A Secure Error-Resilient Lossless Source Coding Scheme Based on Punctured Turbo Codes

A. Payandeh, M. Ahmadian, and M. R. Aref

***Abstract*—In this paper, we develop a new error-resistant secure lossless source coding scheme for discrete sources. Recent results indicate that the same turbo principle that provides sub-optimal strategy for channel coding, can be used to obtain efficient source coding schemes. We extend the source turbo-coding idea to include security for transmission of the compressed data and also to ensure the lossless recovery of information at the receiver. Compression and security are achieved by adapting the random puncturing strategy to the statistics of the source. Lossless compression is guaranteed by finely puncturing the encoded data of a parallel-concatenated turbo code while verifying the integrity of the source information at the source encoder. Simulation results show that the proposed scheme can obtain a compression rate close to the source entropy. For the same block length, this scheme yields better compression rates in comparison with the Lempel-Ziv universal source coding. The proposed scheme has a robust and error-resistant performance over noisy channels for E_b / N_0 >5 dB.**

***Index Terms*—Error-resistant source coding, secure source coding, source turbo-coding, security.**

I. INTRODUCTION

THE FOCUS of source coding is to remove redundancy in the data, whereas, the purpose of channel coding is to add redundancy in a controlled fashion to combat errors in the channel. Source coding and channel coding are essentially information-theoretic dual problems [1]. This duality can be used for designing new data compression schemes based on channel codes.

Weiss [2] and Jelinek [3] presented lossless/lossy compression algorithms using binary linear codes. Ancheta [4] used a syndrome former as a source encoder. Given a good channel encoding and decoding algorithm, the same, can be used as a source code, where the channel decoding algorithm is used as a source encoding algorithm. This idea was investigated by McEliece [5]. He illustrated the usefulness of employing a channel code as a source code using perfect block codes. More advanced source coding schemes with better results have been studied based on the more powerful channel codes defined by very sparse graphs (e.g., turbo codes, low-density parity-check codes, etc.). It has been shown that the iterative decoding algorithm based on belief propagation (BP) principle (e.g., turbo decoding, sum-product algorithm) is suitable for source coding. Garcia-Frias and Zhao [6] showed compression scheme for binary memoryless sources using punctured turbo codes. Also in [7], they extended their compression scheme to correlated binary sources. In [8], a lossless source coding scheme based on turbo codes has been presented. Caire, *et al.* [9] presented a fixed length compression algorithm for binary sources with memory based on the LDPC codes. Matsunaga and Yamamoto [10] treated a lossy coding problem of a memoryless binary source with probability distribution $p(0) = p(1) = 1/2$. They proved a coding theorem for lossy compression using LDPC codes. A lossy compression algorithm based on layered LDPC-coding approach has been proposed by Wadayama [11].

The application of LDPC codes to compression problem of an equiprobable memoryless binary source with side information at the decoder has been reported in [12]. The LDPC codes can be used to compress close to the Slepian-Wolf limit [13] for correlated binary sources. The achievable compression performance of the LDPC codes for the Slepian-Wolf coding problem has been analyzed in [14], [15].

Adding security to a source coding algorithm is an attractive idea since it could reduce the overall processing cost of providing secure compressed data. Witten and Cleary [16] proposed a number of methods for combining encryption and compression. In their proposal, the statistical model of source is used as the encryption key: so that the transmitter and the authorized receiver only know the details of the model. This scheme is called model-based. An alternative approach proposed by Irvine, Cleary, and Rinsma-Melchert [17] is coder-based scheme.

We extend the source turbo-coding idea to include security for transmission of the compressed data. This scheme is based on random puncturing of encoded sequence.

The paper is organized as follows. In Section II, the problem is described and an error-resistant secure source coding algorithm is discussed for discrete sources. Computer simulation results of the proposed method are given in Section III. Finally, conclusions and future works are discussed Section IV.

Manuscript received June 27, 2005; revised December 18, 2005.

This work was supported in part by the Applied Science Research Association (ASRA), Tehran, I. R. Iran..

A. Payandeh and M. Ahmadian are with the Electrical Engineering Department, K. N. Toosi University of Technology, Tehran, I. R. Iran (email: payandeh@ee.kntu.ac.ir, mahmoud@eetd.kntu.ac.ir).

M. R. Aref is with the Electrical Engineering Department, Sharif University of Technology, Tehran, I. R. Iran (e-mail: aref@sharif.ac.ir).

Publisher Item Identifier S 1682-0053(06)0388

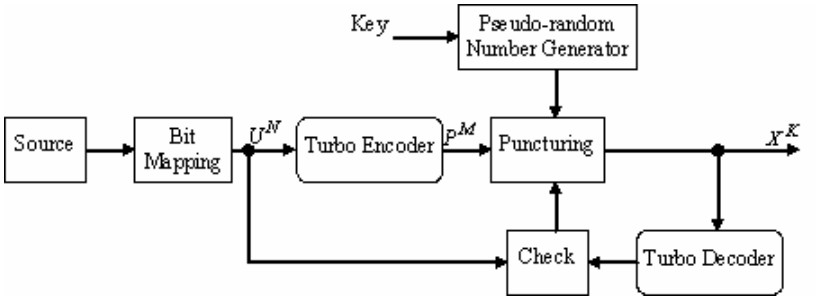

Fig. 1. Secure source turbo-encoder for discrete source.

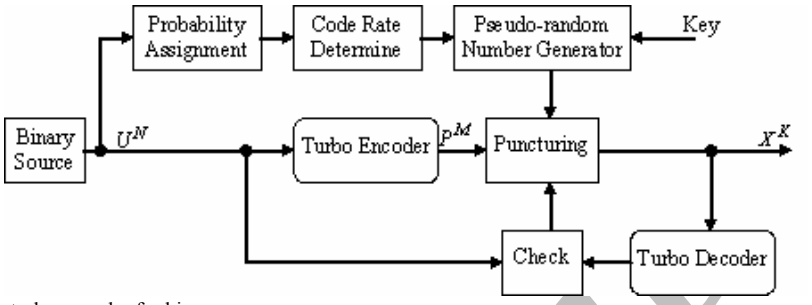

Fig. 2. Adaptive secure source turbo-encoder for binary source

II. Proposed Scheme Structure

A secure source coding scheme is one that provides both data secrecy and data compression in one process. Combining these two steps into one may obtain faster and more efficient implementations. If a simple scheme which combines compression, security, and error control is developed, it would be widely adopted for use in wireless communication.

Let U be a discrete source with entropy $H(U)$. Consider a block U^N of source symbols. Since the entropy of the source is $H(U)$, the source coding theorem states that a source block can be perfectly reconstructed from a binary codeword X^K of length $K \cong NH(U)$, for sufficiently large N. We show how to generate this codeword in order to achieve compression on the one hand and protect the source information against an unauthorized user on the other hand.

Fig. 1 displays our secure source encoding scheme. Each source symbol is first mapped to a binary codeword. A turbo encoder then takes a block U^N of binary sequence and delivers a block P^M of code bits, which are punctured to achieve the desired compression rate. Finally, to guarantee that the decoder is able to decode the input sequence without errors (lossless source coding), the compressed codeword X^K is decoded using a turbo decoder, resulting in the recovered source sequence $\hat{U}^N$. As long as the reconstructed sequence is identical to the source sequence, the compressed codeword is transmitted. If the integrity test fails, the random puncturing of turbo-encoded bits is repeated again. The proposed puncturing scheme is based on the concept of pseudo-random numbers generating for selecting K bits from M turbo-encoded bits. As stated in [6], in order to recover the source sequence perfectly, the randomly chosen puncturing pattern has to be saved and transmitted to the receiver. In [8], each parity sequence is interleaved, written line by line into a square matrix and the bits are erased column wise, and only the index of the last punctured column has to be known to reconstruct the source block. In our scheme, it is enough to send the number of iterations required to find the decodable puncturing pattern for the receiver.

In summary, if the source entropy $H(U)$ is known, the proposed secure source turbo-encoding algorithm has the following steps:

1. Calculate $K \cong NH(U)$, N is the input block length.
2. Let $i = 1$.
3. Encode the source block U^N with a turbo encoder and store the output block P^M.
4. Puncture the encoded block using K pseudo-random numbers.
5. Decode the compressed block X^K.
6. Check for errors. If the decoded block is error free, send the compressed block and i to the receiver, and go to step 1 for a new source block.
7. Let $i = i + 1$ and go back to step 3 for selecting a new random puncturing pattern.

If the source entropy $H(U)$ is unknown, the adaptive secure source turbo-encoding algorithm for binary memoryless time-variant sources (Fig. 2) will have the following steps:

1. Count the number of zero and one symbols in input block U^N, and calculate occurrence probabilities of symbols $P_0^{'}, P_1^{'}$.
2. Calculate: $K \cong NH^{'}(U) = -N\{P_0^{'}\log_2(P_0^{'}) + P_1^{'}\log_2(P_1^{'})\}$.
3. Let $i = 1$.
4. Encode the source block U^N with a turbo encoder and store the output block P^M.
5. Puncture the encoded block using K pseudo-random numbers.
6. Decode the compressed block X^K.
7. Check for errors. If the decoded block is error free, send (X^K, K, i) to the receiver, then go back to step 1 for a new source block.
8. Let $i = i + 1$. If $i \le \max\_iter$ ($\max\_iter$ is the maximum iterations for searching the decodable puncturing pattern), then go back to step 5 for selecting a new random puncturing pattern.

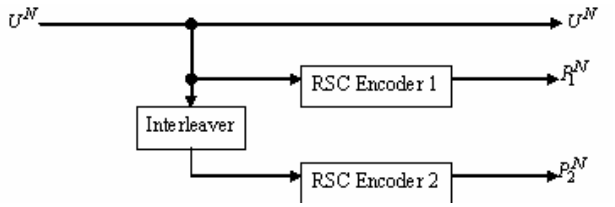

Fig. 3. Rate 1/3 turbo encoder structure used for simulation.

9. Let $K = K + L$ (L is a constant value) and go back to step 3.

In this case ($H(U)$ is unknown), we obtain a compression rate close to the entropy. The compression rate depends on the values of max_*iter* and L.

What remains to be studied is the security of the system. Data compression reduces the redundancy of the data, hence making the encrypted data less sensitive to statistical cryptography and dictionary attacks. The security of this system is based on two computationally hard problems, which are exhaustive search on the key space and turbo decoding of a random punctured sequence. There are two basic attacks on this cryptosystem:

1. Decoding attack–The attacker may try to recover the message U^N directly from the intercepted ciphertext X^K. In the case of successful decoding attack, the plaintext attack is recovered but the cryptosystem remains intact. Turbo decoding of a compressed sequence without knowing the puncturing pattern is the basic problem to be solved. The decoding problem of a general linear code is a NP-complete problem [18]. The number of possible puncturing patterns is as follows

$$\binom{M}{K} = \frac{M!}{K!(M-K)!} . \quad (1)$$

The average work factor to get the plaintext from the intercepted ciphertext is

$$W = C_D(N,I)\binom{M}{K/2}, \quad (2)$$

where $C_D(N,I)$ is the complexity of turbo decoding, which depends on input block length N and number of decoding iterations I. It is expected that if the length of turbo-encoded block P^M is sufficiently large, then this attack will be infeasible. Suppose we choose $M = 1000$ and $K = 400$, then there will be about

$$\binom{M}{K} \cong \frac{\sqrt{2\pi M}M^M}{2\pi\sqrt{K(M-K)}K^K(M-K)^{M-K}} \cong 10^{292} \quad (3)$$

possible puncturing patterns. For source turbo-coding we require block lengths of about $N = 10^4$ to 10^6 [8], therefore the decoding process is more complex without knowing puncturing pattern.

2. Trapdoor attack – The cryptanalyst may try to find the key of pseudo-random number generator from intercepted sequences (ciphertext-only attack) or U^N and X^K pairs (known/chosen-plaintext attack). In ciphertext-only attack, the attacker is assumed to have several ciphertexts, and do not have the private key. Thus, he/she tries to guess key by brute force. Let the size of key be $n = 100$ bits. Then, the private key $2^n = 2^{100} \cong 1.27\times10^{30}$ possible combinations, for linear number generator. If the cryptanalyst employ a 100 MIPS computer to conjecture the key, the computational load is then $2^{100}/(100\times10^6\times60\times60\times24\times365) \cong 4.02\times10^{14}$ years. This is, indeed, a very long time. Thus, our cryptosystem is secure for ciphertext-only attack.

TABLE I
RATE 1/2 RSC COMPONENT CODES USED IN SIMULATED SCHEME

Constraint Length	Feedforward Generator	Feedback Generator
3	$1+D^2$	$1+D+D^2$
4	$1+D^2+D^3$	$1+D+D^2+D^3$
5	$1+D^4$	$1+D+D^2+D^3+D^4$

The other attacks are the known/chosen-plaintext attacks. They are more religious than the ciphertext-only attack. In these attacks, the attacker is assumed to have obtained several plaintext and ciphertext pairs, and all of these pairs share a common key. In these cases, the outlaw can analyze these pairs to obtain the common key, and correctly decrypt the next ciphertext if the sender still encrypts his next original message by it. There are several types of cryptanalytic known/chosen-plaintext attacks against algebraic-code cryptosystems discussed in [19], [20]. These attacks are performed based on the linearity of the system. They will not be applicable for nonlinear coding scheme. The trapdoor attacks seem to be hopeless if the structure of pseudo-random number generator is nonlinear and its period is sufficiently large, because there are so many possibilities for the key. In particular, suppose we choose a nonlinear feedback shift register of length $n = 100$, then there will be about $2^{2^n} = 2^{2^{100}}$ possible keys.

These attacks are performed under the assumption that there is no error occurs in the channel. The presence of channel errors introduces additional level of data security to this system. Compared with other cryptosystems, this scheme has the advantage of high-speed encryption and decryption with high security. It would be a challenge indeed to find cryptanalytic attacks to break this system.

III. SIMULATION RESULTS

To underline the effectiveness of the proposed secure source coding scheme, we present a set of numerical results in this section. A data block generated by a binary memoryless source is encoded through a rate 1/3 turbo encoder. The turbo encoder is a parallel concatenation of two identical recursive systematic convolutional (RSC) component codes (Fig. 3). Table I shows the rate 1/2 component RSC codes used in the simulated schemes. For simplicity, we use a 100-stage linear feedback shift register (LFSR) for pseudo-random number generator. Details of source turbo-decoding are explained in [8].

Fig. 4 shows the results obtained for this secure source turbo-code for different source block lengths $N = 10^3, 10^4, 10^5, 10^6$ and max_*iter* $= 20$. As, we can see from Fig. 4, the code rate is approximately close to the source entropy rate. It is obvious that increasing the input block length in this scheme, results in a better compression rate.

The compression rate of this scheme in different source probabilities for different constraint lengths is shown in

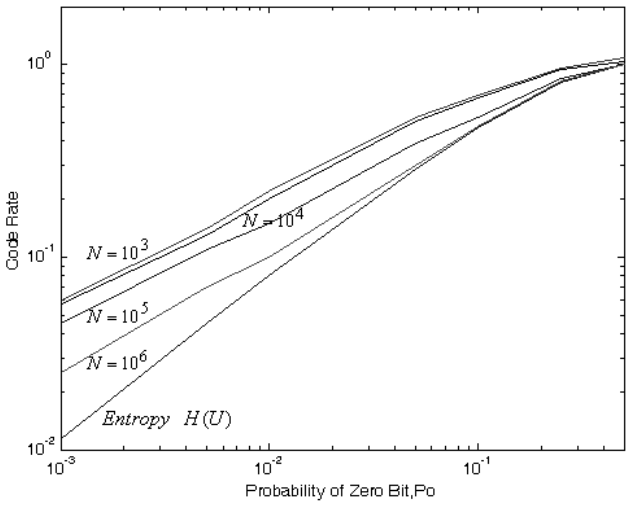

Fig. 4. Code rate as a function of the occurrence probability of zero bit for interleaver lengths $N=10^3,10^4,10^5,10^6$ (constraint length=3, $\max\_iter=20$).

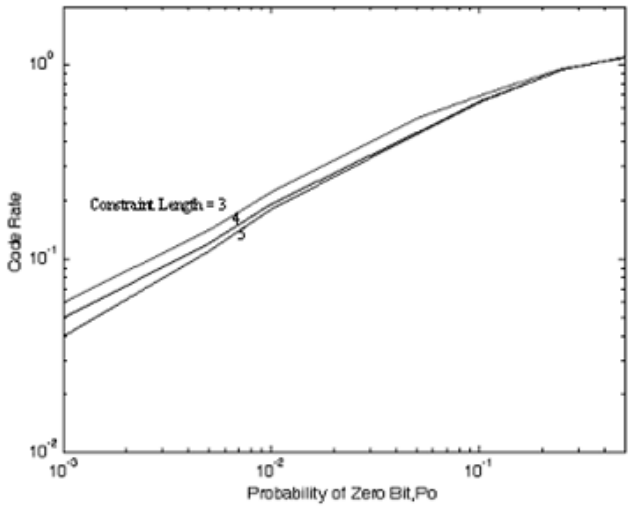

Fig. 5. Code rate as a function of the occurrence probability of zero bit for constraint lengths 3, 4, 5 ($N=10^4$, $\max\_iter=20$).

Fig. 5. In this figure, it can be seen that as the constraint length increases, the performance of the code also increases, resulting in higher compression.

The comparison of this source turbo-coding scheme with Lempel-Ziv 77 algorithm [21] shows (Fig. 6) that for the same block length, the proposed scheme can yield better compression rates.

To validate the error-resistant of proposed scheme, Fig. 7 shows bit error rate (BER) performance results for the simulated secure source turbo-code and unpunctured turbo code over an additive white gaussian noise (AWGN) channel. The performance of the proposed scheme is bad for low E_b/N_0 (<5 dB), but for higher E_b/N_0 (>5 dB) becomes good.

IV. Conclusions

A secure source coding scheme combines data encoding and data encryption into one process and enables the system to compress source sequence as well as conceal information from unauthorized user simultaneously. In this paper, a secure lossless compression algorithm for discrete

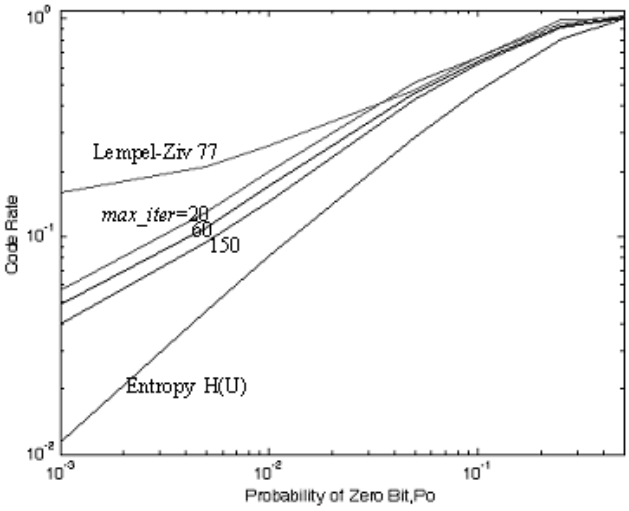

Fig. 6. Average compression rate for the proposed scheme compared to Lempel-Ziv algorithm and entropy as a function of the occurrence probability of zero bit ($N=10^4$).

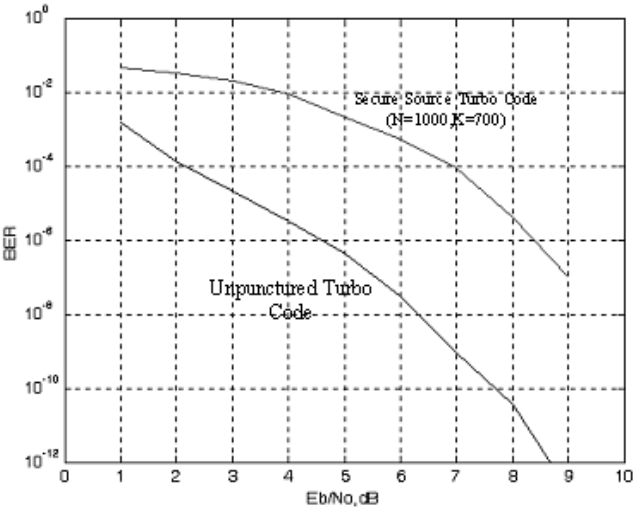

Fig. 7. BER performance of the proposed scheme compared to an unpunctured turbo code (constraint length = 3, $\max\_iter=20$).

memoryless sources was presented. This algorithm was based on random puncturing turbo-encoded blocks. We investigated possible cryptanalytic attacks against the scheme. The cryptanalysis results showed that the proposed scheme has a good practical and theoretical security. Simulation results showed that we could compress the source information close to the entropy of the source. Also, this secure source coding scheme could obtain a robust and error-resistant performance over AWGN channels.

Future research will mainly be focused to extend the proposed secure source turbo-coding scheme to the sources with memory.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York, John Wiley and Sons, 1991.

[2] E. Weiss, "Compression and coding," *IRE Trans. Information Theory*, vol. 8, no. 3, 256-257, Apr. 1962.

[3] F. Jelinek, "Tree encoding of memoryless time discrete source with fidelity criterion," *IEEE Trans. Information Theory*, vol. 15, no. 5, pp. 584-590, Sept. 1969.

[4] T. C. Ancheta, Jr., "Syndrome-source coding and its universal generalization," *IEEE Trans. Information Theory*, vol. 22, no. 4, pp. 432-436, Jul. 1976.

[5] R. J. McEliece, *The Theory of Information and Coding: Encyclopedia of Mathematics and Its Applications*, Addison-Wesley Publishing Company, Inc., 1977.
[6] J. Garcia-Frias and Y. Zhao, "Compression of binary memoryless sources using punctured turbo codes," *IEEE Communications Letters*, vol. 6, no. 9, pp. 394-396, Sept. 2002.
[7] J. Garcia-Frias and Y. Zhao, "Compression of correlated binary sources using turbo codes," *IEEE Communication Letters*, vol. 5, no. 10, pp. 417-419, Oct. 2001.
[8] J. Hagenauer, J. Barros, and A. Schaefer, "Lossless turbo source coding with decremental redundancy," in *Proc. International ITG Conference on Source and Channel Coding*, pp. 333-339, Erlangen, Germany, Jan. 2004.
[9] G. Caire, S. Shamai, and S. Verd'u, "A new data compression algorithm for sources with memory based on error correcting codes," in *Proc. IEEE Information Theory Workshop*, Paris, France, pp. 291-295, 2003.
[10] Y. Matsunaga and H. Yamamoto, "A coding theorem for lossy data compression by LDPC codes", *IEEE Transactions on Information Theory*, vol. 49, no. 9, pp. 2225-2229, Sept. 2003.
[11] T. Wadayama, "A lossy compression algorithm for discrete memoryless sources based on LDPC codes," in *Proc. of the 3rd Asia-Europe Workshop on Information Theory*, pp. 1-22, Kamogawa, Japan, Jun. 2003.
[12] A. D. Liveris, Z. Xiong, and C. N. Georghiades, "Compression of binary sources with side information at the decoder using LDPC codes," *IEEE Communication Letters*, vol. 6, no. 10, pp. 440-442, Oct. 2002.
[13] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Information Theory*, vol. 19, no. 4, pp. 471-480, Jul. 1973.
[14] T. Murayama, "Statistical mechanics of data compression theorem," in *Proc. IEEE Int. Symposium on Information Theory*, pp. 245, Lausanne, Switzerland, 2002.
[15] J. Muramatsu, T. Uyematsu, and T. Wadayama, "Low density parity check matrices for coding of correlated sources," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3645-3654, Oct. 2005.
[16] I. H. Witten and J. G. Cleary, "On the privacy afforded by adaptive text compression," *Computers and Security*, vol. 7, no. 4, pp. 397-408, Aug. 1988.
[17] J. G. Cleary, S. A. Irvine, and I. Rinsma-Melchert, "On the insecurity of arithmetic coding," *Computers and Security*, vol. 14, no. 2, pp. 167-180, Mar. 1995.
[18] E. R. Berlekamp, R. J. McEliece, and H. C. A. van Tilborg, "On the inherent intractability of certain coding problems," *IEEE Trans. Information Theory*, vol. 24, no. 3, pp. 384-386, May 1978.
[19] R. Struik and J. van Tilburg, "The Rao-Nam scheme is insecure against a chosen-plaintext attack," in *Proc. Crypto'87*, pp. 445-457, 1987.
[20] K. N. Nam, *Complexity Analysis of Algebraic-Coded Cryptosystems*, Ph.D. Dissertation, Tech. Report NSA-1-85, Center for Advanced Computer Studies, University of Southwestern La, 1995.
[21] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Information Theory*, vol. 23, no. 3, pp. 337-343, May 1977.

Ali Payandeh received B.Sc. and M.Sc. degrees in electrical engineering from Tarbiat Modarres University, Iran, in 1991 and 1994, respectively, and the Ph.D. degree in electrical engineering from K. N. Toosi University of Technology, Iran, in 2006.

From 1991 to 1995, he was a faculty member in the Department of Electrical Engineering at Malek Ashtar University of technology, Iran. Since 1996, he has been a Director of Research at the Applied Science Research Association (ASRA), Iran, where he has involved in research for secure satellite communications. His research interests include information theory, coding theory, secure communications and satellite communications.

Mahmoud Ahmadian was born in Tehran, Iran on April 15, 1953. He received the M.Sc. degree in electrical engineering from the University of Tehran, Iran, in 1976, and the Ph.D. degree in electrical engineering from the University of Manchester, UK in 1996.

Since 1979, he has been with K. N. Toosi University of Technology, Tehran, Iran, as a faculty member. He has taught electronics, communication theory, digital communications, information theory, and coding courses at K. N. Toosi University of technology. He founded the cryptography and coding lab. at this university in 2003 and is currently supervising research activities in the related fields. His research interests include secure communications, error control coding, coded modulation and multifunctional coding.

Mohammad Reza Aref was born in Yazd, Iran in 1951. He received B.Sc. degree in electrical engineering from the University of Tehran, Iran in 1975, and the M.Sc. and Ph.D. degrees from Stanford University, US, in 1976 and 1980, respectively.

He is now a Professor in the Department of Electrical Engineering at Sharif University of technology, Iran. He wrote three books in electrical engineering subjects. He published about 115 papers in international journals and conferences. His research interests include information theory, estimation theory and coding theory. He has supervised more than 60 M.Sc. and Ph.D. thesis in the related fields.

Prof. Aref is a member of the Editorial Boards of the Iranian Journal of Electrical and Computer Engineering.