Square-free words with square-free self-shuffles

James D. Currie & Kalle Saari
Department of Mathematics and Statistics
University of Winnipeg
515 Portage Avenue
Winnipeg, MB R3B 2E9
Canada
j.currie@uwinnipeg.ca, kasaar2@gmail.com

Submitted: July 27, 2013; Accepted: XX; Published: XX
Mathematics Subject Classifications: 68R15

Abstract

We answer a question of Harju: For every $n \geq 3$ there is a square-free ternary word of length n with a square-free self-shuffle.

1 Introduction

Shuffles of words are natural objects of study in combinatorics on words, and a variety of interesting problems have been posed. (See [5], for example.) Recently, self-shuffles of words have been studied. (See, for example [7, 8] which independently show that it is NP-complete to decide whether a finite word can be written as a self-shuffle.) If a word w is factored as

$$w = \Pi a_i = \Pi b_i,$$

where $a_i, b_i \neq \epsilon$, then we call

$$\Pi(a_i b_i)$$

a **self-shuffle** of w. For example, letting $w = 01101001$, $a_1 = 011$, $a_2 = 01$, $a_3 = 001$, $b_1 = 01$, $b_2 = 1010$, $a_3 = 01$, we get the self-shuffle of w

$$011\underline{01}01\underline{1010}001\underline{01}.$$

(Here the b_i have been underlined for ease of reading.) The notion of a self-shuffle equally applies to infinite words, and in [3] it is shown that the Fibonacci word has a self-shuffle which is equal to the Fibonacci word; similarly, it is shown that the Thue-Morse word is equal to one of its self-shuffles.

The recent note of Harju [4] poses this problem:

Problem 1.1. For every $n \geq 3$ is there a square-free word of length n with a square-free self-shuffle?

In this paper we answer this question in the affirmative; in fact the desired square-free words can be found over a ternary alphabet. In what follows, we freely use the usual notions of combinatorics on words. A standard reference is [6].

2 Long finite square-free words with square-free self-shuffles

Consider a square-free word $u \in \{0,1,2\}^*$ such that neither of 010 and 212 is a factor of u, and u is of the form

$$u = 0120w_0\Pi_{i=1}^{m}(aw_i)2012 \tag{1}$$

where $m \in \{0,1,2,3\}$, the $w_i \in \{0,1,2\}^*$ and $a = 2021020$. We will show later that such words u of length n exist for all large enough $n \equiv 3 \pmod 4$.

Let $b = 2021201020$. Let $\bar{u}$ be the word

$$\bar{u} = 0120w_0\Pi_{i=1}^{m}(bw_i)2012. \tag{2}$$

The longest prefix of b not containing 212 is 2021, which is also a prefix of a. The longest suffix of b not containing 010 is 1020, which is also a suffix of a. It follows that any factor of $\bar{u}$ not containing 010 or 212 is itself a factor of u.

Now consider the self-shuffle w of $\bar{u}$ given by

$$w = \bar{u}2^{-1}020^{-1}\bar{u} = 0120w_0\Pi_{i=1}^{m}(bw_i)20102120w_0\Pi_{i=1}^{m}(bw_i)2012. \tag{3}$$

The prefix of w of length $|\bar{u}|-1$ is a prefix of $\bar{u}$, while the prefix of w of length $|\bar{u}|$ has suffix 010. The suffix of w of length $|\bar{u}|-1$ is a suffix of $\bar{u}$, while the suffix of w of length $|\bar{u}|$ has prefix 212. It follows that the only factors of w not containing either 010 or 212 must themselves be factors either of $\bar{u}$ or of 1021; by the previous paragraph, they are factors of u or of 1021, and in particular are square-free. At this point we will mention that many arguments can be shortened by noting that the definitions of a, b, u, $\bar{u}$ and w are invariant under the operation combining reversal with the substitution $k \to 2-k$ on each letter. Particular words u and $\bar{u}$ need not be invariant under this operation, but they are sent to words of the same form.

Lemma 2.1. *Consider a square-free word u of the form (1) and let $\bar{u}$ and w be defined as in (2) and (3). Fix j, $0 \leq j \leq m$, and let a word U be obtained from $\bar{u}$ by replacing some j occurrences of b by a. Let W be obtained from w by making the analogous replacements. Thus $W = U2^{-1}020^{-1}U$. Then U and W are square-free. In particular, words $\bar{u}$ and w are square-free.*

Proof. Suppose not. Consider a word U obtained from $\bar{u}$ such that one of U and W contains a square, and such that $m-j$ is as small as possible.

We deal first with the case where $m-j=0$. In this case, $U=u$ is automatically square-free, and any factor of $W=w$ not containing 010 or 212 is square-free. Let yy be a factor of $W=w$, $y\neq\epsilon$. Thus, one of 010 or 212 is a factor of yy.

If $|y|_{010}\geq 1$ then $|yy|_{010}\geq 2|y|_{010}\geq 2$; however, $|yy|_{010}\leq |w|_{010}=1$. It follows that in fact $|y|_{010}=0$. Similarly, $|y|_{212}=0$. Now if 010212 is a factor of yy, then depending on how 010212 is distributed between the two copies of y, at least one of 010 and 212 must be a factor of y. This is impossible, so that 010212 is not a factor of yy. It follows that yy must be a factor of one of $0120w_0\Pi_{i=1}^{m}(aw_i)201021$ and $102120w_0\Pi_{i=1}^{m}(aw_i)2012$. (These are, respectively, the longest prefix and the longest suffix of w not containing 010212.)

Suppose then that yy is a factor of $0120w_0\Pi_{i=1}^{m}(aw_i)201021$. (The other case is similar.) Then 212 is not a factor of yy, forcing 010 to be a factor of yy. However, 010 must not be a factor of y, so that, depending on how 010 is split between copies of y, we can write $y=p0=10s$ or $y=p01=0s$, where s must be a prefix of 21, p a suffix of $0120w_0\Pi_{i=1}^{m}(aw_i)2$. However, $y=p0=10s$ is impossible; if $s\neq\epsilon$, then the word on the right-hand side of this equation ends in 1 or 2, while the left-hand word ends in 0; if $s=\epsilon$, $p=1$, which is not a suffix of $0120w_0\Pi_{i=1}^{m}(aw_i)2$. Again, $y=p01=0s$ forces $s=21$, since the left-hand word ends in 1; however $p01$ doesn't end in 21.

This shows that $m-j=0$ is impossible. We now have $m-j>0$, so that multiple copies of 010 and 212 appear in W. It will be useful to work out the distances between occurrences of 010, that is, the minimum value of $|010v|$ such that $010v010$ is a factor of W. From the definition of W, any word $010v$ such that $010v010$ is a factor of W is at least as long as a word of the form $01020w_i20212$, $01020w_m2$ or $0102120w_020212$. From the definition of u, factor $020w_m2012$ of aw_m2012 is square-free, and doesn't contain 010 or 212. This implies that w_m has prefix 1 and suffix 0. However, $w_m\neq 10$ or else aw_m would contain $0w_m$, which starts with 010. In particular $|w_m|\geq 3$, and $|010v|\geq |01020|+3+|2|=9$, and $|v|\geq 6$. From (3) we see that this argument also guarantees that any factor $010v010$ of U will also have $|v|\geq 6$.

Suppose yy is a square in W or in U, $y\neq\epsilon$. Suppose now that $|y|_{010}>0$. Note that 010 occurs in W or U in one of only two possible contexts: either $20212\underline{010}20$ or $02\underline{010}2120$. Observing the 3 characters to the left of an occurrence of 010 is enough to identify this context. If the 3-character string to the left is 212, then the context is 2021201020; if the 3-character string is not 212, then the context is 020102120 (since w_m ends in 0.) Similarly, examining the three characters to the right of an occurrence of 010 establishes its local context. Let us write $y=p010s$. Then $010sp010$ is a factor of W or U and $|sp|\geq 6$, so that at least one of $|p|$, $|s|\geq 3$. This establishes the local context of a certain occurrence of 010 in both copies of y, and these contexts must be the same. Since the local context 20102120 only occurs exactly once in W, and never in U, both local contexts of 010 in y are as a factor of b. Similarly, if $|y|_{212}>0$, then 212 appears in a local context coming from b. In fact, this argument shows that $|yy|_{010212}=0$; if $|yy|_{010212}=1$, then at least half of the occurrence of 010212 lies inside one copy of y, so that an occurrence of 010 or of 212 in y comes from 20102120, which is impossible. Therefore, if yy is a factor of W, we conclude that yy is a factor of one of $U2^{-1}$ and $0^{-1}U$, the longest prefix and suffix, respectively, of W not containing a 010 or 212 coming from 20102120; however, this prefix and suffix are themselves factors of W, so that we see that yy must be a factor of U.

We have shown that any occurrences of 010 in y arise as factors of b. Write $b' = 20212$, $b'' = 20$, so that $b = b'010b''$. We are thus saying that any occurrence of 010 in y is preceded (in W) by b' and followed by b''. Suppose $|y|_{010} \geq 1$. Write $y = p010s$. Suppose $|y|_b = 0$. Then either $|p| < |b'|$ or $|s| < |b''|$. If $|p| < |b'|$, write $W = xyyz$. Then b' must be a suffix of both xp and yp. Let σ be the common suffix of x and y such that $\sigma p = b'$. Replacing y by $\sigma y \sigma^{-1}$, we have a square yy in W such that $|y|_b = 1$. The case where $|s| < |b''|$ is similar; in either case, if $|y|_{010} > 0$, then adjusting yy cyclically if necessary, we can assume that $|y|_b > 0$. Now, replacing b's in y (and hence in U) by a's yields a square in a word of the form of U, with the same m, but larger j. This contradicts the minimality of $m - j$.

From now on, we can assume that $|y|_{010}$, $|y|_{212} = 0$ and yy is a factor of U. If $|yy|_{212010} > 0$, then depending on how 212010 is split between the copies of y, at least one of $|y|_{010}$ and $|y|_{212}$ is non-zero. We conclude that $|yy|_{212010} = 0$. By the same argument as earlier, any factors of U not containing 010 or 212 are square-free. It follows that at least one of $|yy|_{010}$ and $|yy|_{212}$ is non-zero. Without loss of generality (up to reversal and 2-complementation) suppose that $|yy|_{010} > 0$. Since $|y|_{010} = 0$, we must be able to write $y = p0 = 10s$ or $y = p01 = 0s$ where p is a suffix of 12 (since $|y|_{212} = 0$.) If $y = p0 = 10s$, each of $p = 12, 2, \epsilon$ is seen to be impossible. If $y = p01 = 0s$, then p begins with 0, which is also impossible.

We conclude that W and U, and hence w and $\bar{u}$, cannot contain a non-empty square yy. □

As promised, we now show that words of the form $u = 0120w_0\Pi_{i=1}^{m}(aw_i)2102$ of length n exist for all large enough $n \equiv 3 \pmod 4$.

The Thue-Morse word is the sequence $\mathbf{t} = \mu^{\omega}(0)$ where $\mu(0) = 01$, $\mu(1) = 10$. Word $\mathbf{t}$ is well-known to be overlap-free. From the definition of $\mathbf{t}$ it is clear that $\mathbf{t} \in \{01, 10\}^*$. On occasion it is useful to add 'bar lines' to a factor of $\mathbf{t}$ indicating the parsing of $\mathbf{t}$ in terms of 01 and 10. These bar lines always split any occurrence of 00 or 11; viz, 0|0 or 1|1, not |00| or |11|. It is proved in [1, Lemma 4] that $\mathbf{t}$ contains a factor of the form $10x01$ of every length greater than or equal to 6.

Consider the word $\mathbf{s}$ obtained from the Thue-Morse word by counting 1's between subsequent 0's. Thus if we write

$$\mathbf{t} = \Pi 01^{s_i},$$

then

$$\mathbf{s} = \Pi s_i.$$

It is well-known that $\mathbf{s}$ is square-free. It is also well-known and easily verified that neither of 010 and 212 is a factor of $\mathbf{s}$.

Lemma 2.2. *Word* $\mathbf{s}$ *contains a factor of the form* $0120x2012$ *of every length* $n \equiv 3$ *(mod 4),* $n \geq 23$.

Proof. A factor of $\mathbf{s}$ of the form $z = 0120x2012$ corresponds to a factor $v = 00101100y0110010110$ of $\mathbf{t}$. For clarity, add 'bar lines' to v:

$$v = 0|01|01|10|0y01|10|01|01|10.$$

The number of 0's in v is one more than the length of z, giving $|z| = |v|_0 - 1 = (|v| - 1)/2$.

$$\begin{aligned}
& \mathbf{s} \text{ contains a factor of form } z \text{ of length } k \\
\Rightarrow\ & \mathbf{t} \text{ contains a factor of form } v \text{ of length } 2k+1 \\
\Rightarrow\ & \mathbf{t} \text{ contains a factor of form } 10|01|0y'0|10|01 \text{ of length } k+1 \\
\Rightarrow\ & k \text{ is odd and } \mathbf{t} \text{ contains a factor of form } 10|0y''1|10 \text{ of length } (k+1)/2 \\
\Rightarrow\ & (k+3)/2 \text{ is even and } \mathbf{t} \text{ contains a factor of form } 10\hat{y}01 \text{ of length } (k+1)/4
\end{aligned}$$

The result follows. □

The words z of the last lemma begin and end in the form desired for u. We will now show when z is long enough, word $a = 2021020$ is a factor of z at least 5 times. Although the first and last occurrences of a may overlap with the prefix 0120 or suffix 2012 of z, there will be at least three other occurrences of a in z, so that for any $m \in \{0, 1, 2, 3\}$ we can write z in the form

$$z = 0120w_0\Pi_{i=1}^{m}(aw_i)2012,$$

as desired.

Lemma 2.3. *Suppose that* $02102v02102$ *is a factor of* $\mathbf{s}$*, but that 02102 is not a factor of* $2102v0210$*. Then* $|02102v02102| \leq 41$*.*

Proof. A factor 02102 of $\mathbf{s}$ corresponds to a factor $0|01|10|10|01|10|$ of $\mathbf{t}$. Such factors of $\mathbf{t}$ occur precisely in the context $01|10|01|10|10|01|10|01 = \mu^2(0011)$. A factor $02102v01202$ of $\mathbf{s}$ such that 02102 is not a factor of $2102v0210$ corresponds to a factor $(011)^{-1}\mu^2(0011u0011)(01)^{-1}$ of $\mathbf{t}$ which does not contain 0011 as an internal factor. Word $\mathbf{t}$ is concatenated from $\mu^4(0) = 0110100110010110$ and $\mu^4(1) = 1001011001101001$, and each of these contains a factor 0011. In addition, concatenating suffix 0 and prefix 011 of $\mu^4(0)$ produces a factor 0011; so does concatenating suffix 001 and prefix 1 of $\mu^4(1)$. We therefore see that the longest factor $0011u0011$ of $\mathbf{t}$ with no internal 0011 is the word $00110010110|10010110011$, of length 22.

We have determined that $0210v0210$ corresponds to a factor

$$z = (011)^{-1}\mu^2(0011u0011)(01)^{-1}$$

of $\mathbf{t}$ where $|0011u0011| \leq 22$. Because $\mathbf{s}$ is obtained from $\mathbf{t}$ by counting 0's and z begins and ends with 0,

$$|02102v02102| = |z|_0 - 1.$$

Every second letter of $\mu^2(0011u0011)$ is a 0, so that

$$\begin{aligned} |z|_0 &= |\mu^2(0011u0011)|_0 - |011|_0 - |01|_0 \\ &= |\mu^2(0011u0011)|/2 - 2 \\ &= 2|0011u0011| - 2 \\ &\leq 2(22) - 2 \\ &= 42. \end{aligned}$$

We conclude that $|02102v02102| \leq 41$. □

Corollary 2.4. *Any factor of* $\mathbf{s}$ *of length 40 contains 02102 as a factor.*

Corollary 2.5. *Any factor of* $\mathbf{s}$ *of length 42 contains* $a = 2021020$ *as a factor.*

Proof. The word 02102 cannot be preceded by 1 or 0 in s; It follows that 02102 can only be preceded by 2 in $\mathbf{s}$. Similarly, 02102 is only followed by 0. Any length 42 factor v of $\mathbf{s}$ contains 02102. Extending v before and after by one character then forces a to be a factor. □

Corollary 2.6. *Any factor* z *of* $\mathbf{s}$ *of the form* $0120x2012$ *of length at least 134 can be written in the form*

$$z = 0120w_0\Pi_{i=1}^{m}(aw_i)2012.$$

Proof. Since $134 = |0120| + 3(42) + |2012|$, the result follows by the previous Corollary. □

Theorem 2.7. *For every* $n \geq 143$ *there is a square-free word* $u \in \{0,1,2\}^*$ *of length* n *which permits a square-free self-shuffle.*

Proof. We note that $|b| - |a| = 3$. Given $n \geq 143$, let m be least such that $n - 3m \equiv 3 \pmod 4$. We have $|n - 3m| \geq 143 - 3(3) = 134$. By Lemma 2.2 there is a factor u of $\mathbf{s}$ of the form $u = 0120x2012$, $|z| = n - 3m$. By Lemma 2.6, word u has the form

$$u = 0120w_0\Pi_{i=1}^{m}(aw_i)2012.$$

Letting

$$\bar{u} = 0120w_0\Pi_{i=1}^{m}(aw_i)2012$$

gives a word $\bar{u}$ of length n, and by Lemma 2.1, both $\bar{u}$ and the self-shuffle

$$w = \bar{u}2^{-1}020^{-1}\bar{u}$$

of $\bar{u}$ are square-free. □

3 Short square-free words with square-free self-shuffles

It is well-known that $\mathbf{s}$ is the fixed point of $2 \mapsto 210$, $1 \mapsto 20$, $0 \mapsto 1$.

Lemma 3.1. *For every n with $3 \leq n \leq 200$, there exists a ternary square-free word with a self-shuffle that is also square-free.*

Proof. The following claims can be checked computationally[1].

For each n with $29 \leq n \leq 200$, $\mathbf{s}$ has a factor w of length $|w| = n$ such that the shuffle $p_1 p_2 s_1 s_2$ is square-free, where $w = p_1 s_1 = p_2 s_2$. Furthermore, the lengths of s_1 and p_2 can be restricted to satisfy $1 \leq |s_2|, |p_1| \leq 3$.

For each n with $3 \leq n \leq 28$ except for $n = 10$, there exist a ternary square-free word w with a square-free self-shuffle $p_1 p_2 s_1 s_2$ as above. The difference with the above is that we cannot always take w to be a factor of $\mathbf{s}$ and the lengths of s_1 and p_2 cannot be restricted as much.

Finally, for $n = 10$, one can take the square-free word $w = 0102120102$, which has the following square-free self-shuffle:

$$0102\underline{0}12\underline{102}0102\underline{120102}.$$

□

Combining this with the result of the previous section solves Harju's problem:

Theorem 3.2. *For every $n \geq 3$, there exists a ternary square-free word of length n having a square-free self-shuffle.*

Acknowledgements

The authors wish to thank the meticulous work of the reviewers.

References

[1] A. Aberkane & J. D. Currie, There exist binary circular 5/2+ power free words of every length, *Elec. J. Comb.* **11** (2004), R10.

[2] F.-J. Brandenburg, Uniformly growing k-th power-free homomorphisms, *Theoret. Comput. Sci.* 23 (1983), 69–82.

[3] Émilie Charlier, Teturo Kamae, Svetlana Puzynina, & Luca Q. Zamboni. Self-shuffling words, arXiv:1302.3844 (2013)

[4] T. Harju, A note on square-free shuffles of words, *LNCS, WORDS 2013*. To appear.

[5] D. Henshall, N. Rampersad, & J. Shallit, Shuffling and unshuffling, *Bull. EATCS*, **107** (2012), 131–142.

[1] An IPython notebook showing these computations can be found in http://users.utu.fi/kasaar/square-free_shuffles.ipynb

[6] M. Lothaire, Combinatorics on Words, Encyclopedia of Mathematics and its Applications 17, Addison-Wesley, Reading, 1983.

[7] S. Buss, M. Soltys, Unshuffling a square is NP-hard, arXiv: 1211.7161 (2013).

[8] R. Rizzi and S. Vialette. On recognizing words that are squares for the shuffle product, *CSR 2013, LNCS 7913* (2013), 235–245.