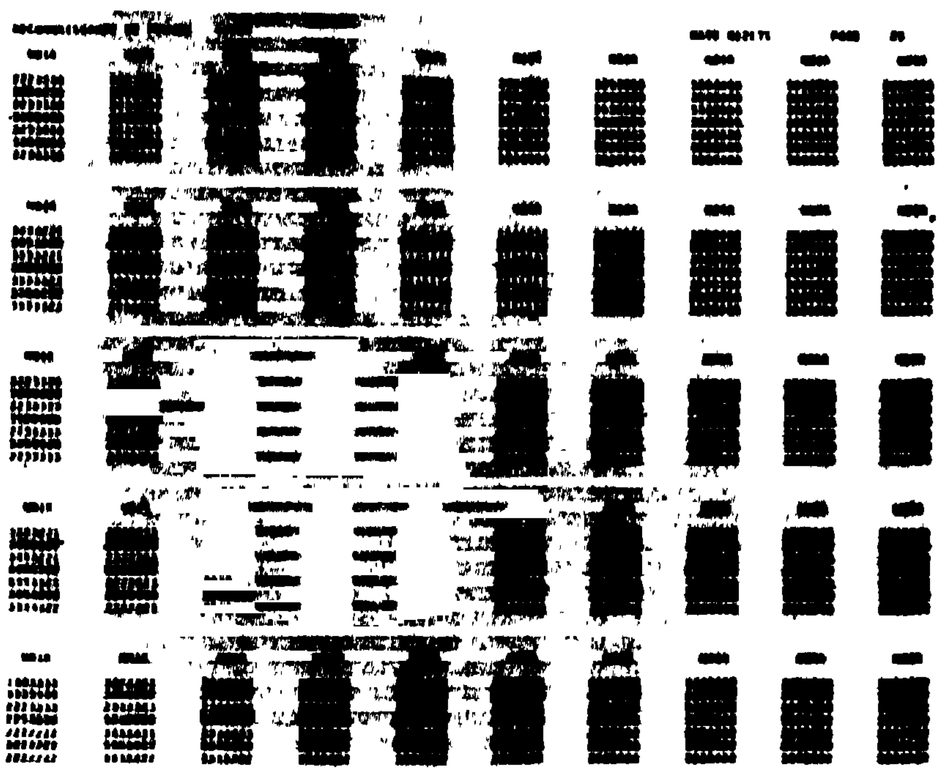


Fig. 4 - Sample of the "Training set" for BLACK

A list of operators is implemented such that the operators are ordered according to a decreasing value of $I[\mathcal{E}, \Omega]$. Thus the starting list $\mathcal{L}_0$ is made of the 49 operators $\mathcal{E}_T(\nu)$.

1 - Two operators are picked up randomly, their attributes are added. The new operator is inserted in $\mathcal{L}$, the lowest I value operator is discarded.

After a number of iterations (2 minutes on UNIVAC 1108), an operator $\mathcal{E}$ is on top of $\mathcal{L}$.

Table II gives the coefficients $\alpha(\nu)$ such That $: \mathcal{E} : X \to a = \sum_{\nu} \alpha(\nu) a_T(\nu)$; $I[\mathcal{E} ; \Omega]$ is equal to 103.

	2					
		1			1	
					2	
	1			1		
			2	1		

I = 103

Table II

		1		1	2	
		2			2	
		2	1	1		

I = 107

Table III

2 - Taking the first eleven operators of $\mathcal{L}_0$, their attribute addition two by two generates 55 new operators. The starting list now consists of 104 elements. Applying the same process as before, after 120 iterations an operator $\mathcal{E}$ is obtained on top of $\mathcal{L}$. Table III gives the coefficients $\alpha(\nu)$ of this $\mathcal{E}$, the mutual information $I[\mathcal{E} ; \Omega]$ is now equal to 107.

3 - Using the multiplication operation

$$\mathcal{E} : X \to a = \prod_{\nu} [a(\nu)]^{\beta(\nu)}$$

With the method of 2., the best $\mathcal{E}$ has an $I[\Omega ; \mathcal{E}] = 75$. The $\beta(\nu)$ are given by Table IV. The performance is barely better than the center cell performance.

		1		1	1	
		2	3		1	
		1	1			
			2	1		

I = 75

Table IV of $\beta(\nu)$

					1	
			1	1	1	
		2	3	2	1	
	1	1	2			

I = 112

Table V of $\alpha(\nu)$

4 - Using the 104 initial list $\mathcal{L}_0$, see 2, the associated operators are selected so that the sum $I[\Omega ; \&_1] + I[\Omega ; \&_k]$ is maximum, $\&_1$, $\&_k$ having not been selected before. The new operator & is obtained by the addition of the two $\&_1$, $\&_k$ attributes. Table V gives $\alpha(\nu)$ for the operator & on top of $\mathcal{L}$ after iterations.

The mutual information is $I[\Omega ; \&] = 112$.

This operator & is quite similar to a "digital filter" such as an experimenter would provide after examination of the proposed problem.

The performance of this last operator was tested. Table VI gives the frequency of occurence of & attributes a versus the $\&_{id}$ attributes on the training set.

ω \ a	0	1	3	7	10	12	13	18	21	23	28	29	32	33	35	36
BOUND		8	6	10	8	2	17	13	13	13		4	2	3		1
BLACK						2				1	10	9	12	10	15	41
WHITE	69	15	8		2	6										
a_c													N			

Table VI

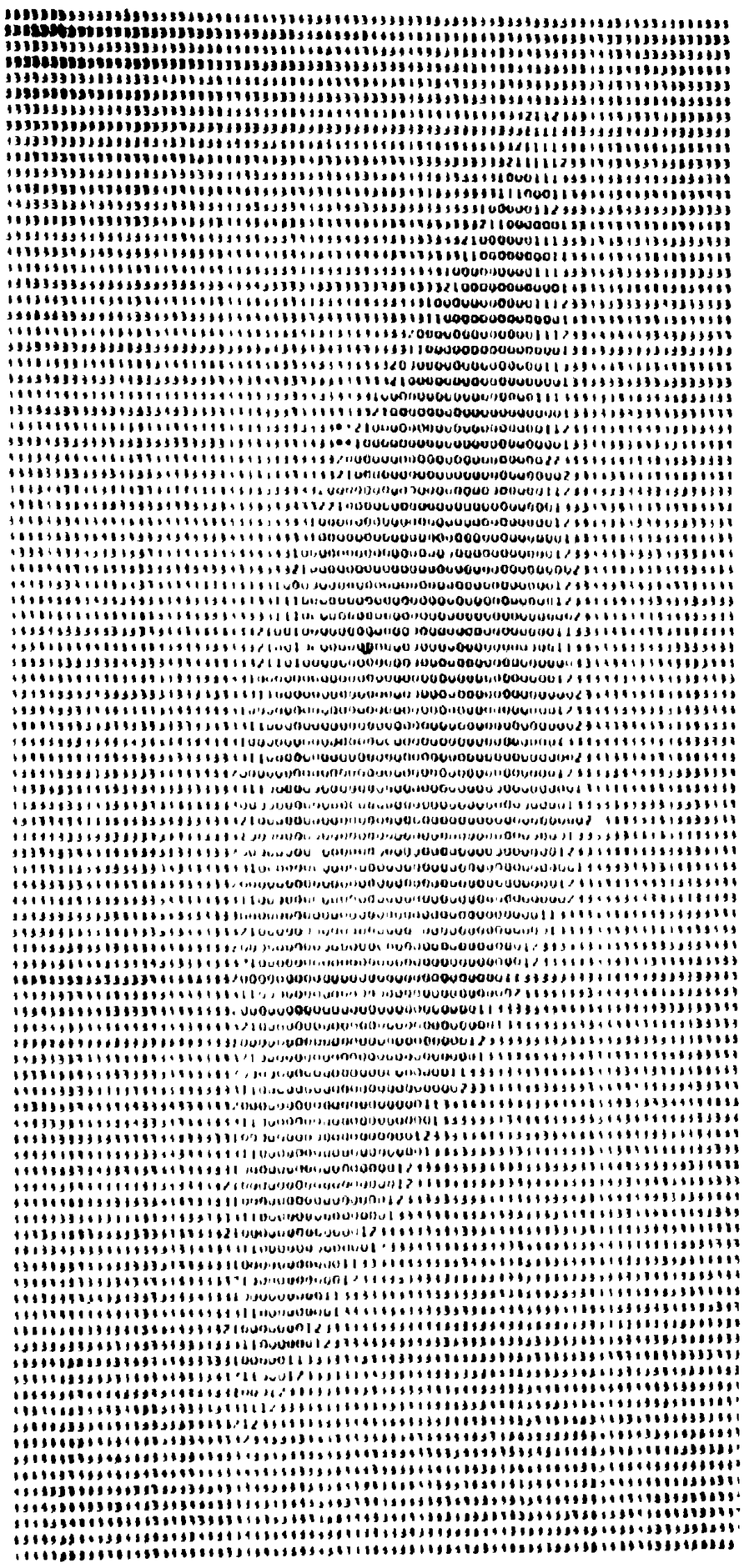


Fig. 5-a : data of the § III experiment

Fig. 5-b : results of the § III experiment

Examination of Table VI leads to choosing the thresholds between 3 and 7 and 23 and 28. A new operator classifier &$_c$ of attribute a_c is implemented such that A_c = { , . , N}.

The couple &$_c$ o & of sequentially applied operators is an approximation of &id. The discrepancy rate with &id is close to 11 %. Application of this machine generated operator is shown on Fig. 5-b. The results are satisfactory.

III.3 CONCLUDING REMARK

Modern biochemists believe that an evolutionary process has built up the nucleotide strings by random variations followed by performance tests. Life did not allow the uninteresting or unfit biomolecules to survive, cf. Monod J. (12).

Our results have a similarity to this point of view. One wonders if brain organisation may not have gone through a similar evolution : random formation of new operators, selection of the best by their performances. This point of view is also tempting in practical Pattern Recognition, though the process of building up at random the highest-order operators seems somewhat improbable.

Acknowledgement : The authors wish to thank a number of graduate students, who during 69-71 have participated to this study : MM. Davencens, Huet, Heude, Raabe, Sabah, Poussin, Sechet.

REFERENCES

(1) Ash R. B. :"information Theory". Interscience Publishers. John Wiley & sons (1967).

(2) Becker P. W. : "Recognition of Patterns". Polyteknisk Forlag, Copenhaguen (1968).

(3) Benzecri J. P. : "Theorie de l'Information et Classification". ISUP. Paris (1969).

(4) Benzecri J. P. : "Statistical Analysis as a tool to make Patterns emerge from data" pp. 35-74 in Methodologies of Pattern Recognition by Watanabe M. S. - Academic Press. (1969).

(5) Bongard M. : "Pattern Recognition". Spartan books (1970).

(6) Bush R. R. and Estes W. K. : "Studies in Mathematical Learning Theory". Stanford Univ. Press. - Calif. (1968).

(7) Fano R. M. : "Transmission of Information". MIT press. (1963).

(8) Fu K. S. ; "Sequential Methods in Pattern Re cognition and Machine Learning". Academic Press. (1968).

(9) Evans T. G. : "Grammatical Inference Techniques in Pattern Analysis". Third International Symposium on Computer and Information Sciences". Miami Beach, Fla. Dec. 18-20(1969).

(10) Lewis P. M. : "The Characteristic Selection Problem in Recognition Systems". IRE Trans. Infor. Theory 8_, pp. 171-178 (1962).

(11) Masson C. G. : "Hierarchical Clustering of Data with a Mutual Entropy Criterion, its Structural Feature Extraction Ability". IEEE Conference Record of the symposium on feature extraction and selection in Pattern Recognition p. 155 ; Argonne National Laboratory, Argonne Illinois, U. S. A. Oct. 5-7-(1970).

(12) Monod J. : "Le hasard et la necessite". Editions du Seuil. Paris (1970).

(13) Picard : "Theorie des Questionnaires". Gauthier-Villars. Paris (1965).

(14) Shaw A. C. : "The formal Description and Parsing of Pictures". SLAG report n° 84, Stanford Linear Accelerator Center, Stanford Univ., Calif. (March 1968).

(15) Terrenoire, Faure and Chesnais : "Aide au diagnostic en Toxicologic". Congres d'Informatique. Toulouse (March 1970).

(16) Tou J. T. and Heydron R. P. : "Some approaches to optimum feature extraction" in "Computer and Information Sciences". Academic Press. N. Y. 1967.

(17) Uhr L. and Vossler Ch. : "A pattern recognition program that generates and evaluates its own operators" en "Computer and Thought" by Feigelbaum and Feldman, Mc. Graw Hill, 1963.

(18) Uhr L. and Jordan S. : "The Learning of Para meters for Generating Compound Characterizes for Pattern Recognition". Proceedings of the IJCAI 1969. May 7-9. Washington D. C.