Chinese Language Steganography using the Arabic Diacritics as a Covered Media

Ahmed Chalak Shakir
Harbin Institute of Technology,
School of Electronics and
Information Technology,
Heilongjiang, Harbin, China.
University of Kirkuk, college of
Science, Computer Science
Department, Kirkuk, Iraq

Gu Xuemai
Professor
School of Electronics
And Information Technology
Harbin Institute of
Technology, 150001, China

Jia Min
Doctor
School of Electronics
And Information Technology
Harbin Institute of Technology,
150001, China

ABSTRACT

Extensive use of digital media like text, image, audio, and video on the Interne generated a requirement for providing traffic security. For carrying out confidential communication over public networks, it was found that simply concealing the contents of a message using cryptography was not adequate. In this study the new procedure in steganography is developed by using the diacritics-Harakat- of Arabic language as a covered medium to hide the Chinese stroke text. So that in the Arabic language, the diacritics-Harakat- which are used to represent vowel sounds are not useful when writing and sending the documents because the receiver can clearly understand the text without needing the Harakat. The proposed approach uses eight different diacritical symbols in Arabic to hide binary bits in the original cover media. The embedded data are then extracted by reading diacritics from the document and translating them back to binary. Two diacritics are used to hide one Unicode character for strengthen the power of the security. The dictionary of English-Chinese and Chinese-English are stored in both sides. Finally the diacritics are saved in the covered media so that the receiver can use it for getting the original message.

General Terms

Security, steganography and Algorithms.

Keywords

Arabic language diacritics; Stroke in Chinese language; text file format; and Unicode

1. INTRODUCTION

Semitic languages such as Arabic and Hebrew are not as much studied as English for computer speech and language processing. In recent years, Arabic in particular has been receiving tremendous attention. Typically Arabic text is presented without short vowels and other diacritic marks that are placed either above or below the graphemes. The process of adding vowels and other diacritic marks to Arabic text can be called diacritization or vowelization. Vowels help define the sense and the meaning of a word. It also shows how it should be pronounced. However, the use of vowels and other diacritics has lapsed in modern Arabic writing [3].

Security and secrecy of information has always been important to people, organizations and governments [1]. Since ancient times, people and nations seek to keep some information secure. Steganography is the approach of hiding the very existence of secret messages, hence securing them. It has gained much importance today, in the era of communications and computation [2]. The prevalent language for communication on the Internet is English. This may be a result of the Internet's origins, as well as English's role as the lingua franca (A lingua franca is any language widely used beyond the population of its native speakers). The Internet's technologies have developed enough in recent years, especially in the use of **Unicode** that good facilities are available for development and communication in most widely used languages. Today on Internet a hidden exchange of information has been an important issue since old times and the issue of information security has gained special significance [4]. One of the main concerns in this field is the ability to privately exchange information and hides the data of interest throughout the transmission process. As a way to hide the exchange of data, steganography has gained a wide interest among researchers and security specialists. So, it is the science of forming hidden messages such that the intended recipient is the only party aware of the existence of the message. This is usually done by embedding the private data in a cover media without destroying the meaningfulness of this media [6]

2. CHARACTERISTICS OF ARABIC LANGUAGE

The Arabic language, written from right to left, is based on an alphabetical system that uses 28 basic letters. Unlike English, Arabic does not differentiate between upper and lower case or between written and printed letters. Moreover, Arabic language uses different symbols as diacritical marks, or simply diacritics which are also known as Harakat. The main eight diacritic symbols are shown in TABLE I. Other diacritic marks also exist but are outside of the scope of this paper.

Table 1. The eight main Arabic diacritics

Fatha	ـَ	Kasrah	ـِ
Dhammah	ـُ	Sukkon	ـْ
Shaddah	ـّ	Tanween Fath	ـً
Tanween Kasr	ـٍ	Tanween Dham	ـٌ

Just like most of the diacritics based written languages, the main purpose of using diacritics in Arabic languages is to alter the pronunciation of a phoneme or to distinguish between words of similar spelling. Nonetheless, the use of diacritics in the text is optional in written Standard Arabic. In this work we utilize this characteristic to define a stenographic scheme for hiding binary data within Arabic text [2].

3. MATERIAL AND METHOD

There is no alphabetic in Chinese language, instead Chinese script is made up of characters. Different from the alphabetic script which is spelled out of letters, Chinese characters are written in various strokes [5]. For that we cannot save all the Chinese characters which are written in various strokes in the predefined table. So, the messages characters are automatically transferred to capital English language alphabetic depending on the hidden preinstalled dictionary. The new method is defined for hiding information written in Chinese language in the Arabic text file. Having eight different diacritics symbols, for empowering the security; each two diacritics can be used to carry one letter, number, or special character. Each letter or diacritic is 16-bit because of the using of the Unicode. For that, two tables are used:

Table of the diacritics: embedded in the covered media (Arabic text file that contains the hidden information that must be sent). This table is not changed according to the change of the language of the text that we want to hide, it always remains unchanged.

Table 2. Diacritics table

◌ً◌ً	◌ً◌ٌ	◌ً◌ٍ	◌ً◌َ	◌ً◌ُ	◌ً◌ِ	◌ً◌ّ	◌ً◌ْ
◌ٌ◌ً	◌ٌ◌ٌ	◌ٌ◌ٍ	◌ٌ◌َ	◌ٌ◌ُ	◌ٌ◌ِ	◌ٌ◌ّ	◌ٌ◌ْ
◌ٍ◌ً	◌ٍ◌ٌ	◌ٍ◌ٍ	◌ٍ◌َ	◌ٍ◌ُ	◌ٍ◌ِ	◌ٍ◌ّ	◌ٍ◌ْ
◌َ◌ً	◌َ◌ٌ	◌َ◌ٍ	◌َ◌َ	◌َ◌ُ	◌َ◌ِ	◌َ◌ّ	◌َ◌ْ
◌ُ◌ً	◌ُ◌ٌ	◌ُ◌ٍ	◌ُ◌َ	◌ُ◌ُ	◌ُ◌ِ	◌ُ◌ّ	◌ُ◌ْ
◌ِ◌ً	◌ِ◌ٌ	◌َ◌ٍ	◌ِ◌َ	◌ِ◌ُ	◌ِ◌ِ	◌ِ◌ّ	◌ِ◌ْ
◌ّ◌ً	◌ّ◌ٌ	◌ّ◌ٍ	◌ّ◌َ	◌ّ◌ُ	◌ّ◌ِ	◌ّ◌ّ	◌ّ◌ْ
◌ْ◌ً	◌ْ◌ٌ	◌ْ◌ٍ	◌ْ◌َ	◌ْ◌ُ	◌ْ◌ِ	◌ْ◌ّ	◌ْ◌ْ

The table has 64 inputs , 8 different diacritics and we take two diacritics for carrying one element (element = letter, number, or special characters), so that and by tacking each diacritic with the other eight diacritics, the number of states that satisfy this condition is 64 state and each state represents two diacritics, as shown in table 2. This is stored as two- dimensional array.

Table of elements: this table is stored as one dimensional array and contains all the English alphabetic, the number from 0 to 9, and other special characters that are used in the English language, all the information written in Chinese language must be converted automatically to English language then compiled and understood by this table. This table is not send within the text file that contains both the covered media and the hidden information; instead it must be already in the side of the receiver as shown in table 3. Take with the regard that the number of elements in this table is 50, and the other 14 double diacritics are used for special using not for hiding elements. The elements are elective by us and are the most used in Arabic language.

Table 3. Elements table

?	$	6	A	I	Q	Y	
)	!	7	B	J	R	Z	
(	0	8	C	K	S		
,	1	9	D	L	T		
؟	2	Δ	E	M	U		
:	3	}	F	N	V		
.	4	{	G	O	W		
%	5	“	H	P	X		

4. HIDING METHOD

In Chinese language each character is a syllable. A character could be a word or part of a word. The total number of Chinese characters is estimated over fifty thousand [5]. So that it is not powerful and not efficient to save the table that contains fifty thousand characters. For that resin the message that is written in Chinese language would be automatically translated into English language because our work is depends on the alphabetic and the Chinese doesn’t have the alphabetic.

4.1 The Method

- Let X is the array that contains the information in English language that is to be hidden and has n elements. Y is the text file in Arabic language that is used as a covered media to hide the information and has r elements. Hiding X in Y using the diacritics and elements tables takes the following steps:

- X is read to determine the length of it which is n.

- Divide X into two halves. If n is even so the values of two halves are also even. But if the n is odd then value of first half is odd and the value of second half is even.

- Let F (first half) is the X from 1 to n/2 if n is even or from 1 to (n+1/2) if n is odd.

- Let L (second half) is the X from [(n/2) +1 to n] if n is even, or from [F+1 → n] if n is odd. The square brackets means the first and last numbers between these brackets are inside the interval.

(Example: if the length of the information that we want to hide is 51, then F= [1 →26] = 26 and L= [27 → 51] = 25

Noting that, the character (Δ) represents the space bar. Each character in X which is as in table 3 is represented by two diacritics as in table 2 according to their indexes.

For example, H is represented by (◌ْ◌َ), $ is represented by (◌ً◌ٌ), and the procedure is continue until the last character (Z) is represented by (◌ٌ◌ّ) which has the index of 50 (starting from 1). The other diacritics in table 3 are used for special using which are:

◌ٍ◌ّ , ◌َ◌ّ , ◌ُ◌ّ , ◌ِ◌ّ , ◌ّ◌ّ , ◌ْ◌ّ , ◌ً◌ْ , ◌ٌ◌ْ , ◌ٍ◌ْ , ◌َ◌ْ , ◌ُ◌ْ , ◌ِ◌ْ ,

◌ّ◌ْ , ◌ْ◌ْ.

4.2 Hiding F in Y

- The first element in F is putted as a diacritics on the first alphabetic in Y.
- The procedure in step 1 is repeated until F is embedded in Y.
- A special diacritics ,which are used for this purpose only (-◌ُّ◌ْ -) are used after w in F to separate between the diacritics that are used for hiding F and the table of diacritics that is embedded in Y.

4.3 Storing Y in table 2

- The first two-pair diacritics (◌ُّ◌ً-◌ً◌ً) of table 2 are stored after the special diacritics (-◌ُّ◌ْ -).
- For empowering the approach, the table 2 is not stored in Y, in order. Instead of that the diacritics are stored according to the table order then two diacritics which don't mean anything putted among the alphabetic which means 4-2-4-2-4-2-4…..4 diacritics are used for two elements and two diacritics are represent nothing of table 3. This procedure for embedding the table 2 in the Y is continued by putting four diacritics from the table and putting two diacritics between them randomly not refers to anything only for attenuating the intruder, this is repeated until all diacritics in the table 2 except 14 elements that are used for special using are embedded in Y which require 172 characters (50 elements in table 3 and each one requires two diacritics to be represented. So that 50*2=100. Between each four diacritics there is a random two diacritics which are 51. Finally 51+100=151 characters needed for representing the table 2.
- The first element in L is putted as a diacritic on the element directly after the table 2. The operation in this step is continued until all elements in L are embedded in the second half of Y.
- If there are other text in the Y and not used because the text to be hidden is completed, the remaining alphabetic in Y is filled by the diacritics that doesn't have any meaning.

Example:
Let us hide the following message that denoted by X.

哈尔滨工业大学，电子与信息技术学院

Which translated to English as follows "Harbin Institute of Technology, School of electronics and Information Technology"; to the text file denoted by Y and that don't have any diacritics, so the Y that contains or hides X must be send as following:
The length of X with space = 79 characters. (n=79), 79 is odd, then F= 1 to [n+1/2] = [1→40] =40 and L= [41 to 79] = [41→79] = 39

اَّلَّغةّ اَلّصُيّنْيَّةِ مُنِ اّلّلُغاُتْ اُلَعَرَيْقِةً جَدًاِ وَاَلِتُّيَ لُاَتنِفِصَلّ عَنْ اِلاُقَتُصُاِدُ، وُهِيٌّ تَكِتسُبُ جُزِءِاّ لُاّيُسْتِهاِنً بِهُاّ مَنِ قِوِتَهّاٌ عِنُ طِرُيِقِ اّقتِصّاْدّهَاَ . وَاَلُنّمُوَ اِلاَقّتِصَاُدَيِ اّلّسَرّيَعْ وّاْلُحَثُيُثْ فَيُ اُلْصَيُنِ يُرَاْفُقُهُ اُقبْالُ عّاُلْمْيَ عِلَىِ تُعِلَمِ اِلِلَغةٌ اِلُصِينِيِةِ فّيِ اِمْرِيَكّاً وّاُورُوّباِ وَبلِدّاُنِ اّلشّرّقّ اّلْاِقْصْىً. وّمُاًّ مِعّهِدً كَونِفُوْشّيّوِسِ اُلِتُيّ تُنّفِتّحِ سِنُوِيّاّ فُيُ كَلِ اّنُحّاَءُ اِلّعَالْمّ اْلاْ خّيْرِ شْاُهْدَ عْلِىْ اُزْدَيّاْدّ اّلّراّغُبّيَنّ فِيّ تُعّلَمِ اْلِلِغّةِ الِتُيَ لِاِيّمُكِنً اُنْ تُتّاُفِسُهُاُ اَيُةِ لُغُةٌّ اَخَرْىَ فّيَ اِلَعُاُلَمَ مَنِ حَيُثَ اَلِاِقْدِمّيِةِ. حِيُثِ اَنِ تِّعِلُمِّ اَلّْلُغُةّ اُلصُيُنُيَةُ لِاُيُفُتَحَ لْكَ فَقّطِ اَبُوَاَبَ اِلَمُسَتَقْبَلْ وّاَنِمْاُ يْمَنْحِكْ اُلْقَدّرْةّ عّلّيِ اّلُتّجَوّاِلّ فُيّ اَلِافْ اِلّسِنِواُتِ مَنِ عِمِرٌ اِلَحُضْاُرّةُ اِلُصُيُنَيُةِ اُلْعُرَيقْةَ وّحِكَمَتُهَاَ اَلِخَاُلَدَةِ. تْعِدّ لِغةِ هُاِنَ اِلِلِغْةِ اَلصُيْنّيّةٌ اِلُنُمْوَذُجِيُهُ ، حُيَثً اْنً اّلَنِاَطُقَوَنً بِهُاٌ اَكَثَرْ مْنْ مّلْيِاْرُ فْرَدْ. وِحْوُاْلَيّ 95 بْاّلّمّاِئّةُ مّنَ اّلِشّعُبّ اَلصْيِنّي فِيِ اُلبَرِ اِلصُيِنِيُ وْفُيّ تُايُوُاُنَ يُتِكُلُمُوَنَ اْلّلّغَةِ اَلُصَيِنَيِةَ. فُضَلًا عِنْ ذِلّكِ نِجِدُ عِدَدِاِ كِبُيِرًاٌ مْنٌ اُلّمُجِمُوُعُاَتٌ اِلّصُيُنّيَّةْ فَيّ كَلِ جَنُوَب شَرِقَ اُسَيَاْ ، وْتحّدْيِدْاُ فْيَ سْنِغْاُفْوَرّةْ، وّاّنّدنِوّسّيَاّ، وِمّاْلّيَزِيْاِ وّتِاِيِلْنِدَ. مِجِمِوٌعِاَتُ كْبُيّرُةِ مُنُ اُلَصُيِنُيُيُنَ تَعْيَشّ فَيِ اَجِزَاُءَ اَخَرِىَ مُنَ اَلِعْاِلّمِ كِمِاُ فِيَ اِوِرِوٌبِاً وٌاْمُرّيُكِاٌ اُلْشَمُاِلُيُةٌ وَاِلَجنَوُبُيْهِ فِيَ جَزْرِ هَاُوِاّيِ.

5. RETRIVING THE PLAIN TEXT FROM THE COVERED MEDIA-TEXT FILE-

The same transformation is certainly will happen at the receiver side to retrieve Chinese message. To extract the information X from the text file Y, the following steps must be done:

I. All the diacritics should have been taken, starting from first alphabetic in Y, this operation is continued alphabetic by alphabetic till the special diacritics -◌ّ◌ْ- would be reached which means that the diacritics which are coming after this one are the table 2.

II. The operation of extracting the table of diacritics is continued until crossing 151 characters.

III. Now table 2 is gotten from the text file. Also we already have table of elements (table 3), so by making a comparison between the index of table 2 and table3 (in instance: ◌ً◌ُ = د ، ◌ُّ◌ُ = ذ etc..) getting the meaning of each two diacritics as elements.

IV. In step 1 the diacritics that represent F are gotten, by making the matching of diacritics in F and ones in table 2 then because of we know that which two diacritics represent which letter, it is easy now to find the element (more clearly to find F).

V. Repeating the operation in step 1 on the second half of Y to get L.

VI. The operation in Step V is repeated until we reach one of the special diacritics (◌ْ◌ّ) that represent nothing and refers to that all diacritics coming after these two diacritics are not referring to the elements of the plain text because it completed.

5.1 Extraction the original text steps:

- The text would be red to find F, the two diacritics are extracting from the first character until reaching the special diacritics that are used for separating the F from the table which is (◌ُّ◌ْ).
- After these special diacritics the table is saved which is in green color, the procedure of reading and returning the diacritics is continued till overcoming 172 characters.
- The text also is red after the table to find the L. the reading is continued till reaching the special diacritics which is used for separating L from the rest of the Y that don't mean anything because the X is completed.

- The F and L are compacted to get X by make a comparison as follows: the diacritics on the first element in F are compared with the table gotten until find a matching. The index of this matching is now compared with table 3 till find a matching. The result of last matching represents the first element in X.
- The procedure in step 4 is done for all F and L to find and extract X.

6. CONCLUSIONS

The paper proposed a novel algorithm for using the Harakat of Arabic language in steganography of Chinese language which are represented by Unicode. Each diacritic has its own location in the Arabic alphabetic, but in this study the concentration is on the hiding of information in diacritics so that the locations of the diacritics are not so important and sufficient for using in the communication. The table of the elements are in the receiver side and doesn't in the covered media that would be sent. For that reason it is impossible for the intruder to discover the table of elements which is in English (Unicode), regardless of the illegal use of the diacritics location.

7. ACKNOWLEDGMENTS

Authors would like to thank Harbin Institute of Technology, School of Electronics and Information Technology, China, for supporting this work research.

8. REFERENCES

[1] Ahmed Ch. Shakir. “Stego Encrypted Message in any Language for Network Communication Using Quadratic Method”. Journal of Computer Science 6 (3): 320-322,. doi:10.1109/MC.1998.10029, (2010) www.scipub.org

[2] A. Gutub, Y. Elarian, S. Awaideh, and A. Alvi. “Arabic Text Steganography Using Multiple Diacritics”. Computer Engineering Department, King Fahd University of Petroleum & Minerals, Saudi Arabia. https://eprints.kfupm.edu.sa/832/3/C_W.pdf (2008).

[3] Imed Zitouni *, Ruhi Sarikaya, “Arabic diacritic restoration approach basedon maximum entropy models”. IBM T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, United States, doi:10.1016/j.csl.2008.06.001. (2008).

[4] J. Memon, K. Khowaja, and H. Kazi. “Evaluation of Steganography for Urdo /Arabic Text”. Journal of Theoretical and Applied Information Technology. http://www.jatit.org/volumes/fourth_volume_3_2008.php (2008)

[5] L. Zengyin, A. Raymond, and B. Abdulhadi. “ A concise Course of Chinese”.Dar EL-CHIMAL printing publishing distributing, Tripoli, Lebanon- P.O.Box 57. Book, 1’st edition, www.daraalchamal.com,(2009).

[6] M. Aabed, S. Awaideh, A. Elshafei, and A. Gutub. “ Arabic Diacritics Based Steganography”. ICSPC. IEEE International Conference:756-759. doi: 10.1109/.4728429 (2007).