Special factors and images of Arnoux-Rauzy words

Michelangelo Bucci Alessandro De Luca

Dipartimento di Matematica e Applicazioni "R. Caccioppoli"

Università degli Studi di Napoli Federico II

via Cintia, Monte S. Angelo, 80126 Napoli, Italy

{micbucci,alessandro.deluca}@unina.it

Abstract

In this paper we prove the following result. Let s be an infinite word on a finite alphabet, and $N \geq 0$ be an integer. Suppose that all left special factors of s longer than N are prefixes of s, and that s has at most one right special factor of each length greater than N. Then s is a morphic image, under an injective morphism, of a suitable standard Arnoux-Rauzy word.

1 Introduction

Factor complexity is a common theme in the combinatorial analysis of finite and infinite words. Being the function counting distinct blocks (factors) of each length, it is one of the most natural measures of complexity of a word. A famous theorem by Morse and Hedlund [9] characterizes ultimately periodic sequences as the ones having bounded complexity.

Sturmian words have the lowest possible unbounded complexity ($n + 1$ factors of each length n). They make up one of the most studied family of infinite words, not just because of their theoretical interest (see [3] for a general introduction, or [2] for a recent survey). From the definition, it follows that Sturmian words are on a binary alphabet, and have exactly one *left special* factor of each length n (a factor is left special if it is a suffix of at least two distinct factors of length $n + 1$).

As is well known, a first natural generalization of Sturmian words for alphabets with an arbitrary number of letters was introduced by Arnoux and Rauzy [1]. A recurrent word $s \in A^\omega$ is *Arnoux-Rauzy* (or *strict episturmian*, see below) if it has exactly one left special factor and one right special factor per length, that can be immediately preceded (resp. followed) in s by all letters of the alphabet.

A remarkable property of Sturmian words, shared by Arnoux-Rauzy words, is their *closure under reversal*: if $w = a_1a_2 \cdots a_n$ is a factor of an Arnoux-Rauzy word s with $a_i \in A$ for $i = 1, \ldots, n$, then $\tilde{w} = a_na_{n-1} \cdots a_1$ is a factor of s too.

This led Droubay, Justin, and Pirillo [6] to a generalization: an infinite word is *episturmian* if it has at most one left special factor per length, and is closed under reversal. The further generalization of ϑ-*episturmian* words was recently introduced in [4], by substituting the reversal operator with any *involutory antimorphism* ϑ of A^*. Generalizing even more, by requiring the condition on special factors only for sufficient lengths, ϑ-*words with seed* are obtained (see [4]).

All such words have a *standard* counterpart, where the unique left special factors correspond to prefixes of the infinite word. For instance, a ϑ-*standard word with seed* is any infinite word s which is closed under ϑ and such that any sufficiently long left special factor of s is a prefix of it. For all the above classes, standard words are good representatives, in the sense that an infinite word s belongs to one of such classes if and only s has the same set of factors of a standard word of that class (see [6, 4]).

Our main result shows that, in the standard case, even when the further step of dropping the "closure under some ϑ" requirement is made, the large class of words thus obtained retains a strong link with Arnoux-Rauzy words. More precisely, we will prove the following.

Theorem 1. *Let $s \in A^\omega$ satisfy the following two conditions for all $k \geq N$, where $N \geq 0$:*

1. *any left special factor of s having length k is a prefix of s,*
2. *s has at most one right special factor of length k.*

Then there exists $B \subseteq \operatorname{alph} s$ and a standard Arnoux-Rauzy word $t \in B^\omega$ such that s is a morphic image (under an injective morphism) of t.

2 Basic definitions and results

In the following, A will denote a finite alphabet, A^* the free monoid of words over A, and A^ω the set of infinite words over A.

Let s be a finite or infinite word. A *factor* of s is any finite word v such that $s = uvw$ for suitable words u, w. We denote by s^f the first letter of s; if $s \in A^*$ we denote by s^ℓ its last letter. With $\operatorname{Fact} s$, $\operatorname{Pref} s$, and $\operatorname{Suff} s$ we denote respectively the set of factors, prefixes, and suffixes of s. The *factor complexity* of s is the function $c_s : \mathbb{N} \to \mathbb{N}$ defined by $c_s(n) = \operatorname{card}(A^n \cap \operatorname{Fact} s)$ for all $n \geq 0$.

Let $v \in \operatorname{Fact} s$. A factor w of s is called a *right* (resp. *left*) *extension* of v in s if v is a proper prefix (resp. suffix) of w. If $|v| = n$, the *right* (resp. *left*) *degree* of v in s is the number of its distinct right (resp. left) extensions of length $n + 1$. Finally, v is called a *right* (resp. *left*) *special* factor of s if its right (resp. left) degree is at least 2, i.e., if there exist two distinct letters a and b such that va and vb (resp. av and bv) are factors of s. If a factor of s is both left and right special, then it is called *bispecial.*

A *complete return* to v in s is any factor of s containing exactly two occurrences of v, one as a prefix and the other as a suffix. If $z = wv$ is a complete return to v, then w is called a *return word* to v (cf. [7]).

An infinite word s is *recurrent* if each of its factors has infinitely many occurrences in s; it is *uniformly recurrent* if the gaps between consecutive occurrences of any factor are bounded. Equivalently, s is uniformly recurrent if for all factors v there are finitely many distinct return words to v in s.

Given any prefix p of an infinite word s, there exists a unique factorization of s by means of the return words to p in s. By mapping each return word to a different letter of a suitable alphabet, and then applying such a map to s thus factorized, we obtain a *derivated word* of s with respect to p (cf. [7]). Clearly, s is a morphic image of its derivated words.

For other usual definitions of combinatorics on words not explicitly listed here, we refer to [8].

The following simple lemma is the first basic ingredient for our main result.

Lemma 2. *Let s be an infinite word such that any sufficiently long left special factor of s is a prefix of it. Then s is recurrent.*

Proof. By contradiction, suppose that λw is a prefix of s ending with the rightmost occurrence of w in s. Then all prefixes of s from length $|\lambda w|$ on do not reoccur in s, and so have *no left extensions* in s. By a counting argument, this implies that s has also at least one factor with *more than one* left extension (i.e., a left special factor) for each length $n \geq |\lambda w|$. For sufficiently large n, such a left special factor should be a prefix of s by hypothesis. We have reached a contradiction. □

We need one of the most well-known and useful restatements of the theorem of Morse and Hedlund (cf. [9, Theorem 7.3]):

Theorem 3. *An infinite word s is ultimately periodic if and only if $c_s(n) = c_s(n+1)$ for some $n \geq 0$.*

As a consequence of Lemma 2, we obtain the following specialization.

Proposition 4. *An infinite word s is (purely) periodic if and only if it has no left special factor of some length n.*

Proof. If $s = p^\omega$ with $p \in A^*$, then s has no left special factors of length $|p|$. Conversely, assume that s has no left special factor of length n. This implies

$$\operatorname{card}(A^n \cap \operatorname{Fact} s) = \operatorname{card}(A^{n+1} \cap \operatorname{Fact} s) ,$$

so that by Theorem 3, s is ultimately periodic. Clearly s has no left special factor of any length $k \geq n$, thus it trivially satisfies the hypothesis of Lemma 2. Therefore s is recurrent, and hence periodic. □

The following proposition was proved in [5, Lemma 7] under different hypotheses. We report an adapted proof for the sake of completeness.

Proposition 5. *Let s be a recurrent aperiodic infinite word. Then every factor w of s is contained in some bispecial factor of s.*

Proof. Since s is recurrent, we can consider a complete return z to w in s. Writing $z = vw$, it cannot happen that the factor w is always preceded by v in s, otherwise s would be periodic. Thus some suffix of z of length at least $|u|$ must be a left special factor of s. Let $x \in A^*$ be of minimal length such that xw is a left special factor of s. Such a word is trivially unique, and w is always preceded in s by x. In a similar way, there exists a unique $y \in A^*$ of minimal length such that wy is right special in s, and w is always followed by y.

Since xw is left special in s and xw is always followed by y one has that xwy is also left special. Similarly, since wy is right special and always preceded by x, xwy is right special. Hence every factor w of s is contained in some bispecial factor $W = xwy$ of s. □

A recurrent word $s \in A^\omega$ with $A = \operatorname{alph} s$ is an *Arnoux-Rauzy* word if it has exactly one left special factor and one right special factor of each length, of degree $\operatorname{card} A$. Arnoux-Rauzy words are uniformly recurrent (cf. [6]); this was part of the definition in [1]. An Arnoux-Rauzy word s is *standard* if its left special factors are prefixes of s. Thanks to Lemma 2, one does not have to consider recurrence, when checking if a given word is a standard Arnoux-Rauzy word.

3 Proof of Theorem 1

Suppose first that s has no left special factor of some length n. Then s is periodic by Proposition 4, so that it is trivially a morphic image of x^ω for any $x \in \operatorname{alph} s$.

Now let us assume that s has at least one left special factor of each length — exactly one, from length N on. By Lemma 2, s is recurrent, so that by Proposition 5 it has infinitely many bispecial factors, which we denote by $W_0 = \varepsilon, W_1, \ldots, W_n, \ldots$, where $|W_i| \leq |W_{i+1}|$ for all $i \geq 0$. Let j be the least index such that $|W_j| \geq N$. By conditions 1 and 2, W_i is a border of W_{i+1} for all $i \geq j$. The sequence whose n-th term is the (right) degree of W_n for all $n \geq j$ is then non-increasing. Hence there exists $k \geq j$ such that W_n has the same degree of W_k for all $n \geq k$, that is, the above considered sequence is constant from its k-th term on. We set

$$B = \{x \in A \mid W_k x \in \operatorname{Fact} s\} \subseteq \operatorname{alph} s \ ,$$

so that $\operatorname{card} B$ is, by definition, the degree of W_k.

We now consider the return words to $w = W_k$ in s. Let $u_1 w = w v_1$ and $u_2 w = w v_2$ be any two distinct complete returns to w in s, and let us show that $v_1^f \neq v_2^f$. Indeed, let p be the longest common prefix of v_1 and v_2. If $p = v_1$, then $|v_2| > |v_1|$ as

$v_1 \neq v_2$; since $wv_1 = u_1 w$, there is an internal occurrence of w in wv_2, contradicting the definition of complete return. The same argument applies if $p = v_2$. Thus p is a proper prefix of both v_1 and v_2, so that wp is a right special factor of s. Since s has only one right special factor per length, and w is a right special factor of s, it follows that w is a suffix of wp. This implies $p = \varepsilon$, since otherwise there would be an internal occurrence of w in wv_1 and wv_2. Hence $v_1^f \neq v_2^f$ as desired. Since w is also left special in s, using a symmetric argument one can prove that $u_1^\ell \neq u_2^\ell$.

From this it follows that for each $x \in B$, there exists a unique complete return $u_x w = wv_x$ to w in s, such that $v_x^f = x$. We define a morphism $\varphi : B^* \to A^*$ by $\varphi(x) = u_x$. Note that φ is injective, as $\varphi(B)$ is a suffix code having the same cardinality of B.

By definition, we have $s = \varphi(t)$, where $t \in B^\omega$ is a derivated word of s with respect to its prefix w. We remark that, as a consequence of the definition of return words, one has

$$z \in \operatorname{Fact} t \Leftrightarrow \varphi(z)w \in \operatorname{Fact} s \quad \text{and} \quad z \in \operatorname{Pref} t \Leftrightarrow \varphi(z)w \in \operatorname{Pref} s\ . \tag{1}$$

We will prove that t is a standard Arnoux-Rauzy word by showing that t has exactly one right special factor (of degree $\operatorname{card} B$) of each length, and that all left special factors of t are prefixes of it.

Let z_1 and z_2 be any two right special factors of t having the same length. Thus there exist distinct letters $x_1, y_1, x_2, y_2 \in B$ such that $x_i \neq y_i$ and $z_i x_i, z_i y_i \in \operatorname{Fact} t$ for $i = 1, 2$. By (1), this implies $\varphi(z_i x_i)w, \varphi(z_i y_i)w \in \operatorname{Fact} s$. Since for $\alpha \in \{x_i, y_i\}$ and $i = 1, 2$ we have

$$\varphi(z_i \alpha)w = \varphi(z_i)u_\alpha w = \varphi(z_i)wv_\alpha \in \operatorname{Fact} s$$

with $v_{x_i}^f \neq v_{y_i}^f$, it follows that $\varphi(z_1)w$ and $\varphi(z_2)w$ are right special factors of s. By condition 2, either $\varphi(z_1)w \in \operatorname{Suff}(\varphi(z_2)w)$, or vice versa. The word w has $|z_1| + 1 = |z_2| + 1$ occurrences in both $\varphi(z_1)w$ and $\varphi(z_2)w$, and it is a prefix of both, by the definition of return word. Hence we derive $\varphi(z_1)w = \varphi(z_2)w$, so that $z_1 = z_2$ by the injectivity of φ.

If z is a right special factor of t, by the above argument $\varphi(z)w$ is right special in s. Since $|\varphi(z)w| \geq |w|$, we obtain that $\varphi(z)wx \in \operatorname{Fact} s$ for all $x \in B$. Since the only complete return to w in s starting with x is v_x, it follows $\varphi(z)wv_x = \varphi(z)u_x w = \varphi(zx)w \in \operatorname{Fact} s$, so that $zx \in \operatorname{Fact} t$ for all $x \in B$, proving that z has degree $\operatorname{card} B$.

Let now z' be a left special factor of t, and let $xz', yz' \in \operatorname{Fact} t$ for some distinct letters $x, y \in B$. Then $\varphi(xz')w, \varphi(yz')w \in \operatorname{Fact} s$. As $\varphi(x)^\ell = u_x^\ell \neq u_y^\ell = \varphi(y)^\ell$, $\varphi(z')w$ is a left special factor of s. By condition 1, it follows $\varphi(z')w \in \operatorname{Pref} s$ and then $z' \in \operatorname{Pref} t$ by (1).

4 Conclusions

In this paper we have shown that what seems to be a natural (though very wide) generalization of the standard episturmian words, retains a strong connection with Arnoux-Rauzy words. Several promising problems arise from this result; an open question, for example, is to examine whether the converse of Therorem 1 holds, i.e., whether any morphic image of a standard Arnoux-Rauzy word has, from a certain length on, at most one special factor, the left special ones also being prefixes. Though this could also turn out to be false in general, it is possible that a particular class of morphisms, for which this happens, exists.

It would also be interesting to study in detail what happens to Theorem 1 when the hypothesis that left special factors are also prefixes is dropped, that is (in some way) the nonstandard case. By modifying the proof of Theorem 1 suitably, it is not difficult to show the following:

Theorem 6. *If $s \in A^\omega$ is recurrent and has at most one left special factor and one right special factor for all lengths $k \geq N$, then there exist $B \subseteq A$, an injective morphism $\varphi : B^* \to A^*$, and an Arnoux-Rauzy word $t \in B^\omega$ such that $s \in \mathrm{Suff}\, \varphi(t)$.*

This is somehow weaker than the original Theorem 1, as we only get that s is *a suffix of* a morphic image of an Arnoux-Rauzy word. Therefore, any improvement of Theorem 6 would be welcome.

References

[1] P. Arnoux and G. Rauzy. Représentation géométrique de suites de complexité $2n + 1$. *Bulletin de la Société Mathématique de France* **119** (1991) 199–215.

[2] J. Berstel. Sturmian and episturmian words (a survey of some recent result results). In S. Bozapalidis and G. Rahonis (eds.), *Conference on Algebraic Informatics, Thessaloniki*, vol. 4728 of *Lecture Notes in Computer Science* (2007) pp. 23–47.

[3] J. Berstel and P. Séébold. Sturmian Words. In M. Lothaire, *Algebraic Combinatorics on Words*, chap. 2. Cambridge University Press, Cambridge UK (2002).

[4] M. Bucci, A. de Luca, A. De Luca, and L. Q. Zamboni. On different generalizations of episturmian words. *Theoretical Computer Science* **393** (2008) 23–36.

[5] M. Bucci, A. de Luca, A. De Luca, and L. Q. Zamboni. On θ-episturmian words. *European Journal of Combinatorics* (2008, to appear), doi:10.1016/j.ejc.2008.04.010.

[6] X. Droubay, J. Justin, and G. Pirillo. Episturmian words and some constructions of de Luca and Rauzy. *Theoretical Computer Science* **255** (2001) 539–553.

[7] F. Durand. A characterization of substitutive sequences using return words. *Discrete Mathematics* **179** (1998) 89–101.

[8] M. Lothaire. *Combinatorics on Words*. Addison-Wesley, Reading MA (1983).

[9] M. Morse and G. A. Hedlund. Symbolic Dynamics. *American Journal of Mathematics* **60** (1938) 815–866.