Non Uniform Coding for Data Encryption Based on Indivertible Bilinear Stochastic Models

Łukasz Maliński, *Silesian University of Technology*
(**01.10.2009**, dr hab. inż. Ewa Bielińska, prof. Pol. Śl. *Silesian University of Technology*)

Abstract

The application of indivertible elementary bilinear stochastic models in data encryption has been recently proposed. The preliminary result obtained, showed that it is possible to encrypt text information by bilinear stochastic process and restore it (decryption) by dedicated identification algorithm. This paper presents a discussion about improvement of encoding procedure used in that application basing on statistical properties of the estimation procedure applied to elementary bilinear stochastic process.

1. Introduction

The elementary bilinear stochastic models (EB) are the simplest representatives of bilinear stochastic models family. The exploration of the nonlinear stochastic modelling begun in 1978 Granger and Andersen [1] are recognised as the first researchers in this filed. They may also be counted as the first authors who focused their work on bilinear stochastic models [2]. Few years later other contributions to the field of bilinear time series modelling have been done by Subba T. Rao [3], Quinn, Gooijger and Heuts [5]. The last mentioned authors prepared the fundaments for method of moments, which is up to now one of the most common identification algorithm used for bilinear time-series. Also the Least Squares (LS) algorithm has been tested for bilinear stochastic models by Guegan and Pham [6]. Unfortunately, both of this algorithms showed very limited efficiency in estimation of the coefficients of bilinear stochastic models.

First improvement in LS algorithm was made by Bielińska and Nabaglo [7] in 1994. They proposed a simple modification for LS algorithm which stabilised the identification procedure and this way reduced the bias of the estimates obtained by it. This modification was based on enforcing the limit on identification error values, which estimates of stimulation signal in this case. One year later Brunner and Hess [8] discovered the next troublesome feature of the identification of bilinear time series models, studying the shape of the cost function in Maximum Likelihood function (ML). They found that this function may possess global minimum which is hard to access by common optimisation algorithm used in identification. The similar observations for LS algorithm were presented in [9]. Also it was stated that complicity of the cost function depends on the EB model coefficient value.

Further research performed by Maliński [10] concerned the influence of proposed in [7] modification on shape of the LS cost function. The author showed that it is possible to obtain unbiased estimates of EB model coefficient in its entire stability range if the enforced limit on identification error values is correctly selected. Next development by the same author [11] provided with solution how to automatically select the correct value of limit enforced on identification error values.

Finally, the practical application of indivertible EB models (which become identifiable) was presented in [12]. The noticeable precision obtained in identification of indivertible EB models provided with opportunity to use it in data encryption applications. The results were promising, but there are still some problems to be solved. Therefore, further in this paper the discussion will be presented to show if it is possible to improve the efficiency of this encryption application using some statistical features of identification of EB models.

2. Theoretical background

The elementary bilinear stochastic (EB) model (1) is defined as a sum of a single bilinear component and a stimulation sequence $e(t)$:

$$y(t) = e(t) + \beta e(t-k)y(t-l)\,. \tag{1}$$

Typically, some assumptions about a stimulation sequence $e(t)$ must be undertaken. For the purpose of this paper we assume that $e(t)$ sequence will have a Gaussian distribution thus following statistical properties:

$$\begin{aligned} &E\{e(t)\} = 0; \quad E\{e(t)^2\} = \lambda^2; \\ &E\{e(t)e(t-1)\} = 0; \quad E\{e(t)^3\} = 0. \end{aligned} \tag{2}$$

Due to above assumptions the stability (3) and inevitability (4) are defined by following formulas:

$$\beta^2\lambda^2 < 1; \tag{3}$$

$$\beta^2 m'^{(2)}_y < 1, \tag{4}$$

where $m'^{(2)}_y = E\{y(t)^2\}$.

In the practical application proposed in [12], the stable indivertible EB models are used. The stable indivertible bilinear model is the model for which coefficient β, and stimulation sequence variance λ^2 satisfy the stability condition (3), but do not satisfy invertibility condition (4). As shown in [10] and in [11] the biases for estimates obtained for those specific models are the lowest. Therefore, it is possible to assign the certain value of β coefficient to the character and encrypt it by generation the stochastic process using EB model. Then, the original β value can be restored using identification of the EB model with the same structure as original. In result the original assigned information can be restored.

Although, this encryption methodology seems to by trivial and uninnovative, there is a special feature of this approach to data encryption which make it interesting. The stochastic process obtained by simulation of EB model can carry the information but its autorotation function has zero (insignificant) values as reported in [12]. It means that everyone, who will use autocorrelation to seek for information imprinted in data sequence encrypted this way, will fail to find it. This is the main idea of using the indivertible EB models for data encryption.

Although, the dedicated identification algorithm is already developed, it is in publishing process at the time this paper is written, so it will not be presented here and it is not possible to provide the literature reference to it now. However, the results provided in [10] and [11] should be enough to prove that indivertible models are identifiable.

3. Encoding procedure

In the original paper [12] the encoding procedure used assignment of each alphanumeric character to specific value of β coefficient. The assignment has been made using predefined coding table which linked the characters to the precise sub-ranges of the coding range. Those sub-ranges were evenly distributed in the entire coding range so the width of each sub-range was the same.

Now, looking into results provided in [10] and [11] it is possible to come to the conclusion that scatter of identification results decreases along with increasing value of EB model coefficient β. It means that a chance to restore a character assigned to the sub-range at the end of the coding range is significantly larger than a chance to restore the character assigned to the sub-range placed at its beginning.

Due to this conclusion, the idea of the modification has been brought up to replace the evenly distributed sub-ranges with sub-ranges of successively decreasing width. This means that a sub-range at the beginning of the coding rage will be significantly wider than its counterpart from the end of the coding range. In theory it should increase the efficiency of this encryption methodology.

For the purpose of the above proposed modification, the following nonlinear sub-range width function has been proposed:

$$c_j = -4hx^2 + (c_m - c_0 + 4h)x + c_0, \tag{5}$$

where: c_j is the right border of the j sub-range and the left border of the c_{j+1} sub-range, h is the deviation from the linear width function (see fig. 1), c_m is the end of the entire coding range and also the right border of the last ($j = m$) sub-range and finally c_0 is the beginning of the coding range and the left border of the first ($j = 1$) sub-range. The x is the auxiliary variable picked from range <0, 1> which is used to divide the coding range to explicit number of ranges (m). The range <0, 1> of the x variable is always divided evenly.

The example of nonlinear sub-range width function (bold line) for $m = 4$ sub-ranges is presented in figure 1. The h parameter defines the deviation from linear function (thin line) and provides the opportunity to control the function concavity (for $h > 0$) or convexity (for $h < 0$).

It is obvious that by taking the large enough absolute value of h, we can force the function to change monotonicity within the used range of x variable. This is certainly not a desired option, therefore the constraint on the function c first divertive (6) in point $(1, c_m)$ has to be enforced.

$$c' = -4h + c_m - c_0 \geq 0. \tag{6}$$

The solution to (6) provides with the following constraint (7) on the h value:

$$|h_{\max}| \leq \frac{c_m - c_0}{4}. \tag{7}$$

This way we obtained the maximum absolute value of h which satisfy our needs for control the shape of sub-range width function. Moreover, in order to make the work with this function even easier, the parameter α which is used in following formula:

$$h = \alpha h_{\max}, \tag{8}$$

can be introduced.

Using (8) we released the h parameter with α which can be picked from range <-1,1> to ensure the correct monotonicity of the sub-range width function.

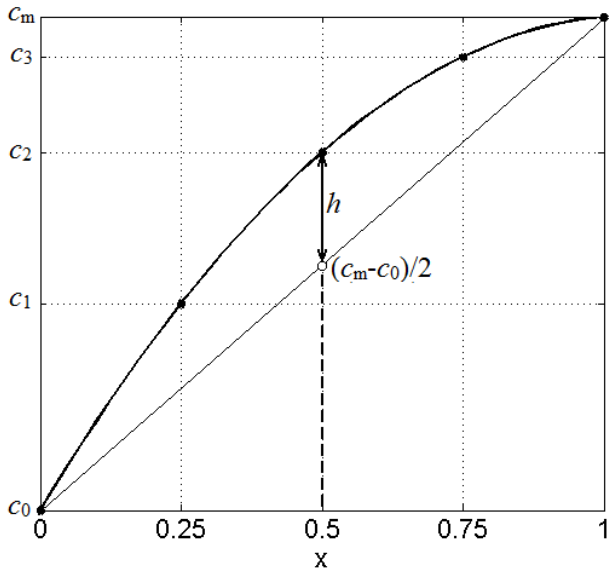


Fig.1. Nonlinear sub-range with function (example for m = 4 sub-ranges).

Finally, in order to obtain the sub-ranges borders (c_j values), the three arguments are needed (c_m, c_0 and α) and formulas (8), (7) and (5) have to be solved in that precise order.

The proposed nonlinear sub-range width function provide us with possibility to assign the characters to the sub-ranges of decreasing ($0 < \alpha \leq 1$) or incising ($0 > \alpha \geq -1$) width. It is also possible to obtain the sub-ranges of equal width (evenly distributed in coding range) using function c and parameter $\alpha = 0$. This will produce the same coding tables as presented in original paper [12].

During the encoding procedure the β value equal to the centre of the j sub-range is assigned to the corresponding alphanumeric character (according to the coding table). Then the encryption is performed by simulation of N samples of bilinear stochastic process using EB model.

In the opposite procedure, first the decryption procedure is executed by acquiring the β value from identification performed on N samples of bilinear stochastic process. Then the decoding begun and obtained β value is compared to the borders of particular sub-ranges. The decoded character is taken from the coding table for j sub-range if identification result is satisfying following condition: $c_{j-1} \leq \beta < c_j$.

4. Simulation results

Initial simulations have been made similarly to [12] using the same information: 'The indivertible elementary bilinear time series models for data encryption', the same coding range, N-values, etc. Moreover, two different experiments were performed:

- the first with the same coding table as in [12] ($\alpha = 0$).
- the second with the modified coding table ($\alpha = 1$).

The performance indices have been computed according to this proposed in [12]:

- Efficiency Ratio – a percentage of successfully decrypted characters, defined as:

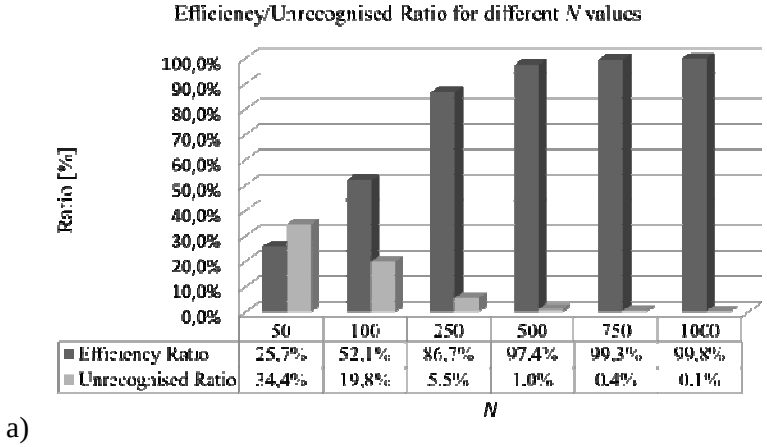
$$ER = \frac{1}{Rn}\sum_{i=1}^{R} s_i \cdot 100\% . \tag{9}$$

- Unrecognised Ratio – a percentage of unrecognised characters (identification results beyond used coding range), defined as:

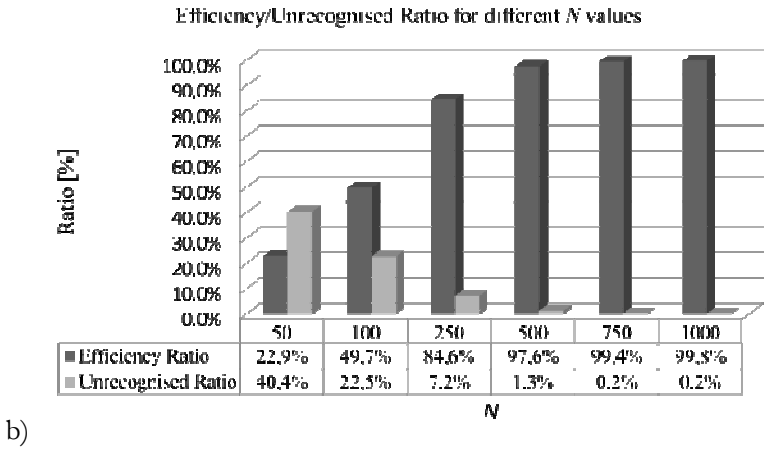
$$UR = \frac{1}{Rn}(Rn - \sum_{i=1}^{R} u_i) \cdot 100\% . \tag{10}$$

In formulas above, the R is the number of independent encryption/decryption runs, n is number of characters in information text, s_i is number of correctly decrypted characters in encryption/decryption run number i, and the u_i is the number of unrecognised characters (identification result beyond the coding range) in encryption/decryption run number i. The results obtained have been summarised in similar graphic representation (Fig. 2) as in [12].

The results obtained for $\alpha = 0$ (Fig. 2a) are very similar to those presented in [12], as it has been anticipated. However, surprising is the fact that the results presented in figure 2b ($\alpha = 1$) clearly shows that proposed modification slightly decreased the efficiency of the decryption. The most significant loss in efficiency has been observed for the lower values of N-values. The more thorough analysis of this unexpected outcome will be presented in the discussion below.



a)



b)

Fig.2. The summarised results of initial simulations: a) for $\alpha = 0$; b) for $\alpha = 1$.

Some changes have been made to improve the simulations efficiency. The first one concerned the information text itself. The original text used in [12] contain almost all letters which are lower case. Therefore, it cover a fragment of the coding range only. The new text information for testing is consisted every character in the coding table and is composed in following arrangement ("_" means "space"): "_0123456789ABC…XYZabc…xyz".

The second one concerns the *N*-values that are tested. Looking at the initial simulations the results obtained for *N*-values of 500 and higher are very similar regardless to the value of α parameter. Moreover, the efficiency of the encryption methodology using those high *N*-values are fairly satisfactory, so there is no immediate need for improvement in this area. Therefore, the following simulations and analysis concern the first three *N*-values (50, 100, 250) only.

The new simulations have been performed for $R = 100$ independent encryption/decryption procedures. As in the previous simulations, the overall efficiency has been evaluated only. This time the efficiency of decryption for every single character in the coding table is also rated. The simulations have been performed for four different α values taken from set $A = \{0, 0.2, 0.5, 1\}$. The overall performance indices are presented in table 1.

Tab.1.

Performance indices of the decryption for different α values.

α	0.0			0.2		
N	50	100	250	50	100	250
ER	23.6	50.1	84.1	24.2	50.5	84.8
UR	33.2	17.4	4.3	34.4	17.2	4.1
α	0.5			1.0		
N	50	100	250	50	100	250
ER	23.7	51.1	85.7	23.5	51.3	85.1
UR	34.0	18.0	4.3	36.2	19.3	5.1

As we can see, the introduction of sub-ranges with successively decreasing width ($\alpha > 0$) has very subtle impact on overall efficiency on decryption procedure. For low value of this parameter ($\alpha = 0.2$) a very little improvement for every *N*-value has been observed. For midrange parameter value ($\alpha = 0.5$) the slight improvement in efficiency has been observed for $N = 100$ and $N = 250$ samples, however this improvement is not satisfactory. Finally, the results obtained for maximum parameter value ($\alpha = 1.0$) present somehow average efficiency.

At this point those results can be interpreted in many different ways. For example, an improvement in efficiency at the beginning of the coding range is

nullified by deterioration in efficiency at the end of the coding range. Thus, it can be concluded that the measurement of the overall efficiency might not be a good idea and more thorough analysis is required. On the other hand, using a coding table with decreasing sub-ranges width might simply have not significant enough impact on the overall efficiency to overcome the randomness of the results which always accompany statistical analysis. Also the α value itself might have to be very precisely chosen in order to observe any improvements at all. This might be concluded from the results obtained for $\alpha = 0.2$.

To cast more light on this unexpected outcome, the following tests, analysing the distribution of successes in decryption for different fragments of the coding range have been performed. The coding table consisting 63 characters has been divided into 8 fragments called Character Groups and labelled as follows:

- "_-6" – this fragment consist "space" and digits from "0" to "6",
- "7-E" – this fragment consist digits "8" and "9" and upper case letters from "A" to "E",
- "F-M" - this fragment consist upper case letters from "F" to "M",
- "N-U" - this fragment consist upper case letters from "N"-"U",
- "V-c" – this fragment consist upper case letters from "V" to "Z" and lower case letters from "a" to "c",
- "d-k"- this fragment consist lower case letters from "d" to "k",
- "l-s" - this fragment consist lower case letters from "l" to "s",
- "t-z" - this fragment consist lower case letters from "t" to "z".

After the simulations, during the decoding procedure Character Efficiency Ratio (CER) defined in (11) has been computed for every character in the coding table:

$$CER = \frac{s_j}{R} 100\%, \qquad (11)$$

where s_j is the number of successful decryptions of j character in the coding table (j = 1, 2, …, 63). In the next step, the CER statistics for every character in particular fragment of the coding table have been averaged. This procedure has been performed for every value of α picked from set A. The results obtained for N = 50, N = 100 and N = 250 are presented in figure 3.

The results presented in figure 3a-c explain the source of problem encountered during computation of overall efficiency of the decryption procedure due to α value. An explicit improvement in efficiency is observed in the first 3 fragments (the beginning of the coding range) and even more explicit deterioration of the efficiency is observed in last 2 or 3 fragments (the end of the coding range).

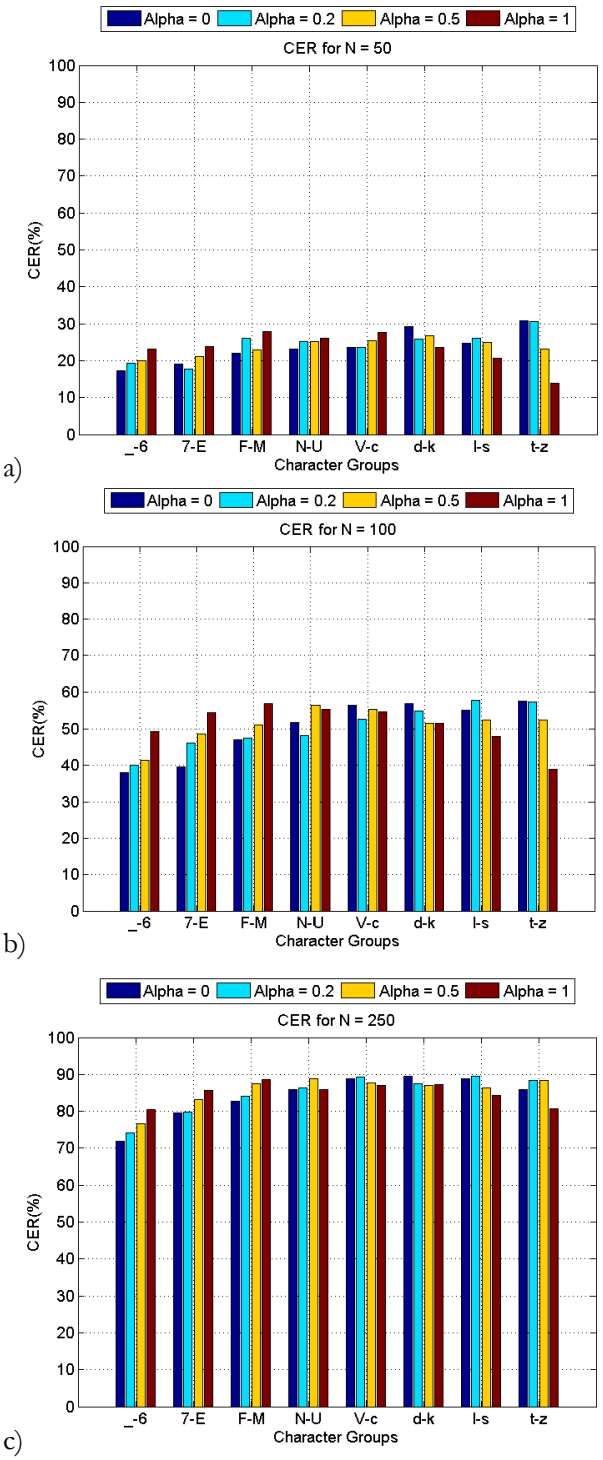


Figure 3. Averaged CER values for a) N = 50; b) N = 100; c) N = 250.

5. Summary

Although the results presented in the previous section are not very optimistic, they seem to explain the lack of expected improvement. As long as, the improvement has not been achieved, the results obtained provide us with interesting remarks. It seems that making further changes to the nonlinear subranges width function is not a good idea. Even if a more complicated shape of this function finally pro-

vides us with a noticeable improvement in decryption efficiency, there is no reason to expect those changes to be significant Therefore, completely different approach to the problem is required in order to push further the development of encryption/decryption methodology based on indivertible EB model.

At this point the possible solution might be seen in reduction of characters in the coding table. Perhaps a two level coding is recurred. First the text information will be coded into transitional code composed of lesser characters and then an encryption with EB model might be performed.

Bibliography

[1] Granger C., Andersen A.: *Nonlinear time series modelling Applied Time series analysis*. Academic Press (1978)

[2] Granger C., A. Andersen.: *An introduction to bilinear time series model,* Vandenhoeck and Ruprecht (1978)

[3] Subba Rao T.: *On the Theory of Bilinear Time Series Models,* Journal of the Royal Statistical Society vol. B44, 244-255 (1981)

[4] Quinn B.: *Stationarity and invertibility of simple bilinear models,* Stochastic Processes and Their Applications vol. 12, 225-230 (1982)

[5] Gooijger J., Heuts R.: *Higher order moments of bilinear time series processes with symmetrically distributed errors,* Proceedings to Second International Tempere Conference in Statistics, 467-478 (1987)

[6] Guegan D., Pham D. T.: *A Note on the Estimation of the Parameters of the Diagonal Bilinear Model by Method of Least Squares,* Scandinavian Journal of Statistics vol. 16, 129-136, (1989)

[7] Bielińska E., Nabagło I. *A modification of ELS algorithm for bilinear time-series model identification,* Zeszyty Naukowe Politechniki Śląskiej: Automatyka, vol. 108, 7-24 (1994)

[8] Brunner A., Hess G.: *Potential problems in estimating bilinear time-series models,*. Journal of Economic Dynamics and Control vol. 19, 663-681 (1995)

[9] Maliński Ł., Bielińska E.: *Statistical Analysis of Minimum Prediction Error Variance in the Identification of a Simple Bilinear Time-Series Model,* Advances in System Science, Academic Publishing House EXIT, 183-188 (2010)

[10] Maliński Ł.: *On identification of coefficient of indivertible elementary bilinear time-series model,* Proceedings XIV Symposium: Fundamental Problem Of Power Electronics Electromechanics and Mechatronics PPEEm, 194-196 (2011)

[11] Maliński Ł.: *The Evaluation of Saturation Level for SMSE Cost Function in Identification of Elementary Bilinear Time-Series Mode,* 17 International Conference on Methods and Models in Automation and Robotics (2012)

[12] Maliński Ł.: *Indivertible Elementary Bilinear Time-Series Models for Data Encryption,* 18th International Conference on System Science, Wrocław (2013)

Author:

MSc. Łukasz Maliński
Silesian University of Technology
ul. Akademicka 16
44-100 Gliwice
tel. (032) 237 19 04
fax (032) 237 21 27
email: *lukasz.malinski@polsl.pl*