Information Processing Letters 113 (2013) 628–633

Contents lists available at SciVerse ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Universal homophonic coding scheme using differential encoding and interleaving

D.R. Simões [a], J. Portugheis [b,*], V.C. da Rocha Jr. [a]

[a] *Communications Research Group – CODEC, Department of Electronics and Systems, Federal University of Pernambuco, Recife, 50740-550, Brazil*
[b] *Faculty of Technology, State University of Campinas, Limeira, 13484-332, Brazil*

ARTICLE INFO

Article history:
Received 15 February 2010
Received in revised form 17 April 2013
Accepted 17 May 2013
Available online 27 May 2013
Communicated by A. Tarlecki

Keywords:
Homophonic coding
Universal homophonic coding
Source coding
Cryptography

ABSTRACT

A universal homophonic coding scheme is introduced which combines properties of differential encoding and homophonic symbol interleaving. The use of a uniform random interleaver is discussed and it is argued that asymptotically, when the interleaver length grows to infinity, the homophonic coding scheme generates a sequence of independent and uniformly distributed binary digits. For practical reasons, an interleaver with a complexity that grows linearly with its length is proposed. As a consequence, both the generation and the recovery of the digits produced by the proposed universal homophonic coding scheme are easy to implement. Validation tests for this new universal homophonic coding scheme were performed using the statistical test suite IR 6483 available from the *National Institute of Standards and Technology*, USA, and the corresponding computer simulation results are presented. The statistical tests demonstrate that the performance of the proposed universal homophonic coding scheme is at least as good as that of other published schemes, however with a significant reduction in plaintext expansion.

1. Introduction

This paper deals with the encoding of messages by a technique called homophonic substitution [1] or homophonic coding in such a way that statistical cryptanalytic attacks are thereafter more difficult when the encoded message is encrypted. The encoding removes redundancy and thus makes statistical attacks less efficient, at the cost of some plaintext expansion. Although homophonic coding is not an encryption scheme itself it is however a relevant technique used in cryptography for fending statistical attacks. Statistical attacks in cryptography basically exploit the ciphertext in a secret-key cryptosystem, looking for statistical deviations from a completely random sequence, i.e., deviations from a sequence of independent and uniformly distributed (i.u.d.) symbols. In its classical form, previous knowledge of the plaintext statistics is required before homophonic coding can be applied, i.e., classical homophonic coding is source dependent.

In 1988, a pioneering paper by C.G. Günther [2] introduced a classical (i.e., source dependent) homophonic coding scheme in which the codewords representing homophones can have variable length, and assumed the plaintext to be a sequence of independent and identically distributed (i.i.d.) symbols. Günther's scheme transforms a plaintext sequence into the output sequence of a truly random source, producing what is known as perfect homophonic coding [1], i.e., the transformed plaintext sequence is an i.u.d. sequence. In [3], the authors proposed a homophonic coding scheme in which codewords have fixed length and the plaintext sequence is not necessarily an i.i.d. sequence. However, a possible difficulty for the use of these techniques is the requirement of *a priori* knowledge of the information source statistics. In many practical situations *a priori* information about the source statistics is not available and thus the use of a source specific homophonic

[*] Corresponding author.
E-mail addresses: drsdaniel82@gmail.com (D.R. Simões), jaime@ft.unicamp.br (J. Portugheis), vcr@ufpe.br (V.C. da Rocha).

0020-0190/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ipl.2013.05.010

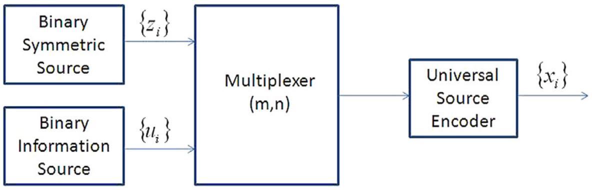

Fig. 1. Universal homophonic encoding scheme proposed by Massey, in which $\{z_i\}$, $\{u_i\}$ and $\{x_i\}$ denote, respectively, the sequences of symbols produced by the binary symmetric source, by the information source and by the universal source encoder.

coding scheme in these cases can be very inefficient. Contrasting with classical homophonic coding, Massey [4] suggested a universal (i.e., source independent) homophonic coding scheme which is based on a universal source coding scheme. Massey's scheme was investigated in [5] and validation tests were performed using the statistical test suite IR 6483 [6] published by the National Institute of Standards and Technology (NIST). The result of the tests in [5] indicated some statistical imperfections in the resulting homophonic sequence and hinted that improvements could still be made. A first improvement to Massey's scheme was proposed in [5] which also employed an auxiliary random source and in addition employed an interleaver. Fionov [7] also proposed a universal homophonic coding scheme, however differently from the schemes considered in this paper, variable-length coding and a small probability of error in decoding are admitted.

In this paper we propose a universal homophonic coding scheme which employs block differential encoding of source symbols in combination with symbols from an auxiliary random source and an interleaver. We stress however that this is not an encryption scheme but rather an encoding procedure to be applied to a message before being encrypted. It turns out that the introduction of block differential encoding reduces plaintext expansion and the resulting universal homophonic coding scheme performs statistically as well as the scheme suggested in [5], i.e., in terms of the randomness of the digits produced. Similarly to [5], recovery of the auxiliary random source digits at the decoder is not required by the proposed scheme. In fact, the scheme proposed earlier in [5] may be viewed as a special case of the proposed universal homophonic coding scheme where the length of the block is equal to two.

The paper organization is as follows. In Section 2 we review Massey's scheme [4] and describe the scheme introduced in [5], discussing advantages and drawbacks of both schemes. In Section 3, we present a new universal homophonic coding scheme using differential encoding combined with an interleaver and in Section 4 we describe the NIST statistical test suite [6] in detail, present simulation results and discuss the advantages and limitations of the proposed universal homophonic coding scheme.

2. Previous universal homophonic coding schemes

The universal homophonic coding scheme suggested by Massey [4] is illustrated in Fig. 1. In Fig. 1 a binary information source U emits symbols u_i, $i = 1, 2, \ldots,$ and a random generator, represented by the block labeled binary symmetric source (BSS) emits symbols z_i. The multiplexer considered is a device which successively outputs blocks of length $m+n$ symbols, containing m symbols from the BSS, followed by n symbols from U. The output from the multiplexer is fed as input to an appropriate universal source coding scheme. Since the multiplexing operation introduces non-stationarity, the multiplexer output sequence no longer in general represents a discrete stationary and ergodic source. However, blocks of multiplexed symbols, the length of which are multiples of $m + n$, produce a cycle-stationary process.

The original sequence $u_1, u_2, \ldots, u_i, \ldots$ produced by the information source U is recovered from the sequence $X = (x_1, x_2, \ldots, x_i, \ldots)$ produced at the output of the universal homophonic encoder, without knowledge of the random generator employed, by processing X by the respective universal source decoder and then by discarding those random bits that were originated from the BSS. In [5] statistical tests for Massey's universal homophonic coding scheme were conducted in order to measure the randomness of its output binary sequence by trying different information sources and various combinations of values for (m, n). The results of these tests, a sample of which is presented in Appendix A (Table 4), indicated that some improvements could still be made.

By time interleaving two sources with distinct statistics, the scheme of Fig. 1 relies on the universal source encoder to perform the rather hard task of learning the statistics of a source which alternately assumes a probability distribution during the time of m consecutive input symbols and changes to another probability distribution during the time of the next n input symbols. Moreover, an additional difficulty for the universal source encoder is the fact that the m bits coming from the BSS cannot be compressed. Even when the encoder operates with blocks the length of which is a multiple of $m+n$, the tests described in [5] showed a non-satisfactory statistical behavior. That was the main motivation for the proposal of the universal homophonic coding scheme shown in Fig. 2.

The binary information source U in Fig. 2 emits symbols u_i, $i = 1, 2, \ldots,$ and is assumed to be stationary and ergodic [4]. As before, the block labeled binary symmetric source (BSS) denotes a random generator with output Z emitting binary symbols z_i which are i.u.d. The multiplexer considered is a device which successively outputs blocks of length 2 symbols, containing one symbol z_i from the BSS and one symbol equal to $u_i \oplus z_i$, where $\oplus$ denotes the exclusive-or operation. The output of the multiplexer is fed as input to an appropriate interleaver. The scheme illustrated in Fig. 2 avoids the difficult situation illustrated in Fig. 1 of the changing statistics at the multiplexer output. Referring to Fig. 2, since the sequence $\{z_i\}$ is i.u.d. it follows that the sequence $\{u_i \oplus z_i\}$ is also i.u.d. Furthermore, we notice that

$$\begin{aligned} P_{ZY}(z_i, u_i \oplus z_i) &= P_Z(z_i) P_{Y|Z}(u_i \oplus z_i | z_i) \\ &= P_Z(z_i) P_U(u_i). \end{aligned} \tag{1}$$

It follows from (1) that, in general, Z and $U \oplus Z$ do not satisfy the condition required for statistical independence,