2076572

Ideographic characters encoding

The present invention relates to apparatuses for encoding ideographic characters and an important application thereof is in the field of typing texts and transmission by telex (Registered Trade Mark).

The invention and the problems which it solves will be better understood if some indications are given about ideogram languages, taking Chinese as example, which is spoken by a quarter of mankind.

Chinese is a monosyllabic ideogram language for which several approximate phonetic representations exist (pin-ying, Wade, Ecole Française d'Extrême-Orient transcriptions), which have the common drawback of not taking homonymies into account, whereas ideograms distinguish them perfectly well. There exists in Chinese 412 different sounds, which may be pronounced with 4 intonations, which makes practically 1500 different sounds. Now, there exist about 49 000 Chinese characters and 5000 to 8000 of them are in fairly current use, which means an average of 20 homonymes per sound or voice utterance. Some of the characters are not very much used and it has been ascertained that the 2000 most frequent characters cover about 95% of current texts.

Different apparatuses have been used or at least developed and exist at this time for encoding Chinese characters. Some are purely electromechanical ("Computers and People", Vol. 27, May 1978). Other types (IEEE Transactions on systems... SMC3 No. 3, May 1973) include an encoding system which involves the breaking down of each character into components and subcomponents which are juxtaposed, the first approach being however a phonetic breakdown of the character between an initial consonant and a final vowel. The keyboard used is a conventional typewriter keyboard. A similar approach is described in UK Patent Appln. No. 2,033,633, to which reference may be made.

The Japanese language, also an ideogram language, has a slightly different structure and writing. It comprises 140 syllabic characters, grouped into two kinds,

"cursive" and "square", representing the sounds of 70 syllables in two different ways; it comprises moreover Chinese characters, amounting to 1850 only since 1945, but no tones.

The problems met with for encoding Japanese characters are nonetheless substantially the same as in the case of Chinese. Korean language raises similar problems.

It is an object of the invention to provide an encoding device which allows high-rate encoding, is efficient (i.e. requires few bits for representing a character), is flexible and can be readily constructed from commercially available components, at an acceptable price, although naturally higher than that of purely electromechanical machines.

According to an aspect of the invention there is provided an apparatus for encoding ideographic characters, comprising a mass memory for storing characters each having an address in said memory, display means associated with said mass memory for simultaneously displaying all characters of a set in said memory, means associated with the memory for causing display of one said set consisting of the most frequent characters when the apparatus is in stand-by condition prior to actuation of a key, a selection keyboard having a plurality of selection keys each manually actuatable for selection of one of the characters of said set displayed on the display means, whereby a unique code representing the selected character is delivered, and a coding keyboard havint a plurality of coding keys each carrying an identification marking and arranged for possible identification of at least a first and a second character components, said keyboard and memory cooperating whereby actuation of coding keys causes display of a memory page containing the characters or most frequent characters defined by at least said first and second character components selected by actuation of coding keys and subsequent action on one of the selection keys results in delivery of a code corresponding to a selected character amont the displayed characters.

It is another object of the invention to provide

a device for easy encoding, so that a Chinese-speaking person, even a beginner, may use the device after a few minutes of explanation, and with automatic apprenticeship (repeated use of the machine rapidly increasing the encoding speed).

For that, the invention may use a phonetic approach, and not a semantic approach or an approach based on the shape of the characters.

This approach uses the fact that each of the monosyllabic sounds forming the set of Chinese characters may be regarded as composed of an initial, a final or termination, and a tone. For example, in the Mandarin dialect, which is the official language in the Chinese People's Republic and in Taiwan, there are 21 initials, 35 finals and 4 tones which may be combined so as to form the syllables represented by characters, while complying with certain rules. There exist in fact characters whose sound corresponds only to a termination and two or three sounds which cannot be pronounced with an initial.

According to another aspect of the invention, there is provided an apparatus for encoding ideographic characters, comprising a mass memory for storing characters each having an address in said memory, display means associated with said mass memory for simultaneously displaying all characters of a set stored in a page of the memory, means associated with the mass memory for causing display of a page containing most frequent characters when the apparatus is in stand-by condition prior to actuation of a key, a selection keyboard having a plurality of selection keys each manually actuatable for selection of one of the characters of said set displayed on the display means whereby a unique code representing the selected character is delivered, and a coding keyboard having a plurality of coding keys each carrying an identification marking and arranged for possible identification of each phonetic initial character component and each phonetic final character component as well as of a tone if necessary, said keyboards and memory cooperating whereby actuation of coding keys causes display of a memory page

containing the characters or most frequent characters defined by the initial and final character components, and possibly the tone, associated with the actuated coding keys and subsequent action on one of the selection keys results in delivery of a code corresponding to a selected character among the displayed characters.

The code is advantageously formed by the address in the memory broken down into the number of the memory page and the number of the characters in the page, although other codes are possible, for example the Chinese telegraphic code, the translation into a language other than Chinese of the meaning of the character, or any combination. The code will in general be provided in electrical form, but it may also be optical, acoustic or mechanical.

The approach thus used is very natural for it uses an old and widespread method for explaining the pronunciation of a character, called "fan-qie". This method consists in taking two characters of known sounds, and in pointing out that the character of unknown pronunciation as the initial of one and the final of the other. If for example the character to be explained has the pronunciation "mu, (fourth tone)", two well-known and frequent characters are used, for example "ma, first tone" and "zhu, fourth tone", and "ma-zhu-mu" is pronounced.

The coding keys of the keyboard will bear characters -- preferably very frequent ones -- which exhaust all the possible initials and finals or their transcription.

The code may in all cases be transmitted remotely or used locally, for example for operating a printer which will recognize the code and print the corresponding character. Decoding will in all cases be very simple since it is sufficient to assign a character to each digital code, which may be effected for example by means of read-only memory forming a lexicon.

The invention is further described hereinafter, by way of example, with reference to the accompanying drawings, in which:

Figure 1 is a simplified diagram, showing the principal components of the device.

Figures 2 and 3 show a possible arrangement of the keys of a device in accordance with Figure 1, for encoding Chinese characters, the indications borne by the keys being in Latin characters in Figure 2 and in ideograms in Figure 3.

Figures 4 and 5 show a possible method of displaying ideograms on the display means of the device of Figure 1.

Figure 6, similar to Figure 2, shows another embodiment of the keyboard.

Figure 7 is a simplified flow chart showing the succession of successive encoding operations.

The simplified diagram of Figure 1 shows only the main components of the device and their connections, without separating data and address bus lines.

The encoding device illustrated may be regarded as comprising an input keyboard 10, a data processing system whose essential element is a read-only memory 11 and display means 12 which will be successively described.

The input keyboard, using a phonetic approach, is provided with coding keys set out as shown in Figure 2 or Figure 3. These keys comprise :

- A set of initial or final keys 13, whose number is sufficient to exhaust all the possible initials. On each key is inscribed an ideogram, preferably among the most frequent ones, representative of the corresponding initial or final in the "fan qie" presentation (Figure 3) or its transcription into Latin characters (Figure 2) or others, for example Cyrillic. All the phonemes chosen must exist in the four tones and one at least of the meanings in the four tones corresponds advantageously to frequent characters: use may particularly be made of the phonemes ma, bo, ke, shi, cai, fei, dao, chou, tan, gen, li, jia, piao, niu, xian, qiang, rang, hong, zhu, sui, zuo, in the pin-ying transcription, carried by twenty-one keys. It can be seen that the choice is made so that the keys having a different initial carry characters also having terminations which

are all different.

- A set of keys 14 carrying phonemes used without initial, i.e. "er" and "weng", which keys carry the indication of the phoneme in Latin characters (Figure 2) or as ideograms (Figure 3); these phonemes are not used in all the tones: those which are lacking are shown in brackets in Figure 2.

- A set of keys 15, corresponding only to phonemes containing finals not appearing on keys 13; for the sake of clarity, these keys 15 have been separated from keys 13 by a dash-dot line in Figure 2; these finals are further symbolized on the keys by frequent ideograms (Figure 3) or their pin-ying transcription. The keys may carry the symbolization in the four tones, as indicated in the figures, although this is not necessary. Generally, keys 15 carry characters in the four tones and at least one of these meanings corresponds to a frequent character.

- Four tone keys 16_1,... 16_4 for choosing the appropriate tone.

The keyboard carries furthermore selection and printing keys 17, a key 18 whose operation indicates that the character to be encoded does not comprise an initial, an error key 19 and a key 20 for calling the next page whose role will appear further on.

The mass memory 11 is a read-only memory organized by pages, each page containing a given number of ideograms coded in a matrix. Other methods for storing the characters are possible: mechanical, optical or storing a set of programmes for calculating matrices of dots. Most of the characters may be satisfactorily represented on a 16 x 16 matrix. Figures 4 and 5 show respectively the representation, on such a matrix, of two ideograms having both the pronunciation "mū".

Memory 11 may in particular be organized in pages of 256 eight-bit bytes each comprising 8 characters each represented on a 16 x 16 matrix and to each of which is assigned a code formed by the address of the page and the order number of the character in the page.

The ideograms are shared out among the memory pages so as to locate a given ideogram by a step which is statistically the shortest possible. For that, the homonymes are grouped together : all the characters corresponding to a given phoneme, for example "mù", are grouped together in successive memory pages n, n + 1, ..., the number of which depends on the number of homonymes, the characters being ranked in the order of decreasing frequency.

The display means 12 may be either on an apparatus separate from keyboard 10 (Figure 1), or carried by this latter (Figures 2 and 3). They are provided so as to display simultaneously the whole contents of a memory page, i.e. eight characters. In the case considered, this result may be obtained with 4096 dots in the case of 16 x 16 matrices taking into account the necessary gaps. The selection keys 17 enable the character which is to be encoded to be chosen from the eight characters visible on means 12; if this character does not appear, depressing key 20 causes the next page to be displayed.

It should be noted that the actions on key 20 do not cause an increase in the number of bits of the code, defining the page, but solely an incrementation of the page number register.

The encoding device is completed by a central processing unit 21 equipped with its working memories 22 and 23, a unit 24 for printing or transmitting the code selected in mass memory 11 and two registers, an input register 25 and a page register 26.

It will be assumed in what follows that mass memory 11 comprises 512 pages each having 256 eight-bit bytes. The page register 26 may in this case comprise 9 page representation bits (portion 27) and 3 bits for locating the character in the page (portion 28). The input register, associated with the keyboard, may comprise 5 initial representation bits, 6 final representation bits and 2 tone bits.

One of the pages in the memory may preferably contain a set of the most frequently used characters in the ideographic language. Such a set approximately represents 11% of the characters formed in a non-specialized text. They may typically consist of:

的.一.是.我.3.有.國.不

The data processing system may be programmed for that page to be

displayed as soon as the apparatus is energized, before any key is actuated. If then the character to be encoded appears on that page, a single operation (depressing a selection key) is required for encoding. That page is again displayed after each actuation of a selection key, i.e. each time the apparatus is in stand-by condition.

The device is used as shown in the logic flow chart of Figure 7. Before any action on a key, the eight most frequent characters appear on the screen; if the character to be coded is one of these eight characters, action on the corresponding key 17 transfers the code to apparatus 24. For the other cases, it will be explained while assuming that the character to be encoded has the pronunciation "mù" and the representation 目

In this case, the operator depresses successively key 13 marked "ma" (Figure 2) or (Figure 3), then key 13 marked "zhu" and finally key 16_4 (fourth tone).

In other cases, the final will have to be searched for among keys 15. The 13-bit word thus generated is transferred to register 25 which presents it to the computing unit 21. This latter determines the number of the first page where are to be found the characters with pronunciation "mù" and stores it in portion 27 of the page register 26. This page contains the most frequent character representations of the sound "mù", as shown in Figure 2.

In the illustrated example, the character sought for appears on the page, in the fourth position : it is then sufficient to depress the key 17 marked "4" in order to write the figure 4 in binary form in portion 28 of the page register 26 and to feed the corresponding code (address in memory 11) to apparatus 24.

The arrangement of keys 17 corresponds to the arrangement of the characters on the screen : the operator has no need to count the characters or the keys in order to choose the appropriate key. After action on one of the keys 17, the device is brought back to stand-by condition, in which it is sufficient to again actuate a key 17 to encode one of the very frequent characters in the particular page which is displayed at the beginning of operation.

Other situations may occur and correspond to different results of the tests shown schematically at 30 and 31 (Figure 7):

- The operator discovers an error after display on 12: there may be no display because no character having the pronunciation defined

by the initial, the final and the tone exists in the memory; the characters displayed are manifestly false, which reveals an error in operating the keyboard. In this case, the operator depresses error key 19, which resets the registers and restores the beginning of the sequence.

- The character sought for is not on the first memory page. The operator then depresses the next page key 20, which increments the register 27 by a unity and causes further addressing. The computing unit may be provided to cause a particular message to appear after calling up the last memory page corresponding to a given pronunciation. Return to stand-by may then occur automatically after a time delay.

It can be seen that the approach adopted ensures maximum rapidity of access to the character sought for. If this character possesses few homonymes, it will appear on the screen as early as the first page, and the action on the "selection-printing" key causes printing of the character. If the character sought for is an infrequent homonyme and does not appear, action on the "next page" key 20 causes the eight following characters to appear, until the desired character appears and can be printed. About 1960 out of the 4096 possible characters in this embodiment are visible in the first page. Since they will be the most frequent characters, recourse to the following pages will not be very frequent since the 2000 most frequent characters of the Chinese language comprise about 95% of those appearing in the texts.

The embodiment shown in Figure 6 uses the same coding principle, but the keyboard does not comprise tone keys. On the other hand it comprises four separate sets of coding keys, each set corresponding to a tone.

The keyboard of Figure 6 (where the parts already shown in Figure 2 bear the same reference number) comprises 256 character keys grouped into four sets of 64 keys having the same tone, designated by 32_1, 32_2, 32_3 and 32_4. Only a few of the keys are shown for the sake of clarity. In each set, for example set 32_3 a fraction of which is shown on a larger scale in Figure 6a, we find :

- 21 keys 33 for representing initials and finals (similar to keys 13 of Figure 2),
- 14 keys 34 corresponding to finals not appearing among keys 33 (similar to keys 15 of Figure 2).

The 4 x 29 = 116 additional keys carry the most frequent characters which are not among the characters visible in the four coding zones (keys 36), for Chinese and western numerical characters, and punctuation signs.

In the embodiment shown, each key carries two inscribed characters. Access to the characters of the upper row takes place by means of a key 37 similar to the "capital letter" key of a western typewriter or the "numerical" key on teleprinters. The upper characters are chosen according to their frequency and the nearness of their sound to the corresponding character in the lower line. Action on each of these keys causes a different memory page to appear on the screen, and another page by actuating key 37. Finally, the keyboard comprises a conventional space bar 38, and possibly additional keys, for example when the device is to fulfill other functions, such as recording and invoicing, or when it is provided for textual treatment involving page-setting, impression, making ready, correction of errors, etc...

This construction of the keyboard will advantageously lead to storing the same character in several pages, each of the codes thus determined being valid, and in providing the computing unit so as to cause the first page containing the character designated on the key to appear as soon as a first key 33 is actuated to select an initial. It is in fact possible for the character sought for to appear in this case immediately among the eight characters displayed on the display means.

An advantageous distribution may be explained with reference to a very frequent character having numerous homonymes, such as "lì". The most frequent of the "lì"s will then appear in the "normal" position. Action on this key will cause the most frequent character to be displayed and seven other characters which will be the third, the fourth ... up to the ninth in frequency; the second in frequency will appear when the keyboard is in the "capital letter" position. Action on this capital letter key will cause this character to appear with its following 10th, 11th... 16th characters in the order of frequency.

With this arrangement, 512 characters or signs are visible to the operator on the keyboard and a little more than 2000 on the first pages. Now 400 characters are sufficient for printing 73% of a

Chinese text and 2000 for printing 96.5%. The result, after a period of apprenticeship, is that the very great majority of the characters of a current text only require action on one key or two keys (one for coding, one for printing), which leads to a very high typing speed compared with the present systems.

The keyboard may be completed by indicator lights beside keys 36, which indicate to the operator who has sought for, then selected a ter by successive action on two keys 33 and/or 34, and possibly one or more actions on key 20, even though this character figured among keys 36, that a much more rapid approach is available to him by the lighting up of indicator lights suitably disposed so as to indicate the approach which would have been the most rapid. Thus the use of the encoding device is ergonomic for it helps in memorizing the positions of the characters. The use of the "next page" key 20 decreases with use. In this arrangement, where the same character may appear several times in several pages, each of the codes thus determined must be valid, which implies an appropriate structure of 24 (Figure 1).

Simplified low cost embodiments not shown are also possible : for example the keyboard may comprise a group of initial keys and a separate group of final keys each carrying a phonetic code (transcription into Latin or other characters). The coding approach appears perfectly on such a keyboard.

The device is directly transposable to Japanese, in which the finals may be mono- or disyllabic and each Chinese character has two pronunciations one of which is monosyllabic or at most disyllabic.

A simple and practical encoding device uses a keyboard comprising 70 coding keys on each of which is inscribed a square syllabic character and additional keys for the figures and typographical signs. Three keys make it possible to indicate if it is desired to write a square syllabic character, a cursive syllabic character or a Chinese character.

In the case of syllabic characters, the character is encoded by depressing a single key. Each coding key of the keyboard also carries a Chinese character; successive action on the "Chinese character" key and on a coding key of the keyboard causes the memory page in which the desired character appears to be displayed on the

screen. Again the character on the page is chosen by acting on the appropriate selection and printing key. The characters not appearing in clear on the keyboard are, as in the previous cases, coded by successive action on two coding keys which cause appearance of the memory page containing the character whose initial and final correspond to the two keys. The case of homonymes is treated as for Chinese.

But a much smaller memory is sufficient, containing only 2048 characters, that is to say, for a 16 x 16 coding matrix (32 eight-bit bytes), a memory of 64 K8-bit bytes is sufficient.

It has been assumed up to here that each character is coded on a 16 x 16 matrix (Figures 5 and 6). But this choice is not indispensable. The characters may be represented on a different matrix depending on their complexity : a part of the characters may in particular be represented on 8 x 8 matrices, another part on 12 x 12 matrices and another part on 24 x 24 matrices by grouping together the characters whose graphism is the simplest in the same page thus resulting in pages of 144 8-bit bytes (instead of 256), possibly comprising 18-8 or 2 characters respectively.

The characters with the most complex graphisms may also be grouped together in special pages where they are coded on a 32 x 32 matrix; each page of 256 8-bit bytes then contains two characters.

In these cases, the "selection-printing" keyboard must be able to be adapted to the different numbers of characters per memory page. Since the size of a page remains fixed, which considerably simplifies the software, the number of printing keys would in all cases correspond to the maximum number of characters per page. This result may be reached by means of the software of the computer unit 21 which only assigns a selection function to keys 1 and 2 in the case of a two-character display, to keys 1 to 4 in the case of a four-character display and so on. In the case where only 16 x 16 and 32 x 32 matrices are provided, keys 1 and 2 may have a different colour so as to individualize them.

Moreover, methods may be conceived for displaying the characters by other means : electro-optical, electromechanical, or else by stocking not the characters themselves but simple algorithms for calculating material representations of the characters, thus economizing on the size of the memory.

Numerous modified embodiments are possible. For instance, the apparatus may include an auxiliary RAM or REPROM memory having a content corresponding to one page. Prior to encoding a text, the eight more frequently occuring characters are determined and written in the memory. Then the system is programmed for that page to be displayed each time the apparatus returns to stand-by condition, after actuation of an encoding key. That arrangement is of particular interest when dealing with texts relating to a specialized field.

When it is desired to reduce the size of the memory to a minimum, it may be sufficient to store in a permanent (non volatile) memory:

- the most frequent characters;
- those characters which cannot be reproduced for printing of two simpler characters at the same location in succession;
- possibly, a RAM may be reserved for temporary storage of some additional characters which frequently occur in a particular text.

Then an additional key is required which is actuated for indicating that two successive codes should be processed as defining the same character.

* C L M S *

28

CLAIMS:

1. An apparatus for encoding ideographic characters, comprising a mass memory for storing characters each having an address in said memory, display means associated with said mass memory for simultaneously displaying all characters of a set in said memory, means associated with the memory for causing display of one said set consisting of the most frequent characters when the apparatus is in stand-by condition prior to actuation of a key, a selection keyboard having a plurality of selection keys each manually actuatable for selection of one of the characters of said set displayed on the display means, whereby a unique code representing the selected character is delivered, and a coding keyboard having a plurality of coding keys each carrying an identification marking and arranged for possible identification of at least a first and a second character components, said keyboards and memory cooperating whereby actuation of coding keys causes display of a memory page containing the characters or most frequent characters defined by at least said first and second character components selected by actuation of coding keys and subsequent action on one of the selection keys results in delivery of a code corresponding to a selected character among the displayed characters.

2. An apparatus for encoding ideographic characters, comprising a mass memory for storing characters each having an address in said memory, display means associated with said mass memory for simultaneously displaying all characters of a set stored in a page of the memory, means associated with the mass memory for causing display of a page containing most frequent characters when the apparatus is in stand-by condition prior to actuation of a key, a selection keyboard having a plurality of selection keys each manually actuatable for selection of one of the characters of said set displayed on the display means whereby a unique code representing the selected character is delivered, and a coding keyboard having a plurality of coding keys each carrying an identification marking and arranged for possible identification of each phonetic initial character component and each phonetic final character component as well as of a tone if necessary, said keyboards and memory cooperating whereby actuation of coding keys causes display of a memory page containing the characters or most frequent characters defined by the initial and final character components, and possibly the tone, associated with the actuated coding keys and subsequent action on one of the selection keys results in delivery of a code corresponding to a selected character among the displayed characters.

29

3. Apparatus according to claim 2, wherein the coding keyboard comprises a set of keys each assigned to a different initial, a set of keys each assigned to a different final, and means for tone selection.

4. Apparatus according to claim 3, wherein the coding keyboard has four tone keys permitting determination of the tone.

5. Apparatus according to claim 3, wherein the coding keyboard comprises four separate sets of coding keys, each set corresponding to a tone for allowing the determination of the tone.

6. Apparatus according to any one of claims 3 to 5, wherein the coding keyboard further comprises a key the actuation of which indicates that the phoneme to be codified does not comprise an initial.

7. Apparatus for encoding Chinese characters according to any one of claims 3 to 6, wherein the coding keyboard comprises further keys for the phonemes "er" and "weng" which have no initial in the four tones.

8. Apparatus according to any one of claims 3 to 7, wherein on each initial coding key is inscribed an ideogram having a pronunciation element corresponding to the element codified by the key, typically selected among the most frequent characters having the same sound, or its phonetic transcription in a Western language.

9. Apparatus according to claim 8, wherein each of the keys assigned to an initial carries a character having a different ending from that of the characters of the other initial coding keys.

10. Apparatus according to claim 8 or 9, wherein the sounds corresponding to said characters carried by the keys have meanings in each of the four tones, one at least of these meanings corresponding to frequent characters.

11. Apparatus according to any one of claims 3 to 10, further comprising computer means for causing the appearance on the display means, in response to the action on a first coding key with an indication of the tone, of the contents of a memory page containing the most frequent characters having the pronunciation corresponding to that

* C L M S *

coding key.

12. Apparatus according to claim 11, wherein the coding keyboard has a "next page" key for replacing display of one memory page with display of the next page containing homonymes of the first one, and possibly other characters and possibly a message to appear if the page is the last one containing homonymes.

13. Apparatus according to claim 11 or 12, wherein the coded representation of a character is either the stored page number followed by the serial number of the character in the page or its representation in the Chinese telegraphic code, or its translation into another language, or any combination of the preceding representations.

14. Apparatus according to any one of the preceding claims, wherein the keyboard further includes additional keys corresponding to most frequent Chinese characters.

15. Apparatus according to any one of the preceding claims, wherein the memory is organized in pages each containing the same number of coded characters in accordance with a matrix of predetermined dimensions, for example a 16 x 16 matrix, each page of 256 8-bit bytes then containing eight characters.

16. Apparatus according to any one of claims 1 to 14, wherein the characters are coded in accordance with different matrices depending on their complexity, for example a part of the characters on 8 x 8 matrices, another part by 12 x 12 matrices, and another part by 24 x 24 matrices, and the characters whose graphism is the simplest are grouped in the same page.

17. Apparatus according to any one of claims 3 to 14, wherein each initial key is also assigned to a final among those not associated with a final key, the computer means of the device being provided so as to cause the selection of an initial by the first key actuated and the selection of a final by the second.

18. Apparatus for encoding ideographic characters substantially as hereinbefore described with reference to the accompanying drawings.

The text of the specification has been reproduced by photocopying the applicants original typescript. It may contain a few amendments which are difficult to read. The original typescript containing these amendments may be inspected on the premises of the Patent Office