Theoretical Computer Science 588 (2015) 73–80

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Characterization of some binary words with few squares

Golnaz Badkobeh [a], Pascal Ochem [b]

[a] Department of Computer Science, University of Sheffield, UK
[b] CNRS, LIRMM, Montpellier, France

ARTICLE INFO

Article history:
Received 18 April 2012
Received in revised form 25 March 2015
Accepted 30 March 2015
Available online 13 April 2015
Communicated by G. Ausiello

Keywords:
Combinatorial problems
Repetitions
Avoidability

ABSTRACT

Thue proved that the factors occurring infinitely many times in square-free words over $\{0, 1, 2\}$ avoiding the factors in $\{010, 212\}$ are the factors of the fixed point of the morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$. He similarly characterized square-free words avoiding $\{010, 020\}$ and $\{121, 212\}$ as the factors of two morphic words. In this paper, we exhibit smaller morphisms to define these two square-free morphic words and we give such characterizations for six types of binary words containing few distinct squares.

1. Introduction

Let Σ_k denote the k-letter alphabet $\{0, 1, \ldots, k-1\}$. Let ε denote the empty word. A finite word is *recurrent* in an infinite word w if it appears as a factor of w infinitely many times. An infinite word w is *recurrent* if all its finite factors are recurrent in w. If a morphism f is such that $f(0)$ starts with 0, then the *fixed point* of f is the unique word $w = f^\infty(0)$ starting with 0 and satisfying $w = f(w)$. An infinite word is *pure morphic* if it is the fixed point of a morphism. An infinite word is *morphic* if it is the image $g(f^\infty(0))$ by a morphism g of a pure morphic word $f^\infty(0)$. The *factor complexity* of an infinite word or a language is the number of factors of length n of the infinite word or the language. A pattern P is a finite word of variables over the alphabet $\{A, B, \ldots\}$. A word w (finite or infinite) *avoids* a pattern P if for every substitution ϕ of the variables of P with non-empty words, $\phi(P)$ is not a factor of w. Given a finite alphabet Σ_k, a finite set $\mathcal{P}$ of patterns, and a finite set $\mathcal{F}$ of factors over Σ_k, we say that $\mathcal{P} \cup \mathcal{F}$ *characterizes* a morphic word w over Σ_k if w avoids $\mathcal{P} \cup \mathcal{F}$ and every recurrent factor of an infinite word avoiding $\mathcal{P} \cup \mathcal{F}$ is a factor of w. In other words, $\mathcal{P} \cup \mathcal{F}$ characterizes w if and only if every recurrent word over Σ_k avoiding $\mathcal{P} \cup \mathcal{F}$ has the same set of factors as w. In our results, we do not specify the alphabet size k since Σ_k corresponds to the set of letters appearing in $\mathcal{F}$. A *repetition* is a factor of the form $r = u^n v$ where u is non-empty and v is a prefix of u. Then $|u|$ is the *period* of the repetition r and its *exponent* is $|r|/|u|$. A *square* is a repetition of exponent 2. Equivalently, it is an occurrence of the pattern AA. An overlap is a repetition with exponent strictly greater than 2.

Thue [3,10,11] gave the following characterization of overlap-free binary words: $\{ABABA\} \cup \{000, 111\}$ characterizes the fixed point of the morphism $0 \mapsto 01$, $1 \mapsto 10$. Concerning ternary square-free words, he proved that

- $\{AA\} \cup \{010, 212\}$ characterizes the fixed point of f_3: $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$,
- $\{AA\} \cup \{010, 020\}$ characterizes the morphic word $T_1(f_T^\infty(0))$,
- $\{AA\} \cup \{121, 212\}$ characterizes the morphic word $T_2(f_T^\infty(0))$,

E-mail address: golnaz.badkobeh@gmail.com (G. Badkobeh).

http://dx.doi.org/10.1016/j.tcs.2015.03.044

Original form	Standardized form	Morphic word	Section
$\{AA\} \cup \{010, 020\}$	$\{AA\} \cup \{010, 020\}$	$M_1(f_5^\infty(0))$	3.1
$\{AA\} \cup \{121, 212\}$	$\{AA\} \cup \{121, 212\}$	$M_2(f_5^\infty(0))$	3.1
$(5/2, 2, 8)$	$\{SQ_7\} \cup F_8$	$g_8(f_5^\infty(0))$	3.2
$(7/3, 2, 12)$	$\{SQ_9\} \cup F_{12}$	$g_{12}(f_5^\infty(0))$	3.3
$(7/3, 1, 14)$	$\{SQ_9\} \cup F_{14}$	$g_{14}(f_5^\infty(0))$	3.4
$\{AABBCC, SQ_3\} \cup F'_{cs}$	$\{SQ_3\} \cup F_{cs}$	$g_{cs}(f_5^\infty(0))$	3.5
$(5/2, 1, 11)$	$\{SQ_5\} \cup F_{11}$	$g_{11}(f_3^\infty(0))$	4.1
$(3, 2, 3) \cup F'_3$	$\{SQ_3\} \cup F_3$	$g_3(f_3^\infty(0))$	4.2
$\{AABBCABBA\} \cup \{0011, 1100\}$	$\{SQ_5\} \cup F_q$	$g_q(f_3^\infty(0))$	4.3

Fig. 1. Table of results.

where the morphisms f_T, T_1, and T_2 are given below.

$$\begin{array}{lll}
f_T(0) = 012, & T_1(0) = 01210212, & T_2(0) = 021012, \\
f_T(1) = 0432, & T_1(1) = 01210120212, & T_2(1) = 02102012, \\
f_T(2) = 0134, & T_1(2) = 01210212021, & T_2(2) = 02101201, \\
f_T(3) = 013432, & T_1(3) = 0121021202101 20212, & T_2(3) = 0210120102012, \\
f_T(4) = 0434. & T_1(4) = 0121012021. & T_2(4) = 0210201.
\end{array}$$

To obtain the last two results, Thue first proved that $f_T^\infty(0)$ is characterized by $\{AA\} \cup \{02, 03, 10, 14, 21, 23, 24, 30, 31, 41, 42, 040, 132, 404, 1201, 2012\}$.

In this paper, we prove such characterizations mostly for the binary words considered by the first author [1]. We also obtain smaller morphisms for Thue's words avoiding $\{AA\} \cup \{010, 020\}$ and $\{AA\} \cup \{121, 212\}$ as well as a characterization for words avoiding the patterns $AABBCC$ (i.e., three consecutive squares), $ABCABC$ and a finite set of factors. The results are summarized in Fig. 1. The first column shows the description of the considered language given in the literature. It is either given by forbidden sets of patterns and factors, or by the notation (e, n, m), which means that we consider the binary words avoiding repetitions with exponent strictly greater than e, containing exactly n distinct repetitions with exponent e as a factor, and containing the minimum number m of distinct squares. We use the notation SQ_t for the pattern corresponding to squares with period at least t, that is, $SQ_1 = AA$, $SQ_2 = ABAB$, $SQ_3 = ABCABC$, and so on. These languages actually have an equivalent definition with one forbidden pattern SQ_t and a finite set of forbidden factors. This standardized definition, given in the second column, is more suited for proving the characterization. The third column gives the corresponding morphic word. The fourth column indicates the section containing the corresponding set F_{xx} and morphism g_{xx}.

To define a morphic word $g(f^\infty(0))$, we allow that g is an *erasing* morphism, i.e., that the g-image of a letter is empty. Notice that replacing g by $h_c = g \circ f^c$ defines the same morphic word, and that h_c is non-erasing for some small constant c.

The proofs are obtained by computer using the technique described in the next section. An example of proof by hand is given for Theorem 3. The morphic words in Fig. 1 are gathered according the pure morphic word they are built on. We introduce in Section 3 a pure morphic word $f_5^\infty(0)$ similar to Thue's word $f_T^\infty(0)$ and we characterize some of its morphic images. Section 4 is devoted to characterizations of some morphic images of Thue's ternary pure morphic word $f_3^\infty(0)$.

2. Characterizing a morphic word

A morphism $f : \Sigma_k^* \to \Sigma_k^*$ is *primitive* if there exists $n \in \mathbb{N}$ such that $f^n(a)$ contains b for every $a, b \in \Sigma_k$. We are given a primitive morphism $f : \Sigma_k^* \to \Sigma_k^*$, a morphism $g : \Sigma_k^* \to \Sigma_{k'}^*$, and a finite set of factors $\mathcal{F}_m \subset \Sigma_{k'}^*$. We want to prove that $g(f^\infty(0))$ is characterized by $\{SQ_t\} \cup \mathcal{F}_m$.

We assume that $g(f^\infty(0))$ avoids $\{SQ_t\} \cup \mathcal{F}_m$. This can be checked using Cassaigne's algorithm [5] that determines if a morphic word defined by circular morphisms avoids a given pattern with constants. We refer to Cassaigne [5] for the definitions of circular morphisms, synchronization point, and synchronization delay. We can use an online implementation [4] of this algorithm. We also assume that the pure morphic word $f^\infty(0)$ is characterized by $\{AA\} \cup \mathcal{F}_p$ for some finite set of factors $\mathcal{F}_p \subset \Sigma_k^*$.

We compute the smallest integer c such that $\min\left\{|g(f^c(a))|,\ a \in \Sigma_k\right\} \geqslant t$. This c exists because f is primitive. We can consider the morphism $g' = g \circ f^c$ instead of g since we have $g'(f^\infty(0)) = g(f^\infty(0))$.

First, we check that g' is circular. Then, we compute the set S_l of words v such that there exists a word $pvs \in \Sigma_{k'}^*$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$, where $l = \max\left\{|u|, u \in \mathcal{F}_p\right\} \times \max\left\{|g'(a)|, a \in \Sigma_k\right\}$, $|v| = l$, and $|p| = |s| = 4l$. To do this, we simply perform a depth-first exploration of the words of length $9l$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ and for each of them, we put the central factor of length l in S_l. The running time of this brute-force approach is not so prohibitive precisely because the characterization implies a polynomial factor complexity. Finally, we check that every word in S_l is a factor of $g'(f^\infty(0))$.

This implies that an infinite word over $\Sigma_{k'}$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ is the g'-image of an infinite word $w \in \Sigma_k^*$. Now w is square-free, since otherwise $g'(w)$ would contain a square of period at least t. Also w does not contain a word $y \in \mathcal{F}_p$, because $g'(y)$ is a word of length at most l that is not a factor of any word in S_l. So w avoids $\{AA\} \cup \mathcal{F}_p$, and thus has the same set of factors as $f^\infty(0)$. Thus, every infinite recurrent word over $\Sigma_{k'}$ avoiding $\{SQ_t\} \cup \mathcal{F}_m$ has the same set of factors as $g'(f^\infty(0))$.

The programs we used are available at http://www.lirmm.fr/~ochem/morphisms/characterization.htm.