A Family of Circular Systematic Comma-Free Codes

L. J. Cummings
Faculty of Mathematics, University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

Abstract

Comma-free codes are used to correct synchronization errors in sequential transmission. Systematic comma-free codes have codewords with fixed positions for error correction. We consider only comma-free codes with constant word length $n > 1$. Circular codes use the integers mod n as indices for codeword entries. We first show two easily stated conditions are equivalent to the existence question for circular systematic comma-free codes over arbitrary finite alphabets. For $n > 3$ a family of circular systematic comma-free codes with word length $n = p$, a prime, is constructed, each corresponding to a fair partition of a difference set in $\mathbb{Z}_n$.

1 Introduction

If $q > 1$, let $A = \{a_0, \ldots a_{q-1}\}$ denote a finite set of distinct elements called the alphabet. A block code with codewords of length $n > 1$ is a collection of n-tuples whose entries are elements of A. For notational convenience we write the codewords as $\mathbf{x} = x[0] \cdots x[n-1]$ rather than, say, $\mathbf{x} = x_0 \cdots x_{n-1}$. If $\mathbf{x}$ is any codeword we identify $x[n+k]$ and $x[k]$ for all $k \in \mathbb{Z}_n$, the integers mod n, and refer to $\mathbf{x}$ as a circular codeword. A *circular code* is a collection of circular codewords.

Definition 1 *The mth* overlap *of codewords* $\mathbf{x} = x[0] \cdots x[n-1]$ *and* $\mathbf{y} = y[0] \cdots y[n-1]$ *is a codeword of the form*

$$\mathcal{O}_m(\mathbf{xy}) = x[m] \cdots x[n-1]y[0] \cdots y[m-1],$$

JCMCC 58 (2006), pp. 87-96

where $1 \leq m \leq n-1$. *A code* C *is* comma-free *if for any two codewords* $\mathbf{x},\mathbf{y} \in C$, $\mathcal{O}_m(\mathbf{xy}) \notin C$ *for every* $m = 1, \ldots n-1$.

To establish synchronization; i.e., not permit an incorrect framing of a message stream, one solution is to use comma-free codes. Clearly framing errors are prevented in a noiseless channel by comma-free codes since synchronization is obtained after at most $n-1$ entries are read.

If d denotes Hamming distance, the *index* of any comma-free code C is

$$\rho_C = \min\{d(z, \mathcal{O}_m(\mathbf{xy}))\},$$

where the minimum is over all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in C$ and $1 \leq m \leq n-1$. The index of a comma-free code was first introduced in [10]. Comma-free codes C may also be defined by simply requiring $\rho_C \geq 1$.

There is some confusion in the literature about the term "overlap". On the one hand, in [11] Levenshtein uses the term "splice" rather than "overlap". But Levenshtein & Tonchev [14] and Tonchev [17] use the word "joint" instead. These different terms may result from different translations. Further confusion arises when variable length codewords are considered. Codewords $\mathbf{x} = x[0] \cdots x[n-1]$ and $\mathbf{y} = y[0] \cdots y[m-1]$ are sometimes said to "overlap" if there is $k = 1, \ldots n-1$ such that

$$x[k] \cdots x[n-1] = y[0] \cdots y[k-1].$$

Comma-free codes were first defined in a biology paper [3], and the first mathematical paper about comma-free codes [8] defined "overlap" as we have.

When $\mathbf{x} = \mathbf{y}$ then overlaps $\mathcal{O}_m(\mathbf{xx})$ $m = 1, \ldots n-1$ are cyclic permutations of $\mathbf{x}$. It is useful to rewrite Definition 1 as:

$$\mathcal{O}_m(\mathbf{xy})[i] = \begin{cases} x[i+m] & \text{for } i \in [0, n-m) \\ y[i-n+m] & \text{for } i \in [n-m, n), \end{cases} \tag{1}$$

where $[0, n-m) = \{0, \ldots n-m-1\} \subset \mathbb{Z}_n$ and $[n-m, n) = \{n-m, \ldots n-1\} \subset \mathbb{Z}_n$.

Systematic codes reserve certain positions for error correction. We consider here only correcting of synchronization or "mis-framing" errors in a noiseless channel. For a study of combining both bit error correction and synchronization errors see [9].

Definition 2 $C = C_n(Q_0, \ldots Q_{q-1})$ *is a* systematic circular code *over the alphabet* $A = \{a_0, \ldots a_{q-1}\}$ *with respect to a collection of non-empty disjoint*

sets $Q_0, \ldots, Q_{q-1}$, $q > 0$ *contained in* $\mathbb{Z}_n$, *provided any codeword* $\mathbf{x} \in C$ $\mathbf{x} = x[0] \cdots x[n-1]$ *satisfies* $x[i] = a_j$ *whenever* $i \in Q_j$, $i = 0, \ldots n-1$.

The *redundancy* of a systematic code is $|\cup Q_i|$ where $|\cdot|$ denotes cardinality. The redundancy of a systematic code simply counts the number of positions fixed by the systematic code. Positions not in any Q_i are called information positions and can contain arbitrary entries from A. There can be at most one codeword if $|\cup Q_i| = n$ and $q \geq n$. Otherwise, the number of codewords in a systematic code $C_n(Q_0, \ldots Q_{q-1})$ is bounded by q^t where $t = q - |\cup Q_i|$. $C_n(Q_0, \ldots Q_{q-1})$ can be taken as a t-dimensional finite vector space if A is an appropriate finite field of prime power order.

2 Equivalences

The following theorem gives two easily expressed conditions for a circular systematic code to be comma-free.

Theorem 1 *Let* $C_n(Q_0, \ldots Q_{q-1})$ *be a systematic circular code with codewords of length* n *over a finite alphabet* A. *The following are equivalent:*

(i) $C_n(Q_0, \ldots Q_{q-1})$ *is a comma-free code.*

(ii) For all $m \in \{1, \ldots n-1\}$, *there exists a pair of integers* $i, j \in \{0, \ldots q-1\}, i \neq j$ *such that either*

$$(Q_i + m) \cap Q_j \neq 0 \text{ or } (Q_j + m) \cap Q_i \neq 0.$$

(iii) $\mathbb{Z}_n - \{0\} \subseteq \Delta(Q_0, \ldots Q_{q-1}) = \{a - b \pmod n \mid a \in Q_i, b \in Q_j, i \neq j,\ i, j = 0, \ldots q-1\}$.

Proof: $(i) \Longrightarrow (ii)$. We argue by contradiction. Suppose there is $m = 1, \ldots n-1$ such that for every pair $i, j \in \{0, \ldots q-1\}$, $i \neq j$

$$(Q_i + m) \cap Q_j = \phi. \tag{2}$$

Define a codeword $\mathbf{x}$ for $a \in \mathbb{Z}_n$ by

$$x[a] = \begin{cases} a_i & a \in Q_i \cup (Q_i + m) \\ a_j & a \in Q_j \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Clearly $\mathbf{x}$ is well-defined by (2) and is in $C_n(Q_0, \ldots Q_{q-1})$ by definition.

We claim $\mathcal{O}_m(\mathbf{xx}) \in C_n(Q_0, \ldots Q_{q-1})$ as well. We must check that $\mathcal{O}_m(\mathbf{xx})[a] = a_i$, if $a \in Q_i$. Now,

$$\mathcal{O}_m(\mathbf{xx})[a] = \begin{cases} x[a+m] & a \in [0, n-m) \\ x[a+m-n] & a \in [n-m, n), \end{cases} \tag{4}$$

for all $a \in \{0, \ldots n-1\}$.

If $a \in [0, n-m)$ then $\mathcal{O}_m(\mathbf{xx})[a] = x[a+m]$. Since $a + m \in (Q_i + m), x[a+m] = a_i$, by the definition (3) of $\mathbf{x}$ because $a \in Q_i$. If $a \in [n-m, n)$ then $\mathcal{O}_m(\mathbf{xx}) = x[a+m-n]$. Since indices are from $\mathbb{Z}_n$, $x[a+m-n] = x[a+m]$ and the argument is the same as before. Therefore, $\mathcal{O}_m(\mathbf{xx}) \in C_n(Q_0, \ldots Q_{q-1})$, contradicting (2).

(ii) $\Longrightarrow$ *(iii)*. Let $m \in \{1, \ldots n-1\}$. Then, *(ii)* implies either there will be $a \in Q_i, b \in Q_j$ $i \neq j$ such that $a + m = b$; i.e., $m = b - a$ or there will be $a' \in Q_i, b' \in Q_j$ such that $b' + m = a'$; i.e., $m = a' - b'$. In either case, *(iii)* is satisfied.

(iii) $\Longrightarrow$ *(i)*. Let $\mathcal{O}_m(\mathbf{xy})$ be the *mth* overlap of two codewords $\mathbf{x}$,$\mathbf{y}$ $\in C_n(Q_0, \ldots Q_{q-1})$. Without loss of generality suppose $m = b - a$ where $a \in Q_i, b \in Q_j$. By definition,

$$\mathcal{O}_m(\mathbf{xy})[a] = \begin{cases} x[a+m] & a \in [0, n-m) \\ y[a+m-n] & a \in [n-m, n). \end{cases}$$

If $a \in [0, n-m)$ then

$$\mathcal{O}_m(\mathbf{xy})[a] = x[a+m] = x[b] = a_j,$$

contradicting $a \in Q_i$ because $i \neq j$. Accordingly, $\mathcal{O}_m(\mathbf{xy}) \notin C_n(Q_0, \ldots Q_{q-1})$.

If $a \in [n-m, n)$ then in the same way

$$\mathcal{O}_m(\mathbf{xy})[a] = y[a-n+m] = y[a+m] = y[b] = a_j,$$

again ensuring that $\mathcal{O}_m(\mathbf{xy}) \notin C_n(Q_0, \ldots Q_{q-1})$. Therefore $C_n(Q_0, \ldots Q_{q-1})$ is comma-free.

Clague [1] first claimed the equivalence of *(i)* and *(iii)* for the case $q = 2$. He used the word "synchronous" rather than "comma-free". Since

there are other ways to synchronize a code not discussed here, we prefer the more explicit term which is now standard in the literature. Levenshtein [13] generalized Clague's result to finite alphabets.

3 Comma-Free Codes from Difference Systems of Sets

Comma-free codes over finite alphabets were introduced in [8] but without specifying how the codes were to carry information. The first fixed position or "systematic codes" for synchronization were binary codes proposed by E. N. Gilbert [7] in 1960. Gilbert's binary systematic codes were prefix codes and and he found bounds on the number of code words in the codes given the parameters of codeword length and number of fixed entries used to establish synchronization. Later work by Clague [1] showed that binary systematic comma-free codes could be obtained by bipartitions of difference sets and Levenshtein [13] studied a generalization to finite partitions of perfect difference sets which he called *difference systems of sets.* Levenshtein incorporated (*iii*) of Theorem 1 as part of the definition.

Definition 3 *A* difference systems of sets (DSS) *is a collection of disjoint subsets* $Q_i \subseteq \mathbb{Z}_n$, $i, j = 1, \ldots n-1, i \neq j$ *such that for each* $m = 1, \ldots n-1$ *the equation*

$$m \equiv a - b \pmod{n} \tag{5}$$

has at least one solution in integers a, b *from* $\mathbb{Z}_n$ *where* $a \in Q_i, b \in Q_j, i \neq j$.

If $D = \{d_0, \ldots d_{k-1}\}$ is any difference set in $\mathbb{Z}_n$ then there is always the trival DSS where $Q_i = \{d_i\}, i = 0, \ldots k-1$. If $n = 2$ there is only a trivial DSS since $q > 0$. If $n = 3$ then a DSS with, say, $Q_0 = \{a_0\}, Q_1 = \{a_1\}$ is possible with codewords taken from $\{a_0 a_1 u \mid u \in A\}$. If $q = 2$ with $A = \{a_0, a_1\}$ a resulting code would have two codewords, $a_0 a_1 a_0$ and $a_0 a_1 a_1$ and it is easy to see this code is comma-free since the elements of A are assumed to be distinct. For $q > 2$ and $n = 3$, there is a code $C_3(Q_0, Q_1)$ with codewords, $a_0 a_1 a_0, a_0 a_1 a_1, \ldots, a_0 a_1 a_{q-1}$ and it is comma-free. Wang has compiled exhaustive lists of DSS for the ranges $q = 2, 3, 4, n = 7 \ldots 12$ [18].

If for each m there are precisely $\rho > 0$ solutions to (5) then Levenshtein called the DSS a *perfect difference system of sets.* Partitioning known

(v, k, λ)−difference sets, Tonchev [17] has studied perfect regular DSS where regular means all Q_i have the same cardinality. For $n = mq + 1$ he has constructed perfect regular DSS from the trivial cyclic $(n, n-1, n-2)$ difference set. Using difference sets of quadratic-residue type he has shown that for every prime $n = 2mq + 1 \equiv 3 \pmod 4$ there is a perfect regular DSS with parameters (n, m, q) and index $\rho = (n - 2m - 1)/4$. He has also found examples of DSS from Singer difference sets.

For a systematic code $C_n(Q_0, \ldots Q_{q-1})$ with index ρ over a finite field of prime power order as alphabet Levenshtein [13] gave a lower bound for the redundancy $r_q(n, \rho)$ of $C_n(Q_0, \ldots Q_{q-1})$:

$$r_q(n, \rho) \geq \sqrt{\frac{q\rho(n-1)}{q-1}} \tag{6}$$

which generalized the bound given by Clague [1] for $q = 2$. Levenshtein further showed that the lower bound (6) redundancy is attained if and only if the DSS $C_n(Q_0, \ldots Q_{q-1})$ is perfect and regular.

This bound has been further improved by Hao Wang [18].

If $C_n(Q_0, \ldots Q_{q-1})$ is a systematic comma-free code then $\cup Q_i$ is clearly a difference set in $\mathbb{Z}_n$. For systematic DSS codes the definition of comma-freedom takes a more special form:

A code C is *comma-free* if for all $\mathbf{x},\mathbf{y} \in C$ and for all $m, 1 \leq m \leq n-1$ there exists
$a \in \mathbb{Z}_n$ such that $\mathcal{O}_m(\mathbf{xy}[a]) \neq a_j$ and $a \in Q_j$.

As might be expected, not every systematic comma-free code can be obtained by considering perfect DSS. For example it is easy to see that the $(7, 3, 1)$-difference set $\{1, 2, 4\} \in \mathbb{Z}_7$ contains no subsets Q_0, Q_1 that yield a systematic code but its complement, $\{0, 3, 5, 6\}$, is a $(7, 4, 2)$-difference set that can be partitioned as $Q_0 = \{0, 5\}, Q_1 = \{3, 6\}$ so that (iii) in Theorem 1 is satisfied in $\mathbb{Z}_7$. Further, the $(7, 4, 2)$-difference set $\{0, 3, 5, 6\}$, can be partitioned as $Q_0 = \{0\}, Q_1 = \{3, 5, 6\}$. This is the only partition of $\{0, 3, 5, 6\}$ containing a singleton. All 6 possible partitions of $\{0, 3, 5, 6\}$ of the form $\{a, b\}, \{c, d\}$ are DSS, however. S_q on the under (12)(48).

4 A Family of Systematic Circular Comma-Free Codes

The following definition in the context of systematic codes is due to Hao Wang [18].

Definition 4 *Positive integers $n_0 \dots n_{q-1}$ are a q-partition of the integer $r > 0$ if each $n_i > 0$ and $n = \sum_{i=0}^{q-1} n_i$. A q-partition is fair if there do not exist integers $i \neq j$ such that $|n_i - n_j| \geq 2$. By an abuse of language, we call a collection of disjoint non-empty subsets $\{Q_0, \dots Q_{q-1}\}$ of $\mathbb{Z}_n$ a* fair partition *of $\cup_{i=0}^{q-1} Q_i$ if $\{|Q_0|, \dots, |Q_{q-1}|\}$ is a fair partition of $r = \sum_{i=0}^{q-1} |Q_i|$.*

Equivalently the collection $\{|Q_0|, \dots, |Q_{q-1}|\}$ is fair if $||Q_i| - |Q_j|| \leq 1$ when $i \neq j$. For every difference set $D = \{d_0, \dots d_{k-1}\}$ in $\mathbb{Z}_n$, the trival DSS $Q_i = \{d_i\}, i = 0, \dots k-1$ is a fair partition of D.

Theorem 2 *Let n be a prime $p \geq 3$ and Λ an alphabet with $1 < |\Lambda| = q =\leq \lfloor \frac{n}{2} \rfloor$. There exists a non-trivial systematic circular comma-free code $C_n(Q_0, \dots Q_{q-1})$ over Λ whose sets $Q_0, \dots Q_{q-1}$ are a fair partition of $\cup_{i=0}^{q-1} Q_i$.*

Proof: Define a sequence $\mathcal{X} = \{x_i \mid 0 \leq i \leq \lfloor \frac{n}{2} \rfloor\}$ in $\mathbb{Z}_n$ by the elementary recursion

$$\begin{aligned} x_0 &= 0, \\ x_i &= x_{i-1} + i \quad \text{for } i = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor . \end{aligned} \tag{7}$$

It is easy to see that the sequence $\mathcal{X}$ is also given by $x_{i-1} = \frac{i(i-1)}{2}$ for $i = 1, \dots, \lfloor \frac{n}{2} \rfloor$. The recursion, however, shows that (iii) of Theorem 1 will be satisfied by the Q_i we now define. Define $q = |A|$ subsets $Q_t \subset \mathbb{Z}_n$ by

$$Q_t = \{x_i \mid i \equiv t \pmod q\} \quad \text{for } t = 0, \dots q-1. \tag{8}$$

Note that in this definition, the x_i are integers, not integers mod n. Definition (8) simply distributes the elements of $\mathcal{X}$ into congruence classes mod q according to their indices. Necessarily, successive elements x_i and x_{i+1} from $\mathcal{X}$ are in different Q_t.

If n is odd then the differences

$$1 = x_1 - x_0, \ldots, \frac{n-1}{2} = x_{\frac{n-1}{2}} - x_{\frac{n-3}{2}}$$

are distinct integers less than n and hence will be distinct mod n since n is prime.

Since inverses are unique in $\mathbb{Z}_n$ viewed as an additive Abelian group,

$$n-1 = x_0 - x_1, \ldots, -\frac{n-1}{2} = x_{\frac{n-3}{2}} - x_{\frac{n-1}{2}}$$

are also distinct. No entry of the second list can appear in the first list and vice versa. So the two lists are disjoint. Therefore together they are all non-zero elements of $\mathbb{Z}_n$ listed exactly once.

If n is even then $n-1$ is odd and the argument is similiar.

$$1 = x_1 - x_0, \ldots, \frac{n}{2} = x_{\frac{n}{2}} - x_{\frac{n-2}{2}}$$

and

$$n-1 = x_0 - x_1, \ldots, -\frac{n}{2} = x_{\frac{n-2}{2}} - x_{\frac{n}{2}}$$

also lists the $n-1$ non-zero elements of $\mathbb{Z}_n$ except $-\frac{n}{2} = \frac{n}{2}$.

By Theorem 1 (iii) we see $C_n(Q_0, \ldots Q_{q-1})$ is comma-free.

$C_n(Q_0, \ldots Q_{q-1})$ has redundancy $|\cup Q_i| = \lfloor \frac{n}{2} \rfloor$ by construction.

If an arbitrary element of $\mathbb{Z}_n$ is chosen as x_0 then the resulting sequence is a translate of (7) as is easily seen by simply noting that if we choose as x_0' an arbitrary element $r \in \mathbb{Z}_n$ then, in a finite number of steps, $x_t' = r + 1 + \cdots + t$. Hence, $\mathcal{X}' = \mathcal{X} + r$. It follows that $Q_t' = Q_t + r$ for $t = 0, \ldots, q-1$. This also reflects the fact that translates of difference sets are difference sets.

In Theorem (2) the redundancy is $r_q(n,1) = \lfloor \frac{n}{2} \rfloor$. In the following corollary we use the redundancy as a initial parameter. In a trivial systematic code, each Q_i is a singleton and so for any index ρ, $r_q(n,\rho) = q$. If $C_n(Q_0, \ldots Q_{q-1})$ is a non-trivial systematic code then $r_q(n,\rho) > q$.

Corollary 1 *If $\{n_0, \ldots, n_{q-1}\} \subset \{0, \ldots, n-1\}$ is any set of positive integers satisfying $|n_i - n_j| \leq 1$ then there exists a circular systematic comma-free code $C_n(Q_0, \ldots Q_{q-1})$ with $|Q_i| = n_i$.*

Proof: There exists $m > 0$ such that for all i, $n_i = m, m+1$. Reorder $\{n_0, \ldots, n_{q-1}\}$ if necessary so that the $n_i = m+1$ are first and chose the Q_i as defined in (8).

5 Final Remarks

It is not know whether the bound (6) extends to alphabets other than those of prime power order. If (6) does extend then the family of codes exhibited in Theorem 2 is not best possible.

6 Acknowledgement

The author is indebted to Vladimir Tonchev for pointing out his own work and that of V. I. Levenshtein. The author is further indebted to conversations with Hao Wang.

References

[1] D. J. Clague, *New classes of synchronous codes*, IEEE Trans. Electronic Computers **EC-16**(1967), 290–298.

[2] J. Berstel and D. Perrin, Codes circulaires, Combinatorics on Words, Progess and Perspectives, L. J. Cummings, ed., Academic Press, 1983, 133–165.

[3] F. H. C. Crick, J. S. Griffith, and L. E. Orgel, Codes Without Commas, Proc. Nat. AcadSci. (Physics), **43**(1957), 416–421.

[4] L. J. Cummings, Overlaps in binary systematic codes, *Congressus Numerantium* **171**(2004), 33–39.

[5] L. J. Cummings and M. E. Mays, On the parity of the Witt formula, *Congressus Numerantium* **80**(1991), 49–56.

[6] W.L. Eastman, *On the construction of comma-free codes*, IEEE Trans. Information Theory **11**(1965), 263–267.

[7] E. N. Gilbert, *Synchronisation of binary messages*, IRE Trans. Information Theory **IT-6**(1963), 470–477.

[8] S. W. Golomb, B. Gordon, and L. R. Welch, *Comma-free codes*, Canadian J. Math. **10**(1958), 202–209.

[9] W. E. Hartnett,ed, Foundations of Coding Theory, D. Reidel, Boston, 1974.

[10] B. H. Jiggs, *Recent results in comma-free codes*, Canadian J. Math. **15**(1963), 178–187.

[11] V. I. Levenshtein, *Combinatorial problems motivated by comma-free codes*, Journal of Combinatorial Designs, **12**(2004), 184–196.

[12] V. I. Levenshtein, *Bounds for codes ensuring error correction and synchroniation*, Translation from: Problemy Peredachi Informatsii, **5**(1969), 3-13.

[13] V. I Levenshtein, *One method of constructing quasilinear codes providing synchronization in the presence of errors*, Translation from: Problemy Peredachi Informatsii, **7**(1971), 30-40.

[14] V. I. Levenshtein and V. D. Tonchev, Constructions of difference systems of sets, in "Algebraic and Combinatorial Coding Theory", Eight International Workship Proc., St. Petersburg, Russia, Sept. 2002, pp. 194-197.

[15] R. A. Scholtz, *Maximal and variable word- length comma-free codes*, IEEE Trans. Inform. Theory **IT15**(1969), 300-306.

[16] B. Tang, S. W. Golomb, and R. L. Graham, *A New Result on Comma-Free Codes of Even Word Length*, Canadian J. Math. **39**(1987), 513-526.

[17] V. D. Tonchev, *Difference systems of sets and code synchronization*, Rendiconti del Seminario Mathematico di Messina, Series II, **9**(2003), 217–226.

[18] H. Wang, *A New Bound and Exhaustive Algorithm for Difference Systems of Sets*, (to appear in this volume).