Decipherment of Substitution Cipher using Enhanced Probability Distribution

Apparao Naidu G
Research Scholar
Jawaharlal Nehru Technological
University, Kakinada

Bhadri Raju MSVS
Associate Professor in CSE
S.R.K.R Engineering College
Bhimavaram, AP, India

Vishnu Vardhan B
Professor in CSE
Jawaharlal Nehru Technological
University, Hyderabad, AP, India

Pratap Reddy L
Professor & Head of ECE
Jawaharlal Nehru Technological
University, Hyderabad, AP, India

ABSTRACT
Information theoretic approach for decipherment problems is the recent trend in cryptanalysis. The behavioral transformation of message units is addressed upto certain extent in the encryption process. However the amount of confusion and diffusion in terms of statistical distribution parameters between message and cipher text is a point of interest for cryptanalyst. In the present work we addressed this issue with the help of enhanced probability distribution function. The basic units of any message text are observed to be heuristic in nature depending on the sample. Averaging function is adopted while evaluating the enhanced probabilities of message units. The retrieved efficiency of cipher text only attack on samples of English, Hindi Telugu, Kannada is presented in this paper.

General Terms
Information Security, Cryptography, Cryptanalysis, Language model, Cipher text.

Keywords
Cryptanalysis, probability distribution, Conditional probability distribution, Enhanced probability distribution

1. INTRODUCTION
A cryptographic algorithm and a set of cryptographic keys are the basic elements of a security system. The degree of protection depends on several factors which include the strength of the cryptosystem, its implementation in hardware or software and possible keys. It is unlikely that the secrecy of an algorithm can be maintained for an extended period of time. A policy of keeping the algorithm public is to promote the widespread use as well as comparing the levels of cryptographic strength with other algorithms. The general assumption is that the opponent knows the cryptosystem that is under use. This approach assures the level of security which is dependent on the secrecy of the key only. Thus an essential factor in the design of a cryptographic system is focused on the length of the key, whose size places an upper bound.

Security analysis identifies several attack models [1], which include Cipher text-only attack, Chosen plain text attack, Known plain text attack and Chosen cipher text attack. The complex properties of natural languages play an important role in the formation of message text. A natural language text, in addition to statistical properties has another fundamental property, which may loosely be referred as meaning. For a meaningful message, it is possible to shorten it without destroying the meaning. This property is referred as redundancy. Shannon addressed [2] the redundancy of a language as its uncertainty. Every language is associated with certain amount of redundancy. Large amount of redundancy gives a way to cryptanalyst towards simple mechanisms of decipherment. The basic goal of the cryptanalysis is to determine the key. In case of cipher text-only attack, cryptanalyst is having infinite computational resources. Based on the apriori knowledge the analyst is able to rule out certain keys. But many "possible" keys may remain, out of which one is the correct key and the remaining are spurious keys. Elimination of spurious keys, while determining correct one demands for the knowledge from the statistical behavior of the message text.

The Computing power available today makes brute-force attacks against cryptographic systems less costly and simple. Virtually the existing key size is not offering much protection against brute-force attacks. Although systems allow electronic information to be encrypted using large keys, increase in computing power keep pushing up the size of keys. The increase in length of the key results in improved hardware and software complexity. To protect information adequately, in addition to key complexity, the type (language)of application need to be considered. A generic model applied on Latin text based applications need not be suitable for other languages. An approach for providing adequate security with smaller key size, which is dependent on the complexity of the language, need to be explored. The present work is addressed towards this direction.

2. DECIPHERMENT PROBLEMS

Substitution ciphers represent the basic building blocks of complex and secure ciphers that are most widely used today. Understanding the vulnerability of simple ciphers is important while building more complex ciphers. A large number of techniques [3] are available in the literature to break substitution ciphers, each of them having advantages and disadvantages over one another.

While attacking the cipher models, the goal is to derive the secret key or cipher text under the influence of known algorithm. Different techniques are explored [4, 5] in the literature to find the key, there by decrypting the entire cipher text. Several possible methods to break a substitution cipher include exhaustive search, simulated annealing, frequency analysis, genetic algorithm, particle swarm optimization, tabu search and relaxation algorithm etc.

Ryabko suggested [6] an attack on block ciphers called gradient statistical attack. Analysis of statistical properties of block ciphers is used in the process of cryptanalysis. Applicability of this method to the RC5 cryptanalysis is analysed. Automated attack algorithms are proposed [7] for which human intervention is largely minimized by exploring genetic algorithms. Joe Gester proposed [8] a search algorithm based on likely hood approach. The proposed Genetic algorithm involves an iterative process of finding the fitness of the individuals in the population. If this method is not satisfactory, then attempt is made to search a smaller problem space by restricting the key space to those which are generated by a keyword.

Algebraic cryptanalysis is a method [9, 10] in which cryptanalysis begins by constructing a system of polynomial equations in terms of plaintext bits, cipher text bits and key bits. The factors that play important role are the number of variables, the number of polynomials and the degrees of the polynomials while modeling the system of equations. Another factor is reported as the computing power of cryptanalyst.

Differential cryptanalysis uses probabilistic properties of block ciphers, in which cipher text pairs and their differential properties are explored. An alternative approach while exploring non-linearity in block ciphers is termed as linear cryptanalysis. Both methods are explored [11] extensively forming the foundation for cryptanalysis. Differential cryptanalysis is originally developed to attack Fast data Encipherment Algorithm (FEAL), and later proved its worth when applied to the cryptanalysis of the Data Encryption Standard (DES). One aspect of differential cryptanalysis that appears to be overlooked is the use of several differences to attack a cipher simultaneously.

The role of Cryptanalysis has another face of exploring the weaknesses of the system. Few approaches of cryptanalysis are reported in literature using language characteristics while understanding the strength of cipher system. One such approach deals with frequency statistics. A method is presented [12] for de-ciphering texts in Spanish using the probability of letters in the language. This method is proposed on a mono alphabetic cipher by assigning weights to different letter alphabets of Spanish language and achieved positive results. However 100% success is far reaching due to inconsistencies in the frequency distribution. Samuel W. Hasinoff presented [13] a system for the automatic solution of short substitution ciphers. The proposed system explored n-gram model of English and stochastic local search over all possible keys of the key space. This method resulted in the median of 94% cipher letters with an exact reverse map. Sujith Ravi et al. studied [14] attacking Japanese syllable substitution cipher, with the help of natural language models. They proposed several improvements over previous probabilistic methods, while achieving improved results.

From a natural language perspective, cryptanalysis task can be viewed as unsupervised tagging problem. Language Modeling (LM) techniques are used [15] to rank proposed decipherment. This work mainly attacks on difficult cipher systems that have more characters than English and on cipher lengths that are not solved by low-order language models. Language-model perplexity is related to decipherment accuracy. Jackobsen proposed [16] a method for cryptanalysis of substitution ciphers. In this method the initial guess of the key is refined through a number of iterations. In each step the recovered plain text is used to guess the closeness between the recovered key and the correct key. G W Hart proposed [17] a method which works well on a small sample of text where the probability distribution of letters is far from what is expected. This method reported with better performance on longer and easier cryptograms. However an exponential time is required in the worst case. This method fails under the condition that words in the plain text are in the dictionary.

A deciphering model is proposed by Lee [18] to automate the cryptanalysis of mono alphabetic substitution ciphers while exploring enhanced frequency analysis technique. In the decipherment process of mono alphabetic substitution cipher, monogram frequencies, keyword rules and dictionary are explored one after the other. Knight et al. described [19] a number of natural language decipherment problems that use unsupervised learning. These include letter substitution ciphers, phonetic decipherment, character code conversion and word-based ciphers with emphasis on machine translation. An efficient algorithm that accomplishes a naive application of the Expectation Maximization (EM) algorithm to break a substitution cipher is reported. Ravi and Knight explored [20] low-order letter n-gram model based on integer programming to search over the key space. This proposed method is reported with a study of decipherment accuracy as a function of n-gram order and cipher length.

Shannon proposal of entropy is extended [21] to evaluate the redundancy of a language. Entropy measures the average information produced on each letter of a text in the language where as redundancy measures the amount of constraint imposed on a text in the language because of its statistical nature letters. H Yamamoto [22] presented a survey on different information theoretic approaches in cryptology. The survey reported Shannon's cipher system, Simmons authentication approach, wire tape channel, secret sharing communication system approaches.

Diffie and Hellman introduced [23] another approach to achieve practical security based on computational complexity. Trap door functions and one way functions are explored. Borissov and Lee computed [24] bounds on the theoretical measure for the

strength of a system under known plain text attack. Zhang proposed [25] the key equivocation, which is the conditional entropy of the key given the cipher text and corresponding plain text is considered as a measure of strength of the system.

3. INFORMATION THEORETIC APPROCH

Entropy of a natural language is a statistical parameter that measures the average information of every letter or character in the respective language. It also indicates the uncertainty of various units in the message. Whenever the uncertainty reaches a maximum value, then the receiver has no information about the message that is transmitted. From the information theoretic approach entropy, otherwise visualized as uncertainty plays a vital role in cryptanalysis.

Entropy is evaluated using a series of approximations F_0, F_1, F_2, ……F_n, considering large samples of the language. The n-gram entropy F_n measures the entropy of n successive letters and is given by the expression (1)

$$F_n = -\sum_{i,j} p(m_i, j) \log_2 p(m_i, j) + \sum_i p(m_i) \log_2 p(m_i) \quad (1)$$

Where m_i is a block of n-1 successive letters
$p(m_i, j)$ is the probability of the n-gram m_i, j
j is any arbitrary number following the block m_i

In the above expression, Fn measures the average uncertainty of letter j with n-1 proceeding known letters. Considering long range statistics, the entropy H is given in equation (2)

$$H = \lim_{n \to \infty} F_n \quad (2)$$

For small values of n =1,2,3 in n-grams, H is evaluated using standard unigram, bigram and trigram probabilities. The value of F_0 is equal to $\log_2$ (A) where A represents size of the alphabets in the respective language. The value of A for Telugu is 64 (eliminating unused characters), where as it is 26 for English without spaces. The entropy of Indic scripts is relatively high than the English. Unigram, bigram, trigram entropies for Telugu are 4.70, 3.97 and 3.43 respectively, where as for Kannada these values are 4.76, 3.56, 3.30 and for Hindi 5.09,4.25, 3.47 respectively.

Based on the entropy of the language and the size of n-gram space the redundancy R_L is calculated using the expression (3)

$$R_L = 1 - H_L / \log_2 |P| \quad (3)$$

The redundancy for four different languages Telugu, Kannada, Hindi, English are calculated using above expression where the quantified figures are 19%, 19.2%, 34.6%,75% respectively.

4. UNICITY DISTANCE

Let R_L be the redundancy of the respective language, then for a string of sufficiently large cipher text of length n, the expected number of spurious keys (assuming equiprobable keys). 'S_n' satisfies the relation (4).

$$S_n = \frac{|K|}{|P|^{nR_L}} - 1 \quad (4)$$

Where |P| and |K| are the length of message space and key space respectively. The value of n at which the number of spurious keys S_n is zero is defined as unicity distance U. While Substituting S_n with zero in the above expression, and $n = n_0$

$n_0 \approx \log_2 |k| / (R_L \log_2 |P|)$

Therefore $U \approx H(K) / (R_L \log_2 |P|)$ (5)

U in equation (5) represents the average amount of cipher text needed for an opponent to uniquely determine the key.

As per Shannon's definition, unicity distance doesn't make any deterministic prediction, rather gives probabilistic results. A larger value of unicity distance is a probabilistic measure of increased strength of the crypto system. The above model is adopted for Telugu, Kannada, Hindi and English languages with different key sizes and the computed figures are listed in Table 1.

Unicity distance for English with a 128 bit key is observed as 19. Where as the same measure is found between key sizes of 56 and 64 for Indic scripts viz Telugu, Kannada and Hindi. The similar strength of the crypto system which uses message units of English with a help of 128 bit key size can be achieved with less than 64 bit key on Indic scripts. The above probabilistic measure is based on the assumption that the language units are equiprobability nature. In the real world scenario the statistical behavior of message units these no correlation with the above assumption. It is necessary to understanding and realize the behavioral patterns of message units of every language to strengthen the probabilistic prediction as described by Shannon. In the present work we addressed this issue using cipher text only attack in decipherment processes.

Table.1 Key size vs Unicity Distance

S. No	Key size in bits	Unicity Distance in characters			
		Telugu	Kannada	Hindi	English
1	40	13	14	11	6
2	56	18	19	15	9
3	64	20	22	17	10
4	80	26	27	21	12
5	128	41	43	34	19
6	256	82	87	68	38

5. DECIPHERMENT PROCESS USING CIPHER TEXT ONLY ATTACK

In a conventional cryptographic system, a plain text message 'm' is generated by the sender. An encryption transformation E, which depends on a secret key k, transforms the plain text 'm' to cipher text 'c' using the expression $c = E_k(m)$. Cipher text 'c' is then transmitted to decryption transformation D, which depends

on the secret key k (in case of symmetric key cryptography), which is used to recover the plain text 'm' using the expression $m = D_k(c)$. The assumption is that opponent does not possess 'k' and cannot recover 'm' from 'c' using D. (Note that algorithms D and E are public). For the key 'k' to remain secret, a secure communication channel is needed between the sender and receiver.

The information available for a cryptanalyst is a variable that can be only cipher text, the knowledge of the system (except the key), the algorithm used, the characteristics of the language and language statistics. If 's' represents the information available to an opponent and D_1 represents the process of cryptanalysis, then the deduced information 'm_1' is expressed as $m_1 = D_1(s)$. The coincidence between 'm' and 'm_1' is considered as a measure of strength.

Though the key is generated randomly, since it is fixed for entire message, the mapping function in the encryption algorithm transforms it into a distinct point set in the orthogonal plane. On many occasions almost all characters are present for large size of text. The efficiency of predictable text in the decipherment process is mainly dependent on the complexity of the language.

A general decipherment model to recover plain text from cipher text is illustrated in Fig-1. The plaintext 'm' is enciphered using algorithm E, and a secret key k, resulting in cipher text c which is $E_k(m)$. Deciphering 'c' using algorithm D and the key 'k' results in plain text 'm'. The opponent's initial information is a variable that can be as little as only cipher text or as much as complete knowledge of the system (except for the key). If 'y' represents the apriori knowledge and language statistics available to an opponent and D^1 represents the process of cryptanalysis, then the deduced information, m^1 can be expressed as $m^1 = D^1(y)$. The equivalence between m and m^1 can be taken as a performance measure for the strength of the system.

The statistical parameters like apriori probability distribution of primitive units, characters, words, phrases, sentences etc and the inter relation among these units is the main focus of interest for cryptanalyst. The conditional probability distribution of the above primitive units is additional knowledge of redundancy. In the present work 'm^1' is derived with the help of mapping function between 'c' and 'm' using the above apriori distribution. Nearest neighborhood approach is used as a mapping function while arriving at a decision.

In the present work, we attempted unigram code point probability distribution as apriori knowledge. Many researchers quantified the probabilities of individual code points by increasing the sample size. The code point probability distribution of English, Telugu, Hindi, Kannada is evaluated on sample sizes of 1000000, 3200000, 900000, 1600000 respectively. The quantified figures of these samples may vary with increasing sample size due to heuristic nature of language which is represented in Fig-2. Nine independent characters of English language and their probabilities on various samples varying from 1000 to 110000 are presented above. The similar computations with associated results of code points with regard to Telugu language are observed with the similar nature which can be depicted from the above figure. This nature results in inconsistent variation in the mapping process. We consider average probability distribution as an enhancement while computing the probability of individual code points.

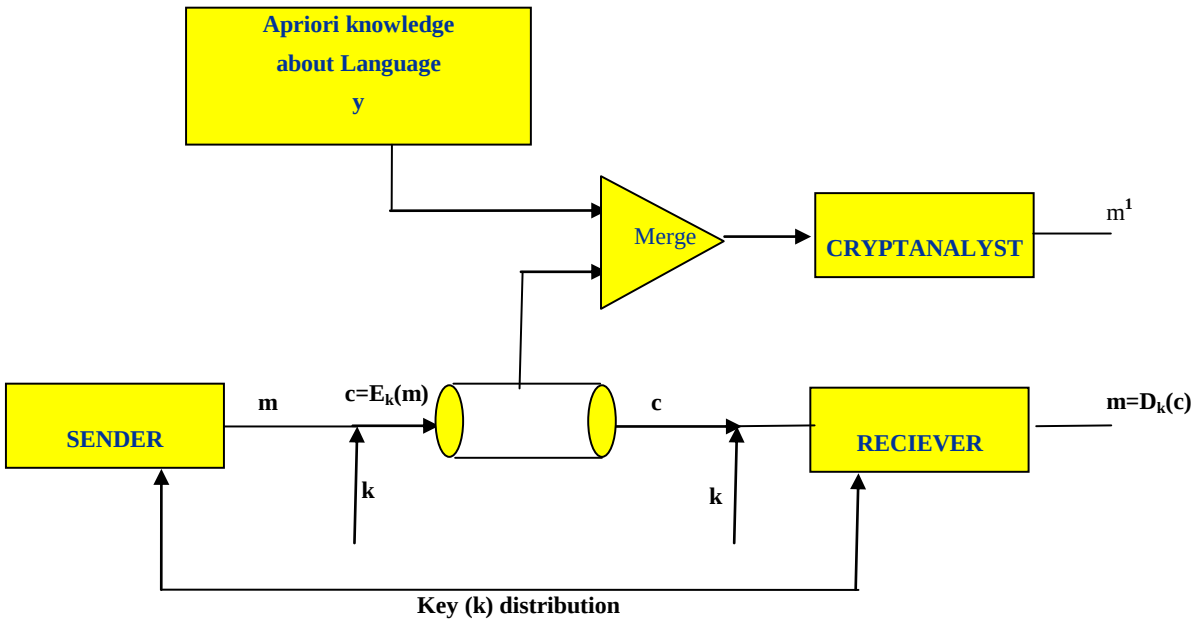
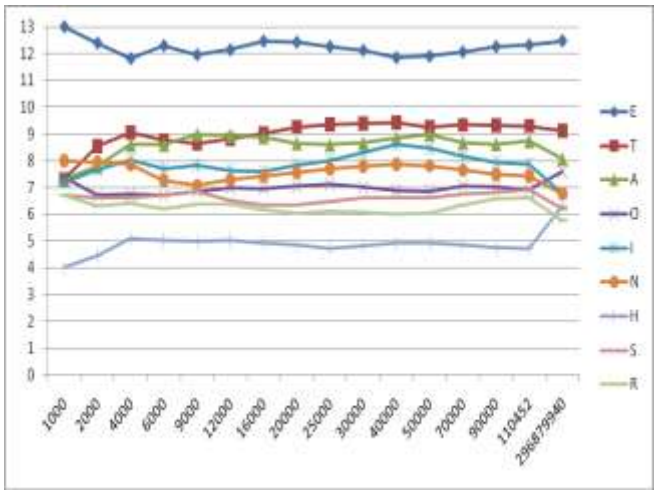
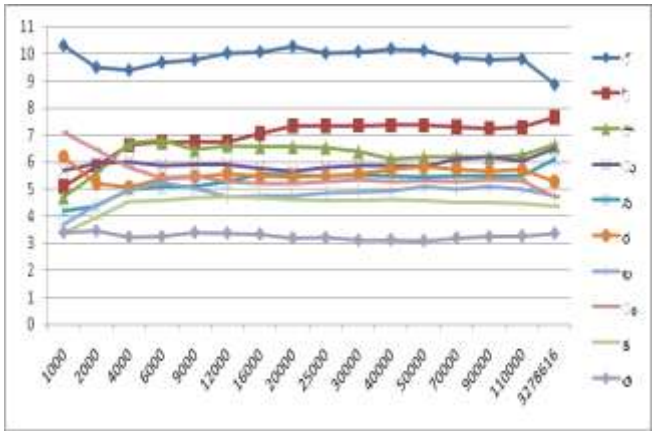


Fig 1. Decipherment Model



a) Probability distribution of 9 independent alphabets of English



b) Distribution of 10 code points of Telugu

Fig.2 Heuristic nature of character code points on various samples

Let p_1, p_2….. p_n represent the probability of code point C in a different samples. Then the average probability distribution of C represented by $p_{avg}(C)$ is given by

$$p_{avg}(C) = (p_1 + p_2 + \ldots\ldots p_n)/n.$$

5. RESULTS

The coincidence between 'm' and 'm^{1}' is treated as retrieval efficiency for the given sample. The probability distribution for the transformed code points is extracted from the text size of 1000. The coincidence is identified with the help of nearest neighborhood technique between the probability distribution of 'm' and 'm^{1}'. This process is extended on 15 different samples of varying size from 1000 to 110000. The probability distribution of code points and the associated enhanced probability is adopted as apriori knowledge in decipherment process. The results are presented in Table.2. Similar computation is carried out using conditional probability of every code point given an independent code point. The pair wise distribution of code points probability and the associated conditional probability is used as apriori knowledge while performing mapping function. The mapping outcome in terms of retrieval efficiencies of conditional probability distributions are presented in Table.2

Retrieval efficiency of English is observed to be relatively larger than the Indic scripts viz Telugu, Kannada and Hindi. In case of probability distribution of character code points the minimum retrieval efficiency of English text is observed as 24.90% with a text size of 2000. In case of Telugu the minimum efficiency is 04.40% associated with text size of 1000 code points. The same is observed as 23.50% with Kannada language. Where as it is 17.72% for Hindi for the text size of 9000.

In case of enhanced distribution, more than half of the eight samples are observed with over and above 65% retrieval efficiency. Quite interestingly the similar phenomenon is observed with Telugu samples where the retrieval text is over and above 50% only. The outcome of Kannada is around 45%. Where as Hindi is observed around 55%.

Adaptation of conditional probability allowed appropriated coincidence with 'E', 'T', 'G', 'U' character code points, which is not the case in the matching pattern of probability alone. Similar improvements are observed with other languages where the number of code points are relatively large.

The conditional probability and the associated enhanced probability of English alphabets results in average retrieval efficiencies of 48% and 54% respectively. The quantified result observed with improved text retrieval with the help of conditional probability distribution. Similar observations are found with other languages also. However the improvement is limited to a maximum of 8% in case of English and Telugu where as its only 2% in case of Hindi. The computed results are the reflective measures of language redundancy which is associated with unicity distance as mentioned in the Table.3.

6. CONCLUSION

Information theoretic approach, proposed by Shannon is the basis for the present work. Unicity distance is considered as measure of index while determining the strength of algorithm associated with language. Apriori knowledge in the form of code point distribution as well as redundancy associated with conditional probability is evaluated from various text samples. A decipherment model is proposed for retrieving the plain text from cipher text using the above knowledge. English language is observed with higher level of redundancy when compared with Indic scripts. The maximum average retrieval efficiency of 54% is observed with English text in case of enhanced conditional probability distribution. Minimum retrieval efficiency is observed with Telugu text. However Hindi, Kannada and Telugu languages are posses relatively same degree of complexity. Extension this model for higher levels of language statistics is in progress.

Table.2 Retrieval efficiency of message text using Probability distribution & Enhanced distribution

Sl. No	Text	English		Telugu		Kannada		Hindi	
	Size of plain text	Probability	Enhanced Probability	Probability	Enhanced Probability	Probability	Enhanced Probability	Probability	Enhanced Probability
		Retrieval %	Retrieval %	Retrieval %	Retrieval %	Retrieval %	Retrieval %	Retrieval %	Retrieval %
1	1000	33.20	32.10	04.40	22.20	23.50	37.30	37.80	26.40
2	2000	24.90	42.55	05.70	39.35	31.45	38.25	38.50	28.60
3	4000	31.50	66.00	17.10	52.42	32.07	44.00	19.12	38.92
4	6000	29.97	67.95	22.51	41.23	27.86	38.03	19.88	37.30
5	9000	35.41	69.66	22.15	37.02	36.79	51.42	17.42	58.68
6	12000	35.93	73.84	26.68	42.92	27.96	47.61	31.55	53.92
7	16000	38.03	67.08	18.09	44.14	32.64	49.11	27.71	56.99
8	20000	35.01	43.94	24.41	50.53	39.43	47.09	29.74	56.37
9	25000	40.40	37.83	23.77	56.48	38.95	66.68	23.83	54.58
10	30000	31.22	65.00	22.28	54.59	33.93	44.61	26.38	56.04
11	40000	42.34	44.81	16.90	70.13	26.73	40.57	27.10	59.24
12	50000	38.33	52.91	08.81	50.68	28.03	35.65	27.89	59.95
13	70000	63.17	31.03	23.60	54.46	26.11	43.01	26.31	56.97
14	90000	48.04	50.38	17.44	56.59	25.58	36.19	25.93	51.43
15	110000	78.59	31.02	18.76	56.84	25.70	36.43	27.22	43.89
	Avg	40.40	51.74	18.17	48.64	30.45	43.73	27.09	49.28

Table. 3 Retrieval efficiency using conditional Probability and Enhanced distribution

Sl. No	Text	English		Telugu		Kannada		Hindi	
	Plain text size	Conditional Probability	Enhanced Probability	Conditional Probability	Enhanced Probability	Conditional Probability	Enhanced Probability	Conditional Probability	Enhanced Probability
		Retrieval %	Retrieval %	Retrieval %	Retrieval %	Retrieval %	Retrieval %	Retrieval %	Retrieval %
1	1000	36.44	38.44	08.61	22.12	29.83	38.94	18.01	38.24
2	2000	38.17	51.07	17.81	29.66	34.07	42.87	25.51	32.32
3	4000	48.39	57.24	23.65	45.74	38.68	45.24	27.33	24.28
4	6000	42.42	56.26	24.72	43.99	29.77	49.58	31.15	33.24
5	9000	48.23	71.51	24.97	41.32	38.05	53.18	26.35	38.93
6	12000	49.65	72.85	27.29	38.24	35.94	46.32	34.71	54.71
7	16000	41.81	65.05	33.89	47.95	42.10	39.48	33.05	60.95
8	20000	47.89	48.49	32.86	50.73	48.42	47.83	27.38	53.48
9	25000	52.76	53.86	27.99	53.28	38.30	57.76	25.00	50.69
10	30000	50.60	53.05	27.27	46.35	39.88	54.02	24.91	51.97
11	40000	48.76	49.59	26.92	52.65	36.07	47.83	25.74	47.44
12	50000	43.42	45.47	26.46	42.80	34.88	44.43	31.33	53.94
13	70000	56.33	44.23	33.93	42.30	30.69	37.79	31.01	48.38
14	90000	57.85	57.26	33.13	38.11	30.33	35.73	32.67	39.93
15	110000	57.35	48.60	29.88	37.58	31.31	36.16	34.55	38.26
	Avg	48.00	54.20	26.62	42.19	35.89	45.14	28.58	44.45

7. REFERENCES

[1] Michael J. Wiener.: The Full Cost of Cryptanalytic Attacks. J. Cryptology Vol. 17, pp 105–124 (2004).

[2] C. E. Shannon.: A Mathematical Theory of Communication. The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656 (1948).

[3] Francois-Xavier Standaert, Gilles Piret, Jean-Jacques Quisquater. : Cryptanalysis of Block Ciphers: A Survey. UCL Crypto group Technical report, CG2003/2 pp. 1-25 (2003).

[4] B. Carter and T. Magoc.: Classical Ciphers and Cryptanalysis. Technical Report (2007).

[5] Nalini N.: Cryptanalysis of Block Ciphers via improved Simulated Annealing Technique. In proceedings of 9th International Conference on Information Technology (ICIT'06), pp 182-185 (2006).

[6] Nalini N.: Cryptanalysis of Block Ciphers via improved Simulated Annealing Technique. In proceedings of 9th International Conference on Information Technology (ICIT'06), pp 182-185 (2006).

[7] Albassall A.M.B., Wahdan A.: Genetic Algorithm cryptanalysis of a Fiestal type block cipher. Proceedings of ICEEC '04, pp. 217-221(2004).

[8] Joe Goester.: Solving Substitution Ciphers with Genetics Algorithm. A Technical Report submitted to university of Rochester, pp 1-8 (2003).

[9] Carlisle Adams.: Designing against a class of algebraic attacks on symmetric block ciphers. J. Applicable Algebra in Engineering, Communication and Computing Vol. 17, pp. 17–27 (2006).

[10] Sean Simmons.: Algebraic Cryptanalysis of Simplified AES. In J. Cryptologia, Vol. 33, pp 305–314 (2009).

[11] Ali Aydın Selçuk.: On Probability of Success in Linear and Differential Cryptanalysis. In J. Cryptology Vol. 21, pp. 131–147 (2008).

[12] Bárbara E. Sánchez Rinza, Diana Alejandra, Bigurra Zavala, Alonso Corona Chavez.: De-encryption of a text in spanish using probability and statistics. In proceedings of 18th International Conference on Electronics, Communications and Computers IEEE 2008, pp 75-77 (2008).

[13] Samuel W. Hasinoff.: Solving Substitution Ciphers. A Technical Report, University of Toronto (2003).

[14] Sujith Ravi and Kevin Knight.: Probabilistic Methods for a Japanese Syllable Cipher. In proceedings of the 22nd International Conference on Computer Processing of Oriental Languages, Lecture Notes in Artificial Intelligence Vol. 5459, pp. 270-281 (2009).

[15] Kevin Knight, Anish Nair, Nishit Rathod.: Unsupervised Analysis for Decipherment Problems. In proceedings of the COLING/ACL ,pp 499-506 (2006).

[16] Thomas Jacobsen.: A fast method for the crypt analysis of substitution ciphers. In J.Cryptologia Vol.19, Issue 3, pp 265-274 (1995).

[17] G.W.Hart.: To decode short cryptograms. In communications of ACM, Vol. 37, Issue 9, pp. 102-108 (1994).

[18] K.W. Lee, C.E. Teh, Y.L. Tan.: Decrypting English Text using enhanced frequency Analysis. In proceedings of National Seminar on Science, Technology and Social Sciences 2006 pp. 1-7 (2006)

[19] Sujith Ravi and Kevin Knight.: Attacking Decipherment Problems Optimally with Low-order N-gram Models. In proceedings of the conference on Empirical Methods in Natural Language Processing, pp. 812-819 (2009)

[20] Sujith Ravi, Kevin Knight.: Attacking Letter Substitution Ciphers with Integer Programming. In J. Cryptologia Vol. 33, Issue 4, pp. 321-334 (2009)

[21] C. E. Shannon.: Prediction and Entropy of Printed English. In the Bell System Technical Journal pp. 53-64(1951)

[22] H Yamamoto.: Information theory in cryptology. J. IEICE transactions, Vol. E 74, No.9, pp. 2456-2464 (1991)

[23] Martin E Hellman.: An Extension of the Shannon Theory Approach to Cryptography. IEEE Transactions on Information Theory, Vol. IT-23, NO. 3, pp. 289-294 (1977)

[24] Yuri Borissov and Moon Ho Lee.: Bounds on Key Appearance Equivocation for Substitution Ciphers. IEEE Transactions on Information Theory Vol. 53, No.6, pp. 2294-2296 (2007).

[25] Zhaozhi Zhang.: A Simplified Method for Computing the Key Equivocation for Additive-Like Instantaneous Block Encipherers In J. Electronic Notes in Discrete Mathematics, 21 pp. 389–391(2005)

[26] M.S.V.S.Bhadri Raju et al.: A Noval Security Model for Indic Scripts – A Case Study on Telugu. International Journal of Computer Science and Security (IJCSS, Malasya), Vol. 3, Issue 4, pp. 303-313 (2009)

[27] M.S.V.S.Bhadri Raju et al. : Effect of Language complexity on Deciphering Substitution Ciphers - A case study on Telugu. International Journal of Security and its applications (IJSIA), Vol. 4 , Issue 1, Science and Engineering Research Society(SERSC), Korea, pp. 11-20 (2010)