Sequence Partitioning and Compression Rate

B.B. Alagoz and H.Z. Alisoy

***Abstract*— In the lossless data compression, the process of splitting a data sequence into appropriate subsequences has a substantial role in improving compression rate. This study theoretically investigates effects of data sequence partition on the overall compression rate of data sets. For this proposes, we show that it is always possible to find a partition of data sequence such that the entropy rate at each subsequence is lower than entropy rate of original sequences. This motivates our work to figure out the overall compression rate of the partitioned data sequences. Then, the effects of sequence partitioning on overall compression rate are discussed to explore an optimal partitioning strategy. Finally, an optimization problem for the optimal partitioning of a data sequences is stated for future works.**

***Index Terms*— Information theory, entropy rate, compression rate, Shannon limit.**

I. Introduction

SHANNON established a fundamental limit to lossless data compression in 1948. This limit was stated depending on entropy rate. He revealed that it was possible to compress information coming from a data source, in a lossless manner, with a compression rate close to the entropy rate of data source by using an coding method [1-2]. The entropy rate indeed depends on the statistical nature of the data sources. As the statistical order of a data source increases, the entropy rate of the source decreases. In this way, a better compression rate can be achievable by applying appropriate coding techniques. Today, the most popular coding methods applied in practice are Huffman Coding and Lempel-Ziv Coding [3-5].

The process of data set partitioning is one of the primary tasks effecting performances of coding schemes. In practice, several coding strategies were practically developed for splitting and coding a data sequence to achieve a lower compression rates [6-10]. For instance, a refined partition providing minimum-entropy basis was used to improve compression rate [6]. In Huffman coding scheme, splitting an original symbol sequence into sub-sequences was shown to give a better compression rate on AR1, ECG and seismic signals at several SNR [7]. In [7], the proposed recursive splitting method splits a symbol sequence into subsequences, such that it makes the symbol probabilities different for each subsequence. Thus, individual Huffman coding of each sequence reduces the total number of bits used for the codewords. All those practical efforts demonstrate us that problem of sequence partitioning is a substantial task in a coding method to improve compression rate. In this study, we address the role of the sequence partitioning on the overall compression rate of the data sequences. In this point of view, we theoretically inquire the relation of overall compression rate with the sequence partitioning regardless of coding methodology.

This paper investigates effects of sequence partitioning on the compression rate. Preliminarily, the study demonstrate that one can always found a data partition that makes entropy rate at each subsequence lower than original data sequence even if the information source model is zero-order model. A zero-order information source model is the worst case for coding methods in term of compression performance, since there is not any statistical link between elements of sequence [1]. An assumption of the zero-order statistical model of data source also implies the case that the coder of sequence is ignorant of the nature of the data source. So, the results of this study are not depended of any coding methods. Our investigation focus on the compressibility of partitioned data sequences regardless of coding method. For this proposes, a brief analysis on the overall compression rate of partitioned data sequences is carried out and the factors, affecting overall compression rate, are discussed. A lower bound for the overall compression rate of a partitioned data sequence is derived in the light of Shannon's compression limit. Moreover, an optimization problem, which is independent of coding methods and statistical feature of data, is put aside for future works.

II. Methodology

A. Basic Definitions

Let an finite set of data be $X=\{x_1, x_2....x_r\}$ and a finite set of symbol (Alphabet) be $A=\{s_1, s_2....s_m\}$, where $m>1$. Data set X is composed of elements of the symbol set A. The set A is commonly called as source alphabet. A binary coding function is given as $\phi(.): X \to \{0,1\}$. In such case, a binary coded sequence can be express as $\{\phi(x_1), \phi(x_2), \phi(x_3)..\phi(x_r)\ldots\}$. Compression rate for a set X, coded by $\phi(.)$, can be defined as

$$R = T/l. \quad (1)$$

Here, T is the total number of elements in the binary coded sequence and l is the number of elements in the data set X.

Entropy rate of a data set formed by the m-symbol set A

B. B Alagoz, was with Inonu University, Department of Electrical-Electronics, Malatya, Turkey (e-mail: baykant.alagoz@inonu.edu.tr).
H. Alisoy, was with Inonu University, Department of Electrical-Electronics, Malatya, Turkey (e-mail: halis@inonu.edu.tr).

was given as $H = \log_2 m$, when the data set X was produced by a zero-order source model. A zero-order source model is assumes that there is not any statistical link between elements of data set X [1]. Entropic volume of a data sequence is defined total entropy contained by a data sequence with length l and expressed as lH.

B. Entropy Rates of Partitioned Data Sets

This section is devoted to theoretically show that any data sequence can be partitioned into subsequences such that entropy rate of each sequence is smaller than the entropy rate of original sequence.

Theorem 1:
Any finite data set, given by $X = \{x_1, x_2 .... x_r\}$ and obtained from a zero-order model source, is split into subsets $X_1, X_2, ..., X_g$, where $g \in [2, r]$. Entropy rate of each subset is equal or lower than entropy rate at X.

Proof:
If X is a finite set, the symbol set of X, denoted by A, has to be finite as well. Let the number of elements in a finite set A denoted by m. The entropy rate of X set is written as $H = \log_2 m$. Since any A_i symbol set of the subset X_i is also contained in the set A, the number of elements in any subset A_i will be equal or lower than m. Hence, the entropy rate at a subset X_i, denoted by H_i, will be equal and lower than H as well. So, the property of $H_i \le H$ is valid for any partitioning of a finite data set.

Theorem 2:
Let an infinite set of data be $X_\infty = \{x_1, x_2 .... x_r ...\}$ and suppose that it is generated by a zero-order source. For a symbol set $A = \{s_1, s_2 .... s_m\}$, There is always one partition of X_∞ such that the entropy rate of the each subset is lower than the entropy rate of X_∞.

Proof:
One can always form a subset from the first $m-1$ elements of X_∞. Lets denote this subset by X_1. For a zero-order source model, entropy rate for the set X_∞ can be given as $H = \log_2 m$. For the subset X_1; since it has $m-1$ elements, number of elements in symbol set of X_1 never becomes larger than $m-1$. So, this specifies a upper bounds for entropy rate of X_1 is written as

$$H_1 \le \log_2 (m-1) \qquad (2)$$

Therefore, $H_1 < H = \log_2 m$, one can state that it is possible to find out a subset of X_∞, whose entropy rate is lower than entropy rate of X_∞.

Let the next $m-1$ elements of X_∞ form the subset X_2 and the following $m-1$ elements of X_∞ form the subset X_3 and so on. Thus, X_∞ is partitioned to $X_1, X_2, X_3, \ldots$ subsets, such that, the entropy rate of each subset X_i is lower than the entropy rate at X_∞. So, $X_\infty = X_1 \cup X_2 \cup X_3 \cup .....$ and all $H_i < H$, one can state that "at least one partition of the data set X_∞ can be always found such that the entropy rate at each subset is lower than entropy rate of the set X_∞".

Definition 1(Excessive partitioning):
When the element number of each subsets is less than element number of symbol set A, such a partitioning of data set X is referred to as excessive partitioning. Excessive partitioning always reduces the compression rate due to Theorem 2.

Definition 2(Heuristic partitioning):
When the number of elements of subsets is adjusted intelligently, these partitioning of data set X is called as heuristic partitioning. Heuristic partitioning may reduce the compression rate of data sequence more than excessive partitioning if generated by non-zero order source models.

Definition 3(Constant length partitioning):
If the number of elements of each subset is equal, this partitioned is referred to constant length partitioning.

Definition 4 (Variable length partitioning):
If the number of elements of subsets is not equal, this partitioning is referred to as variable length partitioning.

Definition 5 (Entropic partitioning):
If a partition of a sequence decreases the overall compression rate of a sequence, this partitioning is called entropic partition. Entropic partition set covers excessive partitioning, heuristic partitioning, constant length partitioning, variable length partitioning if anyone reduces the compression rate of sequences.

Theorem 1 and 2 clearly reveals that the partitioning of a data set can provide a reduction in entropy rate of a data streams regardless of statistical characteristic of the data sequences. For instance, excessive partitioning always reduces entropy ($H_i < H$). However, heuristic partitioning strategies may reduce compression rate more than excessive partitioning. So, we need to figure out overall compression rate of a partitioned data sequence in order to make a direct assessment about impacts of sequence partitioning on the overall compression rate of digital data sequences. In the flowing section, overall compression rate of partitioned sequences are inspected.

C. Overall Compression Rate in a Partitioned Data Sequence

This section is devoted to analyze overall compression rate of partitioned finite data sequences. Firstly, lets figure out the overall compression rate of a finite data sequence D, in the case that it is split into k number of subsequences, denote by

d_i, $i \in [1,k]$. The number of elements in each subsequences d_i is denoted by a_i. Secondly, let assume the binary coding function family denoted by $\phi^j(.)$ used in coding of these subsequences. Assuming that the best binary coding function providing the lowest compression rate for a subsequence is chosen from the $\phi^j(.)$, the overall compression rate of D in this partitioning can be expressed as,

$$R = \sum_{i=1}^{k} w(a_i) \cdot R_i^* , \quad (3)$$

where R_i^* is the compression rate, obtained in the coding of subsequent d_i by mean of the best binary coding function. The term $w(a_i)$ is the size weight of subsequence d_i in the sequence D and expressed as $w(a_i) = a_i / l$. Here, l is total number of elements in the sequence D. (See the appendix for the derivation of Equation (3)) The R_i^* is considered to include an additional bit rate ε_i, which is reserved for the redundant coding. The redundant coding mainly resides in headers of data packs.

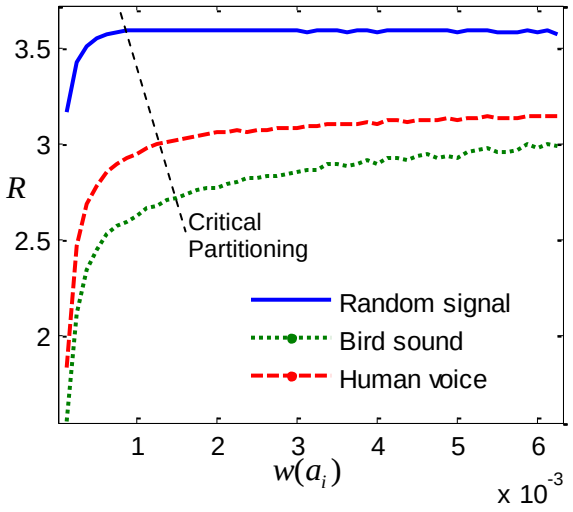

Fig.1. Overall compression rate calculated for equal sized partitioning of data sequences (Random signal, Bird sound and Human voice). Compression rate of pseudo coding function for each subsequences is assumed to $R_i^* = \beta \cdot H_i + \alpha \cdot H_i$.

Example 1:
Fig. 1 illustrates overall compression rates calculated for an constant length partitioning of various time series signals ($a_1 = a_2 = ... = a_k$ and $w(a_1) = w(a_2) = ... = w(a_1) = a_i / l$) in the case of a pseudo coding function, of which the compression rate is assumed as $R_i^* = \beta \cdot H_i + \alpha \cdot H_i$. Here, parameter $\beta > 1$ is used to conform the condition of $R_i^* > H_i$ due to Shannon's limit of lossless compression. We assume to be $\beta = 1.1$ for our pseudo coding scheme. The parameter $\alpha = 0.1$ is stands for the redundant code rates allocated for header to send symbol set and other required codes. The entropy of each subsequent is calculated $H_i = \log_2 m_i$ where m_i is the number of symbol in the subsequence i. In the figure, a critical partitioning line, where compression rate began to decreasing sharply, is shown in overall compression rate plots of time series signals. Bird sound and human voice contains data set generated by a none-order source model. Since there is statistical links between elements, the partitioning and compression rate can decrease for larger subsets compared to random signal set. The random signal is supposed to simulate a none-order source model.

Since $R_i^* > H_i$ according to Shannon's limit of lossless compression [1-2], a lower bound for the overall compression rate can be written as,

$$R > \sum_{i=1}^{k} w(a_i) \cdot H_i . \quad (4)$$

In order to point out effects of sequence partitioning on the compression rate, it will be convenient to express the deviation in the compression rate after a partitioning, which is defined as $\Delta R = R' - R$. Here, R' represents the compression rate in the coding of original sequence without a partitioning and it also satisfies the condition of $R' > H$ due to Shannon's limit. After a sequence partitioning, the deviation in the overall compression rate can be written as: (See the appendix for the derivation of Equation (5)).

$$\Delta R > \sum_{i=1}^{k} w(a_i) \cdot (H - H_i) . \quad (5)$$

If the condition of $H_i \le H$ from Theorem 2 and $w(a_i) \in (0,1]$ is considered, after partitioning sequence, the deviation in compression rate is found as $\Delta R \ge 0$. This noteworthy finding suggests that a partitioning, independently from coding functions and in the absence of a prior statistical knowledge, can reduce the compression rate. In the case that the redundant coding used in headers of data packs are taken into account, the condition of $\Delta R \ge \varepsilon$ should be met to have a better compression after a partitioning. Here, ε represents additional bit rate caused from redundant codes used in applications. In practice, ε is mainly negligible compared to values of R.
Considering Equations (3) and (4), the following substantial remarks can be listed:

i) Overall compression rate of a partitioned sequence strongly depends on the compression rate of each subsequence and their sizes. In fact, the overall compression rate in a partition is a size-weighted average of compression rates of all subsequences. In order to reach a lower overall compression, one should establish an optimal partition strategy such that the subsequences exhibiting a lower compression rates are the larger in size.

ii) It is not a necessity to use one type coding scheme for the coding of all subsequences. The one providing a lowest rate of compression from the coding scheme family can be selected to

code a subsequence and others can be used for other sequence to have a better overall compression rates. So, multi-coding approach can be more effective to improve compression rates.

iii) In order to enhance overall compression rate, the sequence partitioning has to comply with the condition of $\Delta R \geq \varepsilon$.

III. A Discussion on Overall Compression Rate for Optimal Partitioning of Data Sequence

Let assume that a finite length sequence X with entropy rate H is split in subsequence x_i with length of $a_1, a_2, ..., a_k$ and entropy rate H_i. Total entropic volume of X can be defined as, $H\sum_{i=1}^{k} a_i$ and overall entropic volume of partitioned X sequence can be expressed as $\sum_{i=1}^{k} a_i H_i$. Due to partitioning, reduction rate in entropic volume can be written as, $(\sum_{i=1}^{k} a_i H_i) / H\sum_{i=1}^{k} a_i$. In order to improve compression rate of a data sequence by entropic partitioning, the reduction rate in entropic volume should be decreased.

To have better compression rate, the following partitioning rules for entropic partitioning, can be listed:

i) For the data segments having higher entropy rates, form subsequences with a shorter length,

ii) For the data segments having lower entropy rates, form subsequences with a larger length.

iii) Due to Shannon's limit of lossless compression ($R_i > H_i$ and $R > H$), coding method used to compress data is a important factor for the success of compression. That is why, entropic partitioning should be performed by considering coding methods.

Example 2:

A visual example for entropic partitioning strategy is illustrated in Fig. 2. In the figure, original sequence, represented by a large dash rectangle, has the length of $\sum_{i=1}^{5} a_i$ and the overall entropy rate, $H = H_1$. Subsequences are represented by the rectangular areas in different color tones. They have the sizes of a_i and their entropy rates denoted by H_i. In this visual partitioning example, by using Equation (5), the deviation in overall compression rate of original sequence can be expressed as $\Delta R > w(a_2)(H_1 - H_2) + w(a_3)(H_1 - H_3) + w(a_5)(H_1 - H_5) > 0$. This result verifies that a reduction in overall compression rate can be possible by this partitioning. The gray zone with scan lines represents reduced entropic volume as a result of entropic partitioning of original sequence.

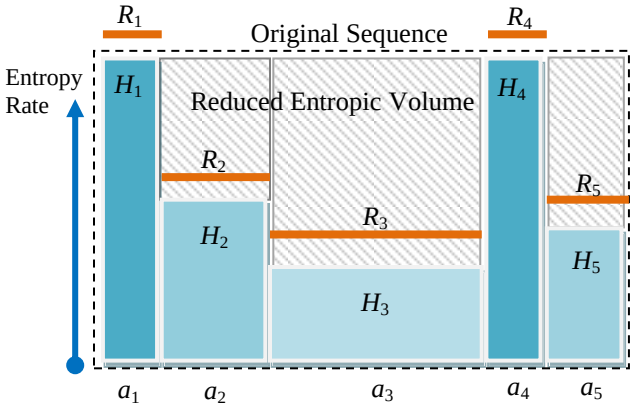

Fig.2. An example of good partitioning reducing entropic volume of original sequence

Entropic partitioning is not dependent of coding methods. A problem of a good entropic partitioning that aims to reduce overall compression rate can be defined as,

$$\Delta R_{opt} = \max_{a_i} \{\sum_{i=1}^{k} w(a_i) \cdot (H - H_i)\}. \quad (6)$$

In the case of a predetermined set of coding schemes, the problem of an optimal partitioning is argued in detail, below:
An optimal partitioning of data sequence can be defined as a partitioning that makes the overall compression rate globally minimum, and it can be simply expressed as

$$R_{opt} = \min_{a_i, R_i} \{\sum_{i=1}^{k} w(a_i) \cdot R_i^*\}. \quad (7)$$

For the practical point of view, the problem of finding an optimal partitioning for a data sequence turns into the problem of finding a_i, R_i^* parameters such that they yields a minimal overall compression rate. An objective function to be optimized can be fashionably written as,

$$E = \sum_{i=1}^{k} (w(a_i) \cdot R_i^*)^2. \quad (8)$$

Overall compression rates obtained for a constant length partitioning (without optimization) and a variable length partitioning (with optimization in accordance with the Equation (8)) is compared in Fig. 3. In this straightforward optimization method, an original data sequence first splits into equal sized subsequences, which is also the case of "without optimization", then, each subsequences shrink or expand toward its neighbors in order to minimize the objective function E given by Equation (8). The figure reveals that variable length partitioning by optimization further decreases compression rate of non-zero order source models.

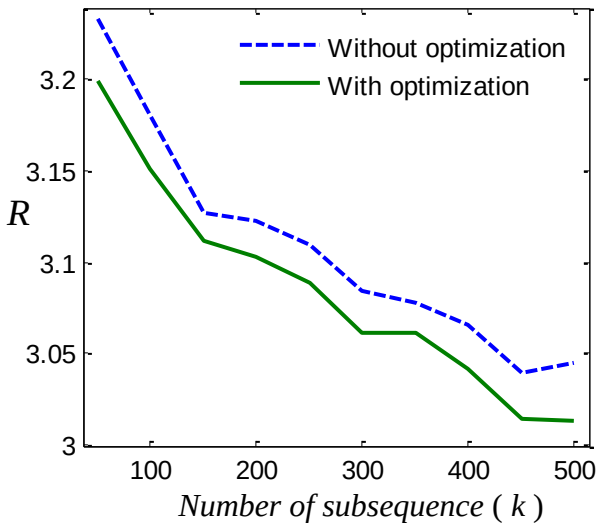

Fig. 3. Overall compression rates calculated for the both case of an equal sized partitioning (without optimization) and a variable size partitioning(with optimization) of data sequences from human voice. Compression rate of the pseudo coding function is assumed as $R_i^* = \beta \cdot H_i + \alpha \cdot H_i$ for $\beta = 1.1$ and $\alpha = 0.1$.

For a faster approximation to the optimal solution for entropic partitioning, a partitioning can be performed subject to a constant bit-length constraint, which is arithmetically defined as,

$$C = a_i \cdot R_i^* , \tag{9}$$

where $C \in R$ is a constant that specify a target length for subsequences in term of bits. A new version of the objective function for an optimal partitioning with a constant bit-length of subsequences can be written as,

$$E = \sum_{i=1}^{k} [(w(a_i) \cdot R_i^*)^2 + (C - a_i \cdot R_i^*)^2] . \tag{10}$$

The solution of this optimization problem was not a preference of this work. We aim to show existence of an optimization problem for a good partitioning, which is applicable in all practical coding schemes in the absence of a prior knowledge about statistical nature of data sequence.

IV. Conclusions

The entropy rate at a data sequence can be simply decreased by splitting it. This enables compression of data sequence regardless of the coding scheme. The property of $\Delta R \geq \varepsilon$ in a partitioning ensures us that the data sequence partitioning decreases the overall compression rate, however, the amount of reduction in the overall compression rate depends on two terms; compression rate of coding methods (R_i^*) and the ratio of subsequences size to original sequence size ($w(a_i) = a_i / l$). In the paper, a theoretical discussion for a general optimal partition strategy, which is applicable to all coding techniques, was given and a corresponding optimization problem to improve overall compression rate is defined. We see that it will be possible to utilize a collaboration of various coding methods in a sequence partitioning problem to reach a better compression performance and referred it as to multi-coding optimal partitioning.

The findings of this theoretical work contribute to comprehension of roles of partitioning in data compression. Bounds of overall compression rates for partitioned sequences is analytically derived (Equation (4)), which can, indeed, be considered as an extension of Shannon's limit of lossless compression rate for the case of sequences splitting. Specifically, for the case of $k = 1$, it yields Shannon's limit for lossless compression.

With a reverse consideration, one can also state that combining data packs with different entropy rates in order to obtain a larger data set can increase overall entropy due to the some waste of entropic volume. Entropy rate of combined data set is determined by the largest entropy rate of data packs.

Appendix

Derivation of Equation (3):

A finite sequence with l elements splits into k number of subsequences. Each subsequence has the lengths of a_i and the compression rate of R_i^*. R_i^* is the best compression of sequence defined as $R_i^* = \min\{R_i^j\}$ for a binary coding function family $\phi^j(.)$. R_i^j is the compression rate in the coding of the subsequence i by $\phi^j(.)$. The total bit numbers used in coding all subsequences by using the best $\phi^j(.)$s from the coding function family can be written as,

$$T = \sum_{i=1}^{k} a_i \cdot R_i^* .$$

Compression rate was defined as $R = T / l$. So, the overall compression rate for a partitioned sequence by using the best coding functions can be written as,

$$R = \sum_{i=1}^{k} \frac{a_i}{l} \cdot R_i^* .$$

When $w(a_i) = a_i / l$ is considered, one obtains the overall compression rate as:

$$R = \sum_{i=1}^{k} w(a_i) \cdot R_i^* .$$

Derivation of Equation (5):

In the coding of a sequence without partitioning, compression rate can be written as $R' > H$ due to Shannon's limit. With using Equation (4) and $R' > H$, the deviation in compression rate, which is $\Delta R = R' - R$, can be written as

$$\Delta R > H - \sum_{i=1}^{k} w(a_i) \cdot H_i .$$

Here, $w(a_i) = a_i / l$ and $l = \sum_{i=1}^{k} a_i$ for a k number partitioning. In this case,

$$\Delta R > H - \sum_{i=1}^{k} w(a_i) \cdot H_i = \frac{\sum_{i=1}^{k} a_i}{l} H - \sum_{i=1}^{k} \frac{a_i}{l} \cdot H_i \,,$$

$$= \sum_{i=1}^{k} \frac{a_i}{l} \cdot H - \sum_{i=1}^{k} \frac{a_i}{l} \cdot H_i = \sum_{i=1}^{k} w(a_i) \cdot (H - H_i) \,.$$

Finally, the Equation (5) is obtained as,

$$\Delta R > \sum_{i=1}^{k} w(a_i) \cdot (H - H_i) \,.$$

REFERENCES

[1] C.E. Shannon and Weaver W., *The Mathematical Theory of Communication*, Illinois, 1949.

[2] T.M. Cover and J.A. Thomas, Elements of Information Theory, Wiley India Pvt Ltd., 1991.

[3] D.A. Huffman, "A Method for the Construction of Minimum Redundancy Codes", Proceedings of the IRE, 40, 1952, pp. 1098-1101.

[4] J. Ziv and A. Lempel, "A Universal Algorithm for Sequential Data Compression", IEEE Transactions on Information Theory, 23, 1977, pp. 337-342.

[5] J. Ziv and A. Lempel, "Compression of Individual Sequences Via Variable-Rate Coding", IEEE Transactions on Information Theory, 24, 1978, pp. 530--536.

[6] R.R. Coifman, M.V. Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection", IEEE Transactions on Information Theory, 38, 1992, pp.713-718.

[7] K. Skretting, J.H. Husoy, S.O. Aase , "Improved Huffman coding using recursive splitting", Norwegian Signal Processing Society Conference (NORSIG-99),Norway, 1999.

[8] T.A. Welch, "A Technique for High-Performance Data Compression", Computer, 1984, pp. 8-18,.

[9] R.M. Hassan and B. Nath, "Data Compression Using Huffman Coding A Novel Approach", International Conference on Applied Computing (IADIS-2005), Portugal,2005.

[10] T. Bonny, J. Henkel, "Instruction Splitting for Efficient Code Compression", Design Automation Conference, 2007. DAC '07. 44th ACM/IEEE, 2007, pp. 646-651.

BIOGRAPHIES

Baris Baykant Alagoz was graduated from University of Istanbul Technical University department of Electronics and Communication Engineering in 1998. He worked for Alcatel Microelectronics and Turkish Telecom several years. He is following PhD at Inonu University of Department of Electrical &Electronics Engineering.

Hafiz Z. Alisoy was graduated from Moscow Technical University department of Electro-Physics Engineering in 1982. He had his PhD degree from USSR Science Academy Physics Institute of P.N. Lebedyev and Doctor of Sciences degree (DSc) from International Ecology-Energy Academy. He became as Full Professor in 1995. He received award of Young Scientist. He works at Inonu University department of Electrical &Electronics Engineering.