

PHONETIC ENCODING SYSTEM OF CHINESE CHARACTERS

5 The subject matter of this invention relates to a
system for encoding Chinese characters phonetically in
which the allotted codes are formed as short as possible
and are suitable as much as possible not only to the
standard pronunciation but also to the pronunciation of the
10 many Chinese dialects, which differ greatly from one other.

Great difference of phonetic rule exists between
Chinese and Occidental languages. A longer spelling
results from adopting the ordinary method of phoneticizing
Chinese characters with Latin letters, e.g. "庄¹(zhuang)",
15 "窗^{1 1/2}(chuang)", "双^{1 1/2}(shuang)", all with 6 letters. What is
more, Chinese contains many greatly differing dialects and
the standard pronunciation is not yet popularized.
Therefore, it is difficult for those from southern-dialect
areas (including overseas Chinese who speak southern
20 dialects) to manage phonetic spelling by standard
pronunciation. Up to now, no phonetic encoding system is
available for shortening phonetic spelling as well as
suitable to various dialects as much as possible.

The invention system for encoding Chinese
25 characters, has been previously filed, number:

GB 2 100 899 A. The present invention takes the aforesaid system a step forward. This system unifies stroke shape encoding and phonetic encoding.

According to the invention, there is provided a method of entering Chinese characters by phonetically encoding them, in which the character is encoded as a two-letter alphabet letter code determined by allotting as the first letter of the code, one of the 21 Latin consonants in the alphabet to represent one of the 21 consonants of the Fan-qui pronunciation of the character and, as the second letter of the code, one of the 5 Latin vowels, one of the 21 Latin consonants or one of the 11 Latin alphabet letters already used to represent one of the 37 vowels of the Fan-qui pronunciation of the character, and in the event that this produces homonymous codes for different characters, supplementing the two-letter code with a number code corresponding to a stroke-shape identification code, and that code is used to enter the Chinese character into character-information processor apparatus.

Thus, the invention uses the 26 Latin letters which stand for 21 consonants and 37 vowels in Chinese and naming each Latin letter by way of reading out the consonant and the vowel, both represented by this same Latin letter, to obtain the pronunciation of the Chinese character which is used for naming the Latin letter.

Stroke shape identification codes are used to differentiate homonyms which could arise in phonetic spelling. Brevity codes may be allotted for frequently used characters.

5 The invention is described in detail as follows:-

(1) The correlation between Chinese characters and Chinese consonants and vowels:-

Referring to the table in Figure 1, Latin letters are listed in the top line, Chinese consonants in the 10second, Chinese vowels in the third, and in the bottom line there are Chinese characters which are the names of their related Latin letters. The 21 Latin letters represent not only the 21 consonants but also the corresponding vowels in the third line. For example, if the phonetic syllable of a 15character is "bb", the first "b" is pronounced according to its original articulation, the second "b" is pronounced according to the articulation of its corresponding vowel "ai". This is what we mean with one letter representing

both a consonant and a vowel. In Figure 1, all consonants have the same role as those in "The Scheme for the Chinese Phonetic Alphabet" but v, w, y, which represent zh, ch, sh, in the Scheme. Vowels represented by 21 Latin letters are shown in Figure 1 in which characters in the fourth line are the names given to the related Latin letters and also the phonetic syllables obtained by reading out the consonants and the vowels accordingly, both represented by the related Latin letters named.

10 The five Latin letters, a, o, e, i, u in Figure 2 are simple vowels.

So far, 21 out of 37 vowels in Chinese are represented by the consonants themselves, another 5, by the five Latin vowels. The remaining 11 will be represented by the Latin letters which have been used for representing vowels, thus all 21 consonants and 37 vowels in Chinese are represented by 26 Latin letters.

For illustrating how the remaining 11 vowels are represented by the already-used letters, it is necessary to point out the theory on which the correlation between consonants and vowels in Chinese is based. The Chinese syllable is of a consonant-vowel structure, i.e. a consonant and a vowel construct a syllable, with the consonant before the vowel. Thus, a Latin letter is in a position to appear twice consecutively to stand for a consonant and a vowel at the same time with the former representing a consonant and the latter, a vowel. Whenever

(109) 5
a Chinese character is represented by the first must be consonant and the second, a vowel. This is the "double spelling with consonant and vowel" for Chinese character syllables.

- 5 Regularly, there is a mutually complementary relationship between a Chinese consonant and a Chinese vowel, e.g. vowel "uan", "uen", "ong(ueng)" can never be combined with consonant "j", "q", "x", whereas, "üan", "ün", "iong(üeng)" can only be combined with "j", "q", "x".
- 10 This law makes it possible to use "z", "e", "s" for representing both "uan", "uen", "ong(ueng)" and "üan", "ün", "iong(üeng)".

Another 8 Latin letters which appear thrice for representation are shown in Figure 3.

- 15 In addition, "er" is pronounced "ㄦ¹²" in Chinese.

Again, use Latin letter "i" for representing the vowel in "資¹¹(zi)", "此¹¹(ci)", "私¹¹(si)" and "知¹¹(zhi)", "吃¹¹(chi)", "詩¹¹(shi)", "日¹¹(ri)".

- In the ultimate, by using 26 Latin letters
- 20 representing all consonants and vowels in Chinese, the foundation is laid for encoding Chinese characters by double spelling with consonant and vowel.

(2) Rules for Spelling:-

- Each Chinese character consists of two letters,
- 25 the former is for consonant, the latter for vowel, as shown in Figure 4. Take "江¹", "江¹²" for example.

The character "江" should be spelt phonetically "jiang" according to "The Scheme for the Chinese Phonetic Alphabet". By the above-explained rule of using one letter representing both consonant and vowel at the same time, the phonetic spelling of the character can be reduced to "jq". The same reason is for character "王" being simplified from "wang" to "uq" in which only vowels appear with vowel "u" taking the place of the lacked consonant. Such a syllable is called zero-consonant syllable.

10 In order to remove a great extent the obstacle to the phonetic spelling of Chinese characters, which arises from the phonetic interference of dialects, this invention adopts a method of pronouncing a character with another two characters which represent the related Latin letters, 15 namely, the new form of Fan-qui (a traditional Chinese phonetic method). By Fan-qui we mean taking the consonant of the first character and the vowel of the second, reading out the consonant and the vowel to get the pronunciation of the third character. Phonetic interference of dialects can 20 be got rid of to great extent this way. There is a close correlation between standard Chinese (spoken language) and various dialects. If one reads out the first and the second characters used for Fan-qui in standard pronunciation, he or she will pronounce the would-be third 25 character in standard manner; if one reads out the first and the second in his or her dialectal sound, he or she will pronounce the third one dialectally. No matter how

7

standard Chinese is different from various dialects, one will get the same character by way of Fan-qui, thus solving the embarrassing question in phonetic spelling and the encoding of Chinese characters, which is due to the
5 different pronunciations of standard Chinese and various dialects.

Figure 5 shows the correlation between standard Chinese and Cantonese, and how the interference of dialectal sound is removed, in which we can see clearly
10 that the character obtained by way of Fan-qui of another two characters in standard pronunciation is the right character obtained by way of Fan-qui of another two characters in Cantonese accent, but this character can be pronounced in standard way or in Cantonese way.

15 Encoding Chinese characters phonetically according to "The Scheme for the Chinese Phonetic Alphabet" is confined by dialectal pronunciations and it is difficult to get a unified phonetic encoding system for Chinese characters. Besides, it is not an easy thing for those who
20 cannot speak standard Chinese to master such a phonetic method. When using the method of Fan-qui with consonant and vowel, those from dialectal areas may easily carry out 90% of all Chinese phonetic spelling in their respective dialectal pronunciation to get the right characters
25 obtained otherwise by way of standard phonetic spelling. For the minor odd characters, one more reference source can be used as a remedy in character-information processors.

q

(22)

(3) Character Shape Identification Codes:-

One can tell a Chinese homonym from others by their different shapes, and the simplest way is to encode Chinese characters by their stroke shapes. The invention of this kind of system has already been claimed. According to statistical data, a stroke shape code of four digits is enough for the identification of characters with same code. The concrete way is, dividing the stroke shapes of a multi-component character into two groups, in each of which two strokes are taken to form a 4-digit stroke shape code; for one-component character, taking four strokes consecutively in the order of from top to bottom and from left to right. Figure 6 exemplifies character shape identification codes of homonyms of syllable "bb(bai)", those out of curves are effective codes, whereas those within are ineffective ones, (i.e. useless codes) which can be omitted in storage or encoding. The character can be differentiated that way.

(4) Brevity Codes:-

By the information theory, briefest codes should be given to the frequently used characters. One Latin letter and a 1-digit stroke shape code can be used for encoding 208 (26x8) Chinese characters. 8 brevity codes are shown in Figure 6.

According to statistics, these 208 characters make up more than 50% of total use frequency of Chinese characters.

②

9

One Latin letter code and two stroke shape codes can be used for encoding 1664 characters, while two letter codes and one stroke shape code can be adopted to encode 5408 characters. That is to say, 7,000 characters can be allotted a 3-digit code.

Statistical data show that 1690 characters make up 97% of total use frequency, and 2393, 99%.

If encoding with Latin letters and digits alternatively (letter before numeral), the ending symbol indispensable to common encoding (including telegraphic coding) can be omitted. By this encoding system, only 2.5 codes are used for each character on the average, whereas 5 codes are used for each telegraphic coding on the average. The former is 50% shorter than the latter. By using this encoding system, character-information processors (including computers, teleprinters, typewriters and dial telephones, etc.) can work easily, accurately, and more rapidly.

This system for encoding Chinese characters phonetically can be used for compiling Chinese dictionaries, codes and character indexes. Thus a dictionary, code or Chinese character index can be compiled by encoding each character as a two-letter alphabet letter code determined by allotting as the first letter of the code, one of the 21 Latin consonants in the alphabet to represent one of the 21 consonants of the Fan-qui pronunciation of the character

and, as the second letter of the code, one of the 5 Latin vowels, one of the 21 Latin consonants or one of the 11 Latin alphabet letters already used to represent one of the 37 vowels of the Fan-gui pronunciation of the character, and in the event that this produces homonymous codes for different characters, supplementing the two-letter code with a number code corresponding to a stroke-shape identification code, sorting the Chinese characters into order according to the resulting codes, and listing the codes in that order to produce the dictionary, code or Chinese character index.

Also this system can be used for establishing Chinese character-information processing system in character-information processor apparatus such as computers, teleprinters, typewriters of large, medium or small size, and dial telephones.

Brief description of the drawings:-

Figure 1: Consonant-vowel table.

Shows the correlation between Latin letters and Chinese consonants, vowels, and the related character names.

Figure 2: Simple vowels table.

5 Shows the correlation between Latin letters and simple Chinese vowels and the related character names.

Figure 3: Letters which appear thrice.

10 Shows the correlation between Latin letters which appear thrice and the represented vowels.

Figure 4: Examples of Chinese phonetic spelling.

15 Gives Chinese characters, consonants and vowels which are used for phonetic spelling and are compared with original Chinese phonetic alphabets.

Figure 5: Fan-qui of Chinese characters.

Shows the characters used for Fan-qui and the third ones obtained by way of Fan-qui in both standard Chinese and Cantonese.

20 Figure 6: Stroke shape identification codes table.

Gives the stroke shape identification codes of the homonyms.

Figure 7: Table of brevity codes of Chinese characters.

25

Gives brevity codes of Chinese characters, composed by one Latin letter and one digit.



12

CLAIMS:-

(1) A method of entering Chinese characters by phonetically encoding them, in which the character is encoded as a two-letter alphabet letter code determined by allotting as the first letter of the code, one of the 21 Latin consonants in the alphabet to represent one of the 21 consonants of the Fan-qui pronunciation of the character and, as the second letter of the code, one of the 5 Latin vowels, one of the 21 Latin consonants or one of the 11 Latin alphabet letters already used to represent one of the 37 vowels of the Fan-qui pronunciation of the character, and in the event that this produces homonymous codes for different characters, supplementing the two-letter code with a number code corresponding to a stroke-shape identification code, and that code is used to enter the Chinese character into character-information processor apparatus.

(2) A method as claimed in Claim 1 in which "double spelling with consonant and vowel" is equal to using two characters for Fan-qui, which are the names of their related Latin letters according to the correlation between standard Chinese pronunciation and various dialectal sounds, when Fan-qui is done in standard pronunciation, the character thus acquired is pronounced standardly and encoded, and when Fan-qui is done in dialectal pronunciation, the character thus obtained has a dialectal pronunciation and is encoded, and despite the difference in



13



pronunciations, the code of the character obtained by two different ways is the same one.

(3) A method as claimed in Claim 1 or Claim 2 in which brevity codes are allotted to frequently used characters.

(4) A method of phonetically encoding Chinese characters substantially as herein described with reference to Tables 1 to 7.

(5) A character-information processor apparatus which uses a phonetic encoding system for Chinese characters as claimed in any preceding claim for Chinese character information processing.
