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 $n \in \mathbb{N}_+$,

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

where $w^{(+)}$ is defined as the shortest palindrome having w as a prefix. The relation between u_n and u_{n+1} can also be explained using morphisms as in [8]. For $a \in \mathcal{A}$, they define the morphism ψ_a by $\psi_a(a) = a$ and $\psi_a(x) = ax$ for $x \in \mathcal{A} \setminus \{a\}$. Let $\mu_0 = Id$ and $\mu_n = \psi_{x_1}\psi_{x_2}\cdots\psi_{x_n}$ for $n \in \mathbb{N}_+$. Moreover, let $h_n = \mu_n(x_{n+1})$. Then

$$u_{n+1} = h_{n-1}u_n,\ n \in \mathbb{N}_+.$$

The above definitions are clarified in the following example.

Example 1. Let $\mathcal{A} = \{a, b, c\}$ and $\Delta(\mathbf{s}) = x_1x_2\cdots = (abc)^\omega$ be directive word of a standard episturmian word $\mathbf{s}$. By the first representation of episturmian words

$$\begin{aligned}
&u_1 = \epsilon,\\
&u_2 = (a)^+ = a,\\
&u_3 = (ab)^+ = aba,\\
&u_4 = (abac)^+ = abacaba,\\
&u_5 = (abacabaa)^+ = abacabaabacaba,\\
&u_6 = (abacabaabacabab)^+ = abacabaabacababacabaabacaba,\\
&\quad\cdots
\end{aligned}$$

and so on. Now, to compute u_i by the morphism representation of episturmian words, one should first compute $h_{i-1} = \mu_{i-1}(x_i)$ and then use $u_i = h_{i-2}u_{i-1}$.

$$\begin{aligned}
&h_0 = \mu_0(a) = a,\\
&h_1 = \mu_1(b) = \psi_a(b) = ab,\\
&h_2 = \mu_2(c) = \psi_a\psi_b(c) = abac,\\
&h_3 = \mu_3(a) = \psi_a\psi_b\psi_c(a) = abacaba,\\
&h_4 = \mu_4(b) = \psi_a\psi_b\psi_c\psi_a(b) = abacabaabacab,\\
&\quad\cdots.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&u_1 = \epsilon,\\
&u_2 = h_0u_1 = a,\\
&u_3 = h_1u_2 = aba,\\
&u_4 = h_2u_3 = abacaba,\\
&u_5 = h_3u_4 = abacabaabacaba,\\
&u_5 = h_4u_5 = abacabaabacababacabaabacaba,\\
&\quad\cdots.
\end{aligned}$$

The words h_i appeared in this example are finite Tribonacci words and their limit, $\xi = \lim_{n\to\infty} u_n$, given by

$$\xi = abacabaabacababacabaabacabacabaabacababacaba\ldots,$$

is known as infinite Tribonacci word or the Tribonacci sequence (to know more about this word and some of its properties, see[15,17]).

From equations $u_{n+1} = h_{n-1}u_n$, $u_1 = \epsilon$ and $u_{n+1} = \overline{u_{n+1}}$, it is concluded that

$$u_{n+1} = h_{n-1}\cdots h_1h_0 = \overline{h_0}\,\overline{h_1}\cdots\overline{h_{n-1}} \tag{1}$$

It is known that for any integer n, h_n is primitive (See Proposition 2.8 of [11]) and so is $\overline{h_n}$. For any integer n define $P(n)$ as the maximum value of i satisfying $i < n$ and $x_i = x_n$; if there is no such i then $P(n)$ is undefined. We have the following lemma.

Lemma 1. (i)

$$h_{n-1} = \begin{cases} u_nx_n & \text{if } P(n) \text{ is undefined,}\\ u_nu_{P(n)}^{-1} & \text{otherwise.}\end{cases}$$

(ii) *If $P(n)$ is defined then*

$$h_{n-1} = h_{n-2}h_{n-3}\cdots h_{P(n)-1}.$$

Proof. (i) See the end of Section 2.1 of [11].
(ii) This is proved by using part (i) and (1). □

It is obvious that $h_{n-1} \lhd h_n$. In addition, by Proposition 2.11 of [11] we have the following lemma.

Lemma 2. (i) *$h_n = h_{n-1}$ if and only if $x_{n+1} = x_n$.*
(ii) *If $x_{n+1} \neq x_n$ then u_n is a proper prefix of h_n.*

Lemma 3. *Let $\Delta(\mathbf{s}) = x_1 \dots x_n \dots$, $x_i \in \mathcal{A}$. Suppose that $x_n = \alpha$ and the letter α has at least one appearance before x_n in $\Delta(s)$. Then we have the following.*

(i) *$h_{n-1} \lhd u_n$ and $\overline{h_{n-1}} \rhd u_n$.*
(ii) *The word $v_{n-1} = u_n(\overline{h_{n-1}})^{-1}$ is palindrome.*
(iii) *$v_{n-1} \rhd u_{n-1}$ and $v_{n-1} \lhd u_{n-1}$.*
(iv) *$u_n \rhd u_{n-1}\overline{h_{n-1}}$.*
(v) *If moreover $x_n \neq x_{n-1}$, then $u_n \rhd (\overline{h_{n-1}})^2$ and $u_{n+1} \rhd (\overline{h_{n-1}})^3$.*

Proof. (i) By Lemma 1(i), $h_{n-1} = u_n u_{P(n)}^{-1}$. So $h_{n-1} \lhd u_n$, which concludes $\overline{h_{n-1}} \rhd \overline{u_n} = u_n$.
(ii) By part (i), there exists a word v_{n-1} satisfying $u_n = v_{n-1}\overline{h_{n-1}}$. Hence by $u_{n+1} = h_{n-1}u_n$, we obtain $u_{n+1} = h_{n-1}v_{n-1}\overline{h_{n-1}}$. But since u_{n+1} is palindromic, from the last equation we conclude that so is v_{n-1}.
(iii) From $u_n = v_{n-1}\overline{h_{n-1}} = u_{n-1}h_{n-2}$ and $|h_{n-1}| \geq |h_{n-2}|$ we conclude that $v_{n-1} \lhd u_{n-1}$, which yields $v_{n-1} = \overline{v_{n-1}} \rhd \overline{u_{n-1}} = u_{n-1}$.
(iv) This is concluded from $u_n = v_{n-1}\overline{h_{n-1}}$ using part (iii).
(v) Using part (iii) and Lemma 2(ii), we get $u_n \rhd (\overline{h_{n-1}})^2$. Combining this with $u_{n+1} = u_n\overline{h_{n-1}}$, we provide $u_{n+1} \rhd (\overline{h_{n-1}})^3$. □

The following representation of directive word is useful for next sections. Let

$$\Delta(\mathbf{s}) = x_1x_2 \cdots = y_1^{d_1}y_2^{d_2} \cdots,$$

where $x_i, y_i \in \mathcal{A}$, $y_i \neq y_{i+1}$ and $d_i > 0$ for $i > 0$. Define the function $g : \mathbb{N} \to \mathbb{N}$ by

$$g(m) = d_1 + \cdots + d_{m-1} + 1.$$

Lemma 4. *With the above definitions, the following statements hold.*

(i) $u_{g(m+1)} = (h_{g(m)-1})^{d_m}u_{g(m)} = u_{g(m)}(\overline{h_{g(m)-1}})^{d_m}$.
(ii) $u_{g(m+1)} = (h_{g(m)-1})^{d_m}(h_{g(m-1)-1})^{d_{m-1}} \cdots (h_0)^{d_1} = (\overline{h_0})^{d_1}(\overline{h_{g(2)-1}})^{d_2} \cdots (\overline{h_{g(m)-1}})^{d_m}$.
(iii) *$u_{g(m)-1}$ is a proper prefix of $h_{g(m)-1}$.*
(iv) *$u_{g(m)} \rhd (\overline{h_{g(m)-1}})^2$ and $u_{g(m+1)} \rhd (\overline{h_{g(m)-1}})^{d_m+2}$.*

Proof. (i) For any integer n with $g(m) \leq n < g(m+1)$ we have $x_n = y_m$ and by Lemma 2(i), $h_{n-1} = h_{g(m)-1}$. Thus for any integer j with $0 \leq j \leq d_m$, we have

$$u_{g(m)+j} = (h_{g(m)-1})^j u_{g(m)} = u_{g(m)}(\overline{h_{g(m)-1}})^j.$$

Particularly for $j = d_m$ the result is provided.
(ii) This is concluded from (1).
(iii) This is obtained from Lemma 2(ii).
(iv) By Lemma 3(v) we obtain $u_{g(m)} \rhd (\overline{h_{g(m)-1}})^2$; Using this and $u_{g(m)+1} = u_{g(m)}(\overline{h_{g(m)-1}})^{d_m}$, we provide $u_{g(m+1)} \rhd (\overline{h_{g(m)-1}})^{d_m+2}$. □

4. z-factorization

Recall that the Ziv–Lempel factorization (z-factorization) of a word $\mathbf{w}$ is $z(\mathbf{w}) = (z_1, \dots, z_m, z_{m+1}, \dots)$, where z_m is the shortest prefix of $z_mz_{m+1} \cdots$ which occurs only once in the word $z_1 \cdots z_m$.

Theorem 5. *Let $\mathbf{s}$ be an episturmian word with directive word $\Delta(\mathbf{s}) = x_1x_2x_3 \dots = y_1^{d_1}y_2^{d_2} \dots$, where $x_i, y_i \in \mathcal{A}$ and $y_i \neq y_{i+1}$, for all $i \geq 1$. The z-factorization of $\mathbf{s}$ is of the form $z(\mathbf{s}) = (z_1, z_2, \dots)$, where $z_1 = x_1$ and $z_k = y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k$ for $k \geq 2$.*

Proof. We prove the result by induction on k. It is easily seen that $z_1 = x_1 = y_1$. Now suppose that the result is true for any $j < k$. Thus we have

$$\begin{aligned} z_1z_2z_3 \cdots z_{k-1} &= y_1\ y_1^{-1}(\overline{h_{g(1)-1}})^{d_1}y_2\ y_2^{-1}(\overline{h_{g(2)-1}})^{d_2}y_3\ \dots\ y_{k-2}^{-1}(\overline{h_{g(k-2)-1}})^{d_{k-2}}y_{k-1} \\ &= (\overline{h_{g(1)-1}})^{d_1}(\overline{h_{g(2)-1}})^{d_2} \cdots \overline{h_{g(k-2)-1}})^{d_{k-2}}y_{k-1} \\ &= u_{g(k-1)}y_{k-1}. \end{aligned}$$

We should conclude that $z_k = y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k$. For this purpose, the two following facts should be proved.

Fact 1. $y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}} \prec u_{g(k)}x_1^{-1}$.

Fact 2. $y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k \not\prec u_{g(k)}$.

We prove these facts in two cases.

Case (i). Suppose that $y_{k-1} = \alpha$ has already appeared in $\Delta(\mathbf{s})$. By Lemma 3 (i), $\overline{h_{g(k-1)-1}} \rhd u_{g(k-1)}$ hence

$$y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}} \rhd u_{g(k-1)}(\overline{h_{g(k-1)-1}})^{d_{k-1}-1}.$$

But the right side, is a prefix of $u_{g(k)} = u_{g(k-1)}(\overline{h_{g(k-1)-1}})^{d_{k-1}}$. This proves Fact 1.

To prove Fact 2, by contrary, suppose that

$$y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k \prec u_{g(k)}. \tag{2}$$

By Lemma 4(iv), $u_{g(k-1)} \rhd (\overline{h_{g(k-1)-1}})^2$ so

$$u_{g(k)} = u_{g(k-1)}(\overline{h_{g(k-1)-1}})^{d_{k-1}} \rhd (\overline{h_{g(k-1)-1}})^{d_{k-1}+2}. \tag{3}$$

From (2) and (3) we conclude

$$y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k \prec (\overline{h_{g(k-1)-1}})^{d_{k-1}+2},$$

which implies that $y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k = w^{d_{k-1}}$ for some $w \sim \overline{h_{g(k-1)-1}}$, but this is possible only if $y_{k-1} = y_k$ which is a contradiction. Hence, Fact 2 is proved in this case.

Case (ii). Suppose that $y_{k-1} = \alpha$ has not appeared before in $\Delta(\mathbf{s})$, hence,

$$\overline{h_{g(k-1)-1}} = y_{k-1}u_{g(k-1)}. \tag{4}$$

Thus Fact 1 is proved as follows

$$y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}} = u_{g(k-1)}(y_{k-1}u_{g(k-1)})^{d_{k-1}-1} \lhd u_{g(k-1)}(\overline{h_{g(k-1)-1}})^{d_{k-1}}x_1^{-1} = u_{g(k)}x_1^{-1}.$$

In order to prove Fact 2, suppose by contrary that

$$y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k \prec u_{g(k)} \tag{5}$$

On the other hand, by (4) and $u_{g(k)} = u_{g(k-1)}(\overline{h_{g(k-1)-1}})^{d_{k-1}}$, we obtain

$$u_{g(k)} = u_{g(k-1)}(y_{k-1}u_{g(k-1)})^{d_{k-1}} \prec (y_{k-1}u_{g(k-1)})^{d_{k-1}+1} = (\overline{h_{g(k-1)-1}})^{d_{k-1}+1} \tag{6}$$

From (5) and (6) we provide

$$y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k \prec (\overline{h_{g(k-1)-1}})^{d_{k-1}+1}$$

which implies that $y_{k-1}^{-1}(\overline{h_{g(k-1)-1}})^{d_{k-1}}y_k = w^{d_{k-1}}$ for some $w \sim \overline{h_{g(k-1)-1}}$, but this is possible only if $y_{k-1} = y_k$, which is a contradiction. This ends the proof. □

Example 2. Using the definition, the z-factorization of the Tribonacci word, ξ, mentioned in Example 1, is obtained as follows

$$z(\xi) = (a, b, ac, abaa, bacabab, acabaabacabac, abaabacababacabaabacabaa, \ldots).$$

To obtain this result using Theorem 5, note that for the Tribonacci word, we have $d_n = 1$, $y_n = x_n$, and $g(n) = n$ for any positive integer n. Moreover, we have $y_{3n-2} = a$, $y_{3n-1} = b$ and $y_{3n} = c$ for any positive integer n. Thus we obtain

$$z_1 = x_1 = a,$$
$$z_2 = a^{-1}\overline{h_0}b = b,$$
$$z_3 = b^{-1}\overline{h_1}c = ac,$$
$$z_4 = c^{-1}\overline{h_2}a = abaa,$$
$$z_5 = a^{-1}\overline{h_3}b = bacabab,$$
$$z_6 = b^{-1}\overline{h_4}c = acabaabacabac,$$
$$\cdots.$$

5. c-factorization

Recall that the Crochemore factorization (c-factorization) $c(\mathbf{w})$ of a word $\mathbf{w}$ is $c(\mathbf{w}) = (c_1, \cdots, c_m, c_{m+1}, \ldots)$ where 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 $c_1 \cdots c_{m-1}$.

Theorem 6. *Let* $\mathbf{s}$ *be an episturmian word with directive word* $\Delta(\mathbf{s}) = x_1x_2x_3 \ldots = {y_1}^{d_1}{y_2}^{d_2}{y_3}^{d_3} \ldots$, *where* $x_i, y_i \in \mathcal{A}$ *and* $y_i \neq y_{i+1}$, *for all* $i \geq 1$. *If* $c(s) = (c_1, c_2, \ldots)$, *then there exist integers* i *and* j *such that* $c_1 \cdots c_k = u_{g(k-j+i+1)}$ *for any* $k \geq j$. *Consequently, we obtain* $c_k = (\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}}$, *for all* $k > j$.

Proof. Let $i = \min\{t : \{y_1, y_2, \ldots, y_t\} = \{1, 2, \ldots, k\}\}$ and $y_i = \alpha$. Since $y_i = \alpha$ has no occurrence in $u_{g(i)}$, we have $u_{g(i)+1} = u_{g(i)}\alpha u_{g(i)}$, hence there exists an integer $j \geq 3$ satisfying $c_1 \cdots c_{j-2} = u_{g(i)}$ and $c_{j-1} = y_i$. Moreover, by Lemma 1(i), we have

$$\overline{h_{g(i)-1}} = y_i u_{g(i)} \tag{7}$$

$$u_{g(i+1)} = u_{g(i)}(y_i u_{g(i)})^{d_i} \tag{8}$$

Now we are going to prove that $c_j = y_i^{-1}(\overline{h_{g(i)-1}})^{d_i}$. Denote the right side by w and note that

$$c_1 \cdots c_{j-1} w = u_{g(i)}\, y_i\, y_i^{-1}(\overline{h_{g(i)-1}})^{d_i} = u_{g(i+1)}.$$

It is clear that $w = u_{g(i)}(y_i u_{g(i)})^{d_i - 1}$ has at least two occurrences in $c_1 \cdots c_{j-1} w = u_{g(i)}(y_i u_{g(i)})^{d_i}$. Thus it is enough to prove that $w y_{i+1} \not\prec u_{g(i+1)}$. By contrary, suppose that $w y_{i+1} \prec u_{g(i+1)}$ so

$$u_{g(i)}(y_i u_{g(i)})^{d_i - 1} y_{i+1} \prec u_{g(i)}(y_i u_{g(i)})^{d_i} \tag{9}$$

Since $y_i \notin Alph(u_{g(i)})$, (9) can happen only if $y_{i+1} = y_i$ which is a contradiction. Thus $c_j = w$ as required.

Now we claim that the following equation

$$c_1 c_2 \cdots c_k = u_{g(k-j+i+1)} \tag{10}$$

holds for any integer $k \geq j$. The statement is true for $\ell = j$ by the above arguments. We proceed by induction on k. Suppose that $k > j$ and that $c_1 c_2 \cdots c_\ell = u_{g(\ell-j+i+1)}$ holds for any integer ℓ with $j \leq \ell < k$. By Lemma 4 (i), it is enough to show that $c_k = (\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}}$. For this, the two following facts should be proved

Fact 1. $(\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}} \prec u_{g(k-j+i+1)}x_1^{-1}$

Fact 2. $(\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}} y_{k-j+i+1} \not\prec u_{g(k-j+i+1)}$

By Lemma 3 (i), $\overline{h_{g(k-j+i)-1}} \rhd u_{g(k-j+i)}$, we provide

$$(\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}} \rhd u_{g(k-j+i)}(\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}-1},$$

which together with $u_{g(k-j+i)}(\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}-1} \prec u_{g(k-j+i+1)}x_1^{-1}$ proves Fact 1.

To prove Fact 2, suppose by contrary that $(\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}} y_{k-j+i+1} \prec u_{g(k-j+i+1)}$. By using Lemma 4 (iv), this concludes that

$$(\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}} y_{k-j+i+1} \prec (\overline{h_{g(k-j+i)-1}})^{d_{k-j+i}+2}.$$

Since h_t is primitive, it has just $d_{k-j+i} + 2$ occurrences in the right side; Thus the last relation implies that $y_{k-j+i+1}$ equals the first letter of $\overline{h_{g(k-j+i)-1}}$, i.e. $y_{k-j+i+1} = y_{k-j+i}$ which is a contradiction. □

Example 3. Using the definition, the c-factorization of the Tribonacci word, ξ, mentioned in Example 1, is obtained as follows

$$c(\xi) = (a, b, a, c, aba, abacaba, bacabaabacaba, cabaabacababacabaabacaba, \ldots).$$

Now, to recompute these c_is using Theorem 6, note that $d_n = 1$, $g(n) = n$ and $y_n = x_n$ for each $n \geq 1$. By the proof of the Theorem 6, $i = \min\{t : \{y_1, y_2, \ldots, y_t\} = \{1, 2, \ldots, k\}\}$ so in this example $i = 3$. It is easy to see that $c_1 = a, c_2 = b, c_3 = a, c_4 = c = x_3$ and $c_5 = aba$, therefore by the definition of j in the proof of the Theorem 6, $j = 5$. Hence, we have $c_k = (\overline{h_{g(k-2)-1}})^{d_{k-2}} = \overline{h_{k-3}}$, for $k > 5$. Therefore,

$$c_6 = \overline{h_3} = abacaba,$$

$$c_7 = \overline{h_4} = bacabaabacaba,$$

$$\cdots.$$

Remark 1. By slight modification of the argument used in the proof of Theorem 6, we find that c-factorization of a standard episturmian word is as follows: $c_1 = y_1$ and

$$c_2 = \begin{cases} y_2 & \text{if } d_1 = 1, \\ y_1^{d_1-1} & \text{otherwise.} \end{cases}$$

For any integer $m \geq 2$, there exists an integer n such that $c_1c_2 \cdots c_m = u_{g(n)}\alpha_n$, where either $\alpha_n = \epsilon$ or $\alpha_n = y_n$. In addition, the next factor, c_{m+1}, is given by

$$c_{m+1} = \begin{cases} y_n & \text{if } \alpha_n = \epsilon \text{ and } y_n \notin \{y_1, \ldots, y_{n-1}\}, \\ (\overline{h_{g(n)-1}})^{d_n} & \text{if } \alpha_n = \epsilon \text{ and } y_n \in \{y_1, \ldots, y_{n-1}\}, \\ y_n^{-1}(\overline{h_{g(n)-1}})^{d_n} & \text{otherwise, i.e. if } \alpha_n = y_n. \end{cases}$$

It is concluded that if $\alpha_n = \epsilon$ and $y_n \notin \{y_1, \ldots, y_{n-1}\}$, then $c_1 \cdots c_{m+1} = u_{g(n)}y_n$; otherwise $c_1 \cdots c_{m+1} = u_{g(n+1)}$. Moreover, setting $k_0 = |Alph(\mathbf{s})|$, it is provided that the values i and j in Theorem 6, satisfy the following equation.

$$j - i = \begin{cases} k_0 - 1 & \text{if } d_1 = 1, \\ k_0 & \text{otherwise.} \end{cases}$$

Remark 2. Considering Theorems 5 and 6 and Remark 1, we conclude that from a point on, the formula $z_k = y_{k-1}^{-1}c_{k+k_0-1-m}y_k$ holds, where $k_0 = |Alph(\mathbf{s})|$ and

$$m = \begin{cases} 1 & \text{if } d_1 = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Remark 3. From Theorem 6, by using Remark 1, we obtain that from a point on, $c_k = (\overline{h_{g(k-k_0+m)-1}})^{d_{k-k_0+m}}$, where k_0 and m are defined as above. Now if $\mathbf{s}$ is characteristic Sturmian, by using definition and notation of Chapter 2 of [13] about standard words and Sturmian words, it is easily proved that $h_{g(p)-1} = s_{p-1}$, for any integer $p \geq 1$. So in this case, by replacing $k_0 = 2$, we conclude that from a point on, $c_k = (\overline{s_{k+m-3}})^{d_{k+m-2}}$. Thus in the case $d_1 > 1$ (resp. $d_1 = 1$) by calculating the first four factors (resp. first three factors), we conclude Theorem 1 of [2] about c-factorization of characteristic Sturmian words.

Acknowledgement

The research of the first author was in part supported by a grant from IPM (No. 89050046). The research of the second author was in part supported by a grant from IPM (No. CS1390-4-07).

References

[1] J. Berstel, Sturmian and episturmian words (A survey for some recent results), LNC 4728 (2007) 23–47.
[2] J. Berstel, A. Savelli, Crochemore factorization of Sturmian and other infinite words, in: Rastislav Kralovic, Pawel Urzyczyn (Eds.), Mathematical Foundations of Computer Science 2006, 31st International Symposium, MFCS 2006, Star Lesn, Slovakia, August 28-September 1, 2006, Proceedings, in: Lecture Notes in Computer Science, vol. 4162, Springer-Verlag, 2006, pp. 157–166.
[3] M. Crochemore, Recherche linéaire d'un carré dans un mot, Comptes Rendus Sci. Paris Sér. 1 Math. 296 (1983) 781–784.
[4] M. Crochemore, C. Hancart, T. Lecroq, Algorithmique du texte, Vuibert, 2001.
[5] M. Crochemore, L. Ilie, W.F. Smyth, A simple algorithm for computing the Lempel-Ziv factorization, in: J.A. Storer, M.W. Marcellin (Eds.), 18th Data Compression Conference, IEEE Computer Society, Los Alamitos, CA, 2008, pp. 482–488.
[6] M. Crochemore, W. Rytter, Text Algorithms, The Clarendon Press Oxford University Press, 1994.
[7] A. de Luca, Sturmian words: structure, combinatorics and their arithmetics, Theoret. Comput. Sci. 183 (1997) 45–82.
[8] X. Droubay, J. Justin, G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, Theoret. Comput. Sci. 255 (2001) 539–553.
[9] A. Glen, J. Justin, Episturmian words (A survey), ITA 43 (3) (2009) 403–442.
[10] J. Justin, G. Pirillo, On a characteristic propery of Arnoux–Rauzy sequences, Theore. Inform. Appl. 36 (4) (2002) 385–388.
[11] J. Justin, G. Pirillo, Episturmian words and episturmian morphisms, Theoret. Comput. Sci. 276 (2002) 281–313.
[12] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Transactions in Information Theory IT-22 (1976) 75–81.
[13] M. Lothaire, Algebraic Combinatorics on Words, Cambridge University Press, London, 2002.
[14] M. Lothaire, Applied Combinatorics on Words, Cambridge University Press, London, 2005.
[15] G. Rauzy, Nombres algébriques et substitutions, Bulletin de la Société Mathématique de France 110 (2) (1982) 147–178.
[16] W. Rytter, Application of Lempel-Ziv factorization to the approximation of grammar-based compression, Theoret. Comput. Sci. 302 (2003) 211–222.
[17] B. Tan, Z. Wen, Some properties of the Tribonacci sequence, European J. Combin. 28 (2007) 1703–1719.