Password in Chinese Language: Strong and Memorable

Lee Kok Wah

Faculty of Engineering & Technology, Multimedia University, Jalan Ayer Keroh Lama,
75450 Bukit Beruang, Melaka, Malaysia

kwlee@mmu.edu.my

Abstract. In computing system, password is dominated by Roman alphabet or Latin alphabet due to character encoding of ASCII. Here, a pronounceable and memorable password policy in Chinese language is proposed. Phonetic encoding of Hanyu Pinyin and symbolic encoding of Sijiao Haoma (or four corner method) are used to create the uniqueness of each Chinese character or Han character for alphanumeric representation. Based on about 70217 Han characters in the Unihan database (CJK ideographs) of Unicode 4.1, each Han character has entropy of 16.1 bits. Five Han characters will satisfy the 80-bit minimum randomness requirement of symmetric key cryptosystem for strong password. Self-created signature-like Han character and passphrase represented by printable ASCII generates shorter, stronger and more memorable password with 27.3 and 85.4 bits per Han character. CJK languages using the Han characters like Cantonese language and Japanese language are applicable via the pronunciation romanization systems of jyutping and rōmaji respectively.

Keywords: Password security, Chinese language, Han character of Unicode, memorability, passphrase generation, cryptography.

1 Introduction

Secret, token and biometrics are the three main approaches of authentication. Authentication is normally required for the control of access to resources. Secret is about something you know like password. Token is about something you have like smart card. Biometrics is about something you are like fingerprint. For the sake of cost and compatibility, secret in the form of password is the most popular authentication approach.

According to Kerckhoffs' law [1], a cryptosystem shall depend 100% on the secrecy of password or key only. In the words of Shannon's maxim, it means "enemy knows the system". This law makes the civilian cryptosystem to have publicly known algorithm except the classified governmental and military information.

If a cryptographic algorithm is securely tested, the required key length in character (L_C) of a password will depend on the factors of number of characters (C), key space (S), secure period (T), guesses per unit of time (G) and probability of guessing (P) [2]. The minimum key length has to be able to resist the brute force attack. The relationships of L_C, C, S, T, G and P are given in Eqs.(1) and (2).

$$S = \frac{GT}{P} \tag{1}$$

$$L_C \geq \left\lceil \frac{\log_2 S}{\log_2 C} \right\rceil \tag{2}$$

In today computing system, the character encoding of ASCII is the most popular code. ASCII has the possible keys of 26 lowercase characters, 26 uppercase characters, 10 digits, 62 alphanumeric characters, 33 non-alphanumeric characters, 95 printable characters, etc. If a password has only the symbols of digits, its specialized name is *passcode*. If a password is long or consists of printable characters, it is named as *passphrase*.

There were once three Data Encryption Standard (DES) challenges as in year 1997, 1998 and 1999. Using the distributed network computing, maximum guesses of 2.45×10^9 keys per second was once recorded. For the latest guesses per computer as at end of year 2005, it is about 1.5×10^7 keys per second. The increment rate follows the Moore Law where computer performance is double for every 18 months. This indicates that strong password has to be longer as time passing by.

Key length in bit (L) means that there are 2^n possible keys for n-bit key. By year 2010, the required key is 80 bits for symmetric key algorithm as announced by U.S. National Institute for Standards and Technology (NIST). Meanwhile, asymmetric key algorithm like RSA needs 1024 bits to be equivalently strong with 80-bit symmetric key algorithm as claimed by RSA Security. The key space varies and depends on the security requirements. The symmetric key algorithm of Advanced Encryption Standard (AES) uses the settings of 128-, 192- and 256-bit keys.

Password choice depends on strength and memorability. Strength depends on key size in bit. Memorability depends on number of secrets. For minimum key sizes at different security levels, it is shown in Table 1 [3]. For short term memory of English-based digit, Miller [4] showed an average of seven items plus or minus two (7 ± 2). The good choice is longer key size in bit and still memorable. Here, we create strong and memorable password in Chinese language. Then Cantonese language is added.

Table 1. Minimum symmetric key sizes for different security levels of protection (# 1 to # 7).

Key Size (bit)		Protection
# 1	32	Individual attacks in "real-time". Only acceptable for authentication tag size.
# 2	64	Very short term protection. Obsolete for confidentiality in new systems.
# 3	72	Short to medium term of protection depending on organization size.
# 4	80	Smallest general purpose level, < 5 years protection.
# 5	112	Medium term protection. About 20 years.
# 6	128	Long term protection. Good, generic application independent recommendation, about 30 years.
# 7	256	Foreseeable future. Good protection against quantum computers.

2 Methods

A good password has to be strong and memorable [5] [6]. The random password with printable ASCII characters is the strongest password but it is poor in memorability [7]. However, password with good memorability tends to be weak password and under the password cracking threats of guessing and dictionary attack [8]. As time lapses, longer key length is needed due to the advance of computer technology. Hence the trend is the strong and memorable passphrase.

The most popular email encryption Pretty Good Privacy (PGP) 9.0 allows a maximum of 255 characters to be the passphrase. There are two ways to generates passphrase. One way is to have an entire phrase or full sentence. Another way is through the acronym as in Table 2 to create a 12-character passphrase.

Table 2. Passphrase examples for types of full sentence and acronym.

Passphrase Type	Example
Full Sentence	Hanyu Pinyin, Zai Jia Sijiao Haoma.
Acronym	HyPy,ZjSjHm.

2.1 Environ Password

Good memorability exists when it is linked to the learnt language. For English language, U.K. government introduced the case insensitive Environ password in October 2005 for short term protection. It has an 8-character key pattern as in Table 3.

Table 3. Environ password.

Form	[consonant - vowel - consonant - consonant - vowel - consonant - digit - digit]
Example	pinray34, yankan77, supjey56, kinkin99

This pronounceable and hence memorable password has an entropy (E) or key length in bit (L) of 34.86 bits per unit of Environ password. It is not 64 bits as the key space involves special arrangement of 62 alphanumeric ASCII characters. For every unit of Environ password, it carries four secrets. These secrets are two syllables and two non-associative digits. The relationship of E, L, L_C and C is shown in Eq. (3).

$$E = L = L_C \log_2 C \tag{3}$$

2.2 Chinese Language Password

For Chinese language, in order to have the Chinese character uniquely represented, the phonetic encoding of Hanyu Pinyin (汉语拼音) pronunciation system is not enough. In Hanyu Pinyin, there are about 415 unique syllables with 22 initials (or onsets) and 39 finals as in Table 4 [9]. Excluding permutation and capitalization, it has conservative entropy of 8.70 bits per syllable. An 80-bit or 128-bit symmetric key will need 10 or 15 syllables respectively. As the average length of Hanyu Pinyin syllables is 3.22 characters per syllable, it will be 33 or 49 characters respectively. This is highly not efficient. A better form of Han character romanization is needed.

Table 4. Phonetic encoding of Hanyu Pinyin in Chinese language (Mandarin-based).

Initial (22)	nil	b	p	m	f	d	t	n
	l	g	k	h	j	q	x	z
	c	s	zh	ch	sh	r		
Final (39)	a	o	e	ê	ai	ei	ao	ou
	an	en	ang	eng	ong	i	ia	io
	ie	iao	iu	ian	in	iang	ing	iong
	u	ua	uo	uai	ui	uan	un	uang
	ueng	ü	üe	üan	ün	-i	er	

N.B.: For romanization, ê and ü can be represented by [oe] and [v] respectively.

Also, to allow a memorable password or passphrase, we have to reduce the amount of secrets to be remembered. There are a total of 70217 Chinese characters or Han characters in Unicode 4.1 [10] (URL: http://www.unicode.org/charts/) for CJK unified ideographs, CJK unified ideographs extension A and CJK unified ideographs extension B. About 200 CJK ideographs are radicals (部首). Hence about 70000 Han characters in the Unihan database (CJK ideographs) are to be uniquely romanized and represented. The entropy of this Unihan is 16.10 bits per Han character. For 80-bit or 128-bit, it needs 5 or 8 Han characters respectively. For unique romanization representation, the symbolic encoding of Sijiao Haoma or four corner method (四角号码) [11] [12] searching system is recommended to assist Hanyu Pinyin.

Four digits (Haoma) of Sijiao Haoma are used to describe the symbolic strokes of a Han character. The upper left number will be the first digit. Upper right number will be the second digit. Lower left number will be the third digit. Lastly, lower right number will be the fourth digit. This system is summarized as in Table 5 as a Chinese poem. Table 6 shows the types of strokes associated with the digits of Sijiao Haoma.

Table 5. Chinese poem for easy memorization of Sijiao Haoma.

横一垂二三点捺
叉四插五方框六
七角八八九是小
点下有横变零头

Table 6. Symbolic encoding of Sijiao Haoma for Han characters.

Stroke name 笔名	Digit 号码	Stroke 笔形
Tou 头	0	亠
Heng 横	1	一
Chui 垂	2	丨 丿 亅
Dian 点	3	丶
Cha 叉	4	十 乂
Chuan 串	5	扌 丰
Fang 方	6	口 囗
Jiao 角	7	乛 厂
Ba 八	8	八 人 入
Xiao 小	9	小 忄

The tone mark (声调) of Hanyu Pinyin and attached number (Fu Hao) (附号) of Sijiao Haoma are optional. The tone mark is numbered as 1, 2, 3, 4 and 5 with corresponding to Yin Ping (阴平), Yang Ping (阳平), Shang Sheng (上声), Qu Sheng (去声) and Qing Sheng (轻声). For the hybrid encoding of Hanyu Pinyin and Sijiao Haoma, the forms of (汉) in Table 7 can be adopted to be the example of Han character romanization. Now, minimum entropy of 16.10 bits per Han character is achieved and it is quite efficient for strong password with good memorability. Here, one Han character carries only one secret.

Table 7. Forms of Han character romanization for (汉).

Form	[Hanyu Pinyin] (Tone Mark) [Sijiao Haoma] (Fu Hao)						
Example	han3714	han43714	han37140	han437140	3714han	3714HAN4	3H7A1N4

3 Randomness Improvement for Chinese Language Password

The combination of 415 Hanyu Pinyin syllables and 10000 Sijiao Haoma numbers are more than enough to encode 70000 Han ideographs. In order to increase the randomness or entropy of Han character, one may consider the permutation and capitalization as in Table 7.

3.1 Self-created Signature-like Han Character

To further increase the entropy so as to have less Han character to fulfill the security requirements of key length in bit like 128-bit AES, the creation of new Han character is a must. This situation happens in real life for the individual name in gaining uniqueness. The created Han character is also signature-like. For Han character creation, it may follow the six methods of Liu Shu (六书) [13]. The Liu Shu includes Xiang Xing (象形), Zhi Shi (指事), Hui Yi (会意), Xing Sheng (形声), Jia Jie (假借) and Zhuan Zhu (转注) [14] [15] [16]. An example of created Han character is shown in Fig. 1 by modifying the Han character of (汉) from [han437140] to [han437141] by adding a horizontal stroke between the upper right corner and lower right corner.

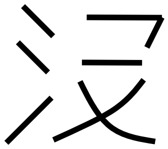

Fig. 1. Example of self-created signature-like Han character by modifying the Han character of (汉) from [Hanyu Pinyin = han4] and [Sijiao Haoma = 37140] to [Hanyu Pinyin = han4] and [Sijiao Haoma = 37141].

Self-created signature-like Han characters enlarge the Unihan key space for Chinese language password to 4,150,000. When tone mark (excluding Qing Sheng) and Fu Hao are included, it becomes 166,000,000 or entropy of 27.31 bits per Han character. This allows the 80-bit or 128-bit symmetric key requirements to demand for 3 or 5 Han characters respectively. The efficiency of Chinese language password is greatly increased.

3.2 Self-created Chinese Language Passphrase

Nevertheless, self-created Han character is not efficient for 256-bit and higher encryption. For this purpose, self-created Chinese language passphrase is needed. A single Han character secret can be made passphrase to satisfy the key size demand.

At least one non-alphanumeric character has to be included together with *capitalization*, *permutation* and *character stuffing*. Character stuffing is like bit stuffing in data communication to enable the syllable length at a fixed value of 6. It is 6 as the maximum syllable length is 6, excluding the tone mark. Adding fixed syllable length, tone mark, Sijiao Haoma, Fu Hao and one non-alphanumeric character together, a string of 13 ASCII characters is obtained as the basic unit for one self-created Chinese language passphrase.

This single Han character Chinese language passphrase will carry two secrets: Han character and non-alphanumeric character. It has better strength and memorability with entropy of 85.41 bits per 13-character string. This type of Chinese language passphrase will need only 1 or 2 Han characters for 80-bit or 128-bit symmetric key respectively. The examples of Chinese language passphrase are shown in Table 8.

Table 8. Forms of self-created Chinese language passphrase for (汉).

Form	No Character Stuffing	With Character Stuffing	Capitalization & Permutation
Example	han4&37140	han4***&37140	37140&HaN4***

3.3 Discussions

For unbreakable encryption, the key size has to be at least the same with message size as in one time password [17]. So far, we use the full Unihan database of 70217 Han ideographs to build the Chinese language password and passphrase. In the Han unification of Unicode, Han ideographs are called as Hanzi in Chinese language, Kanji in Japanese language and Hanja in Korean language.

If only the basic Unihan database of 20924 Han ideographs in the CJK unified ideographs are used, by excluding the CJK unified ideographs extension A and CJK unified ideographs extension B, the entropy will drop from 16.10 to 14.35 bits per Han character. Then more Han characters are required to fulfill the key length requirements. Hence for short, strong and memorable password, self-created signature-like Chinese language password and passphrase are in favourite. The situation of various Chinese key spaces is shown in Table 9.

Table 9. Minimum key lengths for various Chinese key spaces (in Han character).

Database	Key Space	Entropy (bit / Han char.)	Minimum Key Length (in Han character)			
			80-bit	128-bit	192-bit	256-bit
Basic Unihan	20924	14.35	6	9	14	18
Full Unihan	70217	16.10	5	8	12	16
Self-created Han Character	166,000,000	27.31	3	5	8	10
Self-created Cantonese	377,400,000	28.49	3	5	7	9
Self-created Passphrase	95^{13}	85.41	1	2	3	3

4 Applications to Other CJK Languages

Han unification of Unicode builds Han characters database for CJK languages (Chinese, Japanese and Korean). The proposed password and passphrase generation method can be applied to any CJK language using the Han characters by changing the pronunciation romanization system. The symbolic encoding of Sijiao Haoma remains the same for all the Han characters in any CJK language. Below we discuss on its applications for Cantonese language and Japanese language.

4.1 Cantonese Language Password Using Jyutping

Cantonese language is used by a global population of about 80 millions. Being the official language in Hong Kong SAR (Special Administrative Region) and Macau SAR of PRC (People's Republic of China), the regulation works of Cantonese language are done here. It shares majority of the Han characters with Chinese language in Mandarin except those Han characters in the HKSCS (Hong Kong Supplementary Character Set). For HKSCS-2004, it has 4941 Han characters as in year 2004 under ISO 10646 standard. Hence, it is compatible with Unicode which implements the ISO 10646 standard.

There are many Cantonese pronunciation systems. Among them, two systems are romanized and computer friendly. One is standard Cantonese pinyin or HKED (《常用字廣州話讀音表》拼音方案) (「教院式」拼音方案). This is the only pronunciation romanization system accepted by Education and Manpower Bureau of Hong Kong and Hong Kong Examinations and Assessment Authority. Another is jyutping proposed by LSHK (The Linguistic Society of Hong Kong) in year 1993.

Nowadays, regulation works of Cantonese pronunciation for Unicode adopt jyutping system. Han characters in Unicode are matched with jyutping where the lists are downloadable from the URLs of [http://www.iso10646hk.net/jp/index.jsp] and [http://www.info.gov.hk/digital21/eng/structure/jyutping.html].

Table 10. Phonetic encoding of jyutping in Cantonese language.

Initial (20)	nil	b	p	m	f	d	t	n	l
	g	k	ng	h	gw	kw	w	z	c
	s	j							
Final (59)	i	ip	it	ik	im	in	ing		iu
	yu		yut			yun			
	u	up	ut	uk	um	un	ung	ui	
	e	ep	et	ek	em	en	eng	ei	eu
			eot			eon		eoi	
	oe		oet	oek			oeng		
	o		ot	ok		on	ong	oi	ou
		ap	at	ak	am	an	ang	ai	au
	aa	aap	aat	aak	aam	aan	aang	aai	aau

In jyutping system, there are 20 initials (or onsets) and 59 finals as in Table 10. These initials and finals construct about 629 syllables for Cantonese language as compared to 415 syllables for Chinese language in Mandarin. For tone mark, 6 distinct tone contours are used for 9 tones. For completeness, the jyutping has syllables that have no matching Han character. Nevertheless, in the application for Cantonese language password, all jyutping syllables are useful for self-created password and passphrase as in Table 9.

Table 11. Forms of self-created Cantonese language password and passphrase for (汉).

Form	Traditional Chinese (漢)	Simplified Chinese (汉)	With Character Stuffing
Example	hon3&34185	hon3&37140	hon3***&34185

Table 11 shows the examples of Cantonese language password. It is similar to Chinese language password in Mandarin. Sijiao Haoma is exactly encoded. For jyutping, the maximum syllable length is 6 also. Capitalization, permutation and character stuffing in Section 3.2 can be used to generate self-created signature-like Cantonese language password and passphrase. The key space of self-created Han characters in Cantonese can reach 377,400,000 keys or entropy of 28.49 bits per Han characters as in Table 9.

4.2 Japanese Language Password Using Rōmaji

In Japanese language, there are four writing systems: Two syllabaries of Hiragana (平仮名) and katakana (片仮名), one logogram of kanji (漢字), and one romanization of rōmaji (ローマ字). There are a few romanization systems for Japanese language. Here the most widely used Hepburn romanization is adopted for rōmaji. The password generation method for Chinese language password can be used for kanji of the Japanese language via the combination of rōmaji and Sijiao Haoma.

Firstly, obtain the Sijiao Haoma with or without the Fu Hao for the Japanese word in kanji or Han character. Then, the kanji is converted to rōmaji for pronunciation romanization. Character stuffing cannot be used for Japanese language password as the kanji is having variable number of syllables from a minimum of one syllable. For Hepburn romanization, there are about 132 syllables.

Table 12 shows examples of password for kanji. To avoid dictionary attack, one can use self-created kanji as in Section 3.1 to add randomness. Capitalization, permutation and additional non-alphanumeric ASCII character in Section 3.2 can be applied to further enhance the entropy of Japanese language password.

Table 12. Forms of Japanese language password for (大) (だい), (漢) (かん) and (山) (やま).

Form	dai (大) (だい)	kan (漢) (かん)	yama (山) (やま)
Example	dai&40800	kan&34185	yama&22770

5 Conclusions

Through phonetic encoding of Hanyu Pinyin and symbolic encoding of Sijiao Haoma, we can create strong and memorable Chinese language password and passphrase from Unihan database (CJK ideographs) of Unicode. Self-created signature-like Han character and Chinese language passphrase are shorter, stronger and more memorable in term of bit per Han character. In general, it is one secret per Han character.

By changing the pronunciation romanization system, this method can be applied for other CJK languages using the Unihan of Han characters. Japanese language and Cantonese language are the given examples. For the members of Chinese language family, Mandarin and Cantonese languages have pronunciation romanization systems that are computer friendly. For other members like Wu, Min, Jin, Xiang, Hakka, Gan, Hui and Ping languages, a better pronunciation romanization system is needed.

To apply this password generation method, knowing the CJK language, Hanyu Pinyin, jyutping, rōmaji and Sijiao Haoma will be the prerequisites. The proposed method has longer key length in term of ASCII characters, but additional keying time in seconds is tolerable for strong and memorable password.

Acknowledgments. Readers may learn Sijiao Haoma or four corner method from the references like [11] and [12]. There are only few pages of simple rules.

References

1. Schneier, B.: Applied Cryptography. 2nd edn. John Wiley & Sons, New York City, New York, USA (1996)
2. U.S. Department of Defense: Password Management Guideline. DoD Computer Center, Fort George G. Meade, Maryland, USA (1985)
3. Gehrmann, C., Näslund, M. (ed.): ECRYPT Yearly Report on Algorithms and Keysizes. European Network of Excellence in Cryptology (ECRYPT), IST-2002-507932. Katholieke Universiteit Leuven, Leuven-Heverlee, Belgium (2006)
4. Miller, G.A.: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. The Psychological Review 63 (1956) 81-97
5. Gehringer, E.F.: Choosing Passwords: Security and Human Factors. Proceedings of the IEEE International Symposium on Technology and Society (ISTAS 2002), Raleigh, North Carolina, USA (2002) 369-373
6. M-Tech Information Technology: Password Management Best Practices. An Online Guideline, Calgary, Alberta, Canada (2004)
7. Yan, J., Blackwell, A., Anderson, R., Grant, A.: Password Memorability and Security. IEEE Security and Privacy Magazine 2(5) (2004) 25-31
8. Klein, D.V.: "Foiling the Cracker": A Survey of, and Improvements to, Password Security. Proceedings of the USENIX Workshop on UNIX Security, Portland, Oregon, USA (1990) 5-14
9. Popular Book (大众书局): Han Yu Pin Yin Xue Xi Ka (汉语拼音学习卡). Popular Book, Singapore (2003) (in Chinese language)
10. The Unicode Consortium: The Unicode Standard 4.0. Addison-Wesley Professional, Boston, Massachusetts, USA (2003)
11. United Publishing House (联营出版有限公司): Zui Xin Han Yu Da Ci Dian (最新汉语大词典 [修订版]). United Publishing House, Seri Kembangan, Selangor, Malaysia (2001) (in Chinese language)
12. United Publishing House (联营出版有限公司): Xin Han Yu Zi Dian (新汉语字典). United Publishing House, Seri Kembangan, Selangor, Malaysia (2002) (in Chinese language)
13. Huang, X.R. (黄秀如): A History of Dictionaries (词典的两个世界). Net and Books (网路与书), Taipei, Taiwan ROC (2002) (in Chinese language)
14. Xu, S. (许慎): Shuo Wen Jie Zi (说文解字). Chung Hwa Book (中华书局), Hong Kong SAR, China (2001) (in Chinese language)
15. Luo, H.Y. (罗华炎): Jian Ming Han Yu Yu Fa (简明汉语语法). Yakin (雅景), Cheras, Kuala Lumpur, Malaysia (1990) (in Chinese language)
16. Luo, H.Y. (罗华炎): Xian Dai Han Yu Yu Fa (现代汉语语法). Seni Hijau (艺青), Ipoh, Perak, Malaysia (2003) (in Chinese language)
17. Shannon, C.E.: Communication Theory of Secrecy Systems. Bell System Technical Journal 28 (1949) 656-715