International Journal of Scientific Research in Computer Science and Engineering
Research Paper Vol-2, Issue-1 E-ISSN: 2320-7639

Encoding of Hindi Text Using Steganography Technique

Rajesh Shah[1*], Yashwant Singh Chouhan[2]

[1*,2]*Christian eminent college, Indore (M. P.)-India*

Available online at www.isroset.org

Received: 12 Jan 2013 Revised: 22 Jan 2014 Accepted: 10 Feb 2014 Published: 28 Feb 2014

Abstract- Establishing hidden communication and conveying information secretly has been of interest since long past ago. One of the method introduce for establishing hidden communication is Steganography. Methods of Steganography have been mostly applied on image, audio, videos and text files while the major characteristics of these methods are to change in the structure and features so as not to be identifiable by human users. Text document are the best example for this. In this paper we produce a new approach for Hindi text Steganography, which uses letter and its diacritics and numerical code. This method is very useful in Hindi Text and in all other similar Indian Languages.

Keyword - Hindi Text, Cryptography, Steganography, Text Steganography, Text Watermarking, feature coding

I. INTRODUCTION

Information hiding is a general term encompassing many sub disciplines. One of the most important sub disciplines is Steganography [1], [8]. Steganography, is derived from a work by Johannes Trithemus (1462-1516) entitled "Steganographia" and comes from the Greek (óôåãáíü-ò, ãñáö-åéí) defined as "covered writing" [2].It is not enough to simply encipher the traffic, as criminals detect, and react to, the presence of encrypted communications [9].

It is an ancient art of hiding information in ways a message is hidden in an innocent-looking cover media so that will not arouse an eavesdropper's suspicion. Steganography differs from cryptography in the sense that where cryptography focuses on keeping the contents of a message secret, steganography focuses on keeping the existence of a message secret [2,3].

Fig. 1. Types of Steganography

Steganography gained importance because the US and the British government, after the advent of 9/11, banned the use of cryptography and publishing sector wanted to hide copyright marks [9]. Steganography works have been carried out on different media like images, video clips, text, music and sound [5],[18].Among them image steganography is the most popular of the lot. In this method the secret message is embedded into an image as noise to it, which is nearly impossible to differentiate by human eyes [4], [18], [9]. In video steganography, same method may be used to embed a message [5], [14]. Audio steganography embeds the message into a cover audio file as noise at a frequency out of human hearing range [4].

Most difficult kind of steganography is text steganography or linguistic steganography because due to the lack of redundant information in a text compared to an image or audio.

The text Steganography is a method of using written natural language to conceal a secret message as defined by Chapman et al. [18]. Some Image steganographic algorithm with high security features has been presented in [1],[2],[3],[4],[18].

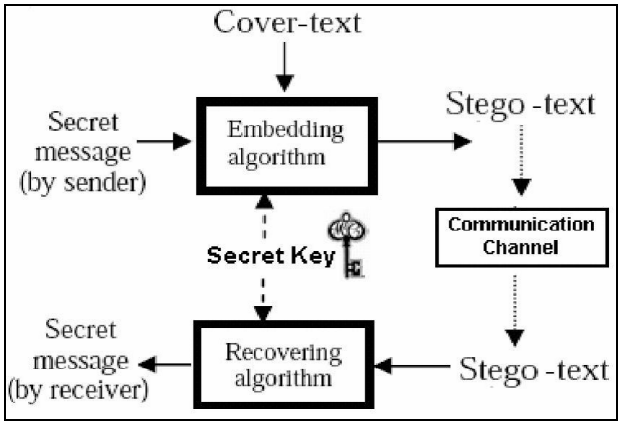

Figure 2: The Mechanism of Text Steganography

Firstly, a secret message (or an embedded data) will be concealed in a cover-text by applying an embedding Algorithm to produce a stego-text. The stego-text will then be transmitted by a communication channel, e.g. Internet or

Corresponding Author: Rajesh Shah

Mobile device to a receiver. For recovering the secret which sent by the sender, the receiver needs to use a recovering algorithm which is parameterized by a stego-key to extract the Novel Text Steganography through Special Code Generation Secret message. A stego-key is used to control the hiding process so as to restrict detection and/or recovery of the Embedded data to parties who know it [10], [18].

II. TEXT STEGANOGRAPHY

Text steganography can be classified in three basic categories [18] -Format-based, Random and statistical and Linguistic.

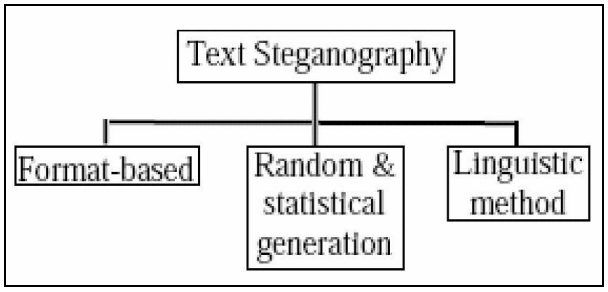

Fig. 3: Three broad categories of text steganography

Format-based: Format-based methods use and change the formatting of the cover-text to hide data. They do not change any word or Sentence, so it does not harm the 'value' of the cover-text. A format-based text steganography method is open space method [6]. In this method extra white spaces are added into the text to hide information. These white spaces can be added after. End of each word, sentence or paragraph. A single space is interpreted as"0" and two consecutive spaces are interpreted as "1". Although a little amount of data can be hidden in a document, this method can be applied to almost all kinds of text without revealing the existence of the hidden data. Another two format-based methods are word shifting and line Shifting. In word shifting method, the horizontal alignments of some words are shifted by changing distances between Words to embed information [7]. These changes are hard to interpret because varying distances between words are very Common in documents. Another method of hiding information in manipulation of white spaces between words and paragraph [8].In line shifting method, vertical alignments of some lines of the text are shifted to create a unique hidden shape to embed a message in it [18].

Random and statistical generation methods: Random and statistical generation methods are used to generate cover-text automatically according to the statistical properties of language. These methods use example grammars to produce cover-text in a certain natural language. A probabilistic context-free grammar (PCFG) is a commonly used language model where each transformation rule of a context free grammar has a probability associated with it [7]. APCFG can be used to generate word sequences by starting with the root node and recursively applying randomly chosen rules. The sentences are constructed according to the secret message to be hidden in it. The quality of the generated stego-message depends directly on the quality of the grammars used. Another approach to this type of method is to generate words having same statistical properties like word length and letter frequency of a word in the original message. The words generated are often without of any lexical value.

Linguistic method:
The linguistic method [18] considers the linguistic properties of the text to modify it. The method uses linguistic Structure of the message as a place to hide information. Syntactic method is a linguistic steganography method where some punctuation signs like comma (,) and full-stop (.) are placed in proper places in the document to embed a data. This Method needs proper identification of places where the sign scan is inserted. Another linguistic steganography method Semantic method. In this method the synonym of words for some pre-selected are used. The words are replaced by their Synonyms to hide information in it.

Existing Approaches: In this sub-section, we present some of the popular approaches of text steganography.

Line Shift: In this method, secret message is hidden by vertically shifting the text lines to some degree [11, 12]. A line marked has two unmarked control lines one on either side of it for detecting the direction of movement of the marked line [11]. To hide bit 0, a line is shifted up and to hide bit 1, the line is shifted down [13]. Determination of whether the line has been shifted up or down is done by measuring the distance of the centroid of marked line and its control lines [11]. If the text is retyped or if a character recognition program (OCR) is used, the hidden information would get destroyed. Also, the distances can be observed by using special instruments of distance assessment [11].

Word Shift: In this method, secret message is hidden by shifting the words horizontally, i.e. left or right to represent bit 0 or 1 respectively [13]. Words shift are detected using correlation method that treats a profile as a waveform and decides whether it originated from a waveform whose middle block has been shifted left or right [12]. This method can be identified less, because change of Distance between words to fill a line is quite common [11, 12]. But if someone knows the algorithm of distances, he can compare the stego text with the algorithm and obtain the hidden content by using the difference. Also, retyping or using OCR programs destroys the hidden information [11, 12].

Syntactic Method: This technique uses punctuation marks such as full stop (.), comma (,), etc. to hide bits 0 and 1. But problem with this method is that it requires identification of correct places to insert punctuation marks [11, 12]. Therefore, care should be taken in using this method as readers can notice improper use of the punctuations [10].

White Steg: This technique uses white spaces for hiding a secret message. There are three methods of hiding data using white spaces. In Inter Sentence Spacing, we place single space to hide bit 0 and two spaces to hide bit 1 at the end of each terminating character [10]. In End of Line Spaces, fixed number of spaces is inserted at the end of each line. For example, two spaces to encode one bit per line, four spaces to encode two bits and so on. In Inter Word Spacing technique, one space after a word represents bit 0 and two spaces after a word represents bit 1. But, inconsistent use of white space is not transparent [10].

Spam Text: HTML and XML files can also be used to hide bits. If there are different starting and closing tags, bit 0 is interpreted and if single tag is used for starting and closing it, then bit 1 is interpreted [13]. In another technique, bit 0 is represented by a lack of space in a tag and bit 1 is represented by placing a space inside a tag [13].

SMS-Texting: SMS-Texting language is a combination of abbreviated words used in SMS [14]. We can hide binary data by using full form of word or its abbreviated form. A codebook is made which contains words and their corresponding abbreviated forms. To hide bit 0, full form of the word is used and to hide bit 1, abbreviated form of word is used [14].

Feature Coding: In feature coding, secret message is hidden by altering one or more features of the text. A parser examines a document and picks out all the features that it can use to hide the information [13]. For example, point in letters i and j can be displaced, length of strike in letters f and t can be changed, or by extending or shortening height of letters b, d, h, etc. [14, 15]. A flaw of this method is that if an OCR program is used or if re-typing is done, the hidden content would get
Destroyed.

SSCE (Secret Steganographic Code for Embedding): This technique first encrypts a message using SSCE table and then embeds the cipher text in a cover file by inserting articles or an with the non specific nouns in English language using a certain mapping technique [15]. The embedding positions are encrypted using the same SSCE table and saved in another file which is transmitted to the receiver securely along with the stego file.

Word Mapping: This technique encrypts a secret message using genetic operator crossover and then embeds the resulting cipher text, taking two bits at a time, in a cover file by inserting blank spaces between words of even or odd length using a certain mapping technique [16]. The embedding positions are saved in another file and transmitted to the receiver along with the stego object.

MS Word Document: In this technique, text segments in a document are degenerated, mimicking to be the work of an author with inferior writing skills, with secret message being embedded in the choice of degenerations which are then revised with changes being tracked [17]. Data embedding is disguised such that the stego document appears to be the product of collaborative writing [14].

1	2	3	4	5	6	7	8	9	10

Cricket Match Scorecard: In this method, data is hidden in a cricket match scorecard by pre-appending a meaningless zero before a number to represent bit 1 and leaving the number as it is to represent bit 0 [15].

CSS (Cascading Style Sheet): This technique encrypts a message using RSA public key cryptosystem and cipher text is then embedded in a Cascading Style Sheet (CSS) by using End of Line on each CSS style properties, exactly after a semicolon. A space after a semicolon embeds bit 0 and a tab after a semicolon
Embeds bit 1 [16].

III. PROPOSED CODING SCHEME FOR HINDI TEXT

Hindi alphabet set consists of 49 symbols – 13 vowels and 36 Consonants.

.

Codes assigned to the alphabets: The vowels are assigned values from 1 to 14 and the Consonants have range 1 to 36.some Diacritics are also and range is 1 to 10 .The code is a decimal number assigned uniquely for every Alphabet of the language. Here we give decimal no to vowels and

Diacritics randomly whereas we give no to consonants according to usages. Some consonant are use lesser than the other have less priority.

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
1	2	3	4	5	6	7	8	9	10

अं	अः	अँ	ऋ
11	12	13	14

Fig 4. Table of Vowels

Fig 5. Table of Diacritics

क	ख	ग	घ	ङ	च	छ	ज	झ	ञ
1	2	3	4	36	5	6	7	8	35

ट	ठ	ड	ढ	ण	त	थ	द	ध	न
9	10	11	12	34	13	14	15	16	33

प	फ	ब	भ	म	य	र	ल	व	श
17	18	19	20	21	22	23	24	25	26

ष	स	ह	क्ष	त्र	ज्ञ				
27	28	29	30	31	32				

Fig 6. Table of Consonants

METHODOLOGY ADOPTED

Encoding: We consider 16 bits for representing every character of Hindi language. The 16 bits are divided into 3 different regions, where

Each region represents a different feature of the character. It can be viewed below Table 7.

4 bits	6 bits	6 bits
Diacritics (region 1)	Vowels (region 2)	Consonants (region 3)

Fig 7. Representation of Characters

The bits in Region 1 represent the Diacritics.
The Region 2 represents vowels.
The Region 3 represents the consonants.

Algorithm for Message Encoding:

- Read character one by one.

For each character:-

- If character is vowel then write code from vowel table and convert into binary value and append zero into r1 & r3 region.
- If character is consonant then write code from consonant table and convert into binary value and append zero into r1 & r2 region.
- If any Diacritics then write code from Diacritics table and convert into binary value append and zero into r2 & r3 region.
- If character is compound statement(have Diacritics then)
- Differentiate the consonant, vowel and Diacritics.
- Write code from table and convert into binary.
- If any region not have any value append zero.
- Convert each binary into 4-bit hexa-decimal no.
- Now append the code of each alphabet into a single string.

For example, let us consider the Hindi word

यहाँ

For each character:-

य:- is a simple alphabet, therefore its code is 22 and is represented as follows Region 3 would have the binary equalent of 22 and the Regions 1& 2 would be zeroes as it does not have any diacritic or vowels.

0000 000000 010110

Now on converting the above binary number into 4 bit Hexa-Decimal
Number, we obtain 0016(16).

हाँ: - is a compound statement. It has both diacritic and a compound. Now we would represent it as follows:

The codes are as follows:

ह Have 29 no from consonants.

आ + ँ has vowel no 2 and diacritic no is also 2.

Step 1: Convert the given Decimal codes into their equalent Binary numbers to fit into their respective regions. (Zeroes haveTo be padded in the left over bits in each region to fill all the bit places of every region)

0010 0010 011101

Step 2: Now convert the Binary number of step 1 into its equalent Hexa-decimal number. Which is 113A (16).
Step 3: Now append the code of each alphabet into a single string.

The Complete Hex representation of the word यहाँ would be 0016113A (16). Now the code is ready to be crypted or embedded in any cover media.

*Decoding:*For decoding the generated hex code back into Hindi text we follow the inverse of the encoding procedure:

Step 1: Convert the hex-string into its equalent binary number. Group 4 bits each to represent a character. Grouping is to be Done from left to right.

Step 2: 16 bit binary number is generated from each 4 bits of Hex-code.

Step 3: Now group these 16 bits into 3 regions as specified in the Methodology. (The bit is append from first r3 and then r2,r1 region respectively).

Step 4: Convert the code in each Region into its corresponding Decimal equalent.

Step 5: Map the Decimal code with its corresponding Character in the chart to get the final Hindi character.

IV. EXPERIMENTAL RESULTS

The proposed scheme provides double security for the text to be transmitted as the original text is encoded first and then it is encrypted using a public or private key for getting transmitted or being hidden in any cover. Any intruder has to be first successful in decrypting the hex-code which is transmitted and then he must be able to de-code into the original text. The fig. (8) Gives an overview of the proposed method's application.

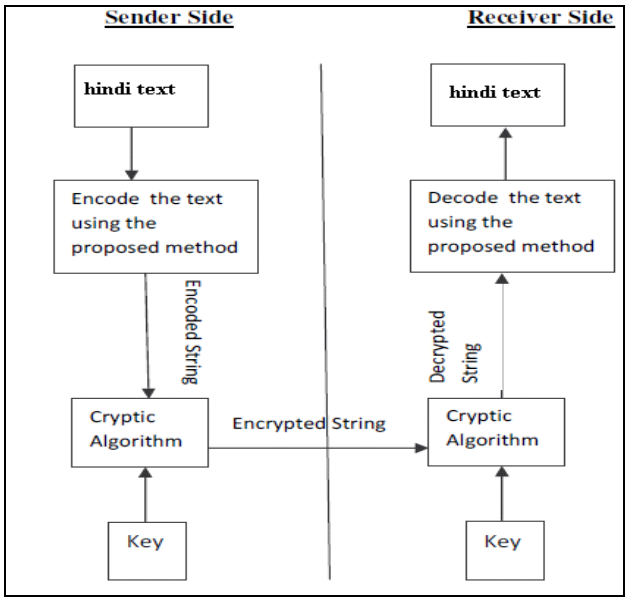

Fig 8. Overview of the proposed method application

The proposed method is implemented using Rijndael algorithm. A key of 60 characters is supplied in-order to encrypt the 00114B5A.Use the encrypted code along with the key in-order to decrypt the message [5]. The maximum key length minimizes the probability to assess the key.

The key considered as input is "this is a test case for implementation of the proposed code" fig. (9)

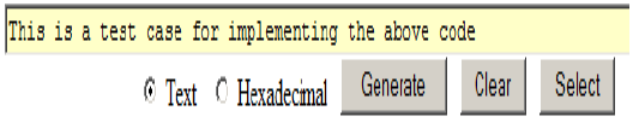

Fig-9: Actual key

The encrypted key is WAFHZ-FADNN-BONGX-TKAFP-KIFFQ-LJCYW-MAPNB-ECGAG-TKASI-FWKZD-BHDQO-OWSZS" as shown in Fig. (10).

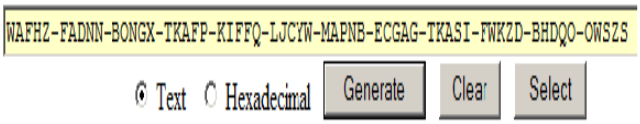

Fig 10:-Encrypted key

The sting to be encrypted is 0016113A (16) and is given in the text box fig. (11)

Fig. 11: String to be encrypted

On encrypting the string, the resultant string is
"ZZZZZ FKIUM ASAHE UAJEM GTHHX AXLUE JJTHW WKHHJ FNMPA TDDMT ENLUH
RDKEX INHGE XJFFO KIKIN JAVQC THJRQ OECVD LWGFI XQKXM SZZZZ YYYYY "Fig. (12)

Fig. 12: The encrypted String

Decryption:-
To decrypt, concatenate the key and encrypted text, so that we see the crypted text fig.(13).

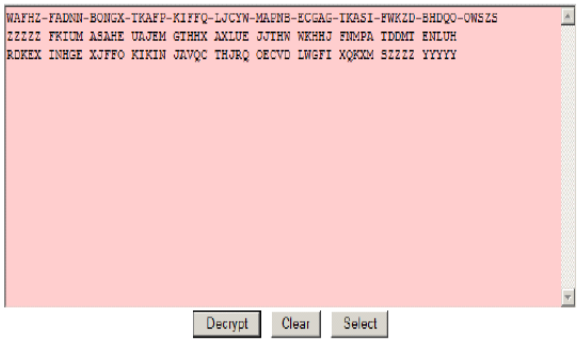

Fig 13:-decrypting
The resultant is "is 0016113A (16)" fig.(14)

Fig. 14: String after encryption

The results have been tested with Rijndael algorithm [5] and were successful in encoding and decoding the text.

V. CONCLUSION

In this paper a novel encryption algorithm using Hindi character set is proposed. The proposed method can be extensively used for all data hiding purposes. The coding scheme developed helps to minimize the bits required for hiding each character along with its diacritic and compound. The generated hex-code is much suitable to be hidden in any cover media. We can hide a larger message within a minimal effort. It ensures multiple security levels for better secrecy of the content.

VI. FUTURE WORK

The code generated by the proposed coding scheme can be used for cryptic transmission of Hindi text using any encryption Algorithm. Another area of application is embedding the encrypted string into any cover media, where the cover can be a image, text of any language, audio or video for Steganographic or Watermarking purposes.

REFERENCES:-

[1]. K. Bennett, "Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hiding Information in Text", Purdue University, CERIAS Tech. Report, **2004**.

[2]. Ross J. Anderson and Fabien A.P. Petit colas, "On the limits of steganography," IEEE Journal on Selected Areas in Communications (J-SAC), Special Issue on Copyright& Privacy Protection, vol. **16** no. **4**, **pp 474-48**1, May**1998**.

[3]. Fabien A. P. Petit colas, Ross J. Anderson and Markus G.Kuhn. "Information Hiding – A Survey", Proceedings of the IEEE,special issue on protection of multimedia content, July **1999**, pp. **1062 – 1078**.

[4]. N.F. Maxemchuk J.T. Brassil, S. Low and L. O.Gorman. Electronic marking and identification techniques to discourage document copying. IEEE Journal on Selected Areas in Communications, **13:1495–1504, 1995**.

[5]. G. Doerr and J.L. Dugelay. Security pitfalls of frame by frame approaches to video watermarking. IEEET transactions on Signal Processing, Supplement on Secure Media, **52:2955–2964, 2004**.

[6]. W. Sweldens. The lifting scheme. A construction of second generation wavelets. SIAM J. Math. Anal., **29:511–546, 1997**.

[7]. N.F.Johnson. And S. Jajodia. Steganography: seeing the unseen. IEEE Computer, **16:26–34, 1998**.

[8]. W. Sweldens R. Calderbank, I. Daubechies and B.L. Yeo. Wavelet transforms that map integers to integers. Appl. Comput. Harmon. Anal.,**5:332–369**, **1998**.

[9]. K. Moon Y. Kim and I. Oh. A text watermarking algorithm based on word classification and inter-word space statistics. In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'**03**), pages **775–779, 2003**.

[10]. K. Rabah, "Steganography-the art of hiding data," Information Technology Journal, vol.**3**, pp. **245-269**, **2004**.

[11]. W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," IBM Systems Journal, vol.**35**, pp. **313-336**, **1996**.

[12]. [12] M. H. S. Shahreza, and M. S. Shahreza, "A new approach to Persian/Arabic text steganography," In Proceedings of 5th IEEE/ACIS Int. Conf. on Computer and Information Science and 1st IEEE/ACIS Int. Workshop on Component-Based Software Engineering, Software Architecture and Reuse, **2006**, pp. **310-315**.

[13]. M. H. S. Shahreza, and M. S. Shahreza, "A new synonym text steganography," Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing, **2006**, pp. **1524-1526**.

[14]. J. Cummins, P. Diskin, S. Lau, and R. Parlett, "Steganography and digital watermarking," School of Computer Science, **2004**, pp.**1-24**.

[15]. T. Y. Liu, and W. H. Tsai, "A new steganographic method for data hiding in Microsoft word documents by a change tracking technique," IEEE Transactions on Information Forensics and Security, vol.2, no.1, pp. **24-30**, **2007**.

[16]. M. Khairullah, "A novel text steganography system in cricket match scorecard," Int. Journal of Computer Applications, vol.**21**, pp. **43-47**, **2011**.

[17]. H. Kabetta, B. Y. Dwiandiyanta, and Suyoto, "Information hiding in CSS: a secure scheme text steganography using public key cryptosystem," Int. Journal on Cryptography and Information Security, vol.1, pp. **13-22**, **2011**.

[18]. A Novel Approach of Secure Text Based Steganography Model using Word Mapping Method(WMM) Souvik Bhattacharyya, Indradip Banerjee and Gautam Sanyal International Journal of Computer and Information Engineering **4:2 2010**.

[19]. Kalavathi Alla, R. Siva Rama Prasad, "An Evolution of Hindi Text Steganography," citing, p.**1577-1578**, 2009Sixth International Conference on Information Technology: New Generations, **2009**.

