Theoretical Computer Science 565 (2015) 90–101

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Vertical representation of $\mathbf{C}^\infty$-words ☆

Jean-Marc Fédou [a], Gabriele Fici [b,*]

[a] Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis, CNRS & Université Nice Sophia Antipolis, 2000, route des Lucioles – 06903 Sophia Antipolis cedex, France
[b] Dipartimento di Matematica e Informatica, Università di Palermo, Via Archirafi 34 – 90123 Palermo, Italy

ARTICLE INFO

Article history:
Received 28 February 2013
Received in revised form 22 May 2014
Accepted 11 November 2014
Available online 18 November 2014
Communicated by G. Ausiello

Keywords:
Kolakoski word
$\mathbf{C}^\infty$-words
Directed acyclic graph
Recursive function
Directed set

ABSTRACT

We present a new framework for dealing with $\mathbf{C}^\infty$-words, based on their left and right frontiers. This allows us to give a compact representation of them, and to describe the set of $\mathbf{C}^\infty$-words through an infinite directed acyclic graph G. This graph is defined by a map acting on the frontiers of $\mathbf{C}^\infty$-words. We show that this map can be defined recursively and with no explicit reference to $\mathbf{C}^\infty$-words. We then show that some important conjectures on $\mathbf{C}^\infty$-words follow from analogous statements on the structure of the graph G.

1. Introduction

Every finite or infinite word w over a finite alphabet Σ of cardinality greater than 1 can be written in a unique way by replacing maximal blocks of consecutive identical letters, called runs, with the single letter having the length of the block as exponent, as in $a^2b^1c^3 = aabccc$. The sequence of exponents of w is called the *exponent word* of w, and is denoted by $\Delta(w)$.

Fixing an integer alphabet Σ (i.e., a finite subset of $\mathbb{N} \setminus \{0\}$), the words over Σ such that their exponent word is still a word over Σ are called *differentiable words*.

Studying the exponent words in the context of symbolic dynamics, Oldenburger [14] showed in 1938 that there exist infinite words that coincide with their exponent word, and that these must be non-periodic. For example, if $\Sigma = \{1, 2\}$ there exist precisely two such words, namely the word

$$\mathcal{K} = 2211212212211211221211121\cdots$$

and the word $1 \cdot \mathcal{K}$. The word $\mathcal{K}$ is known as the *Kolakoski word* [12], although perhaps it should be more properly called the *Oldenburger–Kolakoski word*.

Several longstanding conjectures on the combinatorial structure of the Kolakoski word remain unproved—some of them are in the original 1938 paper by Oldenburger. For example, in his paper Oldenburger asks whether or not there exist recurrent[1] words that coincide with their exponent word. This question has been answered in the affirmative for words

☆ Some of the results in this paper were presented at the *Conference for the 20th Anniversary of LaCIM*, 29–31 August 2010, Montreal, Canada [8]. The second author has been partially funded by ANR EMC (ANR-09-BLAN-0164).
* Corresponding author.
E-mail addresses: fedou@i3s.unice.fr (J.-M. Fédou), fici@math.unipa.it (G. Fici).
[1] Recall that a word is said to be recurrent if all of its factors appear infinitely often.

http://dx.doi.org/10.1016/j.tcs.2014.11.011
0304-3975/

D^0	2121122	21221211221221	2122121122
D^1	1122	12112212	121122
D^2	22	1221	122
D^3	2	2	2

Fig. 1. The $\mathbf{C}^\infty$-words $u = 2121122$, $u' = 21221211221221$ and $u'' = 2122121122$, represented together with their non-empty derivatives.

over binary alphabets in which the two letters have the same parity [3], but it is still open for words over the alphabet $\{1, 2\}$ (see [11]), and in particular thus for the Kolakoski word. It is easy to see that a sufficient condition for the Kolakoski word being recurrent is that its set of factors is closed under complement (swapping of 1's and 2's) (see [7]). In fact, Brlek and Ladouceur [4] proved that it is even sufficient to prove that $\mathcal{K}$ contains arbitrarily long palindromic factors. However, all these properties are still unproven.

This motivates us to study the set of factors of the Kolakoski word.

1.1. The set of $\mathbf{C}^\infty$-words

In order to study the finite factors of the Kolakoski word, the operator Δ is not convenient, since it does not preserve the set of factors. For example, 121 is a factor of $\mathcal{K}$ but $\Delta(121) = 111$ is not. For this reason, we use the operator D, called *the derivative* in [7] but, as an anonymous referee pointed out, yet introduced in [14] under the name of *proper exponent block*, that consists in discarding the first and/or the last letter in $\Delta(w)$ if these are equal to 1. For example, the derivative of 121 is $D(121) = 1$, while the derivative of 12 is ε, the empty word. The set of finite words over $\Sigma = \{1, 2\}$ that are derivable arbitrarily many times over $\Sigma = \{1, 2\}$ is called *the set of* $\mathbf{C}^\infty$*-words*. It is closed under complement and reversal, and contains the set of factors of the Kolakoski word. Thus, one of the most important open problems about the Kolakoski word is to decide whether all the $\mathbf{C}^\infty$-words occur as factors in $\mathcal{K}$:

Conjecture 1. *(See [7].) Every $\mathbf{C}^\infty$-word is a factor of the Kolakoski word.*

Actually, the set of $\mathbf{C}^\infty$-words contains the set of factors of any right-infinite word over $\{1, 2\}$ having the property that an arbitrary number of applications of Δ still produces a word over $\{1, 2\}$. Such words are called *smooth words* [1,4]. Nevertheless, the existence of a smooth word such that the set of its factors is equal to the whole set $\mathbf{C}^\infty$ is still an open question. By the way, we notice that should Conjecture 1 be true, the Kolakoski word would be recurrent (see [7]).

In addition to the aforementioned problems, there is a conjecture of Keane [10] stating that the frequencies of 1s and 2s in the Kolakoski word exist and are equal to $1/2$. Chvátal [6] showed that if these limits exist, they are very close to $1/2$ (actually, between 0.499162 and 0.500838).

Up to now, only few combinatorial properties of $\mathbf{C}^\infty$-words have been established. Weakley [15] started a classification of $\mathbf{C}^\infty$-words and obtained significant results on their complexity function. Carpi [5] proved that the set $\mathbf{C}^\infty$ contains only a finite number of squares, and does not contain cubes (see also [13] and [2]). This result generalizes to repetitions with gap, i.e., to the $\mathbf{C}^\infty$-words of the form uzu, for a non-empty z. Indeed, Carpi [5] proved that for every $k > 0$, only finitely many $\mathbf{C}^\infty$-words of the form uzu exist with z not longer than k. In a previous paper [9], we proved that for any $\mathbf{C}^\infty$-word u, there exists a z such that uzu is a $\mathbf{C}^\infty$-word, and $|uzu| \le C|u|^{2.72}$, for a suitable constant C. In the same paper, we proposed the following conjecture:

Conjecture 2. *(See [9].) For any u, v $\mathbf{C}^\infty$-words, there exists z such that uzv is a $\mathbf{C}^\infty$-word.*

Despite Conjecture 2 being a weaker condition than Conjecture 1, it remains an open question.

1.2. Outline of the results

We find convenient to represent $\mathbf{C}^\infty$-words together with all their non-empty derivatives, as shown in Fig. 1 (we adopt the convention that $w = D^0(w)$ for any word w).

Notice that every $\mathbf{C}^\infty$-word w can be reduced to the empty word with a finite number k of applications of the derivative, and we call the least of such k the *height* of w. For example, the words in Fig. 1 all have height 3.

The sequence of the first letters of the non-empty derivatives of a $\mathbf{C}^\infty$-word w can be encoded into a word $\Psi(w)$ over the alphabet $\Sigma_0 = \{0, 1, 2\}$, that we call the *left frontier* of w. For every $0 \le i < k$, the $(i+1)$th letter of $\Psi(w)$ is 0 if $D^{i-1}(w)$ begins in 122 or 211, or the first letter of the ith derivative $D^i(w)$ otherwise.

Analogously, one can define the *right frontier* of w, $\Psi^R(w)$, as the left frontier of the reversal of w. The pair $[\Psi(w), \Psi^R(w)]$ is called the *vertical representation* of the word w, and allows one to uniquely represent any $\mathbf{C}^\infty$-word by means of a pair of words whose length is logarithmic in the length of the $\mathbf{C}^\infty$-word (Theorem 16).

For example, the vertical representations of the words u, u' and u'' in Fig. 1 are, respectively, $[2122, 2222]$, $[2110, 1010]$ and $[2110, 2222]$.