International Journal of Computer Applications (0975 – 8887)
Volume 30– No.5, September 2011

Extended Encoding of Telugu Text for Hiding Compatibility

Lakshmeeswari G
GITAM Institute of Technology
GITAM University
Andhra Pradesh, India

Rajya Lakshmi D
GITAM Institute of Technology
GITAM University
Andhra Pradesh, India

Lalitha Bhaskari D
College of Engineering
Andhra University
Andhra Pradesh, India

ABSTRACT
Conveying information secretly and establishing a hidden relationship between the message and its counterpart has been of great interest since long time. This paper presents a new scheme for encoding of the Telugu alphabet set, its diacritics and compounds which can be used in the Linguistic Steganography as well as cryptic transmission of Telugu text. This encoding scheme helps its users to have better support in implementing the hiding techniques and this scheme can be extended to other similar Indian Languages.

Keywords: Cryptic Algorithms, Steganography, Telugu text, diacritics, compounds, Information hiding.

1. INTRODUCTION
Information hiding is the ability to prevent certain aspects of a component from being accessible to third party. Since the rise of the Internet, one of the most important factors of information technology and communication has been the security of information[6]. Infact, hiding information while two or three parties communicate is an art. The most common way of secret communication is hiding information in another media(like text,images,audio,video) so as to make it unnoticeable during transmission. The efficiency of the method depends upon the robustness of the hiding technique. Steganography, Digtal watermarking, covert communications deal with different techniques of information hiding where some applications require absolute invisibility of the secret information, some require a larger message(media) for a secret message to be hidden into it and so on. A lot of research is carried out to find robust ways of information hiding, copy right protection, copy protection and secret communication of the same. But a very little work is done with respect to Telugu language and so it has been a source of inspiration in formulating a new encoding technique.

This paper deals with a novel scheme for encoding Telugu text which can be transmitted via network by encrypting it using any of the Crypto algorithms. A new encoding scheme suitable for Telugu language is proposed which can be used for Steganographic as well as Cryptographic applications.

The main idea behind choosing Telugu language as our media is that Telugu is one of the most prominent languages used in the state of Andhra Pradesh, India and a few places around the world. Telugu is the official language of the state of Andhra Pradesh, India. 73 million people use this language around the world, which include countries like Singapore, Phiji, U.S.A[4]. It occupies 15th position amongst the most widely used languages in the world and stands second in India. Telugu is a richly developed language and the biggest linguistic unit in India[4].

Organization of the remaining part of the paper is as follows: Section 2 contains description about the Historic evolution information of Telugu language, Section 3 briefs about the Telugu Literature, Section 4 presents a detailed description of the Methodology of the proposed scheme (both encoding and decoding procedures), Section 5 lists the Experimental Results of the methodology tested with Rijendel algorithm, Section 6 concludes the discussion and Section 7 contains the possible future extensions of the proposed methodology.

2. BASIC INFORMATION OF THE LANGUAGE
Telugu is a language derived from ancient Brahmi script. Brahmi based script is well known for its complex conjunct formations[9]. Telugu, Kannada and Tamil languages are called Dravidian languages[4]. These languages have complicated script. It was also referred to as `Tenugu' in the past.

Telugu script is a combination of vowels, consonants, compounds and diacritics. Almost all Indian languages are built upon the concepts of vowels, consonants, diacritics and compounds and also have their origin from Sanskrit. Telugu comprises of 70% Sanskrit orientation[3]. Therefore the method that has been proposed below for Telugu language would well suit other Indian languages because Sanskrit is one of the base for all the Indian languages.

Telugu is one of the 22 official languages of India. Telugu also has official language status in the Yanam District of the Union Territory of Puducherry. Telugu was called the Italian of the East by Italian explorer Niccolò Da Conti[6].

Telugu script can reproduce the full range of Sanskrit phonetics without losing any of the text's originality. Telugu has made its letters expressive of all the sounds and hence it has to deal with significant borrowings from Sanskrit, Tamil and Hindustani[7].

3. LITERATURE SURVEY
A *crypto system* transforms plain text messages (using a key) to un-intelligible text. *Cryptography* is the study of "secret writing" or cryptograms. Encryption is the process of converting plain text into cipher text. Decryption is the vice-versa of encryption i.e transforming the cipher text to original plain text[9].

Steganography's goal in general is to hide data well enough that unintended recipients do not suspect the steganographic medium of containing hidden data[8]. Steganography is a technique of

hiding information in digital media. In contrast to cryptography, it does not reveal others from knowing that the hidden information even exists. The goal of steganography is to avoid drawing suspicion to the existence of a hidden message. This approach of information hiding technique has become important in a number of application areas like Text, Digital audio, video, and pictures[8].

Many techniques are existing to hide English text into text of different languages like hiding English text in Bengali[2], Hindi[1], Urdu, Arabic text[12], etc. which use different techniques to achieve the same.

The proposed system facilitates stores to store the encoded Telugu text in any cover media. The cover media can be an image, audio, video or plain English text. The main applicative part of this can be in the area of communication between the banker and its customer, between the cops for intelligence purposes. As Telugu is the local language of Andhra Pradesh, application can well suit the security necessities of the organizations requiring any authentication or hidden transmissions.

4. PROPOSED CODING SCHEME FOR TELUGU TEXT

Telugu alphabet set has been consistently modified since its origin and presently it consists of 52 symbols - 16 vowels and 36 consonants. Sanskrit and Telugu alphabets are similar and exhibit one-one correspondence.

4.1 Codes assigned to the alphabets:

The vowels are assigned values from 0 to 15 and the same range is considered for the diacritics also as diacritic is a combination of consonant with an vowel. The codes of compounds and constants are the same as the compound would always be a consonant.

The code is a decimal number assigned uniquely for every alphabet of the language. Just varying the alphabets depending upon the language would make the coding scheme suitable for other languages.

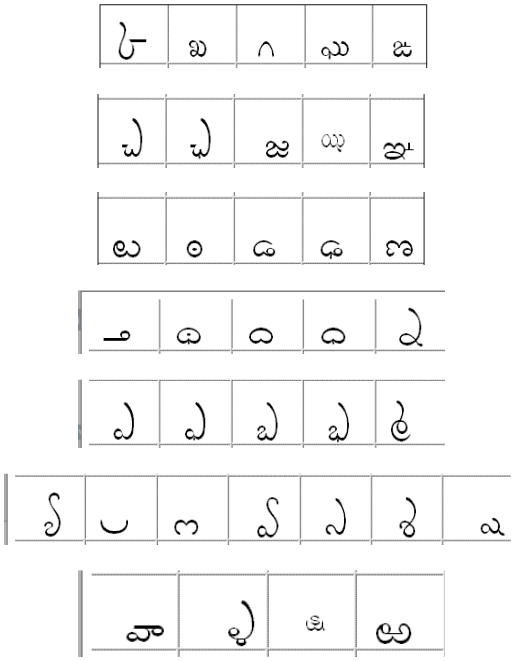
Table 1. Vowels (Achulu)

అ	ఆ	ఇ	ఈ
0	1	2	3
ఉ	ఊ	ఋ	ౠ
4	5	6	52
ఎ	ఏ	ఐ	ఒ
7	8	9	10
ఓ	ఔ	అం	అ:
11	12	13	15

Table 2. Consonants (Hallulu)

క	ఖ	గ	ఘ	ఙ
16	17	18	19	51
చ	ఛ	జ	ఝ	ఞ
20	21	22	23	50
ట	ఠ	డ	ఢ	ణ
24	25	26	27	28
త	థ	ద	ధ	న
29	30	31	32	33
ప	ఫ	బ	భ	మ
34	35	36	37	38
య	ర	ల	ళ	వ
39	40	41	42	43
శ	ష	స	హ	ఱ
44	45	46	47	48
Diacritics (Gunintalu)				
√	ా	ి	ీ	ు
0	1	2	3	4
ూ	ృ	ె	ే	ై
5	6	7	8	9
ొ	ో	ా	్	ం
10	11	12	14	13
ఃౕ				
15				

Table 3. Compounds (Vattulu)



Diacritics have same code as that of the vowels. Similarly, compounds have the same code as that of the consonants.

4.2 Methodology Adopted:

4.2.1 Encoding:

We consider 16 bits for representing every character of Telugu 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 4.

Table 4. Representation of Characters

6 bits						4 bits				6 bits					
15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Region 3						Region 2				Region 1					

The bits in

Region 1 represents the Alphabet

Region 2 represents the Diacritic

Region 3 represents the Compound

For example, let us consider the Telugu word

ఖడ్గం

ఖ is a simple alphabet, therefore its code is 17 and is represented as follows

Region 1 would have the binary equalent of 17 and the Regions 2 & 3 would be zeroes as it does not have any diacritic or compound.

000000 0000 010001

Now on converting the above binary number into Hexa-Decimal Number, we obtain $0011_{(16)}$

డ్గం has both diacritic and a compound. Now we would represent it as follows:

The codes are as follows:

డ is 26

O is 13

ం is 18

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

010010 1101 011010

Step 2: Now convert the Binary number of step 1 into its equalent Hexa-decimal number. Which is

$4B5A_{(16)}$

Step 3: Now append the code of each alphabet into a single string.

The Complete Hex representation of the word ఖడ్గం would be $0\,0114B5A_{(16)}$. Now the code is ready to be crypted or embedded in any cover media.

4.2.2 Decoding:

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

Step 1: Convert the hex-string into its equalent binay 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.

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 Telugu character.

5. 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. (1) gives an overview of the proposed method's application.

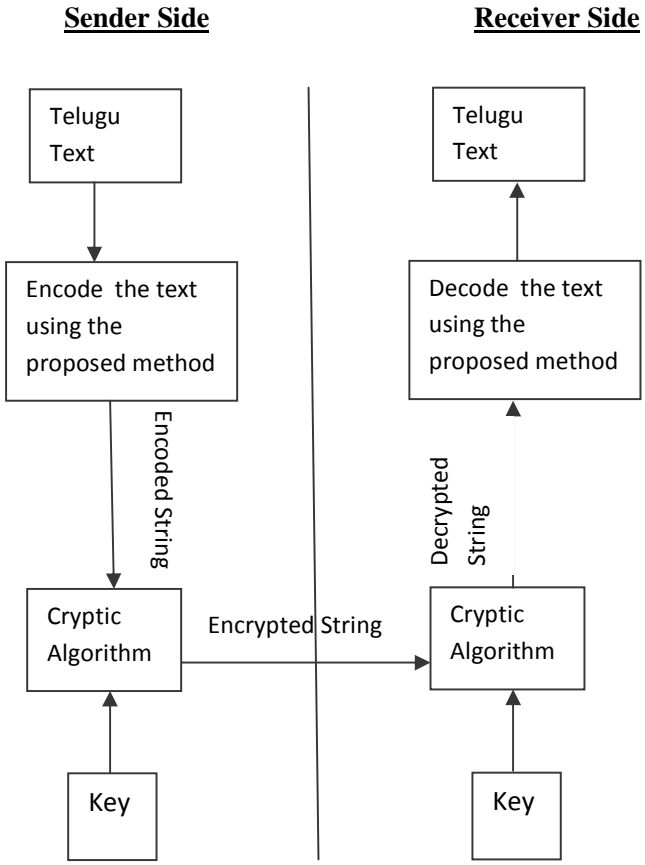


Fig. (1): Overview of the proposed method application

The proposed method is implemented using Rijndael algorithm. An 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. (1)

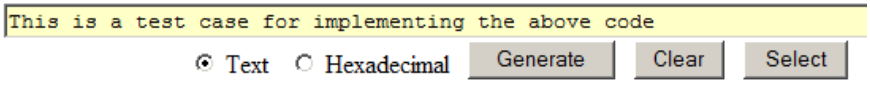


Fig.(2): Actual key

The encrypted key is

" WAFHZ-FADNN-BONGX-TKAFP-KIFFQ-LJCYW-MAPNB-ECGAG-TKASI-FWKZD-BHDQO-OWSZS" as shown in Fig. (3).

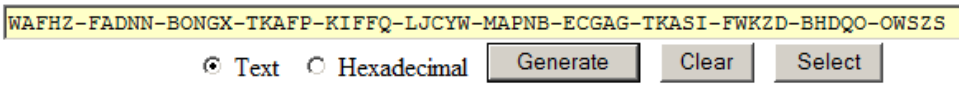


Fig. (3): Encrypted key

The sting to be encrypted is 00114B5A and is given in the text box fig. (4)

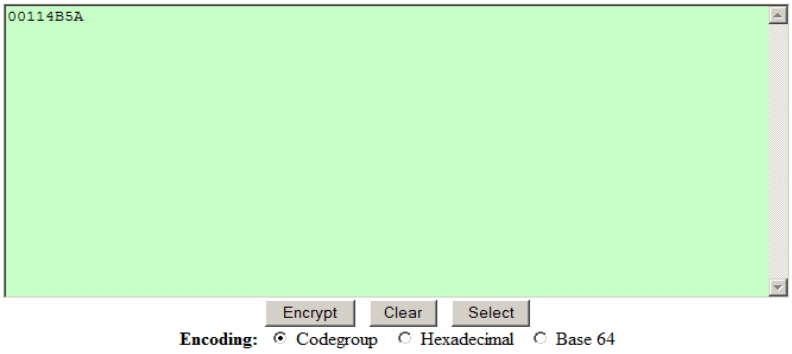


Fig. (4): String to be encrypted

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. (5)

Fig. (5): The encrypted String

To decrypt, concatenate the key and encrypted text, so that we see the crypted text fig.(6).

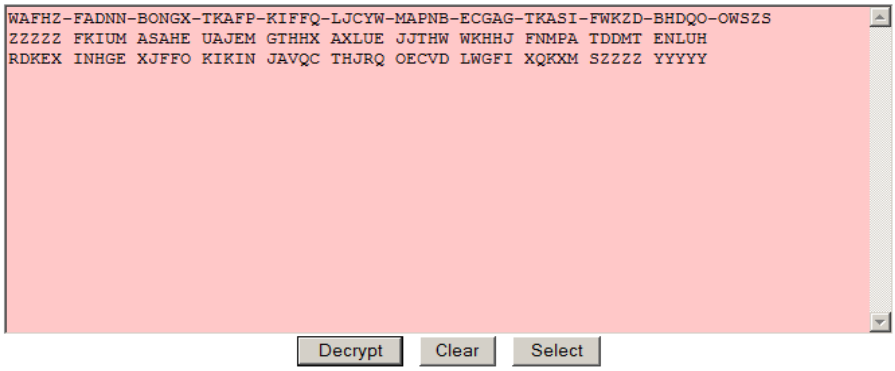


Fig. (6): Decrypting

The resultant is " 00114B5A" fig.(7)

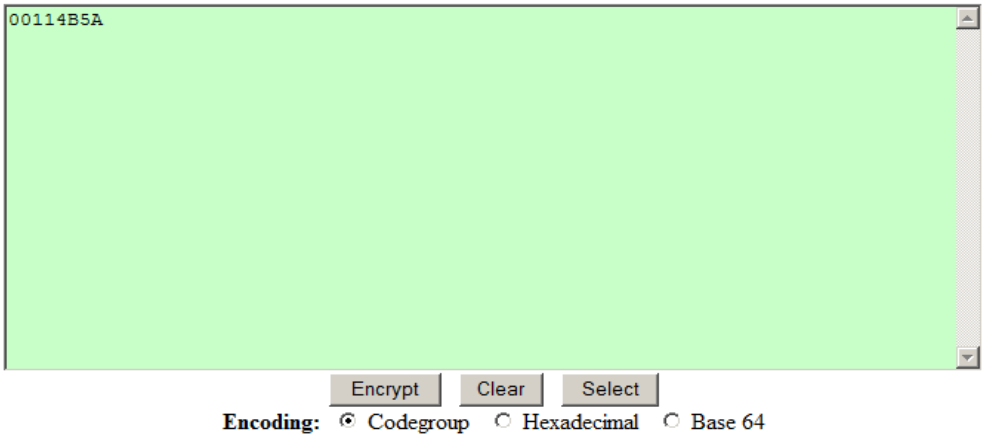


Fig. (7): String after encryption

The results have been tested with Rijndael algorithm[5] and was successful in encoding and decoding the text.**6.**

6. CONCLUSION

In this paper a novel encryption algorithm using Telugu 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.

7. FUTURE WORK

The code generated by the proposed coding scheme can be used for cryptic transmission of Telugu 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.

8. REFERENCES

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

[2] Shri S. Changder.S.Das and Shri D. Ghosh "Text Steganography through Indian Languages Using Feature Coding Method", 2nd International Conference on Computer Technology and Development, 2010.

[3] Compounds have been referenced from the http://www.balmitra.com/languagebook/telugu

[4] An overview of the Telugu language is from www.andhrabulletin.com/Telugu/Telugu_home.php

[5] JavaScrypt Encryption and Decryption, www.fourmilab.ch/javascrypt/javascrypt.html

[6] Telugu language, available at http://en.wikipedia.org/wiki/Telugu_languages

[7] Brief history of the Telugu Language has been extracted from the web reference http://reference.findtarget.com/search/Telugu%20language/

[8] D. Artz, "Digital Steganography: Hiding Data within Data", IEEE Internet Computing, pp75-80, May-Jun 2001.

[9] M. T. Chapman, "Hiding the hidden: A software system for concealing cipher text as innocuous text", Master's thesis, University of Wisconsin-Milwaukee, May 1997.

[10] B. Vishnu Vardhan, L. Pradap Reddy and A. Vinay Babu, "A Model for Overlapping Trigram Technique for Telugu Script", Journal of Theoritical and Applied Information Technology, 2007.

[11] R Chandramouli, N. Menon, "Analysis of LSB Based Image Image Steganography Techniques", IEEE pp. 1019-1022, 2001.

[12] M. Hassan and Shirali Shaherza " A New Approach to Persian and Arabic Text Steganography", fifth IEEE/ACIS International Conference on Computer and Information Science(ICIS COMSAK'06), p.p., 310-315., July 2006.

9. AUTHORS PROFILE

G Lakshmeeswari, is working as Asst. Prof., GIT, Gitam University, Visakhapatnam, AP, India. Her research area is Security. She has 12 years of experience and is now doing her Ph.D in Steganography.

Dr. D Rajya Lakshmi, Prof & HOD, GIT, GITAM University Visakhapatnam, AP, India. Her research interests include image processing, Data mining, Security and Computer Networks. She has 19 years of teaching and research

Dr. D Lalitha Bhaskari, Associate Professor, Dept. of Computer Science and Systems Engineering, Collge of Engineering, Andhra University Visakhapatnam, AP, India. Her research interests include Data Security, Image Processing, Pattern Recognition, Steganography and Digital Watermarking. She has 12 years of teaching and research experience.