JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 35, 871-886 (2019)
DOI: 10.6688/JISE.201907_35(4).0010

A Homophone-based Chinese Text Steganography Scheme for Chatting Applications*

SHIH-YU HUANG[1] AND PING-SHENG HUANG[2]
[1]*Department of Computer Science and Information Engineering*
[2]*Department of Electronic Engineering*
Ming Chuan University
Taoyuan, 333 Taiwan
E-mail: {syhuang; pshuang}@mail.mcu.edu.tw

Text messages can be used as the cover media for data hiding and a form of camouflage for securing secret messages. After data hiding, embedded secret messages can be correctly recovered by data extraction techniques. This paper presents a novel technique for hiding secret information into Chinese-based text messages used for public chat rooms via the selection of homophones. Using the application of chat rooms, users are allowed to generate and correct typing errors. Plausible variations of homophone selection (typing errors) can be adopted as a codebook for hiding secret data. Experimental results have shown that the proposed approach provides an effective way to embed secret data into chat text messages that is not readily detectable. The study concludes that public chat rooms can give a confidential and secure real-time communication channel using the proposed method.

Keywords: Chinese homophones, data hiding, hidden text, steganography, chat rooms

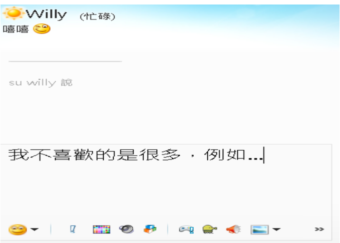
1. INTRODUCTION

Information hiding is developed by hiding secret messages into cover media for secure transmission [1]. The secret messages are camouflaged into the cover media that is indistinguishable to the original copy. Current text steganography schemes are focused on the embedding of text files. Due to the prosperous development of internet technology, the "real-time messages" typed and transmitted in chat rooms become another popular channel for personal communication and information exchange. The reason of using real-time communication systems is that they are more interactive than the email function and the user's daily life will not be interrupted like talking on the phone. Owing to that, current applications of chat rooms such as LINE, Skype, and Google Talk are gradually changing the way of communication between persons. Researches of steganography adopting real-time messages in the chat rooms as a cover media are emerging and rapidly noticed.

According to the definition of linguistic steganography, embedding the secret information into real-time messages belongs to one kind of text steganography. Common approaches of text steganography can be divided into two groups. The first group is based on altering the text format [2-5]. The second group is to change the text content and retain its original meaning at the same time [6-13]. Schemes of the second group are widely employed in embedding the secret information into real-time messages for English applications.

Received June 6, 2018; revised August 2, 2018; accepted October 3, 2018.
Communicated by Berlin Chen.

This paper aims to study text steganography for real-time messages in Chinese chatting rooms. The characteristics of allowing the errata from typing errors, especially for homophone error words, are adopted. In Chinese, homophone words are Chinese characters with same pronunciation but with different meaning. For example, the characters of “坐”, “座”, “做”, and “作” are homophone words. Furthermore, homophone error words are defined as those individual Chinese characters that appear inside Chinese phrases and result in wrong meaning unperceived by readers. In English, the phrase “座位” with two Chinese characters means the noun “seat”. In Chinese, the usage of the wrong character “坐” makes the phrase an inappropriate meaning. However, since those two characters have similar shapes and same pronunciation, the readers is easily confused and treat them the same meaning. In this paper, homophone error words are defined as those Chinese characters appearing in Chinese phrases and making them with wrong meaning. In Chinese chatting rooms, candidate words from Chinese phonetic spelling are needed for users to select for inputting Chinese words. Since this process is tedious, some users tend to be lazy and then carelessly generate a few homophone errors while typing sentences quickly. After those sentences with homophone errors are transmitted to the receiver’s screen, the receiver can still easily understand the meaning of transmitted messages from the content before or after those homophone errors. Fig. 1 shows one example for this situation. The user intends to transmit the message in Chinese: “There are quite a few things (事) I do not like such as …” and this message is wrongly typed as “The things I do not like are (是) many such as …”. The Chinese character “事” is wrongly typed as the character “是” and their speech sounds are the same to each other. Therefore, the receiver can still understand the transmitter’s meaning.



Willy (忙碌)
嘻嘻

su willy 說

我不喜歡的是很多，例如...

Fig. 1. Example of homophone errors in Chinese chatting rooms.

Unlike spelling errors of English words, Chinese homophone errors themselves are still correct words. Like the example in Fig. 1, although “是” is a correct word, “事” should be used instead in this sentence and this error is hard to be detected by any programs. To the best of our knowledge, although there are spelling correction systems that exist during typing Chinese words, there is no correcting system for received Chinese sentences now. In the research literature of Chinese homophone, according to the experimental results from statistical analysis of student composition errors [14, 15], 79.88% of errors belong to homophone words. That is, most of spelling errors in Chinese composition are from homophone errors. Also, Hsieh [16] proposed the error analysis results for Chinese typing input. This paper concludes that a variety of methods are adopted by students for typing Chinese words and the spelling method is the most convenient one to quickly express the student’s meaning. Furthermore, in practice, new types of errors gradually arise after computer spelling input is used. Apart from the original errors from syntax and vocabulary, spelling input errors are also occurred. The test data is a Chinese document used as the homework of typing input in Chinese teaching and there are 33,392 K bytes inside this document. All errors are classified into six types and homophone errors achieve the highest ratio up to 83%. Therefore, according to the aforementioned analysis, homophone errors take most of all errors and they are difficult to be automatically detected by any programs. Therefore, this feature provides the motivation for this paper that it is feasible to use Chinese homophone spelling errors for text steganography in applications of chatting rooms.

The remaining of this paper is organized as follows. Literature survey and related work for text steganography are discussed in Section 2. The proposed schemes of information hiding and extraction are described in Section 3. Experimental results are shown and explained in Section 4. Some conclusions are given in Section 5.

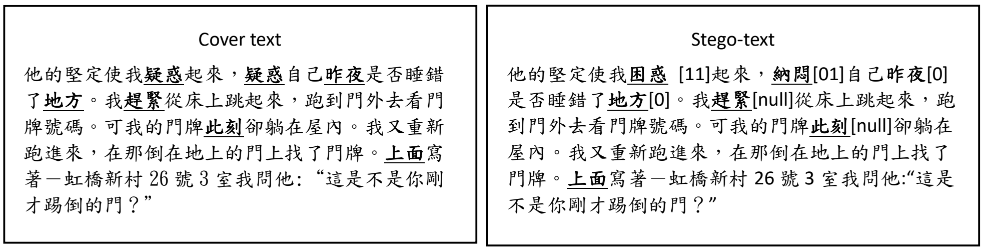
2. RELATED WORK

Using the text for hiding secret message is to adopt the text characteristics in which secret message are embedded. The concept of camouflage is used by text steganography approaches and the main goal is to embed secret message into a cover text. In 2007, a traditional method of text steganography was adopted for secret communication in chat rooms as embedding channels [6]. According to the usual attitude of tending to be lazy in text typing during real-time chatting acronyms and abbreviations are usually used to represent the text meaning. This is adopted for hiding secret message by hiding “0” for abbreviation words and “1” for common words.

In 2009, Liu *et al.* [7] also used the chat rooms as embedding channels and the personal characteristic of generating typing errors during message inputting for information hiding. Misplacing conditions between neighboring alphabets inside a single word are used due to that wrong words always result in an insignificant effect for understanding such as “Guitar” is typed as “Guiatr”. After sorting by using the ASCII code table, the order of alphabets inside each word is obtained and the approach of Matrix coding is further adopted to embed the secret information inside wrong words. However, since English wrong words are adopted for information hiding, this steganography approach is also easily tackled by using English correction software that wrong words are quickly located and corrected.

In Chinese language, synonym words can also be used for embedding secret message [8, 9]. According to one set of fixed synonym words, different synonym words are alternatively changed and used for hiding secret message and specific acronym words are used to hide secret bits of “0” or “1”. Based on this idea [8], another research [9] considers the synonym words appeared nearby and selects specific acronym words for hiding secret message. Take Fig. 2 as an example, the first phrase “疑惑”(doubt) is replaced by “困惑”(confuse) and the second “疑惑”(doubt) is substituted by “納悶”(wonder). When no appropriate synonym words nearby can be found, no secret message will be embedded and the null phrases such as the word “趕緊”(hurry) are inserted.

For those approaches of text steganography mentioned above, all provide the advantages of superb security and fluent text content so that suspicious words are hard to



Cover text

他的堅定使我**疑惑**起來，**疑惑**自己**昨夜**是否睡錯了**地方**。我**趕緊**從床上跳起來，跑到門外去看門牌號碼。可我的門牌**此刻**卻躺在屋內。我又重新跑進來，在那倒在地上的門上找了門牌。**上面**寫著一虹橋新村 26 號 3 室我問他：“這是不是你剛才踢倒的門？”

Stego-text

他的堅定使我**困惑** [11]起來，**納悶**[01]自己**昨夜**[0]是否睡錯了**地方**[0]。我**趕緊**[null]從床上跳起來，跑到門外去看門牌號碼。可我的門牌**此刻**[null]卻躺在屋內。我又重新跑進來，在那倒在地上的門上找了門牌。**上面**寫著一虹橋新村 26 號 3 室我問他:“這是不是你剛才踢倒的門？”

Fig. 2. Examples of Chinese synonym words.

be found. However, the disadvantage is that since the probability of specific synonym words appeared inside one article is low, the hiding capacity of secret message is not high.

In 2010, Chang *et al.* [10] proposed adopting Emotion Icons commonly used in chat rooms for hiding secret information. During the chatting process, users are often adding emotion icons to express their emotion conditions. For example, 🙂 icon and 😠 icon represent the expressions of smile and angry, respectively. In fact, chat rooms have provided enough and detailed emotion icons for different expressions such as 🙂 for smile, 😄 for happy laughing, 😆 for crazy laughing, and 😊 for politely laughing, respectively. In the chatting process, users normally only want to express the status of laughing without caring the degree, any one of 🙂, 😄, 😆, and 😊 can be used. Therefore, this kind of detailed classification is not useful and the motivation of this paper is to adopt this characteristic for information hiding.

With the increasing usage of Internet, a text steganography technique based on HTML documents [11] is proposed in 2011. At first, the secret message is encrypted and then embedded into the HTML Tag and HTML Attribute. Without changing the appearance of HTML documents, the message bits of “0” or “1” are hidden by alternating the orders of Primary Attribute and Secondary Attribute. Since plenty of Tags and Attributes are provided in HTML files, the secret information can be embedded are largely increased.

Based on Genetic algorithms, Mulunda *et al.* [12] presented a method to raise the information amount and security of text steganography in 2013. The embedding order of text steganography is managed by using Genetic algorithms and the information hidden into those positions is not easily detected. Therefore, the security of text steganography can be greatly improved. Furthermore, the cover message is generated from the secret information. Although this is difficult to implement, by combining the cover message with the secret information, the information amount and security level can be both reinforced.

In 2014, an encryption scheme (including HSym, HCod, HNum, and HPhs) and a text hiding technique (consisting of HMea, HAbr, and HEmt) similar to cocktail therapy were proposed by Chandragiri *et al.* [13]. Secret communication can be achieved for online text messaging like Blog information and SMS short text message. The contribution of this method is to integrate the traditional encryption with text hiding schemes.

In 2016, to embed the secret text into the original message, a text steganography algorithm is proposed [17] by dividing each of the secret message alphabet into 4 2-bit pairs. Then those bit pairs are hidden into adequate positions between the bit pairs of the original message. The decryption is done by decoding the position of the secret message in the original message and the secret message is recovered. This is infeasible for on-line chatting applications

Due the rapid progress of deep learning algorithms in recent years, based on a Long Short-Term Memory (LSTM) neural network, a steganographic text generation scheme is proposed [18]. The first step is to arrange token words into bit blocks (shared keys). Then the normal sentence is divided into token words and encoded by referring to shared keys. After training, the LSTM can be used to generate natural texts. This approach has been successfully tested on Twitter and Enron email datasets. However, this is not suitable for Chinese text messages.

3. A HOMOPHONE-BASED CHINESE TEXT STEGANOGRAPHY SCHEME

This section presents a Chinese text steganography scheme based on homophone words. The main purpose of the proposed scheme is to embed the secret message into the cover message used in the chatting rooms for secret communication. The limitation of the proposed approach is that it can not be applied to sentences that already have typos of homophones.

3.1 Main Process of the Proposed Scheme

This paper adopts the characteristic of allowing the errata from typing errors, especially using homophone errors, in chatting rooms. Two modules are included in this proposed system: The first one is the message embedding module in the transmitting end and the second module is the message extraction module in the receiving end, as shown in Fig. 3. One common dictionary of homophone words (DHW) is shared by transmitting and receiving ends for embedding and extracting of secret message.

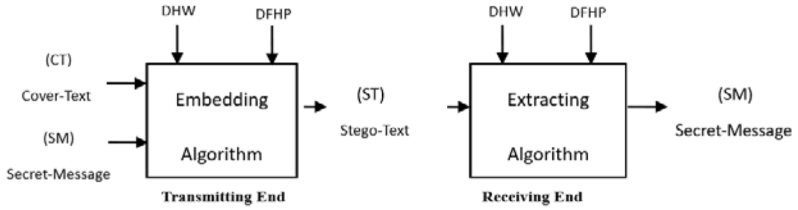


Fig. 3. Embedding and extracting modules for secret message in the proposed system.

In the process of secret communication, the embedding module for secret message will query every Chinese word by referring to the DHW. When the current word is not found in the DHW, this word is treated as stego-text (ST) and transmitted to the receiving end. By contrary, the corresponding homophone word to the secret message founded in the DHW is set to *ST* by the embedding module and transmitted to the receiving end. Assume that the secret message to transmit is ‘010’ and the cover-text (CT) in the transmitting end is “我在市政府中心等您”. Table 1 displays the current DHW. “我” is the first Chinese word in this text that is not appeared in the DHW. Therefore, this word directly becomes the first word in the *ST*. Since the second word “在” in the *CT* appears in the DHW, the embedding process for the secret message is activated. This process will embed different number of bits corresponding to the number of homophone words. This method is similar to the classification of emotion icons mentioned earlier that homophone words are aggregated to one group. Assume that the number of homophone words is N and $n = \lfloor \log_2 N \rfloor$ bits of message can be embedded into this group of homophone words. Take the current ‘ㄗㄞˋ’ as an example, the number $N = 2$ and therefore, $n = 1$. This means the number of bits to embed is one. Therefore, the first bit is extracted from the secret message (SM) by the embedding module and the bit is ‘0’. Then the bit ‘0’ is converted into a decimal number 0. Next, the 0th word is selected from the DHW and the word is “在”. This word becomes the second word of *ST* and is transmitted to the receiving end. Furthermore, since the following three words, “市”, “政”, and “府”, in *CT* are not in the DHW, they are directly become *ST*.

By going further, since the next word in *CT* is “中” that can be found in the DHW, the secret embedding module is activated again. At first, the number N of this group of homophone words is obtained and there are $n = \lfloor \log_2 N \rfloor$ bits of message can be embedded. Now, since $N = 4$ and $n = 2$, two bits can be embedded. Therefore, the second and third bits are extracted from *SM* by the secret embedding module. The extracted message is ‘10’ and the corresponding decimal number is 2. Then the second word “終” is selected from the DHW and transmitted to the receiving end. By following the same procedure, the *CT* can be finally transformed into a new text string *ST* and the result is “我在市政府終心等您”.

Table 1. The current DHW.

spelling	Chinese homophone word	N (number)	n (bits)
ㄗㄞˋ	在，再	2	1
ㄓㄨㄥ	中，衷，終，忠	4	2
ㄗㄨㄛˋ	做，作，座，坐	4	2
ㄧ	一,依,醫,衣,伊,壹,漪,咿	8	3

The secret extracting module works similar to the secret embedding module that every Chinese word is queried by referring to the DHW. If the current word is not found in the DHW, then no secret message is embedded inside this word. By contrary, if the current word appears in the DHW, then the secret extracting module can recover the secret message corresponding to the position of current word in the DHW. Take the same example mentioned above, assume that the message *ST* obtained by the receiving end is “我在市政府終心等您” and Table 4 is the DHW used. “我” is the first Chinese word in *ST* that is not appeared inside the DHW. That is, there is no secret message embedded in this word. On the other hand, since the second word “在” in *ST* appears in the DHW, then the secret extracting procedure is activated. Based on the number of homophone words in the set, the bit length of embedded message is decided by this procedure. That is, the message content can be retrieved by calculating the position of the current homophone word in the DHW. Since the value N is 2 according to the position of “在” in the set of current homophone words set (ㄗㄞˋ), the bit length of embedded message is 1.

Also, owing to that the word “在” appears in position 0 of the homophone word set (ㄗㄞˋ), the content of embedded message is a decimal value of ‘0’. This value is represented by one bit ‘0’ and this is the first secret message. For the following three Chinese words, “市”, “政”, and “府”, in *ST*, since they are not in the DHW, no secret message are embedded. However, the secret extracting module is activated again for the next word “終” that appears in the DHW. According to the query results to this word from the DHW in Table 4, position 2 and $N=4$ are obtained. Therefore, position 2 is decoded by the extracting module as a two-bit secret message ‘10’. For the next three words in *ST*, “心”, “等”, and “您”, no secret message can be extracted because they appear in different homophone dictionaries. After combining all secret messages, the final *SM* is ‘010’ and this proves that the receiving end can successfully obtain the secret message from the transmitting end.

This paper presents a Chinese text steganography scheme based on homophone words. To avoid the conspicuous emergence of errors, this method proposes the usage of frequent homophone phrase (FHP) to enhance the system security. Take the phrase “座位” as an example, this phrase is easily misspelled as “坐位”. In English, the phrase “座位” with two Chinese characters means the noun “seat”. In Chinese, the usage of the wrong character “坐” makes the phrase an inappropriate meaning. However, since those two characters have similar shapes and same pronunciation, the readers is easily confused and treat them the same meaning. In this paper, homophone error words are defined as those Chinese characters appearing in phrases and making them with wrong meaning. If only homophone words are used as the mechanism of embedding secret message, some strange words like “做位” and “作位” will appear in the system and the security is decreased. Therefore, this scheme will adopt a dictionary of frequent homophone phrase (DFHP) to record the corresponding correct and error FHPs. Table 2 displays one part of DFHP. One bit of secret message can be embedded inside every homophone error word and the priority of homophone error words is higher than that of homophone words. When frequent homophone error words appear inside the embedded message, this means the secret message can be embedded. If the current secret message is ‘0’, the correct words will be output as *ST*, by contrary, when the secret message is ‘1’, homophone error words is output to *ST*.

Table 2. Part of DFHP.

Correct	Error
座位	坐位
老闆	老板
忠心耿耿	中心耿耿
甘拜下風	甘敗下風

Table 3. Unicode coding table.

Word	Unicode coded word	Binary word
你	0x4F60	0100111101100000
好	0x597D	0101100101111101
H	0x0048	01001000
I	0x0069	01101001

3.2 The Procedure of Embedding Secret Message

The main purpose of secret communication via chatting rooms is to embed the secret message into one cover chatting message and transmit the stego-message to the receiver. The secret message is encoded by the proposed system using Unicode. Therefore, the information hiding for different languages can be done by the proposed system and another advantage of Unicode is that this is a coding system with fixed length. English words and Chinese words are encoded by 8 bits and 16 bits, respectively. Take the secret message of “Hi你好” as an example, after this message is encoded by Unicode, the obtained code is ‘010010000110100101001111011000000101100101111101’. The corresponding Unicode coding table is listed in Table 3.

The proposed system adopts two carriers for embedding the secret message: First one is the homophone words and second one is the homophone error words. Before transmission, every chatting message has to be segmented and frequent words are extracted from the message. Segmenting system has been previously proposed [19] and a large word database is needed. By considering chatting in real time, since the length of most frequent homophone error words is two, the phrases with two words are used to segment the sentences. Take the sentence of “我在市政府中心等您” for explanation. At first, each word is extracted from this sentence and the sentence becomes “我｜在｜市｜政｜府｜中｜心｜等｜您”. Then each word is matched with those words in the DHW first. Each set of successive two words will be combined as a phrase and then matched with those phrases in the DFHP. The unmatched words belong to one single word. Therefore, the above sentence is converted as “我｜在｜市｜政｜府｜中心｜等｜您”.

For further usage, we call each unit of those segmented phrases as a ‘Token’ and every token is processed sequentially by the following embedding procedure. At first, each token is searched in the DFHP. If this token is found, this means secret message can be embedded inside this token. Then this token or its corresponding phrase is outputted to *ST* when the current secret message is ‘0’ or ‘1’, respectively. After that, the procedure will continue to process the next token. When the following token is not in the DFHP, this token will be searched in the DHW. If this token is found, this means secret message can be embedded inside this token and the available length of embedded message is decided by the number of words inside the homophone set. Assume that N is the number of homophone words for this token and the available length of secret message can be embedded inside this token is $n = \lfloor \log_2 N \rfloor$ bits. After the value n to this token is obtained, n bits of message are retrieved from *SM* and those bits are converted into a decimal value d in which $0 \leq d < N$. Then the dth homophone word corresponding to the token in the set of homophone words is outputted to *ST* and the next token is processed. If this token is absent either in the DFHP or in the DHW, this token is directly outputted to *ST* and the next token is processed. The embedding procedure is continuously performed until all tokens are processed. Fig. 4 displays the flowchart for the embedding procedure in which T represents the current token, T' is the phrase corresponding to T, T'' is the matched homophone word to T, and the first bit of *SM* is s.

3.3 The Procedure of Extracting Secret Message

After *ST* is obtained by the receiving end, the extracting procedure for secret message will be activated. Intrinsically, this procedure is reverse to the embedding procedure for secret message. At first, *ST* is divided into a sequence of tokens using the segmentation mechanism mentioned before. After that, each token will be searched in the DFHP. When this token is found in the dictionary, this means there is secret message embedded. The embedded message of ‘0’ or ‘1’ is decided by the token position appearing in the DFHP. When this token is absent in the DFHP, the extracting procedure will search this token inside the DHW instead. If this token appears in the DHW, this means there is secret message embedded and the length of embedded message is determined by the number of homophone words in the current set. Assume that this number is N and the length of secret message embedded inside the token is $n = \lfloor \log_2 N \rfloor$ bits. Furthermore, the message content is decided by the token position appearing in the homophone word set. For example, if this token appears in the dth position of homophone word set, the decimal number d is converted to an n-bit binary number by the extracting procedure and this binary number is the secret message. Apart from those two cases described above, there is no message embedded inside this token when this token is absent in both the dictionaries of DFHP and DHW. Finally, the extracting procedure will collect all binary numbers and decode them by Unicode. Therefore, the original message can be recovered.

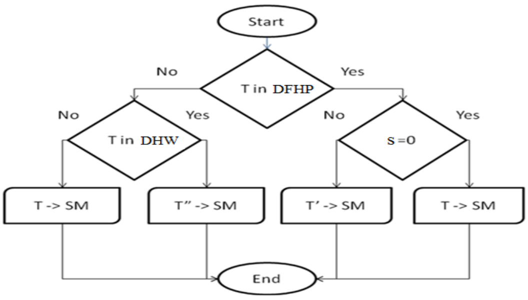


Fig. 4. Flowchart of embedding procedure.

3.4 Dictionary Design

The DHW is the core of secret communication. To reduce the problem of peculiarity resulted by homophone error words in the chatting process, the DFHP is used to increase the system security. The detailed design for the DHW and the DFHP are individually described in the following paragraph.

According to the research results [20], there are around 19486 Chinese words and only 5401 words are frequently used. However, large memory storage space is still needed to retain those words. Therefore, we adopt the information of frequency and percentage for frequently used words provided by《常用國字標準字體表》(Frequently Used Standard Chinese Word Table) [21, 22] to extract most frequently used words from 5401 words and use them as the basic elements in this paper. Those words are classified first and homophone words are grouped into individual sets, such as the homophone word set of "ㄕˋ" includes "是事市使式示視世…". For the case of only one word inside a certain homophone word set, such as only the word "骯" appears in the homophone word set of phoneme "ㄤ" with 1080 words, no message can be embedded into this word set that will not be processed. Table 4 lists a part of three homophone word sets.

Moreover, those frequently appeared words extracted from 5401 words need to be further analyzed. Table 5 demonstrates the distribution of homophone word sets with different extraction ratio. When the top 10% of those 5401 words are extracted as basic elements in the DHW, after classification, there are 93, 15, and 1 homophone word sets that can be used to hide 1, 2, and 3 bits of secret message, respectively. On the other hand, when the top 20% of 5401 words are used as the basic elements of homophone words in the dictionary, there are 187, 50, and 10 word sets that can be used to embed 1, 2, and 3 bits of secret message, respectively. Hence, 247 sets of homophone words can be obtained from the top 20% and the probability of word appearance is around 62% by accumulating the frequency of each word. Thus, the top 20% of 5401 words are extracted for the DHW by the proposed system and used as basic elements. Also, those sets of homophone words with high frequency of appearance are arranged in the front part of DHW that will be adopted for embedding secret message first.

In general, after the words in *CT* are founded inside the DHW, the embedding procedure for secret message will be activated and the secret message can be hidden. At the same time, homophone error words can also be outputted to *ST*. However, too many homophone error words are suspicious that will reduce the security of the proposed system. Therefore, this paper presents a dynamic dictionary design to increase the system safety.

Table 4. Part of three homophone word sets.

ㄕˋ	是事市使式示視世
ㄗㄨㄛˋ	作做座坐
ㄗㄞˋ	在再

Table 5. Ratio distribution of homophone words.

	n=1 (bit)	*n*=2 (bits)	*n*=3 (bits)	Total (sets)	Frequency (Appearance)
10%	93	15	1	109	42.1%
20%	187	50	10	247	62.2%
30%	250	101	16	367	72.5%
40%	313	155	27	495	79.4%
50%	344	197	48	589	83.7%

To lower the appearing frequency of homophone error words, not all words but only a fixed ratio of the DHW can be used by the secret embedding procedure. Assume that $R1$ represents the fixed ratio of homophone word sets in which $0 \leq R1 \leq 1$. When $R1 = 0.1$, this means 10% of the DHW can be used and the top 10% of the DHW will be preferentially used by the proposed system.

The procedure mentioned above can avoid the dense emergence of error words from the embedding of secret message. However, homophone error words will still appear inside some fixed homophone word sets. Therefore, after the message is embedded into a certain homophone word set, this set is only used again after waiting for a period of time. To achieve this goal, each homophone word set will have a time stamp with a valid initial value. This word set can be used only when the time stamp is valid and the time stamp is set to invalid after this word set is used. The invalid time stamp will be recovered to valid after a given time. Note that to maintain a fixed $R1$ value, the next usable word set can only be added after the time stamp of a previous homophone word set is invalid. Similarly, the previous usable homophone word set in the dictionary has to be deleted after the time stamp of a word set is set from invalid to valid.

By adjusting the time stamp for the usage of usable homophone word sets, the word sets used by the proposed system can be dynamically changed. However, since the position of each homophone word in the word set is fixed, only fixed secret message can be generated. To tackle this problem, we propose a scheme to dynamically change the position of each homophone word in the word set. This scheme is to swap the homophone word adopted for embedding secret message with the next word in the word set. Table 6 shows an example for this scheme. Assume that the word "他" is selected to embed the message and the secret message "00" is embedded. Then this homophone word set is modified and the positions of “他” and the next word “她” are swapped. Table 6 (b) displays the homophone word set after swapping. When the word “他” is selected again, the embedded message becomes ‘01’.

Table 6. Part of the DHW.

	00	01	10	11
ㄊㄚ	他	她	它	牠

(a) Before swap

	00	01	10	11
ㄊㄚ	她	他	它	牠

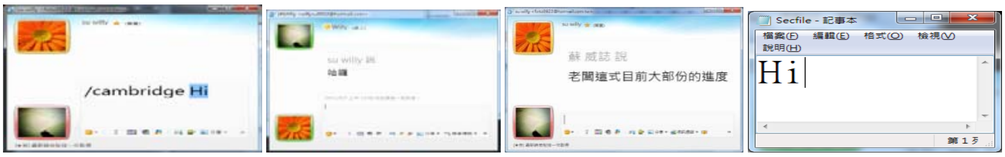
(b) After swap

This paper is based on the FHPs used in the internet [23] and the FHPs with two Chinese words are included. The design concept for the DFHP is similar to that of homophone words. Like the DHW, only part of the DFHP is employed and the time stamp is also used to avoid the emergence of frequent and duplicated FHPs. Meanwhile, the swapping scheme is further adopted to embed dynamic secret message for the same homophone words.

4. EXPERIMENTAL RESULTS

The experiments are conducted on the JavaScript platform of a personal computer. Experimental datasets are selected randomly from Chinese newspapers. It is hard to design a suitable measure to evaluate the message content. Therefore, only three qualitative measures, “OK”, “Strange”, and “Very Strange” are designed. An example is shown to explain the setup steps of the proposed approach.

Fig. 5 (a) demonstrates that user A starts a screen for secret communication in which “/Cambridge” is the Request instruction used by the system to initiate the secret communication and “Hi” is the secret message. Fig. 5 (b) displays the screen of secret communication responded by user B in which “哈囉” is the Acknowledge instruction for replying the secret communication. Fig. 5 (c) reveals the part of chatting content between user A and user B in which ‘0’ and ‘1’ are individually embedded into “老闆” and “部份”, respectively. This is the result of using frequently used homophone error words. Furthermore, Three bits of ‘100’ and 2 bits of ‘00’ are individually embedded into “式” and “的”, respectively. This is the result of using the DHW. After all secret message have been transmitted, user B will display the received secret message encoded by Unicode, as shown in Fig. 5 (d). This is the operating process of this experiment by the proposed approach. The final experimental results are demonstrated in Fig. 6 showing the *CT* and *ST* of chatting message for the example mentioned above. The content of 16 bits of secret message can be transmitted with the stego-message and decoded by the receiver to obtain the secret message “Hi”.



/cambridge Hi

su willy 說
哈囉

蘇威誌 說
老闆這式目前大部份的進度

Hi

(a) (b) (c) (d)

Fig. 5. The process of transmitting chatting message.

老闆這是目前大部份的進度，最近看
了許多論文，不知道實質上有無益
處，近日我會在努力加強學習的。

(a) CT

老闆這式目前大部份的進度，最近看
了許多論文，不知道時質上有無易
處，近日我會在努力加強學習地。

(b) ST

Fig. 6. The content of chatting message.

The system performance affected by the percentage $R1$ used by the DHW is explained in this paragraph. In this experiment, the DFHP is not activated in the beginning, that is, $R2 = 0$. Note that the characteristic of homophone error words allowed in the chatting rooms is adopted by the proposed system to embed secret message. However, since too many homophone error words will result in misunderstanding and doubts to the users for the transmitted message. In this experiment, a measuring scheme to the users for the degree of homophone error word allowance is designed. Also, in the process of secret communication, a third party will monitor and evaluate all message content in the chatting process. When the chatting message is under a normal condition, the result of "OK" is given and the score of "Strange" is graded when the message is a bit strange. Furthermore, when the message content is very weird, the grade of "Very Strange" is provided. This scheme is adopted by the propose system to verify the security of the proposed method. Table 7 lists the experimental results of using different $R1$ values in which the length of transmitted message is around 256 bits and 15 experiments (each experiment corresponds to one person) are conducted for each $R1$ value. When $R1 \leq 0.02$, the message content is normal felt by all monitoring users and only one user will feel strange when the $R1$ value is between 0.05 and 0.1. Furthermore, when the $R1$ value is greater than 0.1, several monitoring users will feel strange or even very strange. Therefore, the setting of $R1 = 0.02$ is an ideal value and there are around 3% of typo errors (Typo rate).

Table 7. The performance analysis of system security to the percentage of *R*1 used for the DHW.

	$R1$=0.01	$R1$=0.02	$R1$=0.03	$R1$=0.04	$R1$=0.05	$R1$=0.1	$R1$=0.2	$R1$=0.5	$R1$=0.8
OK	15	15	14	14	14	14	12	10	7
Strange	0	0	1	1	1	1	3	2	3
Very Strange	0	0	0	0	0	0	0	3	5
Typos rate	4.3%	5.6%	6.9%	7.5%	12.6%	16.9%	23.2%	31.5%	34.5%

In the following paragraph, the embedding capacity of the proposed system is investigated. The capacity is measured by the length of transmission time for the secret message transmitted each time. The experiment adopts $R1 = 0.02$ for the usage of the DHW and six different kinds of Secret Message(SM) and corresponding Cover Text (CT) are used for performance evaluation. Table 8 displays the experimental results when the embedding rate (ER) is defined in *CT* and, in average, 0.18 bits of *SM* can be embedded into one Chinese word. Experimental results show that the value ER is around 18%, that is, around 1.8 bits of *SM* can be hidden into 10 words of *CT*. By considering the typing speed of normal people in chatting conditions and, take the example of 20 words per minute, 108 bits of secret message can be transmitted in 30 minutes and this is around 6 Chinese words.

Table 8. The performance analysis of system capacity to the percentage of *R*1 used for the DHW.

Transcript	SM (bits)	CT (Chinese Words)	Embedding rate
1	213	1579	13.490%
2	582	3680	15.815%
3	831	4505	18.446%
4	1074	5682	18.901%
5	1080	6377	16.935%
6	1413	7482	18.885%
Total	5193	29305	17.720%

Table 9. Performance analysis of system security to the percentage of *R*2 used for the DFHP.

	*R*2=0.2	*R*2=0.4	*R*2=0.6	*R*2=0.8	*R*2=1.0
OK	15	15	15	15	15
Strange	0	0	0	0	0
Very Strange	0	0	0	0	0
Typos rate	0.04%	0.9%	0.15%	0.21%	0.24%

In this paragraph, the effect to the system performance caused by the opening percentage *R*2 of the DFHP usage is discussed. In this experiment, the opening percentage *R*1 is set to 0 and Table 9 demonstrates the performance analysis of system security for different *R*2 values. Experimental results display that no strange conditions in chatting message felt by all monitoring *R*2 users. Table 10 shows the performance analysis of system capacity when $R2 = 1.0$. Experimental results reveal that the ER value is around 0.48%, that is, 1000 words are needed in *CT* to embed 4.8 bits of SM. When there are 6 Chinese words (96 bits) needed to transmit, the transmission time is 1000 minutes by considering the typing speed of 20 words per minute. This proves that the proposed system can have high security level when only using the DFHP. However, since there are few capacities provided by the system, even a few secret messages still need a long time to transmit. Especially, the transmission time takes even longer when the *R*2 value is less than 1.0.

Now the system performance analysis is investigated when the DHW and the DFHP are simultaneously used. In this experiment, different combinations using various *R*1 and *R*2 values are adopted and the *R*1 value is set between 0.02 to 0.2. Furthermore, two *R*2 values are used: $R2 = 0.0$ and $R2 = 1.0$. This means that the DFHP are fully closed or opened. Experimental results are demonstrated in Table 11 which shows that those two *R*2 values (0.0 and 1.0) result in a limited influence for the performance of system security and system capacity. When the *R*1 value is less than 0.1, almost the same performance is provided to system security and system capacity by those two *R*2 values. However, the performance of system security is slightly increased and reduced for that of system capacity. The results are caused by that the homophone error words rarely appear in the process of normal chatting. Therefore, the ideal setting for *R*1 and *R*2 values are 0.02 and 0.0, respectively. This setting will disable the usage of the DFHP. Furthermore, superior system performance for security and capacity can be provided and the system computation time can be reduced.

Table 10. Performance analysis of system capacity to the percentage of *R*2 used for the DFHP.

Transcript	SM (bits)	CT (Chinese words)	Embedding rate
1	6	1579	0.380%
2	16	3680	0.434%
3	15	4505	0.333%
4	15	5682	0.264%
5	41	6377	0.643%
6	50	7482	0.668%
Total	143	29305	0.487%

Table 11. Performance analysis of system security and system capacity affected by the DHW and the DFHP.

	*R*1=0.02 *R*2=0.0	*R*1=0.02 *R*2=1.0	*R*1=0.05 *R*2=0.0	*R*1=0.05 *R*2=1.0	*R*1=0.1 *R*2=0.0	*R*1=0.1 *R*2=1.0	*R*1=0.2 *R*2=0.0	*R*1=0.2 *R*2=1.0
OK	15	15	14	14	14	14	12	13
Strange	0	0	1	1	1	1	3	2
Very Strange	0	0	0	0	0	0	0	0
Typos rate	5.6%	5.8%	12.6%	12.8%	16.9%	17.1%	23.2%	23.3%
Embedding rate	17.72%	18.07%	33.43%	33.72%	43.88%	44.01%	57.17%	57.03%

5. CONCLUSIONS

In the applications of chatting rooms, this paper presents an approach of text steganography based on Chinese homophone words aiming at hiding the secret message into the cover message of chatting rooms. Two characteristics of homophone error words and frequently homophone phrases are adopted to design the corresponding DHW and the DFHP. Those two dictionaries are used for text encoding and embedding the secret message. To lower the appearing frequency of homophone error words and avoid the continuous usage of homophone word sets, this paper has also proposed a scheme for dynamically adjusting the DHW and the DFHP. Experimental results have shown that superior performance of system security and system capacity can be achieved when only 2% of the DHW is dynamically used. The proposed approach is currently a fragile secret embedding scheme. How to enhance its robustness can be considered in the future. Also, the relationship of *SM* expectation and word frequency can be studied.

REFERENCES

1. F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, “Information hiding – a survey,” in *Proceedings of the IEEE*, Vol. 87, 1999, pp. 1062-1078.
2. S. H. Low, N. F. Maxemchuk, J. T. Brassil, and L. O’Gorman, “Document marking and identification using both line and word shifting,” in *Proceedings of the 14th Annual Joint Conference of IEEE Computer and Communications Societies*, Vol. 2-6, 1995, pp. 853-860.

3. D. Huang and H. Yan, "Inter-word distance changes represented by sine waves for watermarking text images," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 11, 2001, pp. 1237-1245.
4. Y. Kim, K. Moon, and I. Oh, "A text watermarking algorithm based on word classification and inter-word space statistics," in *Proceedings of the 7th International Conference on Document Analysis and Recognition*, 2003, pp. 775-779.
5. D. Huang and H. Yan, "Inter-word distance changes represented by sine waves for watermarking text images," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 11, 2001, pp. 1237-1245.
6. M. H. Shirali-Shahreza and M. Shirali-Shahreza, "Text steganography in chat," in *Proceedings of the 3rd IEEE/IFIP International Conference in Central Asia on Internet*, 2007, pp. 1-5.
7. M. Liu, Y. Guo, and L. Zhou, "Text steganography based on online chat," *Signal Processing: Intelligent Information Hiding and Multimedia*, 2009, pp. 807-810.
8. L. Yuling, S. Xingming, G. Can, and W. Hong, "An efficient linguistic steganography for Chinese text," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2007, pp. 2094-2097.
9. X. Zheng, L. Huang, Z. Chen, Z. Yu, and W. Yang, "Hiding information by context-based synonym substitution," *Information Hiding*, LNCS, Vol. 5703, 2009, pp. 162-169.
10. Z. H. Wang, C. C. Chang, T. D. Kieu, and M. C. Li, "Emoticon-based text steganography in chat," in *Proceedings of the Asia-Pacific Conference on Computational Intelligence and Industrial Applications*, Vol. 2, 2010, pp. 457-460.
11. M. Garg, "A novel text steganography technique based on html documents," *International Journal of Advanced Science and Technology*, Vol. 35, 2011, pp. 129-138.
12. C. K. Mulunda, P. W. Wagacha, and A. O. Adede, "Genetic algorithm based model in text steganography," *The African Journal of Information Systems*, Vol. 5, 2013, pp. 131-144.
13. A. Chandragiri, P. A. Cooper, Y. Liu, and Q. Liu, "Implementing secure communication on short text messaging," in *Proceedings of the 2nd International Symposium on Digital Forensics and Security*, 2014, pp. 77-80.
14. C. L. Liu, K. W. Tien, M. H. Lai, Y. H. Chuang, and S. H. Wu, "Phonological and logographic influences on errors in written Chinese words," in *Proceedings of the 7th Workshop on Asian Language Resources*, 2009, pp. 84-91.
15. C. L. Liu, K. W. Tien, M. H. Lai, Y. H. Chuang, and Shih-Hung Wu, "Capturing errors in written Chinese words," in *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, 2009, pp. 25-28.
16. T.-W. Hsieh (謝天蔚), "Error analysis for input spelling in Chinese teaching (中文教學中拼音輸入錯誤分析)," *Chinese Character Teaching and Computer Technology*, Publication supported by a grant from the U.S. Department of Education, 2005.
17. S. S. Iyer and K. Lakhtaria, "New robust and secure alphabet pairing text steganography algorithm," *International Journal of Current Trends in Engineering and Research*, Vol. 2, 2016, pp. 15-21.
18. T. Fang, M. Jaggi, and K. Argyraki, "Generating steganographic text with LSTMs," in *Proceedings of ACL Student Research Workshop*, 2017, p. 17-3,

19. H. Wang, X. Sun, Y. Liu, and Y. Liu, "Natural language watermarking using Chinese syntactic transformations," *Information Technology Journal*, Vol. 7, 2008, pp. 904-910.
20. "注音與國字對照表," http://www.geocities.ws/picmemory/phonic_word.txt.
21. 教育部, "常用國字標準字體表," http://140.111.34.54/files/site_content/M0001/87news/page2-1.htm?open, 1998.
22. 國語推行委員會, "八十七年常用語詞調查報告書," *National Languages Committee*, Taiwan, 1998.
23. 中文期刊部統一用字表, "容易弄錯的詞語," http://mandarin.nccu.edu.tw/data/cuniw.pdf.

Shih-Yu Huang (黃世育) received the B.S. degree in Information Engineering from Tatung Institute of Technology, Taipei Taiwan, Republic of China, in 1988, and the M.S. and Ph.D. degrees from Department of Computer Sciences, National Tsing Hua University Taiwan, in 1990 and 1995, respectively. From 1995 to 1999, he worked in the Telecommunication Laboratories of Chunghwa Telecom Co., Ltd., Taiwan. In October 1999, he joined the Department of Computer Science and Information Engineering, Ming Chuan University, Taiwan. His current research interests are video processing and steganography.

Ping-Sheng Huang (黃炳森) received his BSEE degree from Chung Cheng Institute of Technology, Taiwan, in 1985, the M.S. degree in Computer Science from University of Southern California, United States, in 1990, and the Ph.D. degree in Electronics and Computer Science from University of Southampton, United Kingdom, in 1999. After that, he was with the Department of Electrical Engineering, Chung Cheng Institute of Technology as an Associate Professor until July 2005, then a Professor and Department Head until July 2007. From August 2007, he started to serve as a Professor in the Department of Electronic Engineering, Ming Chuan University, Taiwan until now. His research interests include biometric identification, information hiding, pattern recognition and multimedia applications using FPGA.