(12) UK Patent Application (19) GB (11) 2 158 276 A

(43) Application published 6 Nov 1985

(21) Application No 8409538

(22) Date of filing 12 Apr 1984

(51) INT CL⁴
H03M 11/00

(52) Domestic classification
G4H KU
U1S 2125 2213 2216 2247 G4H

(56) Documents cited
GB 1600841 GB 1398882

(58) Field of search
G4H

(71) Applicants
Jin-Kai Li,
No 140 5th floor, Beijing Normal University, Beijing, China.
Kam-Fu Wong,
22nd floor, Wu Sang House, 655 Nathan Road, Kowloon, Hong King.
Ping Lu,
9 Dai Fhun Street Pai Po Lnb Eftate, New Territories, Hong Kong

(72) Inventor
Jin-Kai Li

(74) Agent and/or Address for Service
Lloyd Wise Tregear & Co,
Norman House, 105–109 Strand, London WC2R 0AE

(54) Phonetic encoding system of chinese characters

(57) 26 Latin letters are used for representing 21 consonants and 37 vowels in Chinese, to spell a character with two Latin letters. A consonant is combined with a vowel, both represented by a Latin letter, to get a character used as a name allotted to the Latin letter. After that, Fan-qui (a traditional Chinese phonetic method) is done with characters used for naming Latin letters. By this way, the obstacle of the phonetic interference of dialects can be removed to great extent.

Stroke shape codes are used to identify homonyms and alternative encoding with letters and digits is adopted to drop the ending symbol. Brevity codes are used for frequently used characters.

GB 2 158 276 A

The drawings originally filed were informal and the print here reproduced is taken from a later filed formal copy.
Chinese Characters in the printed specification were reproduced from drawings submitted after the date of filing, in accordance

Table 1 Consonant-Vowel Table

Latin letters	b	P	m	f	V	W	Y	r
Chinese Consonants	b	P	m	f	zh	ch	sh	r
Chinese Vowels	ai	ei	ao	ou	an	en	ang	eng
Related Chinese Characters	白	陪	矛	否	占	尘	伤	仍
Latin Letters	d	t	n	l	j	q	X	
Chinese Consonants	d	t	n	l	j	q	x	
Chinese Vowels	ie	iao	iou	ian	in	iang	ing	
Related Chinese Characters	跌	挑	牛	连	今	枪	星	
Latin Letters	g	K	h		Z	c	S	
Chinese Consonants	g	K	h		z	c	s	
Chinese Vowels	ua	uo	uei		uan	uen	ong(ueng)	
Related Chinese Characters	瓜	扩	灰		钻	村	松	

Fig. 1

Table 2: Simple Vowels Table

Latin Letters	a	o	e	i	u
Simple Chinese Vowels	a	o	e	i	u
Related Chinese Characters	啊	噢	我鸟	衣	乌

Fig. 2

Table 3: Letters Which Appear Thrice

Latin letters	z	c	s	g	h	o	d	q
Chinese Vowels	üan	ün	iong	ia	üe	ü	uai	uang

Fig. 3.

Table 5: Fan-qui of Chinese Characters

	First Characters	Second Characters	Third Characters
① Chinese Characters	白 +	陪 →	贝
② Standard Chinese (Chinese Phonetic Alphahet)	bai	pei	bei
③ Cantonese (International Phonetic Alphahet)	bak	pui	bui
①	占 +	尘 →	甄
②	zhan	Chen	zhen
③	dzim	tsan	dzan
①	枪 +	星 →	清
②	qiang	xing	qing
③	tsæŋ	siŋ	tsiŋ
①	跌 +	连 →	颠
②	die	Lian	dian
③	dit	lin	din
①	仏 +	扩 →	郭
②	gua	Kuo	guo
③	gwa	kwɔk	gwɔk
①	钻 +	松 →	棕
②	zuan	song	zong
③	dzyn	tsuŋ	dzuŋ

Fig. 5

Table 4: Examples of Chinese Phonetic Spelling

Chinese Characters	江	山	多	娇	汪	洋	哇	呀
Consonants & Vowels	j q	y v	d k	j t	u q	i q	u q	i g
Chinese Phonetic Alphabets	jiang	shan	duo	jiao	wang	yang	wa	ya

Fig. 4

Table 6: Stroke Shape Ideintification Codes

Phouetic Alphchet	Chinese Character	Effective Stroke Shape Codes	Ineffective Stroke Shape Codes
bb (bai)	百	1	301
"	败	2	531
"	白	3	01
"	佰	32	13
"	拜	33	17
"	摆	5	102
"	柏	7	330

Fig. 6

Table 7: Brevity Codes of Chinese Characters

Brevity Codes	Chinese Characters
b_1	不
b_2	北
b_3	保
b_4	部
b_5	把
b_6	比
b_7	本
b_0	别

Fig. 7

SPECIFICATION

Phonetic encoding system of Chinese characters

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 pronunication but also to the pronunciation of the many Chinese dialects, which differ greatly from one another.

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 (zhuang)", " 窗 (chuang)", " 双 (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 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 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.

This invention is characterized by: using 26 Latin letters which stand for 21 consonants and 37 vowels in Chinese; 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.

The invention is described in detail as follows:—

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

Referring to the table in Fig. 1, Latin letters are listed in the top line, Chinese consonants in the second, 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 character 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 Fig. 1, all consonants have the same role as those in "The Scheme for the Chinese Phonetic Alphabet" but v, w, y, which represents zh, ch, sh, in the Scheme. Vowels represented by 21 Latin letters are shown in Fig. 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.

The five Latin letters, a, o, e, i, u in Fig. 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 a Chinese characrer 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.

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". This law makes it possible to use "z", "e", "s" for representing

both "uan", "uen", "ong(ueng)" and "üan", "un", "iong(üeng)".

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

In addition, "er" is pronounced "儿" in Chinese.

Again, use Latin letter "i" for representing the vowel in "资 (zi)", "雌 (ci)", "私 (si)" and "知 (zhi)", "吃 chi)", "诗 (shi)", "日 (ri)".

In the ultimate, by using 26 Latin letters 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, the former is for consonant, the latter for vowel, as shown in Fig. 4. Take "江", "汪" 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 "汪" 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.

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, 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 be got rid of to great extent this way. There is a close correlation between standard Chinese (spoken languag) 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 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 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 different pronunciations of standard Chinese and various dialects.

Fig. 5 shows the correlation between standard Chinese and Cantonese, and how the interference of dialectal sound is removed, in which we can see clearly 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.

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 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 obtained otherwise by way of standard phonetic spelling. For the minor add characters, one more reference source can be used as a remedy in character-information processors.

(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. Fig. 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 (26 × 8) Chinese characters. 8 brevity codes are shown in Fig. 6.

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

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 letyters 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 and character indexes, as well as for establishing Chinese character-information processing system in computers, teleprinters, typewriters of large, medium or small size.

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

Figure 3: Letters which appear thrice. Shows the correlation between Latin letters which appear thrice and the represented vowels.

Figure 4: Examples of Chinese phonetic spelling. 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.

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

CLAIMS

1. A phonetic encoding system of Chinese characters, in which a limited number of notational letters are used for representing consonants and vowels of Chinese characters, and the stroke shape identification codes are used for differentiating homonymous characters.

2. A method as claimed in Claim 1 in which 26 Latin letters are used for representing 21 consonants and 37 vowels in Chinese and any Chinese syllable can be double spelt with consonant and vowel.

3. A method as claimed in the above two Claims in which 21 consonants of Latin letters stand for both 21 consonants and 21 vowels of Chinese characters, 5 Latin vowels for 5 simple Chinese vowels, the remaining 11 Chinese consonants are represented by 11 Latin letters which have already been used for representing Chinese consonants and vowels, thus using 26 Latin letters representing 21 consonants and 37 vowels of Chinese characters.

4. A method as claimed in Claims 1 to 3 in which any Chinese syllable is composed of two letters, the former is a consonant, the letter, a vowel, thus forming the "double spelling with consonant and vowel".

5. A method as claimed in any of the preceding Claims 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; when Fan-qui is done in dialectal pronunciation, the character thus obtained has a dialectal pronunciation; despite

the difference in pronunciations, the character obtained by two different ways is just the same one.

6. A method as claimed in Claim 1 to 5 in which homonymous characters are differentiated by stroke shape codes of Chinese characters, i.e. stroke shape identification codes.

7. A method as claimed in any of the preceding Claims in which brevity codes are allotted to frequently used characters; 3-digit codes are given to majority of Chinese characters; thus it is possible to encode all Chinese characters.

8. A method as claimed in Claims 1 to 7 in which alternative encoding with letter codes and digit codes is used and an ending symbol every two characters can be omitted.

9. Dictionaries, codes and character indexes compiled by a method as claimed in Claims 1 to 8.

10. Character-information processors of large, medium or small size, apparatus such as computers, teleprinters, typewriters and dial telephones using the phonetic encoding system of Chinese characters as claimed in Claims 1 to 8, and adapted for Chinese character information processing.

Printed in the United Kingdom for
Her Majesty's Stationery Office, Dd 8818935, 1985, 4235.
Published at The Patent Office, 25 Southampton Buildings,
London, WC2A 1AY, from which copies may be obtained.