The Spectra of Words

Robin Milner

Cambridge University, Cambridge, UK

Abstract. The k-spectrum of a word is the multiset of its non-contiguous subwords of length k. For given k, how small can n be for a pair of different words of length n to exist, with equal k- spectra? From the Thue-Morse word we find that n is at most 2^k. The construction of this paper decreases this upper bound to θ^k, where $\theta \simeq 1.6$ is the golden ratio; the construction was found, though not published, over thirty years ago. Recently the bound has been further reduced, but remains considerably greater than the greatest known lower bound.

Jan Willem Klop is renowned for his contributions to process algebra, term rewriting, and graphical models of computation; also for his exquisite diagrammatic presentations. Alongside these interests he finds delight in phenomena involving complex illustrations and counterexamples; for example, a single lambda term that takes several pages. He enjoys long words.

One long – indeed infinite – word is the Thue-Morse word. This word, and its finite prefixes of length 2^k, have a rich variety of properties. This paper is about one property that they almost, but not quite, enjoy. There is a sense in which these finite prefixes are not optimal, but appear to be nearly so.

I discovered this phenomenon in 1972, together with a counterexample to the optimality, but did not to publish it. In the last decade or so, the results have been repeated and indeed improved; in my concluding remarks I give a brief survey of some of this more recent work, enough for interested readers to discover the current state of this special problem area. But this festschrift for Jan Willem gives an opportunity to publish the phenomenon as it first appeared to me, in a form which I hope is readily digestible by many who, like me, are not expert in combinatorics.

We are concerned with words a over an alphabet A; for simplicity, we henceforth assume $A = \{0, 1\}$. A word s is a *subword* of a if the members of s occur in a in the correct sequence, not necessarily contiguously. A subword may occur often; for example, 01 occurs once in 010 but five times in 01011. (It is worth noting that there is not universal agreement of terminology; some authors use 'subword' to mean a contiguous occurrence, and in that case 01 appears only twice in 01011.) Denote by $\#(s, a)$ the number of non-contiguous occurrences of s in (i.e. as a subword of) a. Remarkably, the quantities $\#(s, a)$ share many properties with the binomial coefficients; indeed, the latter correspond to the case in which the alphabet A is a singleton. An elegant theory of these quantities $\#(s, a)$ is presented by Sakarovitch and Simon [8]. What little we need of that theory is included here, to make the paper self-contained.

A. Middeldorp et al. (Eds.): Processes... (Klop Festschrift), LNCS 3838, pp. 1–5, 2005.

For any natural number k, the *k-spectrum* of a word a is a function giving, for each word s of length k, the value of $\#(s, a)$. For example, the 2-spectrum of 0110 is

s	$\#(s, 0110)$
00	1
01	2
10	2
11	1

We ask the question: if we know the k-spectrum of a, does this fix a? Certainly not, if a has length at least 2^k. The finite prefixes of the Thue-Morse word provide a counterexample. The prefix a_k of length 2^k is given by

$$a_0, b_0 \stackrel{\mathsf{def}}{=} 0, 1$$
$$a_{k+1}, b_{k+1} \stackrel{\mathsf{def}}{=} a_k\, b_k, b_k\, a_k\ .$$

For example $a_3, b_3 = 01101001, 10010110$. An easy inductive proof shows that a_k and b_k have the same k-spectrum.

One way of presenting this proof depends on a Lemma that will come in useful later. It involves the notion of a *rewriting rule* $u \to v$, which may be used to replace a (contiguous) occurrence of u by v in a larger string. When the rule is understood we shall write $a \to b$, or sometimes $b \leftarrow a$, to mean that a single use of the rule can transform a into b.

Lemma. *Let u, v have the same k-spectrum. Suppose a can be rewritten into b by applying the rewrite rule $u \to v$ once and then applying $v \to u$ once, i.e. $a = cuc'$, $cvc' = dvd'$ and $dud' = b$. Then a, b have the same $k{+}1$-spectrum.*

To see that this provides an inductive proof that the Thue-Morse pair a_k, b_k have the same k-spectrum, assume the property for k, and consider a_{k+1}, b_{k+1}. In the Lemma take a, b to be this pair, and u, v to be a_k, b_k; then the result follows by taking $c = d' = \varepsilon$ (the empty word) and $c' = d = b_k$.

Let us look more closely at this when $k = 3$. The rewriting rule is $0110 \to 1001$, so if we underline the occurrences to be rewritten we have

$$a_3 = \underline{0110}1001 \to 1001\underline{1001} \to 10010110 = b_3\ .$$

Note that the two rewrites are non-overlapping. Now, consider a sequence of pairs a_k, b_k of unequal strings of equal length ℓ_k, where ℓ_k grow *more slowly* than 2^k; a similar inductive proof that a_k and b_k have the same k-spectrum would require that a_{k+1} can be transformed into b_{k+1} by an application of $a_k \to b_k$ followed by an application of $b_k \to a_k$ in which the second rewrite *overlaps the first*! It seemed unlikely that such pairs a_k, b_k could be defined. And it is not easy to think of constructing a sequence of unequal pairs with equal k-spectra other than inductively, applying something like the Lemma to yield the proof.

Therefore it was reasonable to hope to prove that each Thue-Morse pair a_k, b_k is *optimal*, in the sense that no shorter unequal pair has equal k-spectra.