k-Spectra of Strictly Balanced Words

Joel D. Day and Pamela Fleischmann

Department of Computer Science,
Kiel University, Germany
{jda,fpa}@informatik.uni-kiel.de

Abstract. A word u is a scattered factor of w if there exist $u_1, u_2, ..., u_n$, and $v_0, v_1, .., v_n$ such that $u = u_1u_2...u_n$ and $w = v_0u_1v_1u_2v_2...u_nv_n$. We consider the set of length-k scattered factors (k-spectrum) of a given word w, denoted $\mathrm{ScatFact}_k(w)$. We prove several properties of the sets $\mathrm{ScatFact}_k(w)$ in the case of words w over a binary alphabet of length $2k$ for which the number of occurrences of each letter is equal. Such words are called strictly balanced. In particular, motivated by the task of recognising whether a set of words is a k-spectra of some word w, we consider the question of which cardinalities $n = |\mathrm{ScatFact}_k(w)|$ are obtainable for each k. We also consider the task of reconstructing words from their strictly balanced scattered factors.

1 Introduction

A scattered factor of w can be thought of as a representation of w in which some parts are missing. As such, there is considerable interest in the relationship of a word and its scattered factors from both a theoretical and practical point of view. For an introduction, see [3]. On the one hand, it is easy to imagine how, in any situation where discrete, linear data is read from an imperfect input – such as when sequencing DNA or during the transmission of a digital signal – scattered factors form a natural model, as multiple parts of the input may be missed, but the rest will remain unaffected and in-sequence. On the other hand, from a more theoretical perspective, there have been efforts to bridge the gap between the non-commutative field of combinatorics on words with traditional commutative mathematics via Parikh matricies (cf. e.g., [5, 6]) which are closely related to, and influenced by the topic of scattered factors.

One of the most fundamental questions about scattered factors of words and sets of scattered factors in general, is: given a set S of words (of length k), is S the set of scattered factors (or a k-spectrum) of some word w. In general, it remains a long standing goal of the theory to give a "nice" descriptive characterisation of scattered factor sets (and similarly, k-spectra), and to better understand their structure [3]. Another fundamental question concerning k-spectra, and one well motivated in several applications, is the question of reconstruction: given a word w of length n, for what values k does the k-spectrum of w uniquely determined? This question has generally had more success with definitive answers in a variety of cases. In particular, in [1], the exact bound of $\frac{n}{2}+1$ is given in the general case.

Other variations, including for the definition of k-spectra where multiplicities are also taken into account, are considered in [4], while [2] considers the question of reconstructing words from their palindromic scattered factors.

In the current work, we consider the restricted setting of strictly balanced words: words over a binary alphabet $\{\mathtt{a}, \mathtt{b}\}$ with equal numbers of as and bs. We show that the cardinality of their scattered factor sets ranges between $k+1$ and 2^k and we prove for every $k+1 \leq i \leq 3k-2$ whether a k-spectrum of cardinality i exists. Moreover some results between $3k-1$ and 2^k are given. In Section 4 we approach the question of reconstructing strictly balanced words from k-spectra in the specific case that the spectra are also limited to strictly balanced words only. While we are not able to resolve the question completely, we conjecture that the situation is similar to the general case; we show that this bound holds in the case that w contains at most two blocks of bs.

Before we are able to present our results, we need to define the setting of strictly balanced words. We consider words w over an alphabet $\Sigma = \{\mathtt{a}, \mathtt{b}\}$. The number of occurrences of a letter $\mathtt{a} \in \Sigma$ in a word $w \in \Sigma^*$ is denoted by $|w|_{\mathtt{a}}$. The subset of Σ^* which contains only words with equal numbers of occurrences of letters is defined by $\Sigma^*_{sb} = \{w \in \Sigma^* | \, \forall x, y \in \Sigma : |w|_x = |w|_y\}$ and these words are called *strictly balanced*. For example, abaa is not strictly balanced, while abbaba is.

Definition 1. *A word $u = a_1 \dots a_n \in \Sigma^n$, for $n \in \mathbb{N}$, is a* scattered factor *of a word $w \in \Sigma^+$ if there exists $v_0, \dots, v_n \in \Sigma^*$ with $w = v_0 a_1 v_1 \dots v_{n-1} a_n v_n$. Let* $\mathrm{ScatFact}(w)$ *denote the set of w's scattered factors and consider additionally* $\mathrm{ScatFact}_k(w)$ *(full k-spectrum) and* $\mathrm{ScatFact}_{\leq k}(w)$ *(k-spectrum) as the two subsets of* $\mathrm{ScatFact}(w)$ *which contain only the scattered factors of length $k \in \mathbb{N}$ or the ones up to length $k \in \mathbb{N}$.*

We note two obvious, but important symmetries regarding k-spectra: for$w \in \Sigma^*$. $\mathrm{ScatFact}(w^R) = \{u^R \mid u \in \mathrm{ScatFact}(w)\}$ and $\mathrm{ScatFact}(\overline{w}) = \{\overline{u} \mid u \in \mathrm{ScatFact}(w)\}$ hold with the renaming morphism $\bar{\cdot}$. Thus, from a structural point of view, it is sufficient to consider only one representative (here the lexicographically smallest with $\mathtt{a} < \mathtt{b}$) from the equivalence classes.

2 Cardinalities of k-Spectra of Strictly Balanced Words

In the current section, we are interested in the cardinalties of the k spectra, and in the question: which cardinalities are not possible? It is a straightforward observation that not every subset of Σ^k is a k-spectrum of some word w. For example aa and bb can only be scattered factors of a word containing both as and bs, and therefore having either ab or ba as a scattered factor. In fact, for $k = 2$, the sets $\{\mathtt{aa}, \mathtt{ab}, \mathtt{bb}\}$ and $\{\mathtt{aa}, \mathtt{ba}, \mathtt{bb}\}$ are the smallest possible k-spectra of words of length $2k$ in both the general case, and when restricted to strictly balanced words only. Moreover these sets are equivalent in the sense that one is a renaming (or a reversal) of the other. Note that the largest possible set in this case is $\{\mathtt{aa}, \mathtt{ab}, \mathtt{ba}, \mathtt{bb}\}$ which has size $4 = 2k = 2^k$. Our first result generalises the previous observation about minimal-size and maximal-size k-spectra.

Lemma 1. *For all $k \in \mathbb{N}$, the smallest reachable cardinality for any $w \in \Sigma_{sb}^{2k}$ is $|\operatorname{ScatFact}_k(w)| = k+1$, reached exactly for $w = \mathtt{a}^k\mathtt{b}^k$ (up to renaming and reversal), and $\operatorname{ScatFact}_k(\mathtt{a}^k\mathtt{b}^k) = \{\mathtt{a}^r\mathtt{b}^s \mid r+s = k, r, s \in [k]_0\}$ holds.*

Lemma 2. *Let $k \in \mathbb{N}$. Then $w \in \{ab, ba\}^k$ if and only if $\operatorname{ScatFact}_k(w) = \Sigma^k$.*

By the Lemmas 1 and 2, the characterisation for the smallest and the largest closure w.r.t. cardinality of the given set S are given. Now the gap in between will be investigated. Since there does not exist a gap for $k = 2$, assume $k \in \mathbb{N}_{\geq 3}$. The following two statements show that $2^k - 1$ and $2k$ are always reachable and thus the possible cardinalities for $k = 3$ are fully characterised.

1. $|\operatorname{ScatFact}_k(w)| = 2^k - 1$ iff $w \in \{(\mathtt{ab})^i\mathtt{a}^2\mathtt{b}^2(\mathtt{ab})^{k-i-2} \mid i \in [k-2]_0\}$ (in particular $\operatorname{ScatFact}_k(w) = \Sigma^k \backslash \{\mathtt{b}^{i+1}\mathtt{a}^{k-i-1}\}$),
2. $|\operatorname{ScatFact}_k(w)| = 2k$ iff $w \in \{\mathtt{a}^{k-1}\mathtt{bab}^{k-1}, \mathtt{a}^{k-1}\mathtt{b}^k\mathtt{a}\}$

The cardinality of $2k$ is important since there is a gap between $k+1$ and $2k$, i.e. $\forall w \in \Sigma_{sb}^{2k} : |\operatorname{ScatFact}_k(w)| \notin \{k+2, \ldots, 2k-1\}$. This shows that with increasing k the number of possible cardinalities at the *beginning* of the scala from $k+1$ to 2^k decreases: the larger k is the more unlikely it is somehow to find a k-spectrum of a small cardinality. To investigate the second gap we have $|\operatorname{ScatFact}_k(\mathtt{a}^{k-i}\mathtt{b}^k\mathtt{a}^i)| = k(i+1) - i^2 + 1$ for $i \in \left[\lfloor \frac{k}{2} \rfloor\right]$. It is worth noting that this includes all square numbers being at least four: $|\operatorname{ScatFact}_k(\mathtt{a}^{\frac{k}{2}}\mathtt{b}^k\mathtt{a}^{\frac{k}{2}})| = \left(\frac{k}{2}+1\right)^2$ holds for k even. Moreover $|\operatorname{ScatFact}_k(\mathtt{a}^{k-2}\mathtt{b}^k\mathtt{a}^2)| = 3k-3$ holds. This result is important since it will be shown in the following that the cardinalities $2k+1$ up to $3k-4$ are not reachable. In other words $\mathtt{a}^{k-2}\mathtt{b}^k\mathtt{a}^2$ delivers the third smallest cardinality after $k+1$ and $2k$. Contrarily the cardinality $3k-2$ belongs to the word $\mathtt{a}^{k-1}\mathtt{b}^2\mathtt{ab}^{k-2}$.

Proposition 1. *For $k \geq 5$, no word $w \in \Sigma_{sb}^{2k}$ has k-spectrum of cardinality $2k+i$ for $i \in [k-4]$, i.e. between $2k+1$ and $3k-4$ is a cardinality-gap.*

We will end this analysis with the conjecture that in contrast to the first gap, the last gap ends earlier the larger k is. More precisely, if for $k \in \mathbb{N}_{\geq 4}$ and $i \in [k-2]_0$, $w = \mathtt{a}^2\mathtt{b}^2(\mathtt{ab})^{k-3-i}\mathtt{ba}(\mathtt{ab})^i$ holds then $|\operatorname{ScatFact}_k(w)| = 2^k - 2 - i$ follows. Notice that this conjecture implies that indeed similar to the second gap here $4k-4$ is always reached. On the other hand, in contrast to the second gap, the third gap is not of the form $4k-4-i$ for $i \in [k-4]$.

3 Reconstructing Strictly Balanced Words from their k-Spectra

As with the general case, it is easy to see that strictly balanced words of length $2k$ are not uniquely determined by their scattered factors of length k. In the current section we discuss the question of when a strictly balanced word w of length $2k$ is uniquely identified by the set $\operatorname{ScatFact}_{k'}(w) \cap \Sigma_{sb}^{k'}$ for $2k > k' > k$.

Of course if k' is odd then $\mathrm{ScatFact}_{k'}(w) \cap \Sigma_{sb}^{k'} = \emptyset$ for all words w, so in these cases the answer is trivially negative. In the general case, Dress and Erdös [1] showed, that if $\mathrm{ScatFact}_{k+1}(w) = \mathrm{ScatFact}_{k+1}(w')$ holds for $w, w' \in \Sigma^{2k}$ then $w = w'$ follows. If w is strictly balanced we found a straightforward proof for their proposition. However, in both proofs, there is a necessity in some cases to consider scattered factors u consisting mostly of $\mathtt{a}$s or mostly of $\mathtt{b}$s – i.e., that do not belong to Σ_{sb}^*. Thus it remains an open problem whether the same bound of $k+1$ (or in the case that k is even, $k+2$) is sufficient. While we do not resolve the question completely, we conjecture that these bounds do still hold.

Conjecture 1. Let $k \in \mathbb{N}$. Let $k' = k+1$ if k is odd, and $k' = k+2$ if k is even. Let $w, w' \in \Sigma_{sb}^{2k}$ such that $\mathrm{ScatFact}_{k'}(w) = \mathrm{ScatFact}_{k'}(w')$. Then $w = w'$.

It is possible to show that the conjecture holds when there are at most two blocks of $\mathtt{b}$s (by symmetry at most two blocks of $\mathtt{a}$s), i.e. $w \in \mathtt{a}^*\mathtt{b}^*\mathtt{a}^*\mathtt{b}^*\mathtt{a}^* \cap \Sigma_{sb}^{2k}$:

- for k odd, w is uniquely determined by $\mathrm{ScatFact}_{k+1}(w) \cap \Sigma_{sb}^{k}$,
- for k even, w is uniquely determined by $\mathrm{ScatFact}_{k+2}(w) \cap \Sigma_{sb}^{k}$.

References

1. A. W.M. Dress and P. Erdös. Reconstructing words from subwords in linear time. *Annals of Combinatorics*, 8(4):457–462, 2004.
2. Š. Holub and K. Saari. On highly palindromic words. *Descrete Applied Mathematics*, 157:953–959, 2009.
3. M. Lothaire. *Combinatorics on Words.* Cambridge University Press, 1997.
4. J. Manŭch. Characterization of a word by its subwords. In *Developments in Language Theory*, pages 210–219. World Scientific, 1999.
5. A. Mateescu, A. Salomaa, and S. Yu. Subword histories and parikh matrices. *Journal of Computer and System Sciences*, 68(1):1–21, 2004.
6. A. Salomaa. Connections between subwords and certain matrix mappings. *Theoretical Computer Science*, 340(2):188–203, 2005.