Arabic Language Text Steganography Based on Singular Value Decomposition (SVD)

Dr. Hanaa M. Ahmed
Computer Science Department,University of Technology/ Baghdad.
Email:salmanhanna2007@yahoo.com
Maisa'a A. A. Khohder
Computer Science Department, University of Technology/ Baghdad.
Email:maisaa.ali2007@yahoo.com

Received on:13/9/2015 & Accepted on:22/6/2016

ABSTRACT

With the fast development of internet innocent over communication in the network environment has become an important research direction. Steganography means that secret information is embedded into cover data imperceptibly for transmission. Linguistic Steganography covers all the techniques that deal with using written natural language to hide the secret message. This paper, presents a linguistic steganography for Arabic language texts, using Kashida and Fast Fourier Transform on the basis of using a new technique entitled Random Singular Value Decomposition Image as a location to hide a secret message. The proposed approach is an attempt to present a transform linguistic steganography using levels for hiding to improve implementation of kashida, and improve the security of the secret message by using Random Singular Value Decomposition Image. The proposed algorithm achieves typical steganography properties such as capacity, security, transparency, and robustness.

Keywords: Arabic text, Linguistic Steganography, random Singular Value Decomposition, Kashida, Transform Based

INTRODUCTION

Linguistic steganography is focused on applying changes to a cover text so as to embed a secret message, in a way that the changes do not cause any unnatural or ungrammatical text. According to cover, text steganography can be categorized into three groups [1], as depicted in Figure (1), [2]:

1. Format-based method: uses and changes the formatting of the cover-text to hide data [2]. "It involves altering physically the format of text to conceal the information. This method has certain flaws. If the stego file is opened with a word processor, misspellings and extra white spaces will get detected. Changed fonts sizes can arouse suspicion to a human reader" [3]. Format- based methods usually modify existing text for hiding the steganography text, by insertion of space or non-displayed space, as depicted in Figure (2), [4].
2. Random and statistical generation: methods are used to generate cover-text automatically according to the statistical properties of language [2]. "In order to avoid comparison with a known plaintext, steganographers often resort to generating their own cover texts [5], [6]. One method is concealing information in a random sequence of characters. Character sequences method hides the information within character sequences "[5]. 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. A PCFG 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

https://doi.org/10.30684/etj.34.5B.2
2412-0758/University of Technology-Iraq, Baghdad, Iraq

hidden in it". The quality of the generated stego-message depends directly on the quality of the grammars used [2].

3. Linguistic steganography: "specifically considers the linguistic properties of generated and modified text, and in many cases, uses linguistic structure as the space in which messages are hidden. Context Free Grammer(CFG) creates tree structure which can be used for concealing the bits where left branch represents '0' and right branch corresponds to '1'. A grammar normal form (GNF) can also be used where the first choice in a production represents bit 0 and the second choice represents bit 1 [4]. This method has some drawbacks. First, a small grammar will lead to a lot of text repetition. Secondly, although the text is syntactically flawless, there is a lack of semantic structure "[4], [5].

Linguistic Steganography is focused on making changes to a cover text in order to embed information, in such a way that the changes do not result in ungrammatical or unnatural text. Most of the linguistic steganography methods use either syntactic transformations or lexical (semantic) or the combination of both [5], [7].

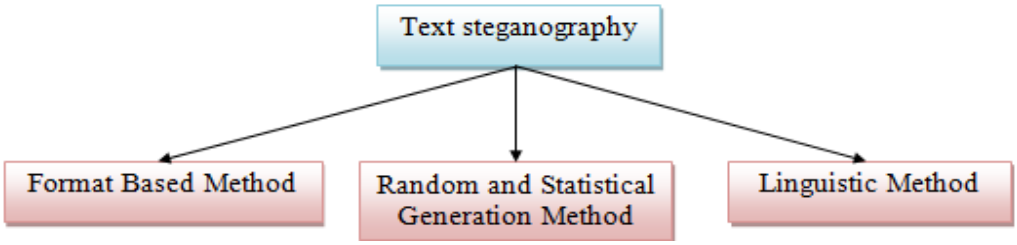

Figure (1): The Types of Text Steganography [1].

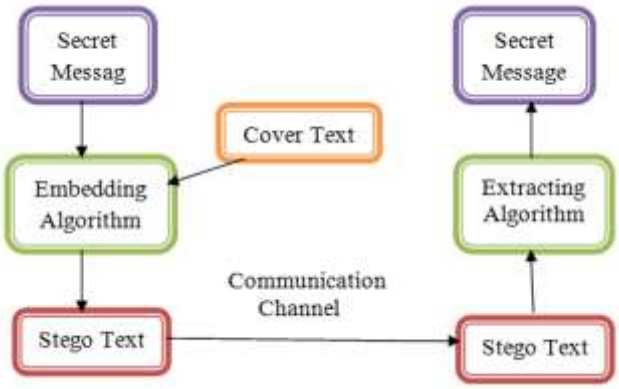

Figure (2): Mechanism of Text Steganography [5].

This paper, proposes layered steganography technique for Arabic language text using Fast Fourier Transform (FFT) and kashida. The proposed approach uses Singular Value Decomposition Image (SVD) to generate random location, to embed the secret message bits using FFT and kashida as a first layer followed by adding kashida characters randomly as the second layer. The proposed algorithm typical steganography properties are capacity, transparency, robustness, and security of the secret message for Arabic text based secure communication.

The other sections of the paper are structured as follows: Section II presents the literature review for kashida based linguistics steganography and explain fundamental used of proposed system. Section III explain the algorithm for proposed system and results and discussions are done in section V, and IV deals with the conclusion.

Literature Review and Fundamentals used in The Proposed System Literature Review

Kashida is a redundant Arabic character which is used to justify the text, without affecting the meaning of words. Researchers are suggested using one kashida as bit zero, and two kashida as bit one, or vice versa.

In 2007, A. Gutub, and M.Fattani, introduced a novel Arabic text steganography technique for Arabic script using letter points and kashida. The technique hides secret information as bits in Arabic letters (cover) by using kashida and points of letters. The technique considers un-point Arabic letters followed by a kashida if the secret bit is (0), and point Arabic letters followed by kashida if secret bit is (1).

Their technique enhanced robustness and security but might have some limitations with capacity of the cover media if the number of secret bits of the secret information is large. This steganography technique is found to be suitable for other languages having similar script to that of Arabic for example Persian and Urdu [5].

In 2009, A. H. Fahd, et al., introduced improving security, and capacity for Arabic text steganography using kashida. The approach hides secret information as bits within Arabic letters (cover) with kashida using three scenarios. The approach discusses maximum number of kashida letters that can be added to the Arabic cover word. Also the researchers evaluated the number of hidden bits that can be embedded in the carrier file and compared the results with diacritics, and kashida methods [6].

In 2010, Adnan Abdul-Aziz Gutub, et al., introduced an improved Arabic text steganography technique for Arabic text using kashida. The approach hides secret information as bits within Arabic letters (cover) by using extension character (kashida). The technique considers one kashida if the secret bit is (0) and two kashidas if the secret bit is (1) after any letter which can hold it. The finishing character is embedded just after the last bit of the secret information, then the kashida is embedded randomly to the remaining text in order to enhance the security of the technique. Also their technique enhanced security, capacity and robustness for Arabic texts based on secure communication [7].

In 2010, A. Ali and F. Moayad, introduced Arabic text steganography technique for Arabic text using kashida with Huffman code. The approach hides secret information as bits within Arabic letters (cover) by using extension character (kashida), and compressed the stego file using Huffman code. The technique considers absence of kashida if the secret bit is (0) and one kashida if secret bit is (1) after any connected letters. Also their technique is applied to other Arabic texts than that based secure communication, with different document formats [8].

In 2013, Ammar Oden, et al., introduced an improved Arabic text steganography technique for Arabic text using variation in kashida. The approach selects one of four scenarios randomly to hide secret information embedded as bits within Arabic letters (cover) by using kashida. The technique considers un-point Arabic letters followed by a kashida if the secret bit is (0), and point Arabic letters followed by kashida if secret bit is (1) as first scenario , and vice versa as second senior. The third scenario is adding kashida after Arabic letters if the secret bit is (1) and (0) otherwise, and vice versa as fourth scenario. Also their technique enhanced security, complexity for Arabic text based secure communication [9].

Fast Fourier Transform and its Inverse

The mathematical formula of Fourier Transform of a time domain function f(x), for real numbers x and y is [10], [11]:

$$F(y) = \int_{-\infty}^{+\infty} f(x)\exp^{[-i2\pi xy]}\,dx \qquad \ldots\ldots\ldots (1)$$

And the mathematical formula of its inverse is [10], [11]:

$$f(x) = \int_{-\infty}^{+\infty} F(y)\exp^{[j2\pi xy]}\,dy \qquad \ldots\ldots\ldots (2)$$

where:
f(x) : Time domain function
F(y): Frequency domain function
x: Argument with units of time
y: Argument with units of frequency
e: Base of natural logarithms
i: Imaginary unit ($i^2 = -1$).

Singular Value Decomposition (SVD)

Singular Value Decomposition technique splits given matrix into a product of orthonormal matrices and a diagonal matrix. The mathematical formula of Singular Value Decomposition is [12].

$$A=USV^T \quad \text{..... (3)}$$

$$A= [u_1\ u_2 \ldots\ldots u_m] \begin{pmatrix} S_1 & 0 & \ldots. & 0 \\ 0 & S_2 & \ldots\ldots & 0 \\ . & 0 & . & 0 \\ . & & 0\ . & \\ 0 & \ldots. & & S_m \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \\ . \\ . \\ V_n^T \end{pmatrix}, \quad \text{.....(4)}$$

Let A be an $m \times n$ matrix. Performing SVD on A factorizes it into a product of orthogonal matrix, diagonal matrix and another orthogonal matrix as:

$$A=USV^T \quad \text{..... (5)}$$

where,
A: original image matrix
U: $m \times m$ product of orthogonal matrix
S: $m \times n$ diagonal matrix
V: $n \times n$ orthogonal matrix

Random Singular Value Decomposition (RSVD)

It is a new technique to generate a set of random positions (x_i, y_j) to apply the embedding algorithm. From decomposing original image (A) using SVD, the result is B. Detecting the non-zeros elements, and converging into nearest integer results in RSVD. Figure (3) is an example of the original image [13].

Figure (3): Original Image Example.

The original image A is decomposed into three matrixes: U of size $m_i \times m_j$ matrix, V of size $m_i \times n_j$ matrix, and D of size $n_i \times n_j$ matrix. The new array B = S*V*D as depicted in Figure (4), indicates random location from original image A.

Figure (4): Random Location Generated Using Rsvd.

Algorithm RSVD:
Input: original image A Output: random number Process: Step1: input A Step2: Apply SVD algorithm to A Step1: $B = U*S*V^T$ Step2: For i= 1 to Length of B Step3: No= B[i]; Step4: No= absolute (No); Step5: While (integer (No)==0) Step6: No =No*10; Step7: RSVD[i]=(integer (No)); Step8: Next End of algorithm

Proposed System

Idea of Proposed System

The proposed approach main idea as depicted in Figure (5) is the embedding, and Figure (6) the extraction. Is to use RSVD as a generated random offset location, to add random kashida characters to the remaining Arabic word texts as a second layer, where the first layer inject the secret message bits in the inverse FFT (LSB of (real (FFT) of selected Arabic text word)), and then one kashida character is applied. The first addition of kashida is for the hiding process of the secret information, while the second addition of the kashida is for confusion purpose to ensure security of the secret message.

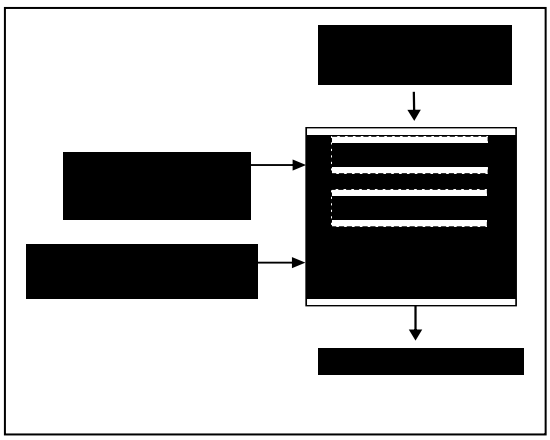

Figure (5) The Proposed Hiding Process.

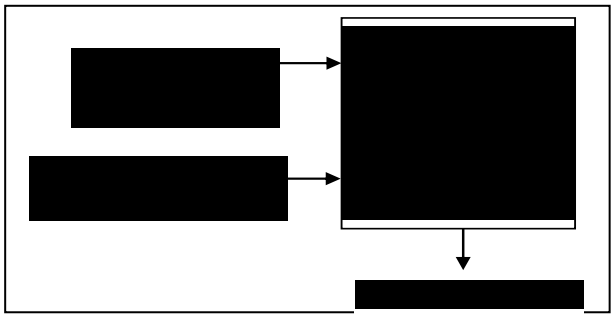

Figure (6) The Proposed Extraction Process.

Embedding process

Embedding Algorithm:
Input: secret message, image A, set of Arabic texts.
Output: stego-cover.
Process:
Step1. Secret message binarization: The secret message is hidden in the form of (0) s, and (1) s, which represent (64) bit Unicode of each character using the hexadecimal representation. N, is the total number of secret message bits. Figure (7) presents the binarization process to secret message. Figure (8) is a simple example of applying binarization process to secret message.
Step2. Generate Random positions: The process of generating random positions, using RVSD, starts by applying SVD algorithm to the input image (A) to generate a sequence of random values C that represent offset of Arabic text words to start the embedding process. The total number of Generate Random positions is(N), where N, is the total number of secret message bits.
Step3. Cover selection: select Arabic text (cover) that can hold input secret message bits.
Step4. Do while not end of Arabic text words
Step5. Embedding layer one: For each secret message bit and Generate Random positions do
Step6. Use C value as offset to next word to embed the secret message bit, into inverse FFT (LSB (real (FFT
(select Arabic text word)))), then apply one kashida if the secret message bit is one or if the secret
message bit is zero.
Step7. End of For.
Step8. Else
Step.9 Embedding layer two: add kashida characters randomly to the remaining Arabic text words
Step10. End of Do.
Step11. End.

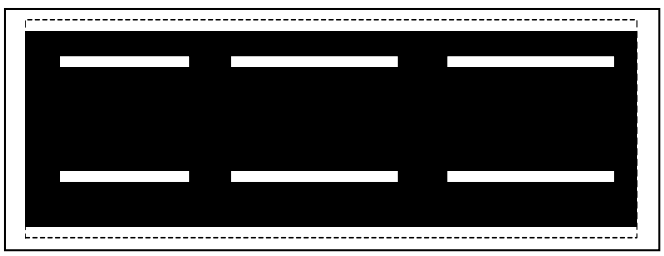

Figure (7): Secret Message Binarization.

Secret Message	---------- Arabic Text ---------- الجمع بين الماء والنار في يديب أصعب
Hexadecimal Representation	---------- Hexa Decimal ---------- 20C7FBADEFDF2020A9FDF22020C7FBEFC7C12020E8C7FBF2C 7D12020BAFD2020FDCFFDA92020C3BEDFA9
Binary Representation	00100000011000111111111011101011011110111111011111010 00000010000010101001111111101111100100010000000100000 11000111111110111110111111100011111000001001000000010 00001110100011000111111110111111001011000111110100001 00100000001000000101110101111110100100000001000001111 11011100111111111110110101001001000000010000011000011 10111110110111111010101001

Figure (8): Secret Message Binarization Example.

Extraction Process

Extraction Algorithm:
Input: secret message, image A, stego cover.
Output: secret message.
Process:
Step1. Generate Random positions: The process of generating random positions, using RSVD, starts by applying SVD algorithm to the input image (A) to generate a sequence of random values C that represent offset of Arabic text words to start the extraction process.
Step2. Loading: Load stego-cover, and Generate Random positions.
Step3. For each Generated Random Positions do
Step4. Use C value as offset to next word to extract the secret message bit, from LSB of select Arabic text word (stego-cover).
Step5. End of For.
Step6. Convert each seven bits into one letter, the result is the secret message.
End.

Results and Discussion

This section discusses cases to ensure the proposed technique security:
Case one: An example of result of applying the proposed technique using embedding layer one, as depicted in Figure (9).

Stego-cover	ملوم كما يجل عن العالم ووقع فعاله فوق الكلام ذراني والقلاة بلا دليل ووجهي والهجير بلا لثام فإني أستريح بذي وهذا وأتعب بالإناخة والمقام عيون رواحلي إن حرت عيني وكل بغام رازحة بغامي فقد أرد المياه بغير هاد سوى عدي لها برق الغمام يذم لمهجتي ربي وسيقي إذا احتاج الوحيد إلى النمام ولا أمسي لأهل البخل ضيفا وليس قرى سوى مخ النعام ولما صار ود الناس خبا جزيت على ابتسام بابتسام وصرت أشك فيمن أصطفيه لعلمي أنه بعض الأنام ما في المقام لذي عقل وذي أدب من راحة فدع الأوطان واغترب سافر تجد عوضا عمن تفارقه وانصب
Secret message	[illegible]
RSVD	
Stego-cover	ملوم كما يجل عن العالم ووقع فعاله فوق الكلام ذراني والقلاة بلا دليل ووجهي والهجير بلا لثام فإني أستريح بذي وهذا وأتعب بالإناخة والمقام عيون رواحلي إن حرت عيني وكل بغام رازحة بغامي فقد أرد المياه بغير هاد سوى عدي لها برق الغمام يذم لمهجتي ربي وسيقي إذا احتاج الوحيد إلى النمام ولا أمسي لأهل البخل ضيفا وليس قرى سوى مخ النعام ولما صار ود الناس خبا جزيت على ابتسام بابتسام وصرت أشك فيمن أصطفيه لعلمي أنه بعض الأنام ما في المقام لذي عقل وذي أدب من راحة فدع الأوطان واغترب سافر تجد عوضا عمن تفارقه وانصب

Figure(9): The Proposed Technique Of Embedding In Layer One.

Cover	الم المتنبي الواقع السياسي الممزق والمهين للعرب في عصره فالدولة مجموعة دويلات الحمدانية بحلب والااخشيدية بمصر والبويهية بفارس والفاطمية بالمغرب والاموية ثم العامرية بالاندلس
Stego-cover using first layer	الم المتنبي الواقع السياسي الممزق والمهين للعرب في عصره فالدولة مجموعة دويلات الحمدانية بحلب والااخشيدية بمصر والبويهية بفارس والفاطمية بالمغرب والاموية ثم العامرية بالاندلس
RSVD	
Stego-cover using proposed technique	الم المتنبي الواقع السياسي الممزق والمهين للعرب في عصره فالدولة مجموعة دويلات الحمدانية بحلب والااخشيدية بمصر والبويهية بفارس والفاطمية بالمغرب والاموية ثم العامرية بالاندلس

Figure(10): The Proposed Technique Of Embedding In Layer Two.

It can be concluded from case two that is visually not easy to find the locations of secret message that is embedded in stego-cover.

CONCLUSION

In this paper a new layer of Arabic language steganography is implemented using the FFT implementation and Kashida as an embedding process, and RSVD as random location generator to embed the Arabic text message in the Arabic text. Some conclusions are presented below:

1. Applying Steganography methods to document (text) files as a cover which is written in Arabic language is difficult, due to the visual sensitivity of Arabic letters to any manner change as in case one.
2. The RSVD is a fast search algorithm, which is improved be used as a means to allocate randomly positions in the cover media (Arabic texts) to perform the embedding operation.
3. Like embedding methods, usually frequency method is harder against attack than time domain method, so using FFT and Kashida as embedding method, improves its security against attack.
4. Algorithm robustness: The proposed algorithm prohibits any change in carrier (Arabic text) during the transmission process since the hidden secret message does not change the cover (Arabic text) file properties such as file size, content, and format during the transmission.
5. Algorithm transparency: The proposed algorithm improves the transparency property by hiding secret message inside the Arabic text using FFT. In addition another layer of hiding is applied using Kashida.
6. Algorithm security: The proposed algorithm improves the security property by hiding secret message inside the Arabic text using FFT and applying kashida as the first layer then applying kashida as a second layer to the remaining Arabic text.

REFERENCES

[1] Salman Hana'a M., " A Natural Language Steganography Technique for Text Hiding Using LSB's", Eng.&Tech. Vol.26,No3,2008.

[2] Hu Xiaoxi, Luo Gang, Yongjing Lu, and Lingyun Xiang, "A Steganography on Synonym Frequency Distribution", Advances in information Sciences and Service Sciences(AISS), Vol.5, no.10, May 2013.

[3] Ching − Yun Chang, and Stephen Clark, "Adjective Deletion for Linguistic Steganography and Secret Sharing", Proceedings of Coling 2012: Technical Papers, pages 493–510, Mumbai, December 2012.
[4] M. K. Kaleem,"An Overview of Various Forms of Linguistic Steganography and Their Applications Protecting Data", Journal of Global Research in Computer Science, Volume 3, No. 5, May 2012.
[5] Adnan Abdul-Aziz Gutub, and Manal Mohammad Fattani," A Novel Arabic Text Steganography Method Using Letter Points and Extensions", International Journal of Computer, Information, Systems and Control Engineering Vol : 1, No:3, 2007.
[6] A.-H. Fahd, G. Adnan, A.-K. Khalid, and H. Jameel, "Improving Security and Capacity for Arabic Text Steganography Using 'Kashida 'Extensions", the IEEE/ACS International Conference on Computer Systems and Applications, 2009.
[7] E. Nyein Chan Wai, and M. Aye Khine," Modified Linguistic Steganography Approach by Using Syntax Bank and Digital Signature",International Journal of Information and Education Technology, Vol. 1, No. 5, December 2011.
[8] Adnan Abdul-Aziz Gutub, Wael Al-Alwani, and Abdulelah Bin Mahfoodh, "Improved Method of Arabic Text Steganography Using the Extension 'Kashida' Character", Bahria University Journal of Information & Communication Technology Vol. 3, Issue 1, December 2010.
[9] A. Ali and F. Moayad, "Arabic Text Steganography Using Kashida Extensions with Huffman Code", *Journal of Applied Sciences,* vol. 10, pp. 436-439, 2010.
[10] A. Odeh, K. Elleithy, and M. Faezipour, "Steganography in Arabic Text Using Kashida Variation Algorithm (KVA) ", Systems, Applications and Technology Conference (LISAT), 2013 IEEE Long Island, 2013, pp. 1-6.
[11] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery, Michael Metcalf," Numerical-Recipes-in-C-Second-Edition.", Cambridge University Press; 2 edition (October) 30, 1992.
[12] K. Mounika, D. Sri Navya Lakshmi, K. Alekya," SVD Based Image Compression", International Journal of Engineering Research and General Science Volume 3, Issue 2, March-April, 2015.
[13] Hanaa M. Ahmed, Maisa'a Abid Ali K.," Arabic Language Text Steganography Based on Microsoft Word Documents", Al–Yarmouk Journal Al-Yarmouk University College Journal No.1, 2015.
[14] A. Monem S. Rahma, H. Bahjat Abdul Wahab, and H. M. ," Hybrid Information Hiding Technique Using Parametric Spline and DFT", Eng. & Tech. Journal ,Vol.28, Eng. & Tech. Journal ,Vol.28, No.4, 2010.
[15] H. M. Salman , "Human Identification Using Normalized Energy Based Spectrum Eigenpalms", JCCE, VOL.10, No1, 2010.