ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

Semantic Analyzer for Audio Steganography

Anamika Sharma[1], Pushpinder Singh[2]

M.Tech Student Department of Computer Science & Engineering, RBIEBT (Kharar, Punjab), India[1]

Assistant Professor, Department of Computer Science & Engineering, RBIEBT (Kharar, Punjab), India[2]

Abstract- In the current internet community, secure data transfer is limited due to its attack made on data communication. So more robust methods are chosen so that they ensure imperceptible and secured data transfer. Audio steganography is the one in which data is hidden inside the audio files in the form of bits spread over the unused bits inside the audio file. This research work shows the implementation of a novel steganography technique for hiding digital data into wav audio files using spread spectrum technique and a semantic analyzer coupled with a lexicon of words. However to maintain more security for data, the steganography is implemented with the help of encryption technique, but in this methodology, the implementation of dictionary system is done where words are stored in the database and each word is assigned with a unique number. These unique numbers of words are used for encoding the words of a secret message that is going to be hidden inside the audio file. The words of a message can be segregated into noun, pronoun, verbs, determiner, adjective, adverb etc for their semantic analysis. Furthermore, the implemented technique used two intermediates to transmit the secret data between communicating parties. In the first intermediate English language message is encoded on the basis of unique numbers and also do semantic analysis of words in the message, after that encoded message is hidden inside the audio file. In the second intermediate the hidden encoded message is recovered from the audio file and decodes the message to get the original information.

Keywords- Audio Steganography, lexicon of words, Security, Spread Spectrum, Semantic Analyzer.

I. INTRODUCTION

Information hiding is a part of information Security. The term hiding here can refer to making the information imperceptible and keeping the existence of the information secret. As a result a technique steganography is used. Steganography is an art and a science of communicating in a way, which hides the existence of the communication. It is also called as covered writing, because it uses a cover of a message for sending any important secret message. Basically, audio steganography is a type of digital steganography that hides digital data into digital audio files such as WAV, MP3, and WMA files. Audio steganography takes advantage of the Human Auditory System (HAS) which cannot hear the slight variation of audio frequencies at the high frequency side of the audible spectrum; and thus, audio steganography can exploit and use this type of frequencies to hide secret data without damaging the quality of the audio file or changing its size. Multimedia data hiding techniques have developed a strong basis for steganography area with a growing number of applications like digital rights management, covert communications, hiding executables for access control, annotation etc.This paper proposes a novel steganography algorithm for hiding digital data into wave audio files using two intermediates to deliver the secret data. In the first intermediate English language message is encoded on the basis of unique numbers and also do semantic analysis of words in the message, after that encoded message is hidden inside the audio file. In the second intermediate the hidden encoded message is recovered from the audio file and decodes the message to get the original information. Basically, in the first intermediate the words of a message are dynamically encoded during the hiding process using a semantic analyzer of lexicon of English words randomly categorized in 9 categories. Further unique number is given to each word in decimal digits. These digits are used to encode the words of the message. This whole procedure is shown in figure 1, 2, 3. As this technique provide more security to data because the whole data is hidden in encoded form. The main objectives of this research are:

- To maintain robustness and imperceptibility of secret data in carrier file during the communication.
- The clarity of digital audio signal should not be harmed.
- To analyze the features of audio file that can be used to implement the high rate data hiding.
- The data received on the receiver side should be semantically correct.

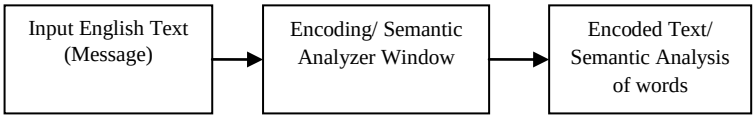

Figure 1 Flow chart for encoding and semantic analysis of message

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

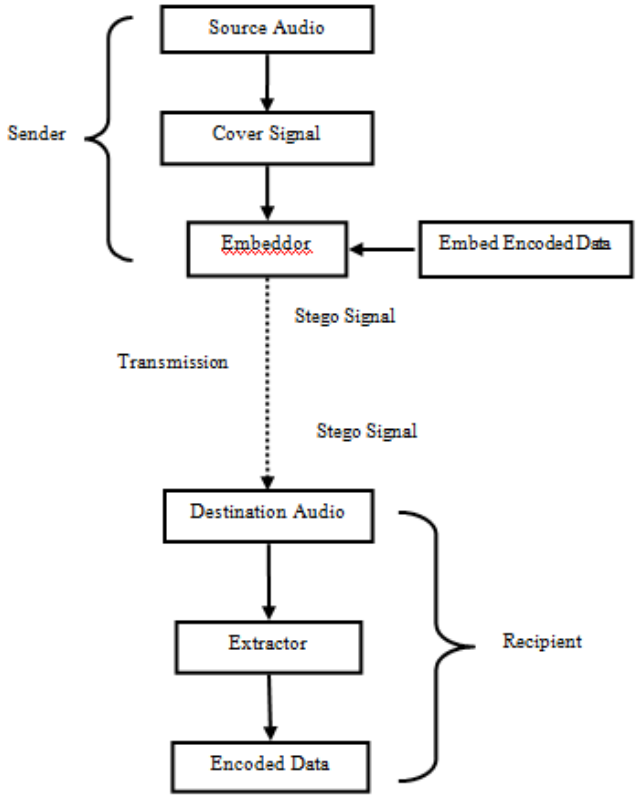

Figure 2: Flow Chart for Audio Steganography

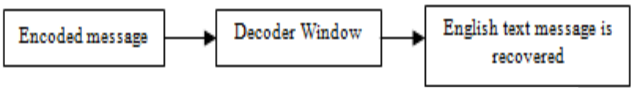

Figure 3: Flow Chart for decoding the message

II. AUDIO STEGANOGRAPHY

Audio steganography is the art and science of hiding digital data such as text messages, documents, and binary files into audio files such as WAV, MP3, and RM files. The output audio file is called the carrier file and is the only intermediate to be sent to the receiver. Imperceptibility is the property of steganography and it refers to the fact that no one apart from the original sender and the intended receiver can suspect the presence of secret data into the carrier file being communicated. Steganography can be achieved by means of three types of techniques: injection, substitution, and generation.

Embedding secret messages in digital sound is usually a more difficult process than embedding messages in other media, such as digital images. In order to conceal secret messages successfully, a variety of methods for embedding information in digital audio have been introduced. In this research work we have used spread spectrum technique.

A. Techniques of Audio Steganography

i) *Injection:* The injection technique implants the data to hide in the insignificant part of the carrier file, which is normally ignored by operating systems and software applications. For example, most computer files comprise what so called an end-of-file marker or EOF for short, which indicates that no more data can be read from a data source. Likewise, an executable file usually ends with an EOF marking the end of binary instructions.

ii) Substitution: The substitution technique substitutes the insignificant bits in the carrier file with the bits of the data to hide. Insignificant bits are those bits that can be modified without damaging the quality or destroying the integrity of the carrier file. For example, in audio files, every unit of sound is made up of a sequence of bits. If the least significant bit of this sequence is modified, its impact is minimal on the perceptible sound so much so that the human ear cannot tell the difference between the original version and the altered one.

iii) Generation: The generation technique reads the data to hide and generates out of them a new set of data. It is a dynamic method of creating a carrier file based on the information contained in the data to hide. For example, one generation technique will take the message to hide and turn its characters into matching audio frequencies that can ultimately make up an audio file.

III. SPREAD SPECTRUM

Spread Spectrum Steganography is a relatively new technology that can provide enhanced levels of security over and above ordinary steganographic techniques. Increasingly, audio media are being used for steganographic purposes. Spread Spectrum Technology forms the basis of Spread Spectrum Steganography. It is a form of Radio Frequency communication. Spread Spectrum techniques intentionally spread the transmitted data signal over a wide frequency range. The bandwidth used is in excess of the minimum bandwidth required for the data being sent. By Increasing the bandwidth improvements in the signal-to-noise performance are obtained. The fundamental idea behind this process is that, in channels with narrowband noise, increasing the transmitted signal bandwidth results in an increased probability that the information received will be correct. The increase in performance for very wideband systems is called the process gain. In order to be considered as a Spread Spectrum system, the system must meet the following criteria:

- The transmitted signal bandwidth is much greater than the information bandwidth.
- Some function other than the information being transmitted is employed to determine the resultant transmitted bandwidth.

A. Spread Spectrum Technique

There are several techniques currently in use for generating Spread Spectrums. These include:

i) Direct Sequence Spread Spectrum (DSSS)
ii) Frequency Hopping Spread Spectrum (FHSS)
iii) Time Hopping Spread Spectrum

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

i) Direct Sequence Spread Spectrum (DSSS): The basic principle behind the Direct Sequence Spread Spectrum (DSSS) technique is the modulation of the RF carrier with a digital code sequence. A two-stage process is used to produce the DSSS. During the first stage, data is spread across the spectrum. This is achieved by dividing the data stream into a symbol stream (small pieces of one bit or more) and then allocating each part of the divided data to a frequency channel across the spectrum. During the second stage, the modulation phase, the DSSS transmitter utilizes a phase varying modulation to modulate each piece of data with a higher data rate bit sequence. DSSS suffers from what is known as a "Near-Far" effect. This effect occurs when an interfering transmitter is much closer to the receiver than the intended transmitter.

ii) Frequency Hopping Spread Spectrum (FHSS): FHSS has an advantage over DSSS in that it is not as affected by the "Near-Far" effect. The basic principle behind the Frequency Hopping Spread Spectrum (FHSS) technique is that the carrier frequency is periodically modified (hopped) across a specific range of frequencies. The frequencies, across which the carrier jumps is the spreading code. Two types of Frequency Hopping signals may be used, slow hopping and fast hopping. With slow hopping, the hopping rate is smaller than the message bit rate, meaning that in one hop, one or more data bits are transmitted. While in fast hopping, one data bit is divided over more than one hop (the hopping rate is greater than the message bit rate).

iii) Time Hopping Spread Spectrum: The third Spread Spectrum technique is Time Hopping. Time Hopping and FHSS are somewhat similar, but in Time Hopping, the transmitted frequency is changed at each code chip time. Time Hopping can be implemented in two ways. In the first technique, each binary is transmitted as a short pulse, known as a chirp. In the second technique for implementing Time Hopping, each chirp has a different duration.

B. Use of Spread Spectrum Technique with Audio Files:

During this research we have analyzed Spread Spectrum that data transmitted in this manner is difficult to detect, can be immune from eavesdropping and jamming, and is very difficult to decode by anyone other than the intended recipient. The techniques described can be used to embed data in audio files relatively easily and these audio files can subsequently be broadcast or passed on using compact discs or other recording media. It would be possible to add another layer of security to the data embedded in the audio file by encrypting it prior to applying the spread spectrum.

IV. SEMANTIC ANALYSIS

Semantic analysis is a statistical model of word usage that permits comparisons of semantic similarity between pieces of textual information. Computational power is increasingly able to analyze more and more complex linguistic structures. Semantic analysis is the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings. More sophisticated approaches use context-free grammars to generate syntactically correct cover text which mimics the syntax of natural text. None of these uses meaning as a basis for generation, and little attention is paid to the semantic cohesiveness of a whole text as a data point for statistical attack. One of the primary goals in text-comprehension research is to understand what factors influence a reader's ability to extract and retain information from textual material. The typical approach in text-comprehension research is to have subjects read textual material and then has them produce some form of summary, such as answering questions or writing an essay. This summary permits the experimenter to determine what information the subject has gained from the text.

To analyze what a subject has learned from a text, the task of the experimenter is to relate what was in the summary to what the subject has read. This permits the subject's representation (cognitive model) of the text to be compared with the representation expressed in the original text. For such an analysis, the experimenter must examine each sentence in the subject's summary and match the information contained in the sentence to the information contained in the texts that were read. Information in the summary that is highly related to information from the texts would indicate that it was likely learned from the text. Nevertheless, matching this information is not easy. It requires scanning through the original texts to locate the information.

A. Use of Semantic Analysis in Steganography

Linguistically naive approaches to the problem use statistical frequency of letter combinations or random dictionary words to encode information. More sophisticated approaches use semantic analyzer to generate semantically correct cover text. It uses the meaning as a basis for generation, and little attention is paid to the semantic cohesiveness of a whole text as a data point for statistical attack. Audio-based information hiding techniques are discussed, providing motivation for moving toward linguistic steganography and steganalysis. Modern steganography is generally understood to deal with electronic media rather than physical objects and texts.

This makes sense for a number of reasons. First of all, because the size of the information is generally quite small compared to the size of the data in which it must be hidden, electronic media is much easier to manipulate in order to hide data and extract messages. Secondly, extraction itself can be automated when the data is electronic, since computers can efficiently manipulate the data and execute the algorithms necessary to retrieve the messages. Also, because there is simply so much electronic information available, there are a huge number

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

of potential cover texts available in which to hide information, and there is a gargantuan amount of data an adversary attempting to find steganographically hidden messages must process. Electronic data also often includes redundant, unnecessary, and unnoticed data spaces which can be manipulated in order to hide messages. In a sense, these data spaces provide a sort of conceptual "hidden compartment" into which secret messages can be inserted and sent off to the receiver.

V. IMPLEMENTATION AND EXPERIMENTAL RESULT

A. The Implemented Algorithm

This paper represents a novel steganography technique for hiding any form of digital data into digital audio files in a random manner. It uses two intermediates to convey the secret data. An uncompressed audio file acting as a carrier files holding the secret data inside the LSBs of its audio samples. The proposed algorithm consists of several steps needed to be executed in sequence to hide some input secret data into an audio WAV file.

1. The secret data are preprocessed so that they become suitable for storage inside the carrier audio file.
2. Store near about 500 words in the database.
3. Make chunks of those words in the form of noun phrase, verb phrase, adverb, articles, helping verb etc. Give unique identification number to each word in each chunk.
4. At front end a GUI is designed in which message is typed by the sender. Firstly it matches the sentence's words with the word in the word in the database, after that it tells which word in the sentence is noun phrase, verb phrase and so on….
5. Then the sentences will be encoded on the base of words unique identification number given to each word.
6. The encode form of the message is stored inside the .wav audio file in the form of bits.
7. The encoded message can be recovered from the same au audio file and that message is decoded on the base of unique identification.

B. Table for Audio steganography

This table represents the various properties if audio file like their size and bitrates and also represents how much data is stored in an audio file according to their size and bitrates. According to this table the amount of data stored in an audio file is directly depend upon the size of the file and its bitrates. As shown in table 1 the file of large size and low bit rate stores large amount of data and the file with large size and high bitrates stores less amount of data.

Table 1: Comparison of amount of data stored in various size of file

Size of the Audio File	Audio Bitrates	Length of Audio File	Amount of Data
175 KB	705 kbps	2 seconds	1019 bytes
227 KB	352 kbps	5 seconds	1063 bytes
290 KB	352 kbps	6 seconds	1512 bytes
569 KB	256 kbps	18 seconds	1625 bytes
796 KB	1411 kbps	4 seconds	950 bytes

C. Graphical Result

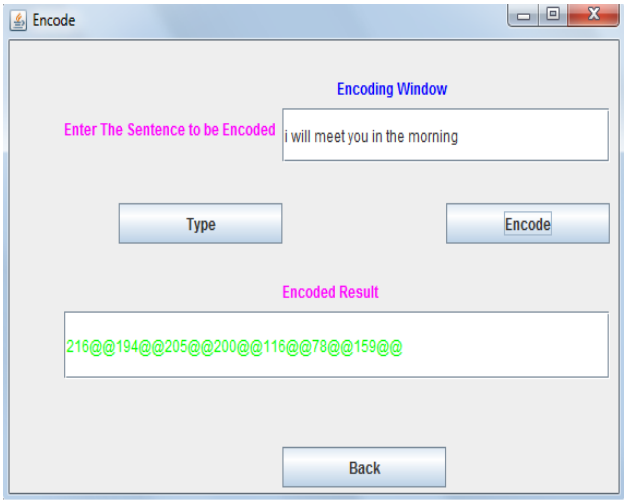

Fig. 4: Graphical Window for Encoding of Message

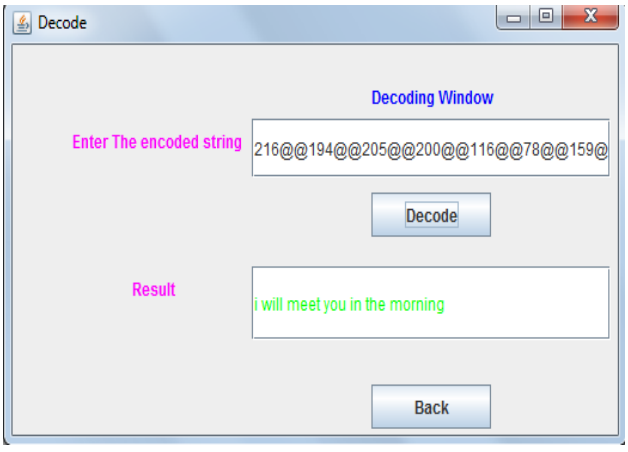

Fig. 5: Graphical Window for Decoding of Message

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

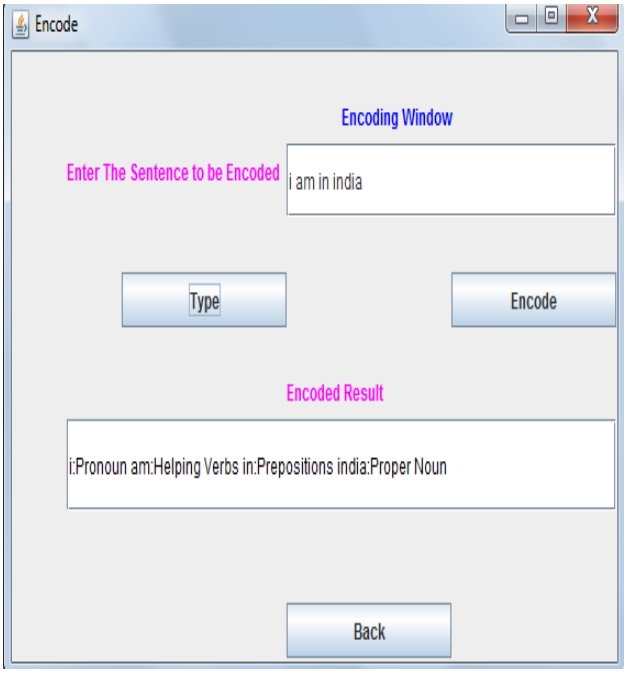

Fig. 6: Graphical Window for Semantic Analysis of Message

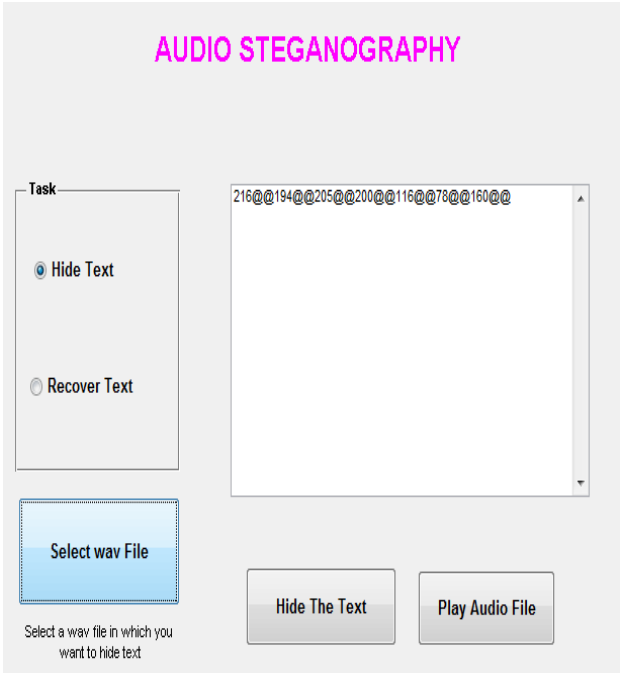

Fig. 7: Graphical Window for Hiding Encoded Text

D. Experimental Result

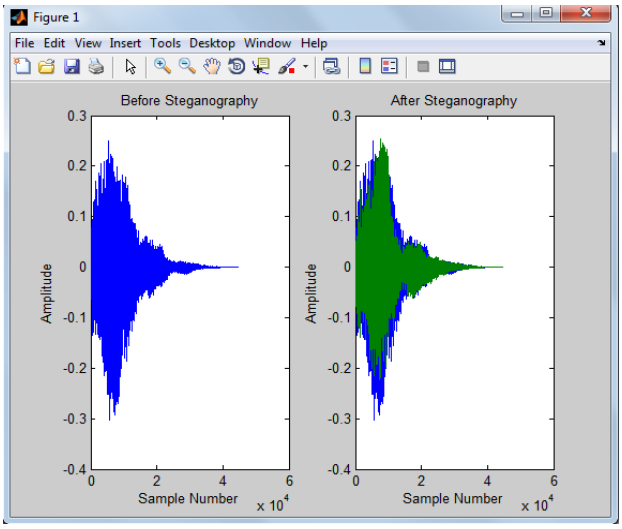

Fig. 8: Graph Shows the Result of Audio Steganography Before and After Steganography

VI. CONCLUSION

Steganography is a powerful tool for data hiding in audio medium when used in conjunction with the encoding of words. A method has been proposed which is easy to implement, transparent, and imperceptible for the human ear. This research work presents a steganography technique for hiding digital data into digital audio files. The carrier audio samples are selected using spread spectrum technique into which bits of the secret data are to be hidden. Spread Spectrum technique is powerful tools for data protection. When used in conjunction with steganography, data is well protected. As Spread Spectrum and steganography develop, there will be an increase in the security that these tools provide. This technique is easy to implement, not that easy to detect, transparent for the human ear, and robust, highly accurate in nature as the recipient can achieve back what they had hidden without any error. The quality of the audio file after steganography is maintained. The key advantage of the technique is the use of two intermediates that complement each other to convey the secret information. As a result, this technique is less susceptible to stego-attacks as third parties often assume that the secret data are hidden in one intermediate and not in two intermediates that together are needed to decode the secret data. Moreover the data is stored in encoded form in audio samples. There is a remarkable increase in capacity of cover audio for hiding additional data and without affecting the properties of the audio file and provide the keys concept for secure data. The main advantage of this method is they are simple in logic and the hidden information is recovered without any error. Thus it succeeds in attaining the basic requirement of data hiding.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 1, January 2014

The future scope of this research work is the implementation of lexicon of words can be done with the help of Context-Free Grammar and parsing.

ACKNOWLEDGEMENT

I would like to thank my guide Er. Pushpinder Singh, Assistant Professor in Department of Computer Science and Engineering at R.B.I.E.B.T. (Kharar), Punjab, India for motivating me to do research work on the topic Semantic Analyzer for Audio Steganography. I would also like to thank my mother, my father and my brother for their continuous support, cooperation, and guidance throughout my M.Tech. work. Moreover I would also like to thank my Professors who were always there at the need of the hour and provided with all the help and facilities, which I required, for my thesis work.

REFERENCES

[1] Peter Wayner, "Disappearing cryptography: information hiding: steganography & watermarking", 3rd Edition, Morgan Kaufmann Publishers, 2009.
[2] Kandel ER, Schwartz JH, Jessell TM, "Principles of Neural Science", 4th edition, McGraw-Hill, 2000.
[3] http://en.wikipedia.org/wiki/Steganography.
[4] Samir K Bandyopadhyay, Debnath Bhattacharyya, Debashis Ganguly, Swarnendu Mukherjee and Poulami Das, "A Tutorial Review on Steganography"
[5] Nick Sterling, Sarah Summers and Sarah Wahl, "Spread Spectrum Steganography"
[6] Krista Bennett, Department of Linguistics "Linguistic Steganography"
[7] Ashwini Mane, Gajanan Galshetwar, Amutha Jeyakumar : " Data Hiding Technique: Audio Steganography using LSB Technique", Vol. 2, Issue 3, May-Jun 2012.
[8] Aho, A.V., Sethi, R. and Ullman, J.D. "Compilers: principles, techniques, and tools", Addison-Wesley Longman Publishing, 1986.
[9] R Sridevi, Dr. A Damodaram, Dr. Svl.Narasimham: "Efficient Method of Audio Steganography by Modified LSB Algorithm and Strong Encryption key With Enhanced Security", © 2005 - 2009 JATIT.
[10] K. Gopalan, "Audio steganography using bit modification", Proceedings of International Conference on Multimedia, vol.1, pp.629-632, 2003.
[11] B. Santhi, G. Radhika and S. Ruthra Reka: "Information Security using Audio Steganography -A Survey", Research Journal of Applied Sciences, Engineering and Technology 4(14): July 15, 2012.
[12] Zameer Fatima and Tarun Khanna: "Audio Steganography Using DES Algorithm", Proceedings of the 5th National Conference; INDIACom-2011.
[13] Mazdak Zamani, Azizah Bt Abdul Manaf, Rabiah Bt Ahmad, Farhang Jaryani, Hamed Taherdoost, Akram M. Zeki: "A Secure Audio Steganography Approach", Copyright © 2009 by the Institute of Electrical and Electronics Engineers.
[14] Mazdak Zamani , Azizah A. Manaf , and Rabiah B. Ahmad: "Knots of Substitution Techniques of Audio Steganography", 2009 International Conference on Computer Engineering and Applications IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore
[15] Prof. Samir Kumar, BandyopadhyayBarnali, Gupta Banik, 'LSB Modification and Phase Encoding Technique of Audio Steganography Revisited', International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2012
[16] Nedeljko Cvejic, Tapio SeppÄanen, Increasing Robustness of LSB Audio Steganography by Reduced Distortion LSB Coding.
[17] Pradeep Kumar Singh, Hitesh Singh and Kriti Saroha, A Survey on Steganography in Audio, Proceedings of the 3rd National Conference; INDIACom-2009.
[18] PETERW. FOLTZ, "Latent semantic analysis for text-based research", Behavior Research Methods, Instruments, & Computers 1996