Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 524 – 531

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

A Method for Text Steganography Using Malayalam Text

Vidhya P.M[a],Varghese Paul[b,*]

[a]School of Computer Science, MG University,Kottayam-686560,India
[b]Division of Information Technology Cochin University of Science and Technology,Cochin-682022,India

Abstract

Recent researches regarding information hiding is mostly concentrating on Linguistic steganography. In this paper, a method to steganography is proposed with an Indian local language, Malayalam. The proposed method consists of custom Unicode based technique with embedding based on indexing, i.e. the original message is encoded to Malayalam text with custom UNICODE values generated for the Malayalam text. The comparison study of the proposed method against an existing method revealed that, the proposed steganography methods is more precise in the encoding process and in the decoding process. The method achieved a precision rate of .95 and decoding rate of .81.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Information hiding; text steganography; UNICODE;Malayalam; index

1. Introduction

Steganography is an approach for information hiding such that the presence of data cannot be detected[1].A secret message is encoded in such a manner that the existence of the information is hidden. Combining with existing communication methods, Steganography can be used to carry out hidden exchanges. Trithemius, the author of Polygraphia and Steganographia, invented the word Steganography and the word is derived from two Greek words

* * Corresponding author. Tel.: 9496560996
E-mail address: vp.itcusat@gmail.com

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.079

steganos, meaning "covered", and graphein, meaning "to write". The initial evidence on Steganography being used to transfer messages is the Heredotous[2] story, which describes about slaves and their shaved heads. The most discussed problem with encryption method is that the cipher texts are very doubtful[3]. When information is concealed in this way, no user can unsure that the text transmits a kind of secure information[4]. The mission of steganography is to create a secret channel (that is a secret communication) in a totally un- detectable way and to evade drawing suspicion on the transporting data. Many steganographic methods have been proposed to embed secret messages using different cover media like image [5, 6, 7], audio[8], video [9] and text. Among the methods, Text steganography is found to be the difficult type of steganography due to the lack of redundant information in a text file as compared with a picture, Markup language or sound file [10, 11].

In this paper, a linguistic steganography is provided for the information hiding, the method is triggered with the help of Unicode and embedding algorithms. The main feature of the proposed approach is the use of local languages as cover text. The use of local language ensures more security to the information exchange as the awareness to the local language is limited. The proposed approach uses two matrices for indexing the alphabets in the common language and the local language. The matrices are loaded with the alphabet letters and their indices in the increasing order. A UNICODE extraction method is designed by the proposed approach to find the Malayalam text corresponding to the English text given. The encoding scheme used by the proposed approach is diagonal index encoding, in which the indices are selected diagonally. The details of the encoding technique are explained in the following sections. The final process is the decoding key generation, which includes the creation of the key that will help in decoding the hidden text or information.

The main contributions of the paper are,

- An Indian local language is used for the steganography method
- Different languages can be used for the steganography as input
- A Unicode based method is used for the steganography method

2. Literature Survey

In this section a number of studies and researches regarding text steganography method is plotted to understand the recent progresses in the steganography domain.

D. Ghosh et al[3] presented a linguistic approach for Steganography through Indian Languages by considering the flexible grammar structure of Indian Languages. The addition of security to the system, as an alternative of hiding the original message the data is converted to an irrelevant binary stream by associating the message bits with the pixel values of an Image. Then, the bits of this binary stream are encoded to some part-of-speech and by creating meaningful sentences starting with a suitable word belonging to the mapped part-of-speech, the proposed method hides the message inside a cover file containing some innocuous sentences. Similarly in receiving side, the algorithm finds the corresponding part-of-speech of the starting word of each sentence and place the bit stream of the mapped part-of-speech to recover the converted message. After comparing these bits with the Image pixels, the algorithm extracted the original message from the cover file. The method exhibits satisfactory result on some Indian Languages like Bengali.

Kalavathi Alla and Ramineni Siva Ram Prasad[12] have proposed a few text based steganographic methods. The methods worked by using the linguistic properties of Telugu language. The first method selects embed position of the secret information in the cover text by using Telugu Ottulu. Based on the two level classification of Ottulu, they are assigned a bit 0 or a bit 1. These symbols embed the secret information in the third character of Telugu cover Text data. It maps a single bit of the data with a Telugu character in the specified manner.

Kalavathi Alla and Ramineni Siva Ram Prasad[13] described a framework of text based steganography for Hindi language. It introduced three approaches for steganography, in Hindi text. First one depicts on classifying the position of matraye of Hindi Characters. Second approach focuses on the classification of Hindi characters by various OCR tools in Hindi Language. Third one portrays on hex katapayadi scheme. Alfonso Muñoz et al [14] have proposed analyses of the usefulness of synonym substitution techniques in Spanish, aiming at a first approximation to their capacity of concealment in Spanish.

The rest of the paper is organized as, the second section contains the literature review of some similar research regarding steganography. The third section gives the motivation behind proposing the new methodology. The fourth section gives the detailed explanation of the proposed text steganography method. The fifth section plots the different experimental analysis regarding the proposed method. Finally, the sixth section gives the conclusion of the method.

3. Proposed Linguistic Steganography in Malayalam Text

In the proposed approach, a linguistic steganography is provided for the information hiding, the method is triggered with the help of Unicode and embedding algorithms. The main feature of the proposed approach is the use of local languages as cover text. The use of local language ensures more security to the information exchange as the awareness to the local language is limited. The proposed method is more specific to the Indian local language, so we are using Malayalam language as the covering text for hiding the information. The proposed approach is shown in fig1 The main features included in the proposed approach are,

- Alphabet index matrix (English & Malayalam)
- UNICODE extraction
- Diagonal index encoding
- Embedding the message

The proposed approach uses two matrices for indexing the alphabets in the common language and the local language as shown in figure 1. The matrices are loaded with the alphabet letters and their indices in the increasing order. A UNICODE extraction method is designed by the proposed approach to find the Malayalam text corresponding to the given English text. The encoding scheme used by the proposed approach is diagonal index encoding, in which the indices are selected diagonally.

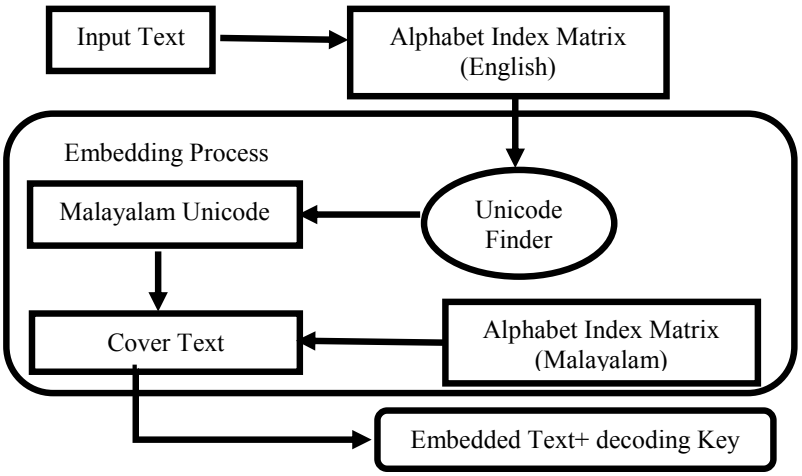

Fig.1. Processing diagram

3.1 Alphabet index matrices (AIM)

We discussed that the proposed approach uses two alphabet index matrices, one for the English alphabet and other for the Malayalam alphabets. The last index of the matrix is also left as zero. The matrix is filled with the letters by their serial order and their index will start with value '0'.

0-0	1-a	2-b	3-c
4-d	5-e	6-f	7-g
8-h	9-i	10-j	11-k

12-l	13-m	14-n	15-o
16-p	0A-q	1A-r	0B-s
1B-t	2B-u	0C-v	1C-w
0E-x	1E-y	2E-z	28-0

Fig 2: Alphabet index matrix

The figure 2 represents the AIM of the English alphabets; the matrix is filled with index values and the letters of the English alphabets. The matrix is used for the diagonal index calculation, which is main embedding scheme of the proposed approach. On the other hand, the Malayalam language has two alphabets, but for the ease of the matrix creation we use a single matrix for all the letters of Malayalam. The total number letters including both alphabets is 56, so an AIM of 8x7 is created for the Malayalam language.

	0D0	**0D1**	**0D2**	**0D3**	**0D4**	**0D5**	**0D6**	**0D7**
0		ഐ 0D10	ഠ 0D20	ര 0D30	ീ 0D40		ൠ 0D60	൰ 0D70
1	□ 0D01		ഡ 0D21	റ 0D31	ു 0D41		ൡ 0D61	൱ 0D71
2	ം 0D02	ഒ 0D12	ഢ 0D22	ല 0D32	ൂ 0D42		ൢ 0D62	൲ 0D72
3	ഃ 0D03	ഓ 0D13	ണ 0D23	ള 0D33	ൃ 0D43		ൣ 0D63	൳ 0D73
4		ഔ 0D14	ത 0D24	ഴ 0D34	ൄ 0D44			൴ 0D74
5	അ 0D05	ക 0D15	ഥ 0D25	വ 0D35				൵ 0D75
6	ആ 0D06	ഖ 0D16	ദ 0D26	ശ 0D36	െ 0D46		൦ 0D66	
7	ഇ 0D07	ഗ 0D17	ധ 0D27	ഷ 0D37	േ 0D47	ൗ 0D57	൧ 0D67	
8	ഈ 0D08	ഘ 0D18	ന 0D28	സ 0D38	ൈ 0D48		൨ 0D68	
9	ഉ 0D09	ങ 0D19	ഩ 0D29	ഹ 0D39			൩ 0D69	൹ 0D79
A	ഊ 0D0A	ച 0D1A	പ 0D2A	ഺ 0D3A	ൊ 0D4A		൪ 0D6A	ൺ 0D7A
B	ഋ 0D0B	ഛ 0D1B	ഫ 0D2B		ോ 0D4B		൫ 0D6B	ൻ 0D7B
C	ഌ 0D0C	ജ 0D1C	ബ 0D2C		ൌ 0D4C		൬ 0D6C	ർ 0D7C
D		ഝ 0D1D	ഭ 0D2D	ഽ 0D3D	് 0D4D		൭ 0D6D	ൽ 0D7D
E	എ 0D0E	ഞ 0D1E	മ 0D2E	ാ 0D3E	◌ 0D4E		൮ 0D6E	ൾ 0D7E
F	ഏ 0D0F	ട 0D1F	യ 0D2F	ി 0D3F			൯ 0D6F	ൿ 0D7F

Fig.3. Unicode chart of Malayalam Text

Figure3 represents the UNICODE chart of the Malayalam text based on this letters we create the AIM for Malayalam language. As mentioned in the above section the main purposes of the AIM is to calculate the diagonal index of the alphabetic letters.

3.2 Diagonal Index Calculation

The proposed approach mainly concentrates on the index of the two AIMs, which would virtue for the embedding of the secret message. The processing diagonal index selection starts from the selection of the secret message, which is to be embedded in the cover text. Consider the secret message is written using the English alphabets with proper writing conventions. Let us discuss the diagonal index calculation through the following example. Let the secret message be, 'king' .Now, we scan each element of the secret message, and then the index of each word is extracted from the English alphabet AIM. According to the AIM, the index labels of the word 'king' will be extracted as, "11 9 14 7". The extracted index will not that much secure, because it can be easily identified. So in order to ensure security to the secret message, we select the index of the diagonal elements of the word 'king' from the AIM

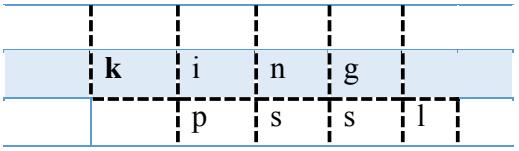

Fig.4. Diagonal indexing

The figure 4 shows the diagonal marking of the proposed approach on the word 'king'. The index of the corresponding value is selected for the extracting the index, the index for 'pnsl' is generated as '16 14 0B 12'. The index values are stored in an array and passed for the UNICODE extraction. The role of the Malayalam AIM comes after the UNICODE extraction. The reverse process of the above process happens for the Malayalam AIM. Here, AIM_ENG represents attribute index matrix for English alphabet, S represents the set words in the secret message, I represents the set of initial index values and diagonal represents the set of index values after diagonal index processing.

Algorithm1. Diagonal indexing
Input: secret message
Output: index values
Step1. **Select** secret message S
Step2. **Select***AIM_ENG*
Step3. **For each** letter in S,
  Find index from *AIM_ENG*
Step4. **Store** index in *I*
Step5. **For each** element in *I*
  Find diagonal values
Step6. **Store** diagonal value in $I_{diagonal}$
Step7. $I_{diagonal}$ ➔ Index values of S
Step8. **End**

3.3 UNICODE extraction

The main association of the proposed approach is with UNICODE and the Text for embedding. The UNICODE play a key role in the process of embedding the secret message in the Malayalam text. On our context, the input

message will be of English language, and the according to the diagonal indexing method, we extract an index value for the input message. The index code is used for triggering the UNICODE for Malayalam text. Let us discuss it in detail. Consider a sample message is given as input for embedding as secret message, according to the proposed method, the initial process will be diagonal indexing. Once the indexes are generated, then it will be selected for UNICODE generation. The proposed approach defines a method to generate UNICODE from the index generated. The index is treated in to ways for generating the UNICODE.

- Splitting the index
- Direct inputting index

The Unicode extraction phase can be analysed through the following diagram shown in figure 5,

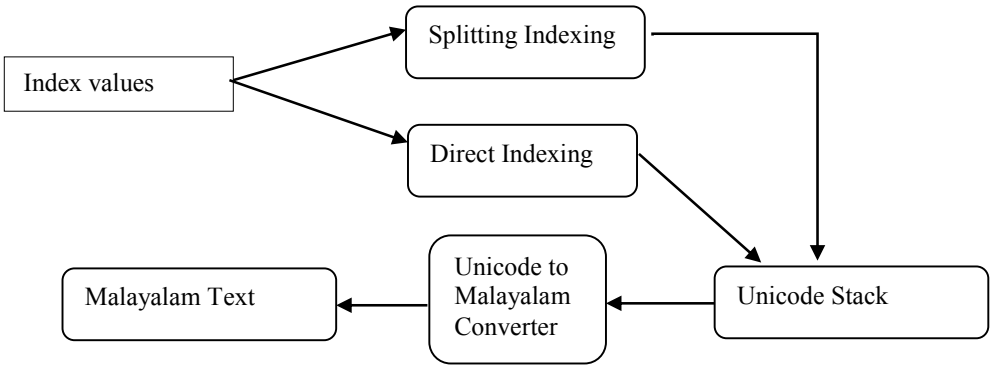

Fig.5. Unicode Extraction

3.4 Embedding the secret message

The embedding of the secret message is the main part in the proposed text steganography method. This phase will work as the extension of the UNICODE extraction phase. In the proposed method, the cover text is also in Malayalam and set of Malayalam cover text are in the database. Since, the cover text is selected randomly from a set of files, a database is necessary for it. The initial step in embedding the secret message is to extract the Malayalam text from the UNICODE. The Malayalam texts corresponding to the secret message are stored in a set M,

$M=[m_1.m_2,m_3\ldots.m_n]$

The values $m_1\ldots m_n$ represents each element in the Malayalam text, which is extracted. Once the Malayalam texts are extracted, the cover text is selected and a string comparison is executed with every element in the set M with every element in the cover text. The word by word comparison is done in order to extract the position of the similar elements in the cover text and the elements in M. The elements similar to that are in M with the cover text are selected and their positions are extracted. Then a diagonal processing is applied in the cover text on the extracted positions, which will give new position for the mapped values of the elements in M. Now the updated positions from the cover text are extracted as the decoding key.

Algorithm: *Embedding_message*
Input : Cover text (CT), Malayalam text set M
Output : Embedded cover text, d_{key}
Step1. **Select** Malayalam text set $M = [m_1, m_2, ..., m_n]$
Step2. **Select** Cover text $CT = [e_1, e_2, ..., e_n]$
Step3. **For each** element **in** M
For each element **in** CT
Compare $((M : m_i), (CT : e_i))$
Step4. **If** $((M : m_i) == (CT : e_i))$
Store position $(CT : e_i)$ in Set P
Else
Forward position
Step5. End **For loop**
Step6. **Select** P
Step7. **Update** position by applying diagonal indexing of elements in CT by position in P
Step8. **Store** updated position in Set d_{key}
Step9. End

4. Comparison Study

In this section, a comparison study of the proposed approach against an existing steganography method is plotted. The proposed text steganography method is based on local languages and UNICODE, while the existing work is based on local languages and the binary encryption. The comparison is done based on the response of both algorithms to encrypt a given secret message in English.

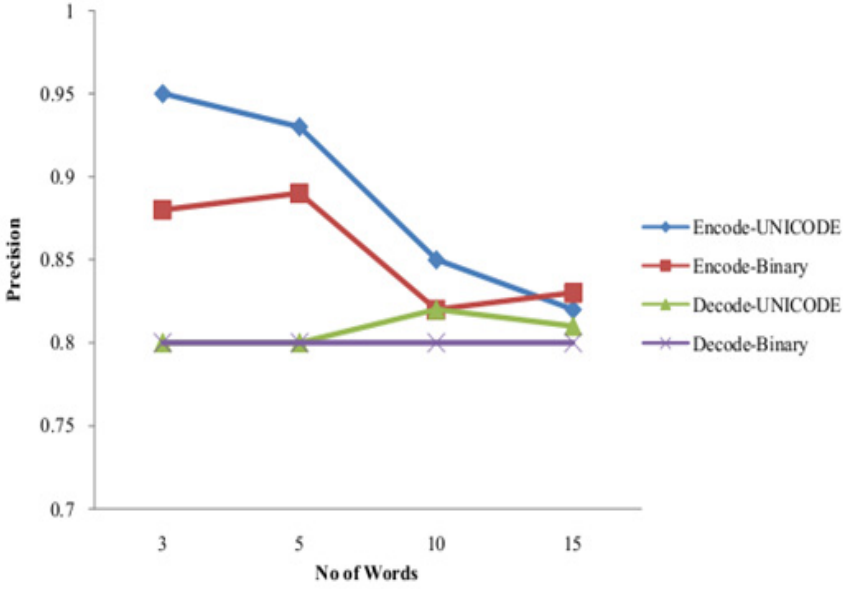

Fig.7. Comparison analysis

The precision in encoding and the decoding is taken into the account for the evaluation. The figure 7 shows the comparison analysis of the proposed and existing steganography methods. The evaluation is executed based on different set of words in the range 3, 5, 10 and 15. The precision in encoding is calculated as the ratio of number of words encoded to the total number of words. The precision in decoding is calculated as the ratio of numbers of words decoded per total number of words. The analysis from the comparison given that, the propose UNICODE based method is precise in encoding the data than the binary based method, while the decoding process is balanced for both methods. The proposed method achieved a higher precision rate of .95 and decoding rate of .81.

5. Conclusion

Steganography is the method of hiding information such that its presence cannot be detected. A secret message is encoded in such a manner that the existence of the information is hidden. We have proposed a method to text steganography with an Indian local language, Malayalam. The proposed method consists of a custom Unicode based technique with embedding based on indexing. After that an embedding algorithm will be designed to mix the encoded original message with the Malayalam text. The experimental study is done to evaluate the efficiency of the proposed approach. The comparison study of the proposed method against an existing method revealed that, the proposed steganography methods is more precise in the encoding process and balanced in the decoding process. The proposed method achieved a precision rate of .95 and decoding rate of .81.

References

1. C .Cachin, An Informaation-Theoretic Model for Steganography, 2nd Information Hiding Workshop, vol. 1525, p. 306-318, 1998
2. Herodotus. The Histories. Penguin Books, London, Translated by Aubrey de Selincourt, 1996.
3. S.Changder, D. Ghosh, N. C. Debnath, Linguistic Approach for Text Steganography through Indian Text, IEEE InternationalConference on Computer Technology and Development, p: 318-322, 2010.
4.J.C.Judge,stegangoraphy:Past, Present, Future,SANS white paper http://www.sams.org/rr/papers/novemeber 30, 2001.
5. J.H.P Eloff, T. Markel and M.S Oliver, An overview of Image Steganography,5th annual information securitySouth Africa Conference.
6. Kim, Bailey, Kevin Curran An Evolution of Image SteganographyMethods, International Journal of Digital Evidence, fall 2003.
7. W. Sweeden, R. Calderbank, I. Daubechies, and B.L. Yeo WaveletTransforms that map integers to integers, Appl, comput, Harmon, Anal p 332-369 1998.
8. K. Gopalan, Audio Steganography Using Bit Modification IEEE International Conference on Accoustics,Speech and Signal Processing, (ICASSP'03), volume 2,p 421-424,6-10 April 2003.
9. G. Doerr and J.I Dugelay, Security Pitfalls of framely frame approaches to video watermarking, IEEE transaction on signal processing:supplemet on secure media 52:2995-2964, 2004.
10. N. Provos and P.Honeyman Hide and Seek Anintroduction to steganography, IEEE Security and Privacy.p. 32-44, May/June 2003.
11. W. Benderr, D. Gruhl, N. Morimoto and A.Lu,Techniques for Data Hiding, IBM System's Journal,Volume 35, Issue 3 and 4, 1996, p 313-336.
12. Kalavathi Alla and Ramineni Siva Ram Prasad, "A New Approach to Telugu Text Steganography", IEEE Symposium on Wireless Technology and Applications, p 25-28, 2011
13. Kalavathi Alla and Ramineni Siva Ram Prasad, A New Approach to Hindi Text Steganography Using Matraye, Core Classification And HHK Scheme, Seventh International Conference on Information Technology, p 1223-1224, 2010.
14. A. Muñoz, J. Carracedo, I. A Álvarez, Measuring the security of linguistic steganography in Spanish based on synonymous paraphrasing with WSD, IEEE International Conference on Computer and Information Technology, p 965-970, 2010.