String attractors of episturmian sequences

L’ubomíra Dvořáková
FNSPE Czech Technical University in Prague, Czech Republic

Abstract
In this paper, we describe string attractors of all factors of episturmian sequences and show that their size is equal to the number of distinct letters contained in the factor.

Keywords: episturmian sequences, Sturmian sequences, string attractors
2000 MSC: 68R15

1. Introduction

String attractors play an important role in the field of data compression. They have been recently defined and studied by Kempa and Prezza [2]: a string attractor of a finite word \( w = w_0w_1 \ldots w_n \), where \( w_i \) are letters, is a subset \( \Gamma \) of \( \{0, 1, \ldots, n\} \) such that each factor of \( w \) has an occurrence containing an element of \( \Gamma \).

In this paper we consider string attractors from the point of view of combinatorics on words. Several results have already been reached in this aspect. Basic combinatorial properties of string attractors (e.g., string attractors of powers, conjugates, etc.) were studied by Mantaci et al. [6]. Moreover, the authors showed that finite standard Sturmian words have an attractor of size 2 containing two consecutive positions and that the size of the smallest string attractor of de Bruijn words grows asymptotically as \( \frac{n}{\log n} \), where \( n \) is the length of the word. The smallest string attractor (of size 4) of a particular factor subset of the Thue-Morse sequence was determined in [4]. Schaeffer and Shallit [8] considered it more natural to study string attractors of prefixes of infinite sequences instead of particular classes of finite factors of those sequences. This was then formalized in terms of string attractor profile function in [7]. The classical notions in combinatorics on words that somehow measure repetitiveness are the factor complexity and the recurrence function. The string attractor profile function builds a bridge between them, as described in [7]. Its behaviour was studied for linearly recurrent sequences and for automatic sequences in [8]. Ibidem, the authors determined the values of string attractor profile function for the period-doubling sequence, the Thue-Morse sequence, the Tribonacci sequence, and the
powers of two sequence (see also \( \mathbb{R} \)). The string attractor profile function of standard Sturmian sequences was determined and its properties for fixed points of morphisms were studied in \( \mathbb{R} \).

In this paper we determine string attractors of all factors of episturmian sequences. In a preliminary paper \[8\], the authors announced the form of string attractor of circularly balanced epistandard words and promised to detail the proof in the full version (using palindromic closures). However to our knowledge, the full version does not contain such results. Our construction of attractors of factors of episturmian sequences is very simple, it is also based on palindromic closures. The size of the obtained string attractors of factors of episturmian sequences is the smallest possible, it equals the number of distinct letters contained in the factor. It provides string attractors for all factors of Sturmian sequences unlike the construction in \[6\], which works only for finite standard Sturmian words. Our attractors are different because they do not contain consecutive positions in general. It is a straightforward consequence of our result that the string attractor profile function is eventually constant for any episturmian sequence.

2. Preliminaries

An alphabet \( \mathcal{A} \) is a finite set of symbols called letters. A word over \( \mathcal{A} \) of length \( n \) is a string \( u = u_0u_1\cdots u_{n-1} \), where \( u_i \in \mathcal{A} \) for all \( i \in \{0, 1, \ldots, n-1\} \). We let \( |u| \) denote the length of \( u \) and \( \pi = u_{n-1}\cdots u_1u_0 \). If \( u = \pi \), then \( u \) is called a palindrome. The set of all finite words over \( \mathcal{A} \) together with the operation of concatenation forms a monoid, denoted \( \mathcal{A}^* \). Its neutral element is the empty word \( \varepsilon \) and we write \( \mathcal{A}^+ = \mathcal{A}^* \setminus \{\varepsilon\} \). If \( u = xyz \) for some \( x, y, z \in \mathcal{A}^* \), then \( x \) is a prefix of \( u \), \( z \) is a suffix of \( u \) and \( y \) is a factor of \( u \).

A sequence over \( \mathcal{A} \) is an infinite string \( u = u_0u_1u_2\cdots \), where \( u_i \in \mathcal{A} \) for all \( i \in \mathbb{N} \). We always denote sequences by bold letters.

A sequence \( u \) is eventually periodic if \( u = vvw\cdots = v(w)^\omega \) for some \( v \in \mathcal{A}^+ \) and \( w \in \mathcal{A}^\ast \). If \( u \) is not eventually periodic, then it is aperiodic. A factor of \( u = u_0u_1u_2\cdots \) is a word \( y \) such that \( y = u_iu_{i+1}u_{i+2}\cdots u_{j-1} \) for some \( i, j \in \mathbb{N} \), \( i \leq j \). If \( i = j \), then \( y = \varepsilon \). In the context of string attractors, the set \( \{i, i+1, \ldots, j-1\} \) is called an occurrence of the factor \( y \) in \( u \). (Usually, only the number \( i \) is called an occurrence of \( y \) in \( u \).) If \( i = 0 \), the factor \( y \) is a prefix of \( u \). If each factor of \( u \) occurs at least twice in \( u \), the sequence \( u \) is recurrent.

The language \( \mathcal{L}(u) \) of a sequence \( u \) is the set of all its factors. \( \mathcal{L}(u) \) is closed under reversal if for each factor \( w \) of \( u \), the language \( \mathcal{L}(u) \) contains also \( \overline{w} \). A factor \( w \) of \( u \) is left special if \( aw, bw \) are in \( \mathcal{L}(u) \) for at least two distinct letters \( a, b \in \mathcal{A} \). The factor complexity of a sequence \( u \) is the mapping \( C_u : \mathbb{N} \to \mathbb{N} \) defined by \( C_u(n) = \# \{ w \in \mathcal{L}(u) : |w| = n \} \). The factor complexity of an aperiodic sequence \( u \) satisfies \( C_u(n) \geq n + 1 \) for all \( n \in \mathbb{N} \). The aperiodic sequences with the lowest possible factor complexity \( C_u(n) = n + 1 \) are called Sturmian sequences. Clearly, all Sturmian sequences are defined over a binary alphabet, e.g., \( \{0, 1\} \).
Kempa and Prezza [2] introduced the notion of string attractor. A *string attractor* (or *attractor* for short) of a word \( w = w_0w_1 \cdots w_{n-1} \), where \( w_i \in \mathcal{A} \), is a set \( \Gamma \subset \{0,1,\ldots,n-1\} \) such that every factor of \( w \) has an occurrence containing at least one element of \( \Gamma \). For instance, \( \Gamma = \{1,3\} \) is an attractor of \( w = 010010 \) (it corresponds to the underlined positions). It is the smallest possible attractor since each attractor necessarily contains occurrences of all distinct letters of the factor.

3. Palindromic closures and episturmian sequences

**Definition 1.** Let \( w \) be a word and \( a \) a letter, then \( (wa)^{(+)} \) is the shortest palindrome having \( wa \) as prefix.

It follows immediately from the definition that \( (wa)^{(+)} = wa \overline{w} \) if \( w \) does not contain \( a \). Otherwise, \( (wa)^{(+)} = vp \overline{v} \), where \( w = vp \) and \( p \) is the longest palindromic suffix of \( w \) preceded by \( a \).

**Example 2.** Let \( w = 000 \), then \( (w0)^{(+)} = w0 = 0000 \) and \( (w1)^{(+)} = w1w = 0001000 \). For \( v = 01101 \), we have \( (v0)^{(+)} = 011010110 \) and \( (v1)^{(+)} = 0110110 \).

**Definition 3.** Let \( \Delta = \delta_0\delta_1\delta_2 \cdots \) with \( \delta_i \in \mathcal{A} \) and define \( w_0 = \varepsilon \) and \( w_{n+1} = (w_n\delta_n)^{(+)} \) for all \( n \in \mathbb{N} \). Then we denote \( u(\Delta) = \lim_{n \to \infty} w_n \), i.e., \( u(\Delta) \) is a unique sequence having \( w_n \) as prefix for each \( n \in \mathbb{N} \), and we call \( \Delta \) the *directive sequence* of \( u(\Delta) \).

**Definition 4.** Let \( u \) be a sequence whose language is closed under reversal and such that for each length \( n \) it contains at most one left special factor. Then \( u \) is called an *episturmian sequence*. An episturmian sequence is *standard* if all left special factors are prefixes.

Sturmian sequences correspond to aperiodic binary episturmian sequences. It is well-known that for each episturmian sequence there exists a unique standard episturmian sequence with the same language. Since we are interested in the language of episturmian sequences, it suffices to consider only standard episturmian sequences [1]. For the study of attractors, the construction of episturmian sequences by palindromic closures seems to be handy. It was introduced by Droubay, Justin and Pirillo [1].

**Theorem 5.** Let \( u \) be a standard episturmian sequence over \( \mathcal{A} \). Then \( u = u(\Delta) \) for a unique sequence \( \Delta = \delta_0\delta_1\delta_2 \cdots \) with \( \delta_i \in \mathcal{A} \).

In the sequel we will always assume without loss of generality that \( \Delta \) is defined over \( \mathcal{A} = \{0,1,\ldots,d-1\} \) and the first letter of \( \Delta \) is 0, the second distinct letter in \( \Delta \) is 1, etc.

**Example 6.** The most famous standard Sturmian sequence is the *Fibonacci sequence*

\[
f = f(\Delta),
\]
where $\Delta = (01)^\omega$. The first six prefixes of $f$ read:

\[
\begin{align*}
    w_0 &= \varepsilon \\
    w_1 &= 0 \\
    w_2 &= 010 \\
    w_3 &= 010010 \\
    w_4 &= 01001010010 \\
    w_5 &= 0100101001001010010.
\end{align*}
\]

4. Attractors of episturmian sequences

In this section we determine attractors of all factors of episturmian sequences. We start with attractors of palindromic prefixes of standard episturmian sequences.

**Theorem 7.** Let $v$ be a non-empty palindromic prefix of a standard episturmian sequence. For every letter $a$ occurring in $v$, denote

\[ m_a = \max \{|p| : p \text{ is a palindrome and } pa \text{ is a prefix of } v\}. \]

Then $\Gamma = \{m_a : a \text{ occurs in } v\}$ is an attractor of $v$ and its size is minimal.

**Proof.** By the definition of palindromic closure, each palindromic prefix $v$ is equal to $w_n$ for some $n \in \mathbb{N}$. We will prove the statement by mathematical induction on $n$. Let us recall that we index positions from 0, i.e., $v = v_0v_1\ldots v_{|v|-1}$.

- For $n = 1$ we have $w_1 = 0$ and its attractor equals $\{0\}$. The longest palindromic prefix of $w_1$ followed by 0 is equal to $w_0 = \varepsilon$ and its length satisfies $|w_0| = 0$.

- For $n \geq 2$ we assume that $w_{n-1}$ has an attractor of the form from the statement. We have $w_n = (w_{n-1}j)^{(+)}$ for some $j \in \{0, 1, \ldots, d - 1\}$. The following three situations may occur:

  1. $w_n = w_{n-1}j$: According to the definition of palindromic closure, this happens only for $j = 0$ and $w_{n-1} = 0^\ell$ for some $\ell \in \mathbb{N}, \ell \geq 1$. The longest palindromic prefix of $w_n = 0^{\ell+1}$ followed by 0 is $w_{n-1} = 0^\ell$. We have thus $|w_{n-1}| = \ell$ and indeed $\{\ell\}$ is an attractor of $w_n$.

  2. $w_n = w_{n-1}jw_{n-1}$: By the definition of palindromic closure, this happens only in case when $w_{n-1}$ contains only letters $0, \ldots, j - 1$. It follows from the form of $w_n$ that the longest palindromic prefix of $w_n$ followed by $i$, where $i \in \{0, \ldots, j - 1\}$, is the same as in $w_{n-1}$. Moreover, the longest palindromic prefix of $w_n$ followed by $j$ is $w_{n-1}$. Let us explain that $\{m_0, \ldots, m_{j-1}, m_j\}$ as defined in the statement is an attractor of $w_n$: each factor of $w_n$ either contains the letter $j$, i.e., it has an occurrence containing the position $m_j = |w_{n-1}|$, or it is contained in $w_{n-1}$, which is a prefix of $w_n$, and by induction assumption it has an occurrence containing $m_i$ for some $i \in \{0, \ldots, j - 1\}$. 


3. \( w_n = w_{n-1}j u \) for some \( u \neq \varepsilon \) and \( u \neq w_{n-1} \): Assume \( w_n \) contains \( k \leq d \) letters. We want to prove that \( \{m_0, \ldots, m_j, \ldots, m_{k-1}\} \), as defined in the statement, is an attractor of \( w_n \). Since the longest palindromic prefix of \( w_n \) followed by \( i \), where \( i \in \{0, \ldots, k-1\} \), \( i \neq j \), is the same as in \( w_{n-1} \), we know by induction assumption that \( \{m_0, \ldots, m_j', \ldots, m_{k-1}\} \) is an attractor of \( w_{n-1} \), where we replaced \( m_j \) by \( m_j' = |w\ell| \), where \( w\ell \) is the longest palindromic prefix of \( w_{n-1} \) followed by \( j \). By the definition of palindromic closure we have

\[
\begin{align*}
    w_n &= \underbrace{w_j \underbrace{w\ell}_{w_{n-1}}} \underbrace{j \underbrace{u}}_{w_{n-1}}.
\end{align*}
\]

(1)

Then each factor of \( w_n \) either has an occurrence containing the position \( |w_{n-1}| \), i.e., crossing the second \( j \) in the expression (1), or is entirely contained in \( w_{n-1} \). In the latter case, it has an occurrence crossing the attractor of \( w_{n-1} \). However, it has no occurrence containing the position \( m_j' = |w\ell| \) because, according to (1), it would then also have an occurrence containing the position \( m_j = |w_{n-1}| \). To sum up, we have proved that each factor of \( w_n \) has an occurrence crossing \( \{m_0, \ldots, m_j, \ldots, m_{k-1}\} \).

\begin{itemize}
    \item \( w_1 \) contains only 0 and its longest palindromic prefix followed by 0 is equal to \( w_0 = \varepsilon \), therefore its attractor equals \( \{0\} \);
    \item \( w_2 \) contains 0, 1 and its longest palindromic prefix followed by 0, resp., 1, is equal to \( w_0 = 0 \), resp., \( w_1 = 0 \), i.e., \( \{0, 1\} \) is an attractor of \( w_2 \);
    \item \( w_3 \) contains 0, 1, 2 and its longest palindromic prefix followed by 0, resp., 1, resp., 2, is equal to \( w_0 = 0 \), resp., \( w_1 = 0 \), resp., \( w_2 = 010 \), i.e., \( \{0, 1, 3\} \) is an attractor of \( w_3 \);
    \item \( w_4 \) contains 0, 1, 2 and its longest palindromic prefix followed by 0, resp., 1, resp., 2, is equal to \( w_3 = 0102010 \), resp., \( w_1 = 0 \), resp., \( w_2 = 010 \), i.e., \( \{7, 1, 3\} \) is an attractor of \( w_4 \);
\end{itemize}

\[\text{Example 8.} \] Let \( \Delta = (012)^\omega \), i.e., \( \mathbf{u} = \mathbf{u}(\Delta) \) is the Tribonacci sequence. We underline the positions of the attractor from Theorem 7 in the first five non-empty palindromic prefixes:

\[
\begin{align*}
    w_0 &= \varepsilon \\
    w_1 &= 0 \\
    w_2 &= 010 \\
    w_3 &= 0102010 \\
    w_4 &= 01020100102010 \\
    w_5 &= 010201001020101020100102010102010.
\end{align*}
\]

Indeed,

\begin{itemize}
    \item \( w_1 \) contains only 0 and its longest palindromic prefix followed by 0 is equal to \( w_0 = \varepsilon \), therefore its attractor equals \( \{0\} \);
    \item \( w_2 \) contains 0, 1 and its longest palindromic prefix followed by 0, resp., 1, is equal to \( w_0 = 0 \), resp., \( w_1 = 0 \), i.e., \( \{0, 1\} \) is an attractor of \( w_2 \);
    \item \( w_3 \) contains 0, 1, 2 and its longest palindromic prefix followed by 0, resp., 1, resp., 2, is equal to \( w_0 = 0 \), resp., \( w_1 = 0 \), resp., \( w_2 = 010 \), i.e., \( \{0, 1, 3\} \) is an attractor of \( w_3 \);
    \item \( w_4 \) contains 0, 1, 2 and its longest palindromic prefix followed by 0, resp., 1, resp., 2, is equal to \( w_3 = 0102010 \), resp., \( w_1 = 0 \), resp., \( w_2 = 010 \), i.e., \( \{7, 1, 3\} \) is an attractor of \( w_4 \);
\end{itemize}
• $w_5$ contains 0, 1, 2 and its longest palindromic prefix followed by 0, resp., 1, resp., 2, is equal to $w_3$, resp., $w_4 = 01020100102010$, resp., $w_2$, i.e., \{7, 14, 3\} is an attractor of $w_5$.

In simple terms, the attractor of $w_n$ is equal to \{|w_{n-3}|, |w_{n-2}|, |w_{n-1}|\} for $n \geq 3$. Moreover, it is not difficult to determine that $|w_n| = \frac{T_{n+3} + T_{n+1} - 3}{2}$, where $T_n = T_{n-1} + T_{n-2} + T_{n-3}$ for all $n \in \mathbb{N}, n \geq 3$, and $T_0 = 0, T_1 = 1, T_2 = 1$. This attractor of $w_n$ is different from the one proposed in \[8\].

We can easily observe that when increasing $n$ by one, either one new element appears in the attractor or one element of the attractor changes. This holds in general for palindromic prefixes of standard episturmian sequences.

Remark 9. Since $w_n$ is a palindrome, the mirror image $\Gamma = \{|w_n| - 1 - m_a : a \text{ occurs in } w_n\}$ of the attractor $\Gamma = \{m_a : a \text{ occurs in } w_n\}$ of $w_n$ is an attractor of $w_n$, too.

Theorem 10. Let $u$ be an episturmian sequence. Each factor of $u$ containing $d$ distinct letters has an attractor of size $d$.

Proof. We will use the fact that there exists a unique directive sequence $\Delta = \delta_0 \delta_1 \delta_2 \cdots$ such that $L(u) = L(u(\Delta))$. If $d = 1$, then the statement evidently holds. Consider a factor $w$ of $u$ containing $d \geq 2$ distinct letters. Let $n \in \mathbb{N}$ be minimal such that $w_n$ contains $w$. Either $w = w_n$, then $w$ has an attractor of size $d$ by Theorem 7. Or $w \neq w_n$. Then there are two cases to be treated:

1. If $w_{n-1}$ does not contain $\delta_n$, then

$$w_n = (w_{n-1} \delta_n)^{(+)} = w_{n-1} \delta_n w_{n-1}.$$ 

Then each occurrence of $w$ contains the occurrence of $\delta_n$: more precisely $|w_{n-1}|$ (otherwise $w$ would be contained in $w_{n-1}$). Consider an arbitrary occurrence of $w$ in $w_n$. Find minimal indices $i$ and $j$, where $0 \leq i, j \leq n-1$, such that $w$ is contained in $w_i \delta_n w_j$. By minimality of $i$ and $j$, it is clear that $w$ contains all proper palindromic prefixes of $w_j$ together with the following letter and all proper palindromic suffixes of $w_i$ together with the preceding letter. Therefore on one hand, the considered occurrence of $w$ in $w_n$ contains the attractor of $w_j$ given in Theorem 7 shifted by $|w_{n-1}| + 1$ and the mirror image of the attractor of $w_5$ given in Theorem 7 shifted by $|w_{n-1}| - |w_1|$ (see Remark 3). Assume WLOG $j \geq i$. On the other hand, each factor of $w$ either has an occurrence crossing $\delta_n$ in $w_n$, i.e., containing the position $|w_{n-1}|$, or it is contained in $w_j$ ($w_1$ is a prefix of $w_j$), in which case it has an occurrence crossing the attractor of $w_j$ shifted by $|w_{n-1}| + 1$. Since $w$ has $d$ distinct letters, $w_j$ has $d - 1$ distinct letters and the attractor of $w_j$ is of size $d - 1$ by Theorem 7. Altogether, it implies existence of an attractor of size $d$ for $w$.

2. If $w_{n-1}$ contains $\delta_n$, then by the definition of palindromic closure we have

$$w_n = (w_{n-1} \delta_n)^{(+)} = w_{n-1} \delta_n \overline{w} = u \delta_n w_{n-1} = u \delta_n w_k \delta_n \overline{w}, \quad (2)$$
where \( u \) is a non-empty word and \( w_k \) is the longest palindromic prefix of \( w_{n-1} \) followed by \( \delta_n \) (\( w_k \) may be empty). Then each occurrence of \( w \) contains both underlined occurrences of \( \delta_n \): more precisely \(|u|\) and \(|w_{n-1}|\) (otherwise \( w \) would be contained in \( w_{n-1} \)). Consider an arbitrary occurrence of \( w \) in \( w_n \). Find minimal indices \( i \) and \( j \), where \( k < i, j \leq n-1 \), such that \( w \) is contained in

\[
v = y\underbrace{\delta_n w_k \delta_n z}_{w_i} = y\underbrace{\delta_n \delta_n z}_{w_j},
\]

where the underlined positions correspond to the ones underlined in (2). By minimality of \( i \) and \( j \), it is clear that \( w \) contains all proper palindromic prefixes of \( w_j \) together with the following letter and all proper palindromic suffixes of \( w_i \) together with the preceding letter. Therefore on one hand, the considered occurrence of \( w \) in \( w_n \) contains the attractor of \( w_j \) given in Theorem 7 shifted by \(|u| + 1\) and the mirror image of the attractor of \( w_i \) given in Theorem 7 shifted by \(|w_{n-1}| - |w_j|\) (see Remark 9). Assume WLOG \( j \geq i \). On the other hand, each factor of \( w \) either has an occurrence crossing the second underlined \( \delta_n \) in \( w_n \), i.e., containing the position \(|\delta_n| + |u| + 1\), or it is contained in \( w_j \) (\( w_i \) is a prefix of \( w_j \)). Hence it has an occurrence crossing the attractor of \( w_j \) shifted by \(|u| + 1\), which is of size \( d \) by Theorem 7.

Restivo et al. [7] defined the string attractor profile function of a sequence \( u \) as a map \( s_u : \mathbb{N} \to \mathbb{N} \) satisfying \( s_u(n) = \) the size of a smallest string attractor of the prefix of length \( n \) of \( u \). By Theorem 10 episturmian sequences have an eventually constant string attractor profile function.

**Corollary 11.** Let \( u = u_0 u_1 u_2 \cdots \) be an episturmian sequence over \( \mathcal{A} \). Then

\[
s_u(n) = \# \{ a \in \mathcal{A} : a \text{ occurs in } u_0 u_1 \cdots u_{n-1} \}.
\]

In particular, the string attractor profile function is eventually constant.

5. Acknowledgements

I would like to thank to Edita Pelantová and Francesco Dolce for their careful reading and several improvement proposals. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the project CZ.02.1.01/0.0/0.0/16_019/0000778.

**References**

[1] X. Droubay, J. Justin, G. Pirillo, *Episturmian words and some constructions of de Luca and Rauzy*, Theoret. Comput. Sci. 255 (2001), 539–553.
[2] D. Kempa and N. Prezza, *At the roots of dictionary compression: String attractors*, in: STOC 2018: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, 827–840.

[3] T. Kociumaka, G. Navarro, N. Prezza, *Towards a definitive measure of repetitiveness*, in: LATIN 2020: Proceedings of the 14th Latin American Symposium on Theoretical Informatics, 207–219.

[4] K. Kutsukake, T. Matsumoto, Y. Nakashima, S. Inenaga, H. Bannai, M. Takeda, *On repetitiveness measures of Thue-Morse words*, arXiv:2005.09524, doi:10.48550/ARXIV.2005.09524 (2020)

[5] S. Mantaci, A. Restivo, G. Romana, G. Rosone, M. Sciortino, *String attractors and combinatorics on words*, arXiv:1907.04660, doi:10.48550/ARXIV.1907.04660 (2019)

[6] S. Mantaci, A. Restivo, G. Romana, G. Rosone, M. Sciortino, *A combinatorial view on string attractors*, Theoret. Comput. Sci. 850 (2021), 236–248.

[7] A. Restivo, G. Romana, M. Sciortino, *String attractors and infinite words*, arXiv:2206.00376, doi:10.48550/ARXIV.2206.00376 (2022)

[8] L. Schaeffer, J. Shallit, *String attractors of automatic sequences*, arXiv:2012.06840, doi:10.48550/ARXIV.2012.06840 (2020)