On a Family of Morphic Images of Arnoux-Rauzy Words

Michelangelo Bucci and 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 [1] 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 [2] for a general introduction, or [3] 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 [4]. An infinite word s is *Arnoux-Rauzy* (or *strict episturmian*, see below) if it is recurrent (i.e., all factors of s occur infinitely often) and it has exactly one left special factor and one right special factor per length, that appear in s immediately preceded (resp. followed) by all letters occurring in s. More detailed definitions will be given in Sect. 2.

A remarkable property of Sturmian words, shared by Arnoux-Rauzy words, is their *closure under reversal*: if $w = a_1 a_2 \cdots a_n$ is a factor of an Arnoux-Rauzy word s with $a_i \in A$ for $i = 1, \ldots, n$, then $\tilde{w} = a_n a_{n-1} \cdots a_1$ is a factor of s too. This led Droubay, Justin, and Pirillo [5] to a generalization: an infinite word is *episturmian* if it has at most one left special factor per length, and is closed under reversal. Episturmian words are recurrent, but have no restriction on the number of letters immediately preceding left special factors. Thus the family of episturmian words strictly contains the one of Arnoux-Rauzy words.

The class of *ϑ-episturmian* words is a further generalization, recently introduced in [6] 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 [6]).

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 if s has the same set of factors as some standard word of that class (see [5, 6]).

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 $n \geq N$, where $N \geq 0$:*

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

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

In the next section we shall give all the formal definitions and preliminary results needed for our proof, which will be given in Sect. 3. For more basics about combinatorics on words, we refer to [7]. For more details on episturmian words and their generalizations, see [3, 5, 8, 9, 6, 10, 11].

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. The identity element of A^* is the *empty word* ε.

Let s be a finite or infinite word. The set of letters occurring in s is denoted by $\mathrm{alph}(s)$. A *factor* of s is any finite word w such that $s = uwv$ for suitable words u, v; if u (resp. v) is the empty word we call w a *prefix* (resp. *suffix*) of s. A *border* of $s \in A^*$ is a word which is both a prefix and a suffix of s. If s is nonempty, we denote by s^f its first letter, and if s is also finite we denote by s^ℓ its last letter[1]. With $\mathrm{Fact}(s)$, $\mathrm{Pref}(s)$, and $\mathrm{Suff}(s)$ we denote respectively the set

[1] This notation should not be confused with powers of a word; no such power occurs in this paper.

of factors, prefixes, and suffixes of s. The *factor complexity* of s is the function $c_s : \mathbb{N} \to \mathbb{N}$ defined by $c_s(n) = \#(A^n \cap \mathrm{Fact}(s))$ for all $n \geq 0$. We remark that the complexity of an infinite word s is a nondecreasing function.

Let $w \in \mathrm{Fact}(s)$. A factor v of s is called a *right* (resp. *left*) *extension* of w in s if w is a proper prefix (resp. suffix) of v. If $|w| = n$, the *right* (resp. *left*) *degree* of w in s is the number of its distinct right (resp. left) extensions of length $n+1$. For all $n \geq 0$, any factor of s of length $n+1$ is uniquely determined by its first letter and by its suffix of length n, or by its last letter and by its prefix of length n. An immediate consequence is the following well-known identity:

$$\sum_{w \in \mathrm{F}_s(n)} \deg^-(w) = c_s(n+1) = \sum_{w \in \mathrm{F}_s(n)} \deg^+(w) \ , \tag{1}$$

in which the operators $\deg^-$ and $\deg^+$ denote the left and right degree respectively, and $\mathrm{F}_s(n) = A^n \cap \mathrm{Fact}(s)$.

We recall that w 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 wa and wb (resp. aw and bw) are factors of s. If a factor of s is both left and right special, then it is called *bispecial*.

A *complete return* to w in s is any factor of s containing exactly two occurrences of w, one as a prefix and the other as a suffix. If $z = vw$ is a complete return to w, then v is called a *return word* to w (cf. [12]).

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 w there are finitely many distinct return words to w 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. [12]). Clearly, s is a morphic image of its derivated words.

The following simple lemma is the first basic ingredient needed 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 there exists a factor w of s having only finitely many occurrences in s, and let λw be the prefix of s ending with the rightmost occurrence of w in s. Then all prefixes of s from length $|\lambda w|$ onward do not reoccur in s, and so have left degree 0.

We claim that this implies that s has also at least one left special factor for each length $n \geq |\lambda w|$. Indeed, for all such n the left sum in (1) has $c_s(n) = \#\,\mathrm{F}_s(n)$ terms. Since the prefix has left degree 0, there must be a term greater than 1 in order to have $c_s(n+1) \geq c_s(n)$ (which is true as s is infinite). By definition, a factor with left degree greater than 1 is a left special factor.

For sufficiently large n, such a factor should be a prefix of s by hypothesis. We have reached a contradiction. □

An infinite word s is *periodic* if it can be written as $s = vvv\cdots = v^\omega$ for some finite word v. An *ultimately periodic* word is an infinite word of the form uv^ω for some $u, v \in A^*$. As is well known (see for instance [13, Lemma 1.4.4]), a recurrent ultimately periodic word is necessarily periodic.

We need one of the most well-known and useful restatements of the theorem of Morse and Hedlund (cf. [1, 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

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

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. $\square$

The following proposition was proved in [10, 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 $|w|$ 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. $\square$

A recurrent word s is an *Arnoux-Rauzy* word if it has exactly one left special factor and one right special factor of each length, all of degree $\#\,\mathrm{alph}(s)$. It is natural to extend this definition to the case of a unary alphabet $A = \{a\}$; the word a^ω is considered an Arnoux-Rauzy word, since it has a unique factor of each length, clearly not special but of degree $1 = \#A$.

Thus we can reformulate the definition as follows: a recurrent word s is Arnoux-Rauzy if for all $n \geq 0$, all factors of length n have minimum left degree (i.e. 1) except one whose left degree is maximum($\#\,\mathrm{alph}(s)$), and the same occurs for right degrees (but the two special factors may be different). By (1), this implies $c_s(n+1) = c_s(n) + \mathrm{alph}(s) - 1$ and then, as $c_s(0) = 1$,

$$c_s(n) = 1 + (\#\,\mathrm{alph}(s) - 1)n \quad \text{for all } n \in \mathbb{N} \ .$$

Arnoux-Rauzy words are uniformly recurrent (cf. [5]); this was part of the definition in [4]. An Arnoux-Rauzy word s is *standard* if its left special factors are prefixes of s.

Example 6. A well-known standard Arnoux-Rauzy word is the so-called *Tribonacci* (or Rauzy) word

$$\tau = abacabaabacababacabaabacabacabaabacabab \cdots$$

which can be obtained as a fixed point of the morphism $a \to ab$, $b \to ac$, $c \to a$ (see [4, 5]).

Remark 7. In order to show that a given infinite word s is a standard Arnoux-Rauzy word, it is sufficient to prove the following two conditions:

1. s has exactly one factor of right degree $\#\,\mathrm{alph}(s)$ for each length,
2. every left special factor of s is a prefix of it.

Indeed, under such hypotheses s is recurrent by Lemma 2. Moreover, in view of (1), by the first condition we derive

$$c_s(n+1) \geq \#\,\mathrm{alph}(s) + c_s(n) - 1 \tag{2}$$

for all $n \geq 0$; by condition 2, all factors which are not prefixes have left degree 1, so that equality holds in (2) and there is one factor of left degree $\#\,\mathrm{alph}(s)$. In conclusion, all factors of length n have left degree 1, except one which has left degree $\#\,\mathrm{alph}(s)$, and the same occurs for right degrees, for all n; hence s is an Arnoux-Rauzy word (standard by condition 2).

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 $x \in \mathrm{alph}(s)$.

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 $|W_j| \geq N$. Since prefixes (resp. suffixes) of left (resp. right) special factors are left (resp. right) special themselves, by conditions 1 and 2 it follows that W_i is a border of W_{i+1} for all $i \geq j$, and 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

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

so that $\#B$ 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 $|w| \geq N$, and w is a right special factor of s, by condition 2 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 $u_x w = w v_x$ 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 as 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 note that, as a consequence of the definition of return words, one has

$$z \in \mathrm{Fact}(t) \Leftrightarrow \varphi(z)w \in \mathrm{Fact}(s) \,, \quad z \in \mathrm{Pref}(t) \Leftrightarrow \varphi(z)w \in \mathrm{Pref}(s) \ . \tag{3}$$

We will prove that t is a standard Arnoux-Rauzy word; it suffices (see Remark 7) to show that t has exactly one right special factor of each length, that each right special factor has degree $\#B$, and finally that all left special factors of t are prefixes of it.

Clearly t is not periodic, as $s = \varphi(t)$ and s is not periodic. Hence t has right special factors of any length. 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 $z_i x_i, z_i y_i \in \mathrm{Fact}(t)$ for $i = 1, 2$. By (3), this implies $\varphi(z_i x_i)w, \varphi(z_i y_i)w \in \mathrm{Fact}(s)$. For $\alpha \in \{x_i, y_i\}$ and $i = 1, 2$ we have

$$\varphi(z_i \alpha)w = \varphi(z_i) u_\alpha w = \varphi(z_i) w v_\alpha \in \mathrm{Fact}(s)$$

with $v_{x_i}^f \neq v_{y_i}^f$, so that $\varphi(z_1)w$ and $\varphi(z_2)w$ are right special factors of s. By condition 2, either $\varphi(z_1)w \in \mathrm{Suff}(\varphi(z_2)w)$, 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 $\zeta := \varphi(z)w$ is right special in s. As $|\zeta| \geq |w|$, the word ζ is a suffix of some W_n with $n \geq k$, so that

$\zeta x = \varphi(z)wx \in \mathrm{Fact}(s)$ for all $x \in B$. Since the only complete return to w in s starting with wx is wv_x, it follows that

$$\varphi(z)wv_x = \varphi(z)u_x w = \varphi(zx)w \in \mathrm{Fact}(s) \ ,$$

so that $zx \in \mathrm{Fact}(t)$ for all $x \in B$, proving that z has right degree $\#B$.

Let now z' be a left special factor of t, and let $xz', yz' \in \mathrm{Fact}(t)$ for some distinct letters $x, y \in B$. Then $\varphi(xz')w, \varphi(yz')w \in \mathrm{Fact}(s)$. 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 $\varphi(z')w \in \mathrm{Pref}(s)$ and then $z' \in \mathrm{Pref}(t)$ by (3). $\square$

4 Concluding Remarks

Theorem 1 shows that what seems to be a natural (though very wide) generalization of the standard episturmian words, retains a strong connection with Arnoux-Rauzy words.

One could ask whether such result could be improved to obtain a full characterization of morphic images of standard Arnoux-Rauzy words. However, the converse of Theorem 1 is false; a simple counterexample is given by the image s of the Tribonacci word under the episturmian morphism (cf. [5, 8]) $f : a \to a$, $b \to ba$, $c \to ca$. Indeed,

$$s = f(\tau) = abaacaabaaabaacaabaabaacaabaaabaacaabaa \cdots$$

does not satisfy condition 1 of Theorem 1 for any N, as aa is a left special factor of s which is not a prefix of it. Nevertheless, this counterexample suggests that the situation could be better in the general (non-standard) case, since the word s, being episturmian, *does* have only one left special factor and one right special factor of each length.

By modifying the proof of Theorem 1 suitably, it is not difficult to show the following:

Theorem 8. *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 8 would be welcome, as well as any step towards the converse (what can be said about special factors of morphic images of Arnoux-Rauzy words?). Having a simple characterization could help in a more general classification of infinite words with low factor complexity.

References

1. Morse, M., Hedlund, G.A.: Symbolic dynamics. American Journal of Mathematics **60** (1938) 815–866

2. Berstel, J., Séébold, P.: Sturmian words. In Lothaire, M., ed.: Algebraic Combinatorics on Words. Cambridge University Press, Cambridge UK (2002)
3. Berstel, J.: Sturmian and Episturmian Words (a survey of some recent results). In Bozapalidis, S., Rahonis, G., eds.: Conference on Algebraic Informatics, Thessaloniki. Volume 4728 of Lecture Notes in Computer Science. (2007) 23–47
4. Arnoux, P., Rauzy, G.: Représentation géométrique de suites de complexité $2n+1$. Bulletin de la Société Mathématique de France **119** (1991) 199–215
5. Droubay, X., Justin, J., Pirillo, G.: Episturmian words and some constructions of de Luca and Rauzy. Theoretical Computer Science **255** (2001) 539–553
6. Bucci, M., de Luca, A., De Luca, A., Zamboni, L.Q.: On different generalizations of episturmian words. Theoretical Computer Science **393** (2008) 23–36
7. Lothaire, M.: Combinatorics on Words. Addison-Wesley, Reading MA (1983)
8. Justin, J., Pirillo, G.: Episturmian words and episturmian morphisms. Theoretical Computer Science **276** (2002) 281–313
9. de Luca, A., De Luca, A.: Pseudopalindrome closure operators in free monoids. Theoretical Computer Science **362** (2006) 282–300
10. Bucci, M., de Luca, A., De Luca, A., Zamboni, L.Q.: On θ-episturmian words. European Journal of Combinatorics **30** (2009) 473–479
11. Fischler, S.: Palindromic prefixes and episturmian words. Journal of Combinatorial Theory, Series A **113** (2006) 1281–1304
12. Durand, F.: A characterization of substitutive sequences using return words. Discrete Mathematics **179** (1998) 89–101
13. de Luca, A., Varricchio, S.: Finiteness and regularity in semigroups and formal languages. Monographs in Theoretical Computer Science. An EATCS Series. Springer-Verlag, Berlin (1999)