CHARACTERIZATION OF A WORD BY ITS SUBWORDS

JÁN MAŇUCH

Department of Mathematics and Turku Centre for Computer Science,
University of Turku, Finland,
E-mail: manuch@cs.utu.fi

We consider what is the amount of subwords of a word needed to completely determine the word. More precisely, we study the maximal length such that all words of this length can be uniquely determined by its subwords of a fixed length. The set of subwords of a fixed length is called a spectrum. Four types of spectrums are analyzed: sparse, factor, sparse with multiplicity and factor with multiplicity. We also study the equivalence between the proper spectrum (subwords of a fixed length) and full spectrum (subwords at most of a fixed length) for each case. In all cases except sparse with multiplicity we give a precise answer. The results about factors and about the equivalence between the proper and full spectrums are new.

1 Introduction

The theory of subwords is a part of combinatorics of words dealing with properties of words related to classical non-commutative algebra. The main problem of this theory can be explained as follows:

> It is easy to compute for a given word the set of its subwords and their multiplicities. The problem is to compute a word knowing a part of the set of its subwords.

An overview of this problem can be found in [1].

In this article we divide the set of subwords of a word w to layers according to the length of subwords. Each such layer is called a *spectrum*, denoted $S(w,k)$, where k is the length of subwords. It is obvious that increasing the length of subwords in a spectrum we get more information. Almost in all cases we prove that knowing the k-spectrum of a word w we can determine the l-spectrum of w for $l<k$ without knowing w. This gives us the equivalence between the proper and full spectrum.

There are several questions we can ask about the spectrums:

Problem 1. Find a suitable characterization of the sets $S(w,k)$.

Problem 2. Define the equivalence relation $\equiv_k$ as follows: $u \equiv_k v \Leftrightarrow S(u,k) = S(v,k)$. What is the index of $\equiv_k$ over an m-letter alphabet?

Problem 3. For which n the words of length n can be uniquely determined by the sets $S(w,k)$?

We consider various types of subwords: factors, sparses, factors with multiplicities and sparses with multiplicities. Roughly speaking, a word u is a *factor* of a word w, if it appears inside w exactly as it is. A *sparse* can appear inside the given word scattered as several factors. In the cases with multiplicities we also consider how many times a word appears inside a given word.

Here are some comments or examples to the problems mentioned above:

1) For example, in the case of sparses, if $S_{\mathrm{S}}(w,k)$ contains both ab and ba, then it contains either aba or bab, where $a,b\in\Sigma$ and $k,|w|\geq 3$.

2) By an easy computation we can conclude that in the case of sparses the index is finite with upper bound $2^{\frac{m^{k+1}-1}{m-1}}$, cf. [1]. It is easy to find similar upper bounds in the other cases.

3) In many cases we can give an answer, see Section 3. In particular, the results in the cases of factors and factors with multiplicities are new. The most complicated are the cases of sparses and sparses with multiplicities. In the first case the answer is known due to [2]. Another proof of the result was given in [1] using a new notion: the subword distance. In the case of sparses with multiplicities there is still a very big gap between the known lower bound (linear) and upper bound (exponential) for the characteristic function. The value of characteristic function $f(k)$ is the maximal length such that all words of length $f(k)$ could be determined by the k-spectrum.

The note is organized as follows: In Section 2 we give necessary definitions. In Section 3 we state results without proof, both known and new ones. In the last two sections we give the proofs or references of these results.

2 Preliminaries

In this section we define factors, sparses and different kinds of spectrums.

Definition 1. A *sparse* s of a word $w \in \Sigma^*$, denoted as $s \preceq_s w$, is any subsequence of letters of w, i.e., there exist $n \geq 0$, $u_1, \ldots, u_n, v_1, \ldots, v_{n+1} \in \Sigma^*$ such that $s = u_1 \ldots u_n$ and $w = v_1 u_1 \ldots v_n u_n v_{n+1}$.

Definition 2. A *factor* f of a word $w \in \Sigma^*$, denoted as $f \preceq_f w$, is any word f for which exist $u, v \in \Sigma^*$ such that $w = ufv$.

Some authors use the name *subword* instead of the name *sparse*. In this paper we mean by a *subword* either a factor or a sparse. Next we give the definition of the notion of *spectrum*.

Definition 3. The *sparse* (resp. *factor*) *k-spectrum* of a word w, denoted as $S_{\mathrm{S}}(w,k)$ (resp. $S_{\mathrm{F}}(w,k)$), is the set of sparses (resp. factors) of length k of the word w. If instead of sets we consider multisets, then we have *k-spectrum with multiplicity*, denoted by $S_{X\mathrm{M}}(w,k)$, where $X \in \{\mathrm{F}, \mathrm{S}\}$. We can also define the *full k-spectrum* as $S_{\overline{X}}(w,k) = \bigcup_{i=1}^{k} S_X(w,i)$, where $X \in \{\mathrm{F}, \mathrm{S}, \mathrm{FM}, \mathrm{SM}\}$.

We represent multisets as formal power series, see e.g. [4]. The coefficient of f in the power series $S_{\mathrm{F}}(w,k)$, denoted $\langle S_{\mathrm{F}}(w,k), f\rangle$, is the number of occurrences of factor f in the word w, if $|f| = k$, and 0, otherwise. Sometimes we say 'X-spectrum' and by Definition 3 it is clear what type of spectrum we have in mind. If we shall talk about a non-full spectrum, we shall sometimes use the name *proper* spectrum to emphasize this fact.

3 Results

In this section we give an overview of the results, both known and new ones. In the following sections we give proofs and references to these results. First, we define

Table 1. Characteristic functions.

case X	result	reference
S, $\overline{\text{S}}$	$f_X(k) = 2k - 1$	[2]
SM, $\overline{\text{SM}}$	$2k - 1 \leq f_X(k) \leq \Phi(k) - 1$	[2], [3]
F, $\overline{\text{F}}$	$f_X(k) = k$	
F*, $\overline{\text{F}}^*$	$f_X(k) = 2k - 1$	[1]
FM	$f_X(k) = k$	
$\overline{\text{FM}}$	$f_X(k) = 2k - 1$	

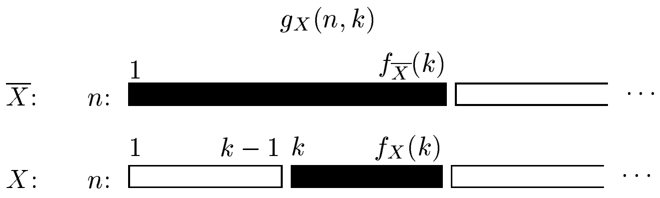
the important notion of the characteristic function. Let us define

$$g_X(n,k) = \begin{cases} 1, & \text{if the words of length } n \text{ are uniquely} \\ & \quad \text{determined by } S_X(w,k), \\ 0, & \text{otherwise.} \end{cases}$$

In other words, $g_X(n,k) = 1$ if and only if $\forall u \neq v \in \Sigma^n : S_X(u,k) \neq S_X(v,k)$. The *characteristic function* is defined as:

$$f_X(k) = \max\{n \mid g_X(n,k) = 1\}\,.$$

Note that Problem 3 is not completely answered even if we know the values of function $f_X(k)$. By the definition of $f_X(k)$, for all $n > f_X(k)$ we have $g_X(n,k) = 0$ and $g_X(f_X(k),k) = 1$. It is intuitive, but not obvious, that if we consider full spectrums, then for any $n < f_{\overline{X}}(k)$ we have also $g_{\overline{X}}(n,k) = 1$. In the case of proper spectrums, since for any $n < k$, $S_X(w,k) = \emptyset$, we have $g_X(n,k) = 0$ for such n. So in this case the intuitive property is that for any $k \leq n < f_X(k)$, $g_X(n,k) = 1$. We can illustrate these intuitive properties as follows:



where the black rectangle means that $g_X(n,k) = 1$.

In the Table 1 below we summarize the old and new results. In the cases with * we do not consider the words of the form $(uv)^k u$ with $k \geq 2$. The function $\Phi(k)$ is defined by the recurrence

$$\Phi(1) = 2, \quad \Phi(2) = 5, \quad \Phi(n+2) = \Phi(n+1) + \Phi(n)\,.$$

4 Properties of spectrums

From Table 1 we can notice that in the cases S, SM and F the result is the same for the proper and full spectrums. The reason for this is that in these cases we can determine the full spectrum of a word long enough from its proper spectrum. For example, cases S and $\overline{\mathrm{S}}$ are equivalent for $n \geq k$, where n is the length of words and k-spectrums are considered. This is obvious in the cases S and F, but if we take into account also multiplicity, the situation gets more complicated. The following example shows that we cannot determine the full FM-spectrum from the proper FM-spectrum even if $n \geq k$.

Example 1.

$$\begin{aligned}&S_{\mathrm{FM}}(aba,2) = ab + ba = S_{\mathrm{FM}}(bab,2), \text{ but}\\ &S_{\overline{\mathrm{FM}}}(aba,2) = ab + ba + 2a + b \neq ab + ba + a + 2b = S_{\overline{\mathrm{FM}}}(bab,2)\end{aligned}$$

For the other cases we have the equivalence by the following theorem.

Theorem 1. *For $X \in \{\mathrm{S},\mathrm{SM},\mathrm{F}\}$ and $n \geq k$ we can determine the full X-spectrum from the proper X-spectrum and vice versa. In particular for each $u, v \in \Sigma^n$:*

$$S_X(u,k) = S_X(v,k) \text{ iff } S_{\overline{X}}(u,k) = S_{\overline{X}}(v,k)\,,$$

and so

$$g_X(n,k) = g_{\overline{X}}(n,k)\,.$$

Proof. The only case which is not obvious is the case why the equivalence of the proper SM-spectrums implies that of the full SM-spectrums. Before we prove the claim in this case let us recall some definitions and theorems from [1].

Definition 4. The *binomial coefficient* of f and g, denoted by $\binom{f}{g}$, is the number of times g is a subword of f. The *infiltration product* $f \uparrow g$ is a mapping to $\mathbb{N}\langle\langle\Sigma\rangle\rangle$, where the coefficient of h in $f \uparrow g$, i.e., $\langle f \uparrow g, h\rangle$, is the number of pairs of subsequences of h that meet the two conditions:

i) they are equal respectively to f and g,

ii) their union gives the whole sequence h.

Now we state, without a proof, Theorem 6.3.18 of [1].

Claim 1. *For all $f, g, w \in \Sigma^*$ we have*

$$\binom{w}{f}\binom{w}{g} = \sum_{u \in \Sigma^*} \langle f \uparrow g, u\rangle \binom{w}{u}\,.$$

We shall need the following consequence of Claim 1.

Corollary 1. *For all $f, w \in \Sigma^*$, $a \in \Sigma$ we have*

$$\binom{w}{f}\left[\binom{w}{a} - \binom{f}{a}\right] = \sum_{\substack{u \in \Sigma^* \\ \pi(u) = \pi(f) + \pi(a)}} \binom{u}{f}\binom{w}{u}\,, \tag{1}$$

where π is the Parikh mapping.

Proof. If in the infiltration product one of the words f and g is a letter, then there are only a few possibilities:

$$\langle f \uparrow a, u \rangle = \begin{cases} \binom{f}{a}, \text{ if } u = f, \\ \binom{u}{f}, \text{ if } \pi(u) = \pi(f) + \pi(a), \\ 0, \text{ otherwise.} \end{cases}$$

So we have by Claim 1

$$\binom{w}{f}\binom{w}{a} = \sum_{u \in \Sigma^*} \langle f \uparrow a, u \rangle \binom{w}{u} = \binom{f}{a}\binom{w}{f} + \sum_{\substack{u \in \Sigma^* \\ \pi(u) = \pi(f) + \pi(a)}} \binom{u}{f}\binom{w}{u},$$

which implies the claim. □

We can finish the proof of Theorem 1.

Proof of Theorem 1 (continued). Assume that we know $S_{\mathrm{SM}}(w,k)$. We want to determine $S_{\overline{\mathrm{SM}}}(w,k)$, i.e., multisets $S_{\mathrm{SM}}(w,l)$, for all $1 \le l \le k$. First, we find 1-spectrum.

For each $a \in \Sigma$, we know $\binom{\binom{w}{a}}{k} = \binom{w}{a^k}$. We denote $|w|_a = \binom{w}{a}$. If $\binom{|w|_a}{k} > 0$, then $|w|_a \ge k$ and we can determine $|w|_a$ from $\binom{|w|_a}{k}$. So assume that $\binom{|w|_a}{k} = 0$ and $|w|_a < k$. We define the function

$$k_a(t) = \sum_{\substack{u \in \Sigma^k \\ |u|_a = t}} \binom{w}{u} = \begin{cases} 0, \text{ if } |w|_a < t, \\ > 0, \text{ if } |w|_a \ge t. \end{cases}$$

Note that for any t we can determine the value of the function. Clearly $|w|_a$ is the maximal $t < k$ such that $k_a(t) > 0$. Hence, we can determine the whole spectrum $S_{\mathrm{SM}}(w,1)$.

Now we show that we can determine $S_{\mathrm{SM}}(w,l-1)$ from the spectrums $S_{\mathrm{SM}}(w,l)$ and $S_{\mathrm{SM}}(w,1)$, where $l \le k$. Inductively we then have the whole full spectrum.

Take $f \in \Sigma^{l-1}$ and $a \in \Sigma$ in (1). Then $\pi(u) = \pi(f) + \pi(a)$ only if $u \in \Sigma^l$, so we can determine the right hand side of (1) from $S_{\mathrm{SM}}(w,l)$. By $S_{\mathrm{SM}}(w,1)$ we also know $m = \binom{w}{a} - \binom{f}{a}$. If $m \ne 0$ for an $a \in \Sigma$, then we can find the value of $\binom{w}{f}$ using Corollary 1. So for all $a \in \Sigma$, we have $\binom{w}{a} = \binom{f}{a}$, i.e., $\pi(w) = \pi(f)$. But then $n = |w| = |f| = l - 1 < k$, which contradicts the assumption $n \ge k$. □

Now we prove the intuitive properties of spectrums mentioned in Section 3.

Theorem 2. *For any $X \in \{\mathrm{S}, \mathrm{SM}, \mathrm{F}, \mathrm{FM}\}$ and any k, n*

i) if we can determine the words of length n from the full k-spectrum $S_{\overline{X}}$, i.e., if $g_{\overline{X}}(n,k) = 1$, then for all $m < n$ we have $g_{\overline{X}}(m,k) = 1$;

ii) if we can determine the words of length n from the proper k-spectrum S_X, i.e., if $g_X(n,k) = 1$, then for all $k \le m < n$ we have $g_X(m,k) = 1$.

Proof. Note that for $X \in \{\mathrm{S}, \mathrm{SM}, \mathrm{F}\}$ it is enough to prove the claim in the case $\overline{X}$ since the case X follows from Theorem 1. Assume to the contrary that there exist such n, k that $g_X(n,k) = 1$ and $g_X(n-1,k) = 0$, i.e.,

$$\forall u \neq v,\ |u| = |v| = n \ :\quad S_X(u,k) \neq S_X(v,k) \tag{2}$$
$$\exists u_0 \neq v_0,\ |u_0| = |v_0| = n-1 \ :\quad S_X(u_0,k) = S_X(v_0,k) \tag{3}$$

The idea is to find different words u_1, v_1 of length n such that $S_X(u_1,k) = S_X(v_1,k)$, which is a contradiction with (2).

Case $\overline{\mathrm{S}}$. It is sufficient to take $u_1 = u_0 a$, $v_1 = v_0 a$ for an arbitrary $a \in \Sigma$ and show that if $s \in \Sigma^{\leq k}$ is an X-subword of u_1, then it is also a subword of v_1. Analogously we have the other inclusion, which implies the equality between spectrums. Since s is a sparse of $u_1 = u_0 a$, either s is a sparse of u_0 or $s = s_0 a$ and s_0 is a sparse of u_0. In the first case, s is also a sparse of v_0, by (3), and then also of v_1. In the second case s_0 is a sparse of v_0 by (3) and hence $s = s_0 a$ is a sparse of $v_1 = v_0 a$. We are done.

Case $\overline{\mathrm{F}}$. Note that if $n > k$, then the words *abab...* and *baba...* of length n have the same full k-spectrums. Hence $g_{\overline{\mathrm{F}}}(n,k) = 1$ implies $n \leq k$, but then clearly for all $m < n \leq k$, $g_{\overline{\mathrm{F}}}(m,k) = 1$.

Case $\overline{\mathrm{SM}}$. Take again $u_1 = u_0 a$, $v_1 = v_0 a$. Since $S_{\overline{\mathrm{SM}}}(u_0,k) = S_{\overline{\mathrm{SM}}}(v_0,k)$ by (3), then clearly also $S_{\overline{\mathrm{SM}}}(u_0,k-1) = S_{\overline{\mathrm{SM}}}(v_0,k-1)$. Trivially,

$$S_{\mathrm{SM}}(za,l) = S_{\mathrm{SM}}(z,l) + S_{\mathrm{SM}}(z,l-1)a\,. \tag{4}$$

Summing this equation over $1 \leq l \leq k$ we get

$$S_{\overline{\mathrm{SM}}}(za,k) = S_{\overline{\mathrm{SM}}}(z,k) + S_{\overline{\mathrm{SM}}}(z,k-1)a\,,$$

which implies

$$\begin{aligned} S_{\overline{\mathrm{SM}}}(u_1,k) = S_{\overline{\mathrm{SM}}}(u_0 a,k) = S_{\overline{\mathrm{SM}}}(u_0,k) + S_{\overline{\mathrm{SM}}}(u_0,k-1)a = \\ S_{\overline{\mathrm{SM}}}(v_0,k) + S_{\overline{\mathrm{SM}}}(v_0,k-1)a = S_{\overline{\mathrm{SM}}}(v_0 a,k) = S_{\overline{\mathrm{SM}}}(v_1,k)\,, \end{aligned}$$

a contradiction.

Case $\overline{\mathrm{FM}}$. We need one definition and one lemma.

Definition 5. The *factor binomial coefficient* of f and g, denoted by $\binom{f}{g}_f$, is the number of times g is a factor of f.

Lemma 1. *If* $S_{\overline{\mathrm{FM}}}(u,k) = S_{\overline{\mathrm{FM}}}(v,k)$ *then we have* $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ *and* $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$.

Proof. It is enough to prove that we can determine letters $c_1, \ldots, c_{k-1}$, i.e., $\mathrm{pref}_{k-1}(w)$, from the spectrum $S_{\overline{\mathrm{FM}}}(w,k)$, where $w = c_1 \ldots c_n$. Analogously we also have the equivalence for suffices. We prove it by induction. Consider that we have already determined letters $c_1, \ldots, c_{l-1}$, where $l \leq k-1$. Note that since $\sum\limits_{z \in \Sigma^l} \binom{w}{za}_f = |c_{l+1} \ldots c_n|_a$, we can determine the value of the function

$$
\begin{aligned}
k_a(l) = \left\{ \begin{matrix} 1, \text{ if } c_l = a \\ 0, \text{ otherwise} \end{matrix} \right\} &= |w|_a - |c_1 \ldots c_{l-1}|_a - |c_{l+1} \ldots c_n|_a \\
&= \binom{w}{a}_f - |c_1 \ldots c_{l-1}|_a - \sum_{z \in \Sigma^l} \binom{w}{za}_f
\end{aligned}
$$

for each $a \in \Sigma$. These values for different a's obviously determine c_l. □

Proof of Theorem 2 (continued). We can take again $u_1 = u_0 a$, $v_1 = v_0 a$, where $a \in \Sigma$. Since $S_{\overline{\mathrm{FM}}}(u_0, k) = S_{\overline{\mathrm{FM}}}(v_0, k)$ we have $\mathrm{suff}_l(u_0) = \mathrm{suff}_l(v_0)$ for all $l \leq k-1$ by Lemma 1. We have

$$
S_{\mathrm{FM}}(u_1, l) = S_{\mathrm{FM}}(u_0, l) + \mathrm{suff}_{l-1}(u_0)a = S_{\mathrm{FM}}(v_0, l) + \mathrm{suff}_{l-1}(v_0)a = S_{\mathrm{FM}}(v_1, l)
$$

for all $l \leq k$ and hence $S_{\overline{\mathrm{FM}}}(u_1, k) = S_{\overline{\mathrm{FM}}}(v_1, k)$, which is a contradiction.

Case FM. In this case we cannot use the ideas of the proofs in the previous cases. The following example shows this.

Example 2. Consider the case $k = 6$. For $n = 7$ we have

$$
S_{\mathrm{FM}}(abababa, 6) = S_{\mathrm{FM}}(bababab, 6)\,,
$$

and all the other spectrums uniquely determine the word. For $n = 8$ we have

$$
\begin{aligned}
S_{\mathrm{FM}}(aabaabaa, 6) = S_{\mathrm{FM}}(abaabaab, 6) = S_{\mathrm{FM}}(baabaaba, 6)\,, \\
S_{\mathrm{FM}}(bbabbabb, 6) = S_{\mathrm{FM}}(babbabba, 6) = S_{\mathrm{FM}}(abbabbab, 6)\,,
\end{aligned}
$$

and all the other spectrums uniquely determine the word. It is clear that it is not enough to add a letter into $u_0 = abababa$, $v_0 = bababab$ in (3), where $n = 8$, to get a pair of words which contradicts (2).

Proof of Theorem 2 (continued). It is obvious that $g_{\mathrm{FM}}(k, k) = 1$. We prove that $g_{\mathrm{FM}}(n, k) = 0$ for all $n > k$. This implies the claim in this case, but also the equality $f_{\mathrm{FM}}(k) = k$ of Table 1.

To prove this we find for each $k \geq 1$ and $n > k$ words $u, v \in \Sigma^n$ such that $S_{\mathrm{FM}}(u, k) = S_{\mathrm{FM}}(v, k)$. Let us take

$$
\begin{aligned}
u &= (ba^x)^l ba^y, \\
v &= \begin{cases} a(ba^x)^l ba^{y-1}, \text{ if } y > 0, \\ a(ba^x)^l, \text{ if } y = 0, \end{cases} \\
\text{where } x = n - k,\ l &= \left\lfloor \frac{n-1}{n-k+1} \right\rfloor,\ y = (n-1) \bmod (n-k+1)\,.
\end{aligned}
$$

Observe that $u = zc_1$ and $v = c_2 z$, where $z \in \Sigma^*$, $c_1, c_2 \in \Sigma$ and $c_1 = c_2$ exactly when $y > 0$. Trivially we have the following equality:

$$
S_{\mathrm{FM}}(zc_1, k) = S_{\mathrm{FM}}(c_2 z, k) - \mathrm{pref}_k(c_2 z) + \mathrm{suff}_k(zc_1)\,,
$$

so if we prove that $\mathrm{pref}_k(v) = \mathrm{suff}_k(u)$, we are done. We can describe $\mathrm{pref}_k(v)$ as follows

$$\begin{aligned}
&\mathrm{pref}_k(v) = a(ba^x)^p \,\mathrm{pref}_q(ba^x), \\
&\text{where } p = \left\lfloor \frac{k-1}{n-k+1} \right\rfloor, \ q = (k-1) \bmod (n-k+1)\,.
\end{aligned}$$

Obviously $\mathrm{pref}_k(v)$ is equal to $\mathrm{suff}_k(u)$ if the equality $q-1=y$ holds in the case $q>0$ and the equality $x=y$ holds in the case $q=0$. Since $q, y \leq x = n-k$, we can rewrite these two conditions into one: $y-q+1 \equiv 0 \mod (n-k+1)$. But we have

$$y - q + 1 \equiv (n-1) - (k-1) + 1 = n - k + 1 \equiv 0 \mod (n-k+1)\,,$$

which completes the proof of Case FM.

We have proved the claim in all cases of Theorem 2. □

5 The characteristic functions

In this section we prove or give a reference of the results on the characteristic functions stated in Table 1. The goal is to find good lower and upper bounds for the characteristic functions. By Theorem 2, if we find n_1 such that $g_X(n_1,k)=1$, then we have the lower bound $f_X(k) \geq n_1$. On the other hand, if we find n_2 such that $g_X(n_2,k)=0$, then we have the upper bound $f_X(k) \leq n_2 - 1$. In the cases of proper spectrums, we require $n \geq k$. Almost in all case we find precise upper and lower bounds. By Theorem 1, we can consider full and proper spectrums at once in all cases except FM.

Cases S, $\overline{\mathrm{S}}$. The lower bound holds by Theorem 6.2.16 of [1]. The upper bound comes from the example, of the words $a^{k-1}ba^k$ and a^kba^{k-1}, which have the same k-spectrums, cf. [1].

Cases SM, $\overline{\mathrm{SM}}$. It is clear that if we know the spectrum with multiplicity, we can find out also the same spectrum without multiplicity. So in this case we have immediately the lower bound from the previous case: $f_{\mathrm{SM}}(k) = f_{\overline{\mathrm{SM}}}(k) \geq 2k-1$. The upper bound follows by considering the following sequences of words, cf. [3],

$$\begin{array}{llll}
u_0 = aba, & v_0 = 1, & w_0 = baa & \\
u_1 = ab, & v_1 = ba, & w_1 = ba & \\
u_{k+2} = w_k u_{k+1}, & v_{k+2} = u_{k+1} u_k, & w_{k+2} = u_{k+1} w_k, & \text{if } k \text{ is even,} \\
u_{k+2} = u_{k+1} w_k, & v_{k+2} = u_k u_{k+1}, & w_{k+2} = w_k u_{k+1}, & \text{if } k \text{ is odd.}
\end{array}$$

Since $S_{\mathrm{SM}}(u_k,k) = S_{\mathrm{SM}}(v_k,k)$ and obviously $u_k \neq v_k$ and $|u_k| = |v_k| = \Phi(k)$, we have the upper bound

$$\begin{aligned}
f_{\mathrm{SM}}(k) \leq \Phi(k) - 1 &= \left(\frac{15+\sqrt{5}}{10}\right)\left(\frac{1+\sqrt{5}}{2}\right)^k + \left(\frac{15-\sqrt{5}}{10}\right)\left(\frac{1-\sqrt{5}}{2}\right)^k - 1 \\
&\approx 1.7 \cdot 1.618^k.
\end{aligned}$$

Table 2. Some values of the function $f_{\mathrm{SM}}(k)$, values of the best known lower and upper bounds, and examples of words of length $f_{\mathrm{SM}}(k)+1$ with the same k-spectrums.

k	$2k-1$	$f_{\mathrm{SM}}(k)$	$\Phi(k)-1$	u, v
1	1	1	1	ab, ba
2	3	3	4	$abba$, $baab$
3	5	6	6	$abbaaab$, $baaabba$
4	7	11	11	$ab^2a^3ba^2b^2a$, $ba^3bab^2a^3b$
5	9	15	18	$ab^2a^5b^3a^4b$, $ba^4b^3a^5b^2a$
6	11	≥ 25	30	

Using the Morse-Thue sequences we would only get the upper bound $f_{\mathrm{SM}}(k) \leq 2^k - 1$, cf. Exercise 6.3.4 of [1]. But still there is quite a big gap between the known lower and upper bounds: the lower bound is linear and the upper bound is exponential. By computer testing we found some first values of the function $f_{\mathrm{SM}}(k)$. The tests needs a lot of memory and time when k and the size of the alphabet increases. But it is enough to perform the testing only on a two-letter alphabet as the following lemma shows.

Lemma 2. *Let $f_{\mathrm{SM}}(k)_d$ be the characteristic function in a d-letter alphabet. Then we have $f_{\mathrm{SM}}(k) = f_{\mathrm{SM}}(k)_d$ for all $d \geq 2$.*

Proof. Let Σ_d be a d-letter alphabet and define the function $g_{\mathrm{SM}}(n,k)_d$ in the obvious way. Since clearly $g_{\mathrm{SM}}(n,k) = g_{\mathrm{SM}}(n,k)_n$, it is enough to prove

$$g_{\mathrm{SM}}(n,k)_d = 1 \quad \text{iff} \quad g_{\mathrm{SM}}(n,k)_2 = 1$$

for all $d \geq 2$. The implication from left to right is obvious. Assume that there are n and k such that $n \geq k$, $g_{\mathrm{SM}}(n,k)_d = 0$ and $g_{\mathrm{SM}}(n,k)_2 = 1$. So there exist $u, v \in \Sigma_d^n$ with the same k-spectrums. This implies $\binom{u}{e} = \binom{v}{e}$ for all $e \in \Sigma_d^k$. Let $\Sigma_2 \subseteq \Sigma_d$ be a two-letter alphabet and let m be any mapping $\Sigma_d \to \Sigma_2$ such that $m(c) = c$ for all $c \in \Sigma_2$. If we take any $w \in \Sigma_d^n$ and any $f \in \Sigma_2^k$, then we have

$$\binom{m(w)}{f} = \sum_{\substack{e \in \Sigma_d^k \\ m(e) = f}} \binom{w}{e},$$

since each occurrence of f in $m(w)$ corresponds to one occurrence of exactly one $e \in \Sigma_d^k$ in w such that $m(e) = f$. Hence, we also have $\binom{m(u)}{f} = \binom{m(v)}{f}$ for any $f \in \Sigma_2^k$ and so $S_{\mathrm{FM}}(m(u),k) = S_{\mathrm{FM}}(m(v),k)$. If $m(u) \neq m(v)$, then we have a contradiction with the assumption $g_{\mathrm{SM}}(n,k)_2 = 1$. But since $u \neq v$, we can choose a mapping m such that $m(u) \neq m(v)$. $\square$

The results on the two-letter alphabet in Table 2 were found by computer.

Cases $\mathrm{F}, \overline{\mathrm{F}}, \mathrm{F}^*, \overline{\mathrm{F}}^*$. The lower bound is trivial. To prove the upper bound it is enough to take words $u = abab\ldots$, $v = baba\ldots$ of length $k+1$. Clearly they have the same k-spectrums. For the bounds in the cases with * we refer to Exercise 6.2.11 of [1].

Cases FM, $\overline{\mathrm{FM}}$. We have already shown the bounds in case FM in the proof of Theorem 2. In the full-spectrum case the lower bound comes from Lemma 1. Indeed, let u, v be two words of length $2k-1$ such that $S_{\overline{\mathrm{FM}}}(u,k) = S_{\overline{\mathrm{FM}}}(v,k)$. By the lemma we have $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ and $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$. Hence the words u and v can differ only in the middle letters. But this is impossible since the equivalence of the spectrums implies $\pi(u) = \pi(v)$. For the upper bound it is enough to take the same pair of words as in case S.

Acknowledgments

The author is grateful to Professor Juhani Karhumäki for useful discussions.

References

1. Lothaire, M., *Combinatorics on words*, Chapter 6: *Subwords* (Addison-Wesley, 1983).
2. Simon, I., *Piecewise testable events*, in: H. Brakhage (ed.), Automata Theory and Formal Languages, LNCS, vol. **33**, 214–222 (Springer–Verlag, 1975).
3. Choffrut, C., Karhumäki, J., *Combinatorics of words*, in: G. Rozenberg and A. Salomaa (eds), Handbook of Formal Languages, Vol. I, 329–438 (Springer, 1997).
4. Salomaa, A., Soittola, M., *Automata-theoretic aspects of formal power series* (Springer-Verlag, New York, 1978).