PATENT SPECIFICATION (11) 1 559 884

1 559 884

(21) Application No. 51622/76 (22) Filed 10 Dec. 1976
(31) Convention Application No. 643 976
(32) Filed 24 Dec. 1975 in
(33) United States of America (US)
(44) Complete Specification published 30 Jan. 1980
(51) INT CL³ G06K 9/00
(52) Index at acceptance G4R 11A 11D 11E 8X

(72) Inventors JOHN JOSEPH HILLIARD, PHILIP JOSEPH MULLAN and WALTER STEVEN ROSENBAUM

(19)



(54) CHARACTER CODE DETERMINING APPARATUS

(71) We, INTERNATIONAL BUSINESS MACHINES CORPORATION, a Corporation organized and existing under the laws of the State of New York in the United States of America, of Armonk, New York 10504, United States of America do hereby declare the invention for which we pray that a patent may be granted to us, and the method by which it is to be performed, to be particularly described in and by the following statement:—

The present invention relates to apparatus for determining in which of two alternative codes a series of characters from a set of characters is printed. The invention has a particular application in the analysis of the output from character recognition equipment.

Most alphabetic characters have differing misread propensities depending on whether they are printed as upper or lower case characters on the document scanned by a character reader. This is readily evident from an examination of the significantly different geometry in the upper and lower case versions of most characters; for example, "A, a"; "E, e"; "G, g"; and so on. It has been discovered that the overall post-processing error correction function in a character recognition apparatus is enhanced by a preprocessing step within that function to determine whether alphabetic characters within a word were printed on the document scanned by the character reader as upper or lower case characters. This preprocessing step enhances the accurancy and reliability of the total overall recognition apparatus post-processing error correction function.

Apparatus embodying the present invention can be seen in its preprocessing role with respect to the error correction apparatus disclosed in United Kingdom Patent Specification No. 1,454,148. The apparatus disclosed in the above specification selects the correct form of garbled input word misread by a character recognition apparatus so as to change the number of characters in the word by character splitting or concatenation. Dictionary words are stored in the system. The vastly different alphabetic character confusion propensities dependent upon whether a given character is printed as an upper or lower case character emphasizes the importance of determining whether an alphabetic character field (i.e., word) output from a character recognition apparatus represents the scan of an upper case or a lower case field printed on the document scanned.

The object of the present invention is to provide improved apparatus for determining in which of two alternative codes a series of characters from a set of characters is printed.

According to the present invention apparatus for determining in which of two alternative codes a series of characters from a set of characters is printed comprises a storage system storing for each character of said set a probability factor representing the probability that said character is printed in one code and the probability that said character is printed in the other code, a read out system for reading out from said storage system for each character of the series the respective probability factor, and a calculator for performing arithmetic calculations on said probability factors, the result of said arithmetic calculation indicating in which of said alternative codes said series of characters is printed.

In order that the invention may be more readily understood, reference will now be made to the accompanying drawings in which:—

Figure 1 displays diagrammatically in block diagram fashion apparatus embodying the invention for discriminating between upper and lower case representations of alphabetic characters.

Figure 2 is a detailed block diagram of one embodiment of the invention wherein a pair of storage devices containing representations of character occurrence probabilities are sequentially accessed by the series of alphabetic characters the output from character recognition apparatus represents a series of alphabetic characters.

Figure 3 is a detailed block diagram of another embodiment of the invention wherein a read-only storage apparatus containing representations of character occurrence probabilities is employed.

Figure 4 displays diagrammatically the operation of the embodiment in Figure 3.

A so-called print convention index (PCI) for a series of characters (a word) developed by apparatus embodying the present invention is a decision function formulation for the purpose of implying upper or lower case character words. For the purpose of such general applications as mail designation reading and text processing, for example, there are three print arrangements to be considered with respect to the PCI operation: i.e., (1) all characters within a word are upper case characters; (2) all characters within a word are lower case characters; and (3) all characters within a word are lower case except for the first character in the word which is an upper case character.

The discrimination between these three print arrangements is not trivial because a misread (including conflicts and rejects) tends to confuse alphabetic character upper and lower case print arrangements as well as character identity. Conflicts and rejects normally contain no print arrangement case inference with respect to the character(s) scanned.

The alphabetic character upper/lower case decision is made on a per field (i.e., word) basis using the PCI formulation set forth in the following equation (1):

$$PCI = \sum_{n=1}^{k-1} \log P_U(\alpha_n) - \sum_{n=1}^{k-1} \log P_L(\alpha_n) \qquad \text{Eq. (1)}$$

where "k" is the number of alphabetic characters in the field; "α_n" is the alphabetic character that has been recognized in the nth position in the field of alphabetic characters; and "n" is an index which is incremented, with respect to the subject field, from right to left (i.e., n=1 at the last, furthest-to-the-right character position in the field). Accordingly, and by way of example, in the field of characters "Smith"; "k"=5; the characters "S" and "h" represents the left-to-right direction of characters; and "n=1" (see above) is associated with the character "h".

An upper case arrangement is implied if the valuce of PCI from equation (1) is greater than 0 (zero), and a lower case arrangement is implied if the value of PCI is equal to or less than 0. This "+" and "−" type assignment is purely arbitrary relative to the operation of the apparatus.

The expression "$P_U(\alpha_n)$" in equation (1) is representative of the probability that the recognition of "α_n" resulted from the scan of an upper case alphabetic character. Correspondingly, the expression "$P_L(\alpha_n)$" in equation (1) is representative of the probability that the recognition of "α_n" resulted from the scan of a lower case alphabetic character. The complete set of "α's" cover all alphabetic upper and lower case characters, and may include conflict recognition rejects as well as the general alphabetic character reject code.

The values for the "$P_U(\alpha_n)$" and "$P_L(\alpha_n)$" expressions are stored in logarithmic form for use in equation (1) after having been calculated from equations (2) and (3), respectively:

$$P_U(\alpha_n) = \sum_{i=1}^{26} P_c[\alpha_n | U(i)] \cdot P[U(i)] \qquad \text{Eq. (2)}$$

$$P_L(\alpha_n) = \sum_{i=1}^{26} P_c[\alpha_n | L(i)] \cdot P[L(i)] \qquad \text{Eq. (3)}$$

where "U(i)" is the "i"th upper case alphabetic character; "L(i)" is the "i"th lower case alphabetic character; "$P_c[\alpha_n | U(i)]$" is the confusion probability of "U(i)" being read as "α_n"; P[U(i)]" is the probability of "U(i)" occurring on the scanned document because of the language (e.g, English) of the word inscribed on the document scanned; "$P_c[\alpha_n | L(i)]$" is the confusion probability of "L(i)"

being read as "α_n"; and "P[L(i)]" is the probability of "L(i)" occurring in the scanned document because of the language in which the word scanned is inscribed on the document.

Table I herein sets forth the frequency of occurrence of alphabetic character letters in the English language so that, for example, the frequency of occurrence of the character "A" is 7.81%; of the character "M" is 2.62%; of the character "T" is 9.02%; and so on through to character "Z" at a frequency of occurrence of 0.09%. The alphabetic upper and lower case occurrence rates "P[U(i)]" and "P[L(i)]", respectively, are derived by applying a 20/80 scaling rule for textual documents and a 70/30 scaling rule for postal documents, respectively, to the frequency data provided in Table I. Accordingly, the respective "P[U(i)]" value for an upper case character in a textural document is 20% of the frequency value in Table I, whereas the "P[L(i)]" value for a textural document lower case character is 80% of the Table I value. On the other hand, the values of "P[U(i)]" and "P[L(i)]" are 70% and 30%, respectively, of the Table I value for a given upper or lower case character, respectively, if the document scanned is in a postal designation reading application. The scaling rules are essentially arbitrary in terms of the exact scaling rule values used, and are empirical from the standpoint of knowledge gained through experience that a great majority of postal designation characters are imprinted in the upper case, whereas a majority of textual document characters are imprinted in the lower case arrangement.

TABLE I.

A/a=7.81%	N/n=7.21%
B/b=1.28%	O/o=8.21%
C/c=2.93%	P/p=2.15%
D/d=4.11%	Q/q=0.14%
E/e=13.05%	R/r=6.64%
F/f=2.88%	S/s=6.46%
G/g=1.39%	T/t=9.02%
H/h=5.85%	U/u=2.77%
I/i=6.77%	V/v=1.00%
J/j=0.23%	W/w=1.49%
K/k=0.42%	X/x=0.30%
L/l=3.60%	Y/y=1.51%
M/m=2.62%	Z/z=0.09%

Table I. Frequency of Occurrence of Letters in the English Language.

Referring to Tables II through VII in the appendix hereto, the Y-axis includes all the upper and lower case alphabetic characters which could be imprinted on the document to be scanned and recognised by the recognition apparatus, whereas the X-axis includes all the possible upper and lower case characters represented by the output from the recognition apparatus. Additionally, the X-axis includes three conflict recognition categories (see Tables VI and VII) which are denoted in the apparatus output as "∂", "?", and "#". Specifically, the conflict recognitions are:

(1) "i/1"—The "i" and "1" recognition circuits within the apparatus are both tuned on; however, the character is more reminiscent of "i" than "1". In the conflict categories this is denoted by "∂".

(2) "1/i"—The "1" and "i" recognition circuits again are both turned on; however, the recognition is more reminiscent of "1" than "i". In the conflict categories, this is denoted by "?".

(3) "N/W"—A recognition is indeterminant; however, characterized by what is interpreted as a strong diagonal tendency. In the conflict categories, this is denoted by "#".

The reject character notation "*" indicates that the recognition was non-conclusive and no character identification could be made.

Tables II through VII give values for each character along the X-axis of the probability that it will have been misread for each of the characters along the Y-axis. By using these Tables to provide the values of $P_o[\alpha_n|U(i)]$ and $P_o[\alpha_n|L(i)]$ for each character α_n in the output from the recognition apparatus and using Table I to provide the value of P[U(i)] and P[L(i)] for each character α_n the values of $P_U(\alpha_n)$ and $P_L(\alpha_n)$ can be calculated for each character α_n in accordance with equations (2) and (3). From the logarithms of these values the PCI for a series of characters

(a word) can be calculated using equation (1). If the PCI so derived is greater than zero, the implication is that the characters scanned are upper case; whereas if the PCI is equal to, or less than, zero, the implication is that they are lower case characters. If the running sum PCI for the several alphabetic characters (including conflicts and rejects) in a word is indicative of upper case characters, the entire word will be treated as of upper case characters. On the other hand, if the running sum PCI for all characters in a word is indicative of lower case characters, the left-most character (i.e., the first character in the field) is to be treated separately from the characters in the remainder of the field which of course are treated as lower case.

Referring to Figure 1, a dual output OCR System 10 is connected to a Bayesian Online Numeric Discriminant (BOND) of the type disclosed in UK Patent Specification No. 1,437,586. OCR System 10 is connected to BOND 11 by the OCR alphabetic character output line 12 and the numeric character output line 13. It will be recognized by those skilled in the art that the BOND 11 is not necessary for the practice of the subject invention. The output of BOND 11 is effectively the alpha/numeric discrimination disclosed in the above Patent Specification. The discrimination is such that the electrical signals pertaining to the series of alpha characters appear along BOND output line 14, whereas the signals pertaining to the series of numeric characters appear on BOND output line 15. A register 16 functions as a storage buffer for the series of alphabetic characters in a field. Under the direction of control 17, register 16 sequentially outputs a series of alphabetic characters to address a probability store 18 on a one-at-a-time character basis. Storage 18 contains both upper and lower case logarithmic occurrence probability constituents of equation (1) herein for each alphabetic character outputted by OCR System 10 through BOND 11. The effective values of both upper and lower case probabilities for a given character together forming a probability factor are outputted to register 19, and directed to logarithmic adder 21 which functionally calculates the sub-totals of all of the upper case logarithmic probabilities and all of the lower case logarithmic probabilities for alphabetic characters in a given field. These sub-totals are used to determine the value of the field or word PCI in accordance with equation (1).

Gate 22 serves as a buffer storage for the numeric character field outputted on line 15 and the alphabetic character field outputted on line 14. Gate 22 also provides an appropriate signal to control 17 as to the position of characters and blanks in the alphabetic and numeric fields. Adder 21 outputs the upper and lower case probability sub-totals to comparator 23 which compares the relative magnitudes of the sub-totals in order to arrive at the value of PCI. If, as stated herein previously, the PCI is greater than zero, the signal on line 24 will cause an upper case character flag to be associated with the alphabetic characters in the field outputted as a series of alphabetic characters from BOND 11 along line 14 through gate 22 and onto output line 25. On the other hand, if the PCI is equal to, or less than, zero, the signal on line 24 will cause a lower case character flag to be associated with the alphabetic character field passing through gate 22 from line 14 to output line 25. It should be recognized that the "greater than" and "less than" zero designations for the value of the PCI are arbitrary in choice, and are not limitations of the invention. They do, however, represent "polarities" indicative of upper/lower case characters.

Referring to Figure 2, dual output OCR system 31 is connected to BOND 32 by OCR output lines 33 and 34. The BOND 32 is connected on its output side by lines 35 and 36 to buffer store 37. As discussed previously in connection with Figure 1, one BOND output line is used to transmit character fields or words flagged as numeric (e.g., line 35), and the other output line is used to transmit character fields flagged as alphabetic (e.g., output line 36). It will be clear to those skilled in this art that it is within the scope of the invention to input directly from an OCR system to the upper/lower case character determining apparatus which is the subject of this invention without of necessity utilizing the BOND apparatus. Thus, and by way of example, the OCR output line 33 (Figure 2) could be connected directly to line 35 which is input to buffer storage 37, and OCR output line 34 could also be connected to line 36, thereby effectively bypassing BOND 32.

The alphabetic character field output from buffer store 37 appears on line 38 for transmittal to alphabetic character shift register 39 and storage address register 41. The numeric output from buffer 37 is directed to numeric character shift register 42 along line 43. The control unit 44 initiates his process and maintains the output sequence of numeric and alphabetic characters identical to the input sequence of the characters to the character determining apparatus. Accordingly, the numeric field character stream is directed to register 42 one character at a time, and the alphabetic

field character stream is directed to registers 39 and 41 in parallel and one character at a time. At the input cell 45 for register 41, a line is connected to blank detector 46 for the purpose of testing the presence of a blank field or word separation "character". Upon detection of a blank, the decision process of the apparatus is activated by control unit 44.

Detection of a blank at input cell 45 of shift register 41 results in the alphabetic field character stream being shifted into register 41 at a rate of one character at a time. Concurrently, each character in the alphabetic character field is sequentially loaded into register 39. Each alphabetic character stored in register 41 provides the storage address for the upper case probablities "log $P_U(\alpha_n)$" in the store 47, and the lower case probabilities "log $P_L(\alpha_n)$" in the store 48. The tables of upper and lower case occurrence probablities are formed as described previously with respect to Figure 1. These probabilities are stored in the storage units 47 and 48. As stated previously, the occurrence probability values in stores 47 and 48 are accessed by the alphabetic character appearing along line 49 from register 41.

The upper case occurrence probability values "log $P_U(\alpha_n)$" stored in output register 51 are sequentially added by adder 52 to the updated "log $P_U(\alpha_n)$" contents of storage register 53. The addition process continues in chain fashion until the sub-total of the upper case occurrence probability values has been calculated for the alphabetic character field stored in shift register 41. The end of the field is detected by testing for the terminating blank at the input cell position 45 of register 41.

Simultaneously, the lower case occurrence probability values "log $P_L(\alpha_n)$", are stored in output register 54, and then sequentially added by adder 55 to the updated "log $P_L(\alpha_n)$" contents of storage register 56. This addition process continues in chain fashion until the sub-total of the lower case occurrence probability values has been calculated for the alphabetic character subfield stored in shift register 41.

The sub-total of the upper case occurrence probabilities for a given field or word stored in storage register 53 are transferred to the register 57, whereas the sub-total of the lower case occurrence probabilities for a given field and stored in storage register 56 are transferred to register 58. The contents of register 58 are inverted in algebraic sign by inverter 59, and added to the contents of register 57 in adder-comparator 61. The output of unit 61 determines the print convention index in accordance with equation (1). If the output from unit 61 is greater than zero, output line 62 will be activated so as to imply that the character field is in upper case. If the output from adder-comparator 61 is less than, or equal to, zero, the output line 63 is activated; the implication being that the character field or word is in lower case. As shown in Figure 2, lines 62 and 63 are input to the upper/lower case identifier generator 64 which, upon command from control unit 44, generates either an upper or a lower case data flag depending upon which of the lines 62 or 63 was activated. This flag is appended to the alphabetic character field in shift register 39. Upon command from control 44, the contents of shift register 39 are output along line 65 to the output register 66 and print convention apparatus output line 67. Accordingly, and by way of summary, the alphabetic character field stored in register 39 will, dependent upon the activation of lines 62 or 63, be output with either an upper or a lower case flag associated therewith. The numeric character field is output as described previously from shift register 42 into output register 66 along the register output line 68. The clear unit 69 under direction of control unit 44 effectively resets storage units 53 and 56, and adder-comparator 61, as the system completes its operation with respect to one field or word, and prepares to receive another grouping of characters.

Referring to Figure 3, the output of OCR system 71 is applied to BOND 72 whose numeric and alphabetic character outputs are directed to buffer store 73 along lines 74 and 75, respectively. As described previously in connection with Figure 2, the numeric character output from store 73 is transmitted along line 94 to register 76, whereas the alphabetic character output from store 73 is directed to registers 77 and 78 via line 95. Input cell 90 is a part of register 78.

Read-only store (ROS) 79 contains therein a probability factor as described below for each upper case and each lower case alphabetic character output from register 78 along line 81. It also contains the probability factor for conflicts and rejects.

The access and use of the read-only store embodiment operation is illustrated in Figure 4. The input from the alpha address register 78 (see also Figure 3) is the OCR alpha field with each possible character represented by a numeric code. This numeric code is assigned to each character by its position in a pre-defined alphabet, starting for example with upper case "A"=1, and continuing sequentially through lower case "z"=52. Additional special characters may also be similarly represented by

higher numeric code numbers. The code is utilized for all OCR post processing to simplify computation. Persons skilled in this art will recognize that any code may be employed for the purpose of access to ROS 79, and that accordingly the code *per se* is not a critical part of this invention.

The example displayed in Figure 4 shows an alpha word or field "MAIm" as having been provided by the OCR 71. This causes storage in the alpha field register 78 of the sequence (1, 9, 39) for the characters "A", "I", "m", respectively. In accordance with equation (1), the first character of the alpha word or field is not included in the determination of the PCI value. The ROS 79 is arrayed in 52 or more sequential positions along a single axis. There is in each storage position a factor which is the precomputed log of the ratio of the probability of a character α_i being an upper case character to the probability of α_i being a lower case character. For the example shown in Figure 4, in indices 1, 9, and 39 are sequentially used to access ROS 79. Accordingly, the factors +2373, +1788, and −2024 are read from ROS 79 and cumulatively summed utilizing the register 83, adder 84, and storage 85, respectively (see also Figure 3).

The cumulative sum in buffer 86 for the example cited in connection with the operation displayed in Figure 4 is +2137. The important characteristic of this sum is its algebraic sign (i.e., the "+" or "−" polarity). A plus (+) sign indicates that the PCI for the word or field is greater than zero with the implication that the original word read by the OCR was upper case. The sign is determined by outputting the value in storage 85 to the buffer 86 and along line 87 to the "+" or "−" polarity detector 88. If the detector senses a positive value, it will generate an upper case flag, and if it senses a negative value, it will generate a lower case flag. Accordingly, the field of characters sequentially transmitted from register 77 via line 97 to output register 88 and output line 89 will have upper or lower case flags consistent with the value of the PCI for the complete word. Accordingly, for the example in Figure 4 discussed herein, the sequence of characters "MAIm" having had a PCI of +2137 will be output from the system output line 89 in Figure 3, as "MAIM". Control unit 92 (Figure 3) effects a control corresponding to that of previously discussed control unit 44 (Figure 2). Correspondingly, clear units 69 (Figure 2) and 93 (Figure 3) perform similar functions within their respective embodiment systems.

TABLE II

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1 A	13				3602	4082			3482	4082		3482	3300	3602		4082		3482		3778
2 B		39	3252	2954	2348	3252												2553		2474
3 C			17		2796		3194				2893	2592							2893	
4 D		3568	2169	53	2963		3568		3568	2964	3268	3268	3568		1362	2664	3568	3268	3568	3092
5 E	4046	4046	3125	4046	17	2446			3721			2704	3553		4046			3421	3721	
6 F						28		2936			2936			2936		2456		2936		2936
7 G			2724		2900		17												3201	
8 H	3509	3509			3509			21	3209			3509	2738	3509		3042		3042	3509	
9 I	3700				3215				108		3700	2328						3700		2411
10 J	3046	3046								30						3046	3046			
11 K											25							2940		
12 L	3482		3482	3958	3357	3958			2731			15			3958				3482	3357
13 M	3252				2775	3432		2778	3252		2554	3721	50	2367		2951		3721	3432	3721
14 N	3700	4000			3388			3388			4000		2268	24		3523		3388	4000	
15 O		3586	1684	1562	3585		3410		3886	3886				3886	50	3194	3586	3886	3586	3886
16 P			3042	3042		2343								3347	3042	11				
17 Q			1794												478		288			
18 R		3959	3959		3638	3041		3348	3959		3168	3959	2743	3638		2904		23	3638	3959
19 S	3922	3922	2890	3328	3922		3922	3922				3922	3638	3922	3554				12	
20 T	3824			3523	3824				2883	3824		3824	3824	3824	3824			3824		20
21 U								3187		2708		2584	3187	3187				3187		
22 V				3087	2783					2783		3087		3087						3087
23 W								2584	3482	3482			2584	2144	3187	3482			3482	
24 X	3000								2697		1522	3000							3000	
25 Y							3168		3168											3168

Please Note: A value of 5000 is ascribed to each blank position in this Table.

TABLE III

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
26 Z																				
27 a	2300						3252					3252			2954					
28 b				2186	2186										2883					
29 c			26				2397								2397					
30 d																				
31 e		3377	2482		2268		3377								2904			3377	3082	
32 f																				
33 g			2335												2335					
34 h																				
35 i	3126								2778			3126								
36 j																				
37 k															2326					
38 l									2660			2417								2221
39 m													2364						2664	
40 n	3250		3250											2432		3250				
41 o			2079	2313											36				2858	3155
42 p									2462			2462				25				
43 q																				
44 r					3260	2955			2955											3260
45 s		2293	2836		2836										2836				33	
46 t		2996			2996							2396								
47 u													2506							
48 v														2270						
49 w														2535						
50 x									2494		2192								2494	
51 y							2582			2582										
52 z																				

Please Note: A value of 5000 is ascribed to each blank position in this Table.

TABLE IV

	U	V	W	X	Y	Z	a	b	c	d	e	f	g	h	i	j	k	l	m
1 A			3602	4082			2500				3478			4081			4081		
2 B								2553			2951		3252		3252				
3 C				3495							3495		3495						
4 D	2870							3092		2867			3268						
5 E			3252		4046		4027				2558		4027	4027					4027
6 F												2932							
7 G								3201			3201		3201					2900	
8 H		3509	3509											2261	3509				
9 I				3700		3700									1448	3700		1594	
10 J	2347															2045			
11 K			3237	3237													2008		
12 L	3958														3114		3959	2358	
13 M		3721	3125	3721							3731			3252	3252		3431	3731	2686
14 N			3092	4000										3996	3693			3693	3693
15 O	2323		3586				3415	3893			3192		3843		3843	3843	3843		3843
16 P					3347	3347									3343				
17 Q							3000												
18 R	3959			3959		3959	3947						3947	3468			2833	3947	3646
19 S								3932					3932		3932		3932		
20 T		3357			3357										2793			2421	3836
21 U	24		2886	2886				3187											
22 V	3087	22	2386		2608														
23 W			40											2788			3487		
24 X	2697			79	3000														
25 Y		2024			20									3170					

Please Note: A value of 5000 is ascribed to each blank position in this Table.

TABLE V

	U	V	W	X	Y	Z	a	b	c	d	e	f	g	h	i	j	k	l	m
26 Z						22													
27 a							25				2783				3084				
28 b								36						1708					
29 c									26	2397	2397								
30 d							2191			40									
31 e			3082				2780				20						3377	3377	
32 f											1870	85							
33 g			2335										73						
34 h														15	2435				
35 i	2822				3126										38			1443	
36 j																600			
37 k				2326													21		
38 l						3620	3263				3263				1564			55	
39 m																			46
40 n										3336			3336		2734				2861
41 o							2857	2857			2857								3158
42 p																			
43 q													1000						
44 r				2955		3260									3260				
45 s							2536				2662		2839						
46 t										2996	2996				2996				
47 u	50	2506					2508			2508				2508	1810				
48 v		32																	1967
49 w		2234	43																
50 x				37	2494													2494	
51 y		1736			1805														
52 z				1435		43												3000	

Please Note: A value of 5000 is ascribed to each blank position in this Table.

TABLE VI

	n	o	p	q	r	s	t	u	v	w	x	y	z	(*)	#	@	?
1 A	4081													2155	2064	3155	4000
2 B														1370	2137	3155	4000
3 C					3495									1830	2432	3155	4000
4 D	2722						3569							1519	2009	3097	4000
5 E					3550		3425							1752	2958	3155	4000
6 F					2455									1700	1824	3155	4000
7 G														1676	2456	3155	4000
8 H	3509						3509							2156	1786	3155	4000
9 I					3215									1518	2237	1013	1936
10 J												3046		1587	1936	2543	4000
11 K	3237													2328	1826	3155	4000
12 L					3482		3458							1944	2638	2482	3301
13 M					3731		3640					3731		1842	1212	3155	4000
14 N	2652													2678	1550	3522	4000
15 O	2688				3843							3843		1668	1827	3155	4000
16 P					3343									2149	2538	3155	4000
17 Q				794										1367	1668	3155	4000
18 R	2867				2448		3646							2114	1687	3522	4000
19 S														1877	3523	3155	4000
20 T					2992		2475					3836		2137	1830	3222	4000
21 U														2042	1632	2886	4000
22 V												3084		2853	1564	3155	4000
23 W												3487		2421	1226	3155	4000
24 X														1788	1034	3155	4000
25 Y												2569		2237	1733	3155	4000

Please Note: A value of 5000 is ascribed to each blank position in this Table.

TABLE VII

	n	o	p	q	r	s	t	u	v	w	x	y	z	(*)	#	@	?
26 Z														1442	1442	3155	4000
27 a	3084											3384		1523	1420	3155	4000
28 b														1407	2402	3155	4000
29 c														2328	1552	3155	4000
30 d	2668													1209	2586	3155	4000
31 e														1678	3046	3155	4000
32 f					1870									1695	1209	3155	4000
33 g				2334										958	2000	3155	4000
34 h	2435													2600	1794	3155	2388
35 i														1637	2087	3155	2794
36 j														300	1000	3155	4000
37 k														1549	2222	3155	4000
38 l														1522	1658	2886	1886
39 m	1550													1448	1925	3155	4000
40 n	21				2861									2009	1760	3155	4000
41 o		36												1864	1706	3155	4000
42 p			25											1522	1103	3155	4000
43 q				156										1000	1000	3155	4000
44 r					22		3260							1939	1765	3155	4000
45 s						33								1524	2301	3155	4000
46 t					2996		24							1662	2143	3155	4000
47 u								50						1428	1736	3155	4000
48 v									32					2155	1425	3155	4000
49 w	2535				2535					43				2276	1208	3155	4000
50 x											37			1830	1522	3155	4000
51 y					2583							38		2023	1715	3155	4000
52 z													43	1639	1639	3155	4000

Please Note: A value of 5000 is ascribed to each blank position in this Table.

WHAT WE CLAIM IS:—

1. Apparatus for determining in which of two alternative codes a series of characters from a set of characters is printed comprising a storage system storing for each character of said set a probability factor representating the probability that said character is printed in one code and the probability that said character is printed in the other code, a read out system for reading out from the said storage system for each character of the series the respective probability factor, and a calculator for performing arithmetic calculations on said probability factors, the result of said arithmetic calculations indicating in which of said alternative codes said series of characters is printed.

2. Apparatus as claimed in claim 1 in which said probability factors comprise for each character a first probability representation of the probability that said character is printed in one code and a second probability representation of the probability that said character is printed in the other code, and said calculator compares the sub-total of the sum of all the first probability representations for the series of characters with the sub-total of the sum of all the second probability representations for the series of characters to produce an output indicating in which of said alternative codes said series of characters is printed.

3. Apparatus as claimed in claim 2 in which said storage system comprises a first storage device for all the first probability representations for all the characters of the set and a second storage device for all the second probability representations for all the characters of the set.

4. Apparatus as claimed in claim 1 in which said probability factors comprise for each character a single probability representation which represents the ratio of the probability that said character is printed in one code to the probability that said character is printed in the other code, and said calculator calculates the sub-total of the sum of all the probability representations corresponding to the series of characters to produce an output indicating in which of said alternative codes said series of characters is printed.

5. Apparatus as claimed in any one of the preceding claims in which each representation is in the form of a logarithm.

6. Apparatus as claimed in any one of the preceding claims in which the two alternative codes are upper case characters and lower case characters.

7. Apparatus as claimed in any one of the preceding claims in which each representation corresponding to a character is derived from the probability that said character will be confused with any other character of the set and the frequency of occurrence of said character in a selected arrangement of printing.

8. Apparatus as claimed in claim 7 in which the selected arrangement of printing is mail designation.

9. Apparatus as claimed in claim 7 in which the selected arrangement of printing is general text.

10. Apparatus for determining in which of two alternative codes a series of characters from a set of characters is printed substantially as described with reference to Figure 1, or Figure 2 or Figures 3 and 4 of the accompanying drawings.

M. J. W. ATCHLEY,
Chartered Patent Agent,
Agent for the Applicants.

Printed for Her Majesty's Stationery Office by the Courier Press, Leamington Spa, 1980.
Published by the Patent Office, 25 Southampton Buildings, London, WC2A 1AY, from which copies may be obtained.

1559884 COMPLETE SPECIFICATION
4 SHEETS *This drawing is a reproduction of the Original on a reduced scale*
Sheet 1

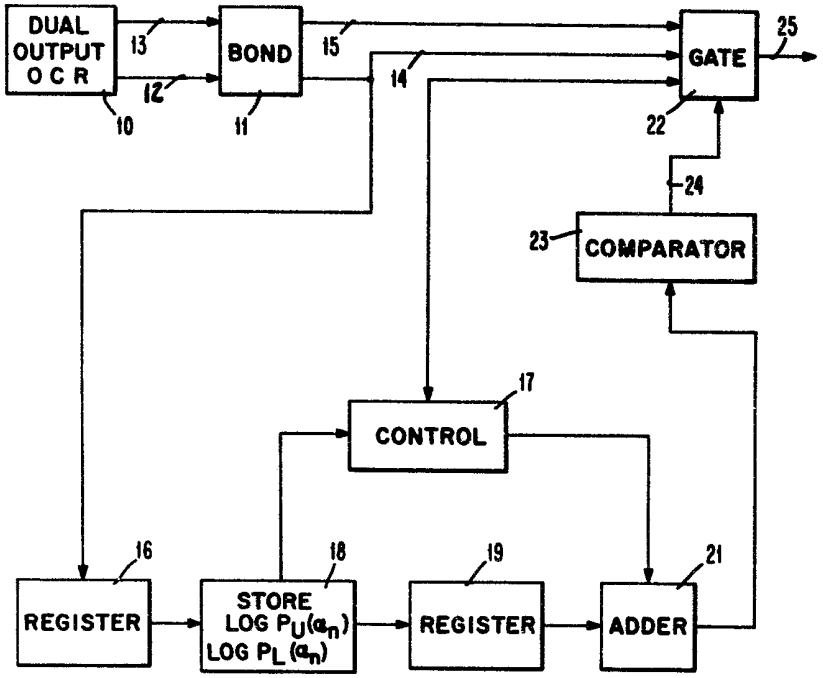


FIG. 1

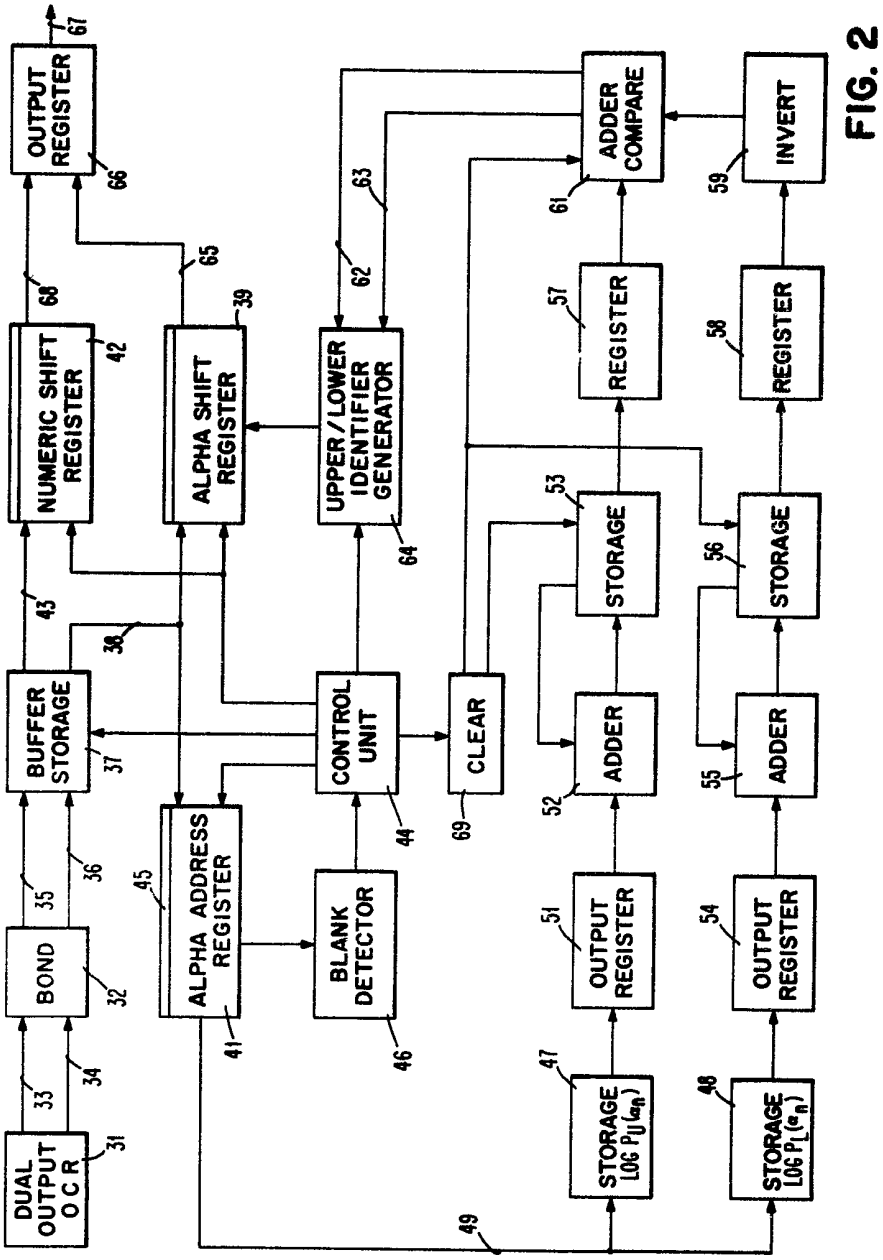

OUTPUT REGISTER
67
66
65
68
NUMERIC SHIFT REGISTER
42
ALPHA SHIFT REGISTER
39
UPPER/LOWER IDENTIFIER GENERATOR
64
62
63
ADDER COMPARE
61
INVERT
59
57
REGISTER
58
REGISTER
53
STORAGE
56
STORAGE
43
38
BUFFER STORAGE
37
CONTROL UNIT
44
CLEAR
69
ADDER
52
ADDER
55
35
36
45
ALPHA ADDRESS REGISTER
41
BLANK DETECTOR
46
51
OUTPUT REGISTER
54
OUTPUT REGISTER
BOND
32
33
34
47
STORAGE LOG $P_U(a_n)$
48
STORAGE LOG $P_L(a_n)$
DUAL OUTPUT OCR
31
49

FIG. 2

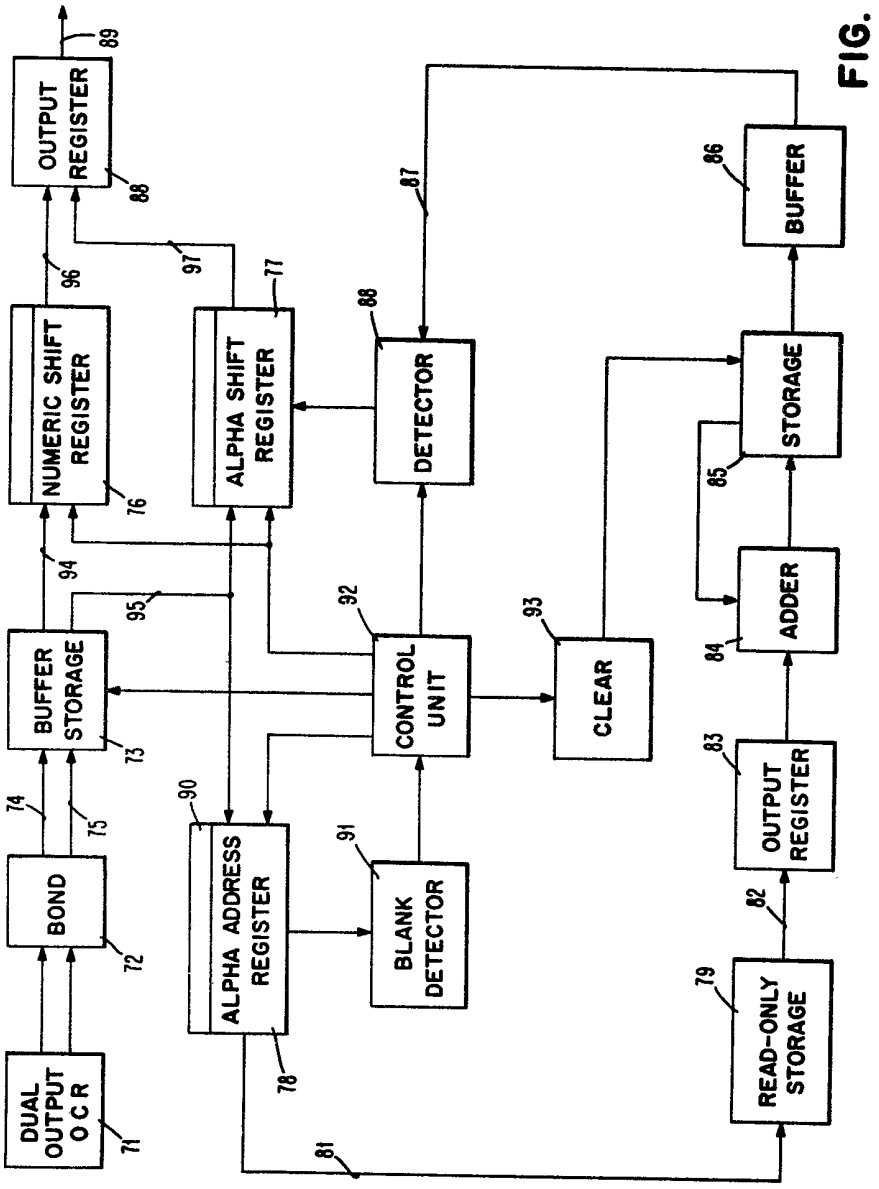

OUTPUT REGISTER
89
88
BUFFER
86
87
NUMERIC SHIFT REGISTER
96
97
ALPHA SHIFT REGISTER
77
DETECTOR
88
STORAGE
85
76
ADDER
84
94
95
92
93
BUFFER STORAGE
CONTROL UNIT
CLEAR
73
OUTPUT REGISTER
83
74
75
90
91
BOND
ALPHA ADDRESS REGISTER
BLANK DETECTOR
82
72
78
79
DUAL OUTPUT O C R
READ-ONLY STORAGE
71
81

FIG. 3

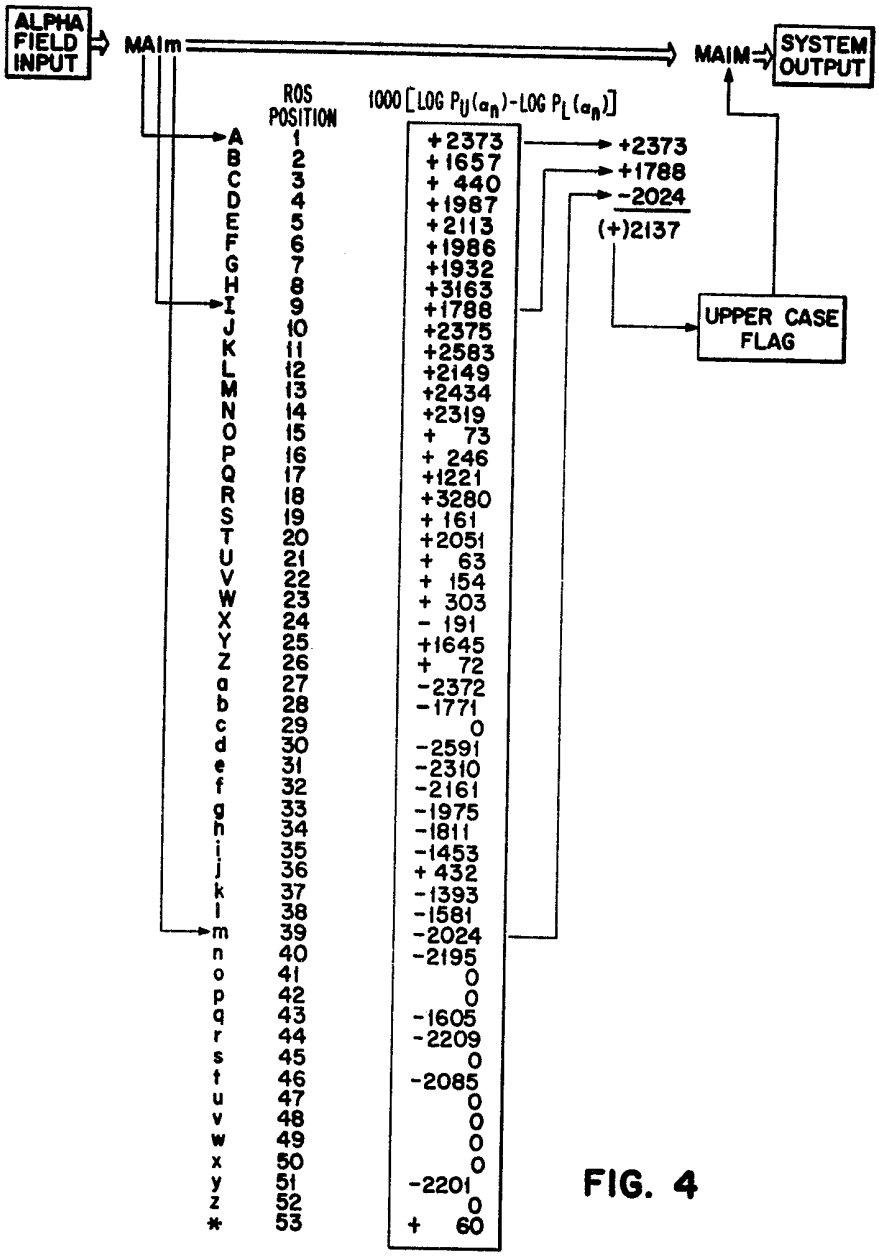

ALPHA FIELD INPUT
MAIm
MAIM
SYSTEM OUTPUT
ROS POSITION
1000 [LOG $P_U(a_n)$ - LOG $P_L(a_n)$]
A 1 +2373
B 2 +1657
C 3 + 440
D 4 +1987
E 5 +2113
F 6 +1986
G 7 +1932
H 8 +3163
I 9 +1788
J 10 +2375
K 11 +2583
L 12 +2149
M 13 +2434
N 14 +2319
O 15 + 73
P 16 + 246
Q 17 +1221
R 18 +3280
S 19 + 161
T 20 +2051
U 21 + 63
V 22 + 154
W 23 + 303
X 24 - 191
Y 25 +1645
Z 26 + 72
a 27 -2372
b 28 -1771
c 29 0
d 30 -2591
e 31 -2310
f 32 -2161
g 33 -1975
h 34 -1811
i 35 -1453
j 36 +432
k 37 -1393
l 37 -1581
m 39 -2024
n 40 -2195
o 41 0
p 42 0
q 43 -1605
r 44 -2209
s 45 0
t 46 -2085
u 47 0
v 48 0
w 49 0
x 50 0
y 51 -2201
z 52 0
* 53 + 60
+2373
+1788
-2024
(+)2137
UPPER CASE FLAG

FIG. 4