Some Remarks on Almost Periodic Sequences and Languages

GABRIEL ISTRATE
Faculty of Mathematics, University of Bucharest
Str. Academiei 14, Bucharest, Romania

1 Introduction

Almost periodicity has been considered in Formal Language Theory in connection with some topics in Symbolic Dynamics (see e.g. [3]). A notorious example of an almost-periodic sequence is the famous Thue-Morse sequence ([4], [7]). Almost periodicity has been considered when dealing with the decidability of monadic theories of unary functions [10]. In [7] some problems concerning this property are raised. For instance, one asks whether or not there exists some almost periodic word \( \alpha \) such that \( \text{Sub}(\alpha) \), the language of its finite factors, is context-free non-regular.

We will answer negatively (even in a stronger form) this question, as well as discussing other related topics.

2 Results

We will use Formal Language Notations from [6], [8]. By \( V^\omega \) we will denote the set of one-sided infinite words over the alphabet \( V \). \( V^{\omega^+} \) will mean its two-sided counterpart. For an infinite \( \alpha \) we denote by \( \text{Sub}(\alpha) \) the language of its finite factors. \( z < w \) will mean that \( z \) is a factor of \( w \). Note that the meaning of almost-periodicity from [10] is less restrictive than the one in [7]. It is this latter version we will deal with:

Definition 1 A sequence \( \alpha \in V^\omega \) is called almost periodic iff for any \( w < \alpha \) there exists a positive integer \( n_w \) such that any factor \( z \) of \( \alpha \) of length at least \( n_w \) has \( w \) as a factor.

Let us move this definition of almost-periodicity from infinite sequences to languages:

Definition 2 A language \( L \subset V^* \) is called

- closed iff \( L \) is closed under taking subwords.
- confluent iff for any \( x \) and \( y \) in \( L \) we may find \( z \in L \) such that \( x < z \) and \( y < z \).
- redundant iff for any \( x \) in \( L \) we may find \( n_x \geq 1 \) such that for any \( z \in L \), \( |z| \geq n_x \) implied \( x < z \).
• almost periodic iff \( L \) is closed and redundant.

### Remarks

(a). If \( L \) is almost periodic then \( L \) is confluent.

(b). The following assertions are equivalent:
- \( L \) is infinite, closed and confluent.
- \( L = \text{Sub}(\alpha) \) for some \( \alpha \in V^*\omega^+ \)

(c). The following assertions are equivalent:
- \( L \) is infinite and almost periodic.
- \( L = \text{Sub}(\alpha) \) for some almost periodic \( \alpha \in V^*\omega^+ \).

(d). If \( \alpha \in V^\omega \) is almost periodic then \( \text{Sub}(\alpha) \) is almost periodic.

**Justification.** To prove the non-trivial part of (b), we enumerate \( L \) as \( x_0, x_1, \ldots \), and define \( y_n \in L \) by \( y_0 = x_0, y_n < y_{n+1}, x_n < x_{n+1} \) (by the confluence of \( L \)). It follows that \( y_0 < y_1 < \ldots < y_n < \ldots \). Taking the (bilateral) limit we find \( \alpha \in V^{\omega^+\omega} \) with the desired properties. Now (c) follows by combining (a) and (b).

**Definition 3** A family \( \mathcal{F} \) of languages avoids almost periodicity iff any almost periodic language in \( \mathcal{F} \) is regular.

Now we may state:

**Theorem 1** Let \( \mathcal{F} \) be a family of languages such that for any infinite \( L \) in \( \mathcal{F} \) one may find \( w \neq \lambda \) such that \( \{w^n | n \geq 1\} \subset \text{Sub}(L) \). Then \( \mathcal{F} \) avoids almost periodicity.

**Proof.** Take \( L \in \mathcal{F} \) infinite and almost periodic and consider the corresponding \( w \). As \( \lim_{n \to \infty}|w|^n = \infty \) it follows (from the definition of almost periodicity) that \( L \subset \text{Sub}(\{w^n | n \geq 1\}) \). On the other hand \( \{w^n | n \geq 1\} \subset \text{Sub}(L) \) implies \( \text{Sub}(\{w^n | n \geq 1\}) \subset \text{Sub}(\text{Sub}(L)) = \text{Sub}(L) = L \), hence \( L = \text{Sub}(\{w^n | n \geq 1\}) \) is a regular language.

**Remark.** Many families of languages, including \( \mathcal{L}_2 \) (the family of context-free languages), \( \mathcal{M}_f \) (the family of matrix languages of finite index [2]), \( \mathcal{SM}_2 \) (the family of simple matrix languages [2]), \( \mathcal{C}, \mathcal{G} \) (the families of external contextual and of generalized contextual languages [5]), \( \mathcal{I}, \mathcal{IS} \) (the families of internal contextual and of internal contextual with choice languages [6]) satisfy the requirements of our lemma.

**Corollary 1** If \( L = \text{Sub}(\alpha) \) for some almost periodic \( \alpha \in V^\omega \) (or \( V^{\omega^+\omega} \)) and \( L \) belongs to one of the above mentioned families of languages then \( L \) is regular.

Another open problem from [7] was the existence of an algorithm deciding whether a given regular language can be written as \( L = \text{Sub}(\alpha) \) for some \( \alpha \in V^\omega \). We still cannot answer this question. However, by restricting ourselves to almost periodic sequences we get a better situation:
Corollary 2 Let \( F \) be a family of languages having the following properties:

- The finiteness problem for \( F \) is decidable.
- \( F \) constructively satisfies the (proof of) Theorem 1 (i.e. there is an algorithm which, given \( L \in F \) infinite, finds the right \( w \) and then tests whether \( L = \text{Sub}\{w^n|n \geq 1\} \)).

Then it is decidable whether \( L \in F \) can be written as \( L = \text{Sub}(\alpha) \) for some almost periodic \( \alpha \in V^\omega \).

Remark. A sufficient condition for the validity of the second condition in the previous corollary is the following:

- one can effectively construct the required \( w \);
- given \( L_1 \in F \) and \( L_2 \in L \) one can effectively check whether \( L_1 = L_2 \).

Corollary 3 The following families satisfy the hypothesis of Corollary 2 (and hence the problem whether an arbitrary language in them can be written as \( L = \text{Sub}(\alpha) \) for \( \alpha \in V^\omega \) almost periodic is decidable):

- the family of regular languages.
- the family of unambiguous context-free languages.

Proof. Both these families of languages satisfy the conditions of the previous remark (for unambiguous context-free languages one uses a result due to Semenov, see [1, 9]).

Open Problem 1 Are Semenov-type results true for the families \( M_f, SM_2 \)?

Open Problem 2 What is the decidability status of this problem for families \( M_f, SM_2, IS \)?

We could add the families \( C, G, I \) to the list from Corollary 3. However, for these families we may state a more precise result.

Theorem 2 Let \( L \) be an infinite almost-periodic language in one of the families \( C, G, I \). Then we can find \( a \in V \) such that \( L = a^* \).

Proof. We will prove our result for the family \( I \) (the other cases are analogous). Let \( L = L(V, B, C) \) be an infinite almost-periodic language in \( I \) such that \( L = \{a\}^* \) for no \( a \) in \( V \). Then \( L \) must include at least two different letters \( a \) and \( b \) (for the only infinite almost-periodic language over a one letter alphabet \( a \) is \( \{a\}^* \)). Consider \( x \), a nonvoid candidate for \( w \) (i.e. \( \{x^n|n \geq 1\} \subset \text{Sub}(L) \)). It follows that \( |x| \geq 2 \). Indeed, suppose that \( x \in V \). As \( \{x^n|n \geq 1\} \subset \text{Sub}(L) \) (from the definition of family \( I \)) it would follow that \( L = \{x\}^* \), which is not the case. As \( L \) is closed under taking subwords, \( a, b \in L \). It follows that \( a, b \in B \) (any context must increase the length by at least two: this follows the same way as \( |x| \geq 2 \)). Take \( w \in V^* \), \( w \neq \lambda \) having minimal length such that \( \{aw^n|n \geq 1\} \cup \{bw^n|n \geq 1\} \subset \text{Sub}(L) \) or \( \{aw^n|n \geq 1\} \cup \{bw^n|n \geq 1\} \subset \text{Sub}(L) \). Clearly there exists such a \( w \) (any nonvoid semicontext is a candidate for \( w \)).
Suppose we are in the first case and, moreover, $w$ does not end in $a$ (if not then exchange $a$ and $b$). Clearly $|w| \geq 2$. As $w^n \in \text{Sub}(L)$ for any $n$ and $L$ is almost-periodic, from the proof of Theorem 1 it follows that $L = \text{Sub}(\{w^n; n \geq 1\})$ hence $aw \in \text{Sub}(L) = L = \text{Sub}(\{w^n; n \geq 1\})$.

As $|w| \geq 2$ and $a$ is not the last letter in $w$, $aw < w^2$, hence $w^2 = zawt$ with $za \in \text{Pref}(w)$, $t \in \text{Suff}(w)$. As $|za| + |t| = |w^2| - |w| = |w|$, it follows that $w = zat$, hence $zaztt = zatzt$ so $(za)t = t(za)$. The equation $w = vu$ has as solutions the system $\{u = \beta^m, v = \beta^n | \beta \in V^*, m, n \geq 1\}$. As $t \neq \lambda$ (otherwise $\alpha$ would have been the last letter of $w$) we have $w = (za)t = \beta^k$ for some $\beta \in V^*, \beta \neq \lambda$ and $k \geq 2$. But then $\{a\beta^n | n \geq 1\} \cup \{b\beta^n | \beta \geq 1\} \subset \text{Sub}(L)$ and $|\beta| < |w|$, contradicting the minimality of $w$. Hence $L = \{a\}^*$. $\square$

Let us return to the problem of testing whether a given regular language $L$ can be written as $L = \text{Sub}(\alpha)$ for some $\alpha \in V^\omega$. The bi-sided version of this problem seems easier to tackle. Indeed, $L = \text{Sub}(\alpha)$ for some $\alpha \in V^{\omega\omega}$ iff $L$ is infinite, closed and confluent. Finiteness and closure are decidable for regular languages: given a regular grammar $G_1$, construct (effectively) a grammar $G_2$ for $\text{Sub}(L(G_1))$ and then test whether $L(G_1) = L(G_2)$. Let us further note that $L$ is confluent iff $\text{Sub}(L)$ is confluent. Now it is straightforward that testing whether a given regular language can be written in the required form is algorithmically equivalent to the problem of testing confluence for regular languages.

**Open Problem 3** Is confluence open for regular languages?

**Note (2022):** The Open Problem 3 has been solved (affirmatively) in Harju, Tero, and Lucian Ilie. "On quasi orders of words and the confluence property." Theoretical Computer Science 200.1-2 (1998): 205-224.

**References**

[1] D. Beauquier and M. Nivat. About rational sets of factors of a bi-infinite word. In *International Colloquium on Automata, Languages, and Programming*, pages 33–42. Springer, 1985.

[2] J. Dassow and G. Păun. *Regulated rewriting in formal language theory*, volume 18. Springer, 1989.

[3] G. Hedlund and M. Morse. Symbolic dynamics. *American Journal of Mathematics*, 60(4):815–866, 1938.

[4] K. Jacobs. Maschinenlenzuge 0-1-folgen. In *Selecta Mathematica I*, pages 1–27. Springer, 1969.

[5] S. Marcus. Contextual grammars. *Revue Roumaine Math. Pures Appl.*, 14(10):1525–1534, 1969.

[6] G. Păun. *Contextual Grammars*. Editura Academiei București (in Romanian), 1982.

[7] G. Păun and S. Marcus. Infinite (almost periodic) words, formal languages and dynamical systems. *Words and Languages Everywhere*, page 341, 2007.

[8] A. Salomaa. *Formal languages*. Academic Press Professional, Inc., 1987.
[9] A. Salomaa. Formal power series: A powerful tool for theoretical informatics. In Proceedings of the 300 years festive Congress of the Hambury Mathematical Association, pages 1033–1048, 1990.

[10] A. L. Semenov. Decidability of monadic theories. In International Symposium on Mathematical Foundations of Computer Science, pages 162–175. Springer, 1984.