Theoretical Computer Science 412 (2011) 5232–5238

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

On z-factorization and c-factorization of standard episturmian words

N. Ghareghani [b], M. Mohammad-Noori [a,c,*], P. Sharifani [a]

[a] Department of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran
[b] School of Mathematics, Institute for Research in Fundamental Sciences (IPM), P.O. Box: 19395-5746, Tehran, Iran
[c] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.O. Box: 19395-5746, Tehran, Iran

ARTICLE INFO

Article history:
Received 15 August 2010
Received in revised form 5 May 2011
Accepted 21 May 2011
Communicated by M. Crochemore

Keywords:
Ziv–Lempel factorization
Crochemore factorization
Standard episturmian words

ABSTRACT

Ziv–Lempel and Crochemore factorization are two kinds of factorizations of words related to text processing. In this paper, we find these factorizations for standard epiesturmian words. Thus the previously known c-factorization of characteristic Sturmian words is provided as a special case. Moreover, the two factorizations are compared.

1. Introduction

Some factorizations of finite words were studied by Ziv and Lempel in a seminal paper [12]. These factorizations are related to information theory and text processing. Several years later, Crochemore introduced another factorization of words for the design of a linear time algorithm to detect squares in a word [3,4,6] and gave a space-efficient simple algorithm for computing the Ziv–Lempel factorization [5]. While these factorizations provide useful information about the structure of repeated factors, they can be computed in a linear time in the length of the word (see for instance [2]). This makes them useful algorithmic tools for finding repeated factors (See Chapter 8 of [14]). Another application of the Ziv–Lempel factorization to the approximation of grammar-based compression is discussed in [16].

The Crochemore factorization (or c-factorization in short) of a word $\mathbf{w}$ is defined as follows. Each factor of $c(\mathbf{w})$ is either a fresh letter, or it is a maximal factor of $\mathbf{w}$, which has already occurred in the prefix of the word. More formally, the c-factorization $c(\mathbf{w})$ of a word $\mathbf{w}$ is

$$c(\mathbf{w}) = (c_1, \ldots, c_m, c_{m+1}, \ldots),$$

where either c_m is the longest prefix of $c_m c_{m+1} \cdots$ occurring twice in $c_1 \cdots c_m$, or c_m is a letter a which has not occurred in $c_1 \cdots c_{m-1}$. The Ziv–Lempel factorization (or z-factorization in short) of a word $\mathbf{w}$ is

$$z(\mathbf{w}) = (z_1, \ldots, z_m, z_{m+1}, \ldots),$$

where z_m is the shortest prefix of $z_m z_{m+1} \cdots$ which occurs only once in the word $z_1 \cdots z_m$. As an example consider $\mathbf{w} = abacabcabacabacacabaa$. The c-factorization and z-factorization of $\mathbf{w}$ are as follows:

$$c(\mathbf{w}) = (a, b, a, c, ab, cab, acab, aca, caba, a),$$
$$z(\mathbf{w}) = (a, b, ac, abc, abacaba, cac, abaa).$$

* Corresponding author at: Department of Mathematics, Statistics and Computer Science, University of Tehran, Tehran, Iran. Tel.: +98 21 22252784.
E-mail addresses: ghareghani@ipm.ir (N. Ghareghani), morteza@ipm.ir, mnoori@khayam.ut.ac.ir (M. Mohammad-Noori), Psharifani@khayam.ut.ac.ir (P. Sharifani).

0304-3975/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2011.05.035

As it is seen c-factorization and z-factorization can be different but there are also some relations between them. In [2], it is shown that if a Ziv–Lempel factor includes a Crochemore factor, then it ends at most a letter after, and a Crochemore factor cannot include a Ziv–Lempel factor. It is concluded that the number of factors of the Crochemore factorization is at most twice the number of factors of the Ziv–Lempel factorization. Also the authors of [2] gave explicit formulas for Crochemore factorizations of some of the well-known infinite words, namely characteristic Sturmian words and (generalized) Thue–Morse words and the period doubling sequence, based on their combinatorial structures.

In this paper, we give explicit formulas for z-factorization and c-factorization of standard episturmian words; thus the previous c-factorization of characteristic Sturmian words in [2] appears as a special case. Moreover, these results reveal a very close relation between two factorizations in the case of standard episturmian words. The rest of the paper is organized as follows. In Section 2, we present some useful definitions and notation. Section 3, is devoted to review the definition and some properties of episturmian words. In Section 4, we study z-factorization of standard episturmian words. Finally in Section 5, we present a result about the c-factorization of standard episturmian words.

2. Definitions and notation

We denote the alphabet (which is finite) by $\mathcal{A}$. As usual, we denote by $\mathcal{A}^*$, the set of words over $\mathcal{A}$ and by ϵ the empty word. We use the notation $\mathcal{A}^+ = \mathcal{A}^* \setminus \{\epsilon\}$. If $a \in \mathcal{A}$ and $w = w_1 w_2 \dots w_n$ is a word over $\mathcal{A}$, then the symbols $|w|$ and $|w|_a$ denote respectively the length of w, and the number of occurrences of letter a in w. For an infinite word $\mathbf{w}$ we denote by $Alph(\mathbf{w})$ (resp. $Ult(\mathbf{w})$) the number of letters which appear (resp. appear infinitely many times) in $\mathbf{w}$ (the first notation is also used for finite words). A word v is a factor of a word w, written $v \prec w$, if there exists $u, u' \in \mathcal{A}^*$, such that $w = uvu'$. A word v is said to be a *p*refix (resp. suffix) of a word w, written $v \lhd w$ (resp. $v \rhd w$), if there exists $u \in \mathcal{A}^*$ such that $w = vu$ (resp. $w = uv$). If $w = vu$ (resp. $w = uv$) we simply write $v = wu^{-1}$ (resp. $v = u^{-1}w$). The notations of prefix and factor extend naturally to infinite words. We say that u is a *right special* (resp. *left special*) factor of w if ua, ub (resp. au, bu) are factors of w for some letters $a, b \in \mathcal{A}$ with $a \neq b$. Two words u and v are *c*onjugate if there exist words p and q such that $u = pq$ and $v = qp$. For a word $w \in \mathcal{A}^*$, the set $F(w)$ is the set of its factors and $F_n(w)$ is defined as $F_n(w) = F(w) \cap \mathcal{A}^n$. These notations are also used for infinite words. If $\mathbf{w}$ is an infinite word, then its *f* actor complexity function (or briefly its *c*omplexity function), is $p_{\mathbf{w}}(n) = |F_n(\mathbf{w})|$. It is easily proved that an infinite word $\mathbf{w}$ is aperiodic if and only if $p_{\mathbf{w}}(n) < p_{\mathbf{w}}(n+1)$ for any positive integer n. The *r*eversal of a word $w = w_1 w_2 \dots w_n$, with $w_i \in \mathcal{A}$ is $\overline{w} = w_n w_{n-1} \dots w_1$. The word w is a *p*alindrome if $w = \overline{w}$. A word $w \in \mathcal{A}^+$ is called *p*rimitive if $m \in \mathbb{N}_+$ and $w = u^m$ implies $m = 1$.

3. Episturmian words

Sturmian words are infinite words which are quite considerable by the number of their different characterizations coming from different mathematical areas, such as geometry, arithmetics and dynamical systems. A simple possible characterization is defining Sturmian words as aperiodic binary infinite words with minimal complexity, i.e. as infinite words $\mathbf{w}$ with $p_{\mathbf{w}}(n) = n + 1$. Hence, a Sturmian word has one right special factor of each length. Also it can be proved that for a Sturmian word $\mathbf{w}$ the set $F_{\mathbf{w}}$ is closed under reversal. A Sturmian word is called *characteristic* (*standard*) if all its left special factors are prefixes of it. A characteristic Sturmian word $\mathbf{w}$ can be computed as the limit of a sequence of words s_n defined recursively by

$$s_{-1} = b,\ s_0 = a,\ s_n = s_{n-1}^{d_n} s_{n-2},$$

where $d_1 \geq 0$ and $d_i > 0$ for $i = 2, 3, \dots$. The sequence $(d_1, d_2, \dots)$ is called the directive sequence and the word $0^{d_1}1^{d_2}0^{d_3}\cdots$ is called the directive word associated to $\mathbf{w}$. As in [2], one may assume $d_1 > 0$ based on a simple observation. The Sturmian word defined by a directive sequence $(0, d_2, d_3, \dots)$ is obtained from the Sturmian word defined by $(d_2, d_3, \dots)$ by exchanging the letters a and b. To see some equivalent definitions and various properties of Sturmian words, see Chapter 2 of [13].

One limitation of Sturmian words is that they are over a binary alphabet. Different characteristic properties of Sturmian words lead to natural generalizations on arbitrary finite alphabet, among which the so-called episturmian words appeared to be the best suited family by the number of properties they share with Sturmian words. This generalization is given and discussed in [8,10,11] based on a construction of Sturmian words given in [7]. In the rest of this section, we study the definition and some properties of episturmian words. For more information the reader is referred to [1,9].

An infinite word $\mathbf{s}$ is episturmian if $F(\mathbf{s})$ is closed under reversal and for any $\ell \in \mathbb{N}$ there exists at most one right special word in $F_\ell(\mathbf{s})$. Then Sturmian words are just nonperiodic episturmian words on a binary alphabet. An episturmian word is *standard* if all its left special factors are prefixes of it. It is well known that if an episturmian word $\mathbf{t}$ is not periodic and $Ult(\mathbf{t}) = k$, then its complexity function is ultimately $p_{\mathbf{t}}(n) = (k-1)n + q$ for some $q \in \mathbb{N}_+$. Let $\mathbf{t}$ be an episturmian word. If $\mathbf{t}$ is nonperiodic then there exists a unique standard episturmian word $\mathbf{s}$ satisfying $F_{\mathbf{t}} = F_{\mathbf{s}}$; if $\mathbf{t}$ is periodic then we may find several standard episturmian words $\mathbf{s}$ satisfying $F_{\mathbf{t}} = F_{\mathbf{s}}$. In any case, there exists at least one standard episturmian word $\mathbf{s}$ with $F_{\mathbf{t}} = F_{\mathbf{s}}$. If the sequence of palindromic prefixes of a standard episturmian word $\mathbf{s}$ is $u_1 = \epsilon, u_2, u_3, \dots$, then there exists an infinite word $\Delta(\mathbf{s}) = x_1 x_2 \cdots,\ x_i \in \mathcal{A}$ called its *directive word* such that for all $n \in \mathbb{N}_+$,

$$u_{n+1} = (u_n x_n)^{(+)}$$