NEW RESULTS AND OPEN PROBLEMS IN REAL-NUMBER CODES

Paulo J. S. G. Ferreira, Dorabella M. S. Santos and Vera M. A. Afreixo

Dep. Electrónica e Telecomunicações / IEETA, Universidade de Aveiro
3810-193 Aveiro, Portugal
phone: +351 234 370 525, fax: +351 234 370 545, email: pjf@det.ua.pt
web: www.ieeta.pt/˜pjf

ABSTRACT

Signal processing techniques (such as the FFT and sampling rate conversion) can be used to locate and correct data errors, as an alternative to error control coding. These error control techniques in the real or complex fields are not subject to stringent restrictions on block length, as it is the case in Galois fields, and can be readily performed in DSPs. Furthermore, they are capable of exploring the redundancy that naturally exists in oversampled signals, rather than introducing redundant data in a different domain and in a separate channel-coding step (the finite field approach).

However, the use of finite-precision real arithmetic raises the question of numerical stability. We study the conditioning of an error correction algorithm in the complex field, in an attempt to understand the effect of the location of the errors and the effect of the error amplitudes. We then consider the parallel concatenation of these codes, which seems to lead to surprisingly stable codes, and formulate a number of open problems concerning them.

1. INTRODUCTION

It is possible to estimate and correct errors using standard digital signal processing techniques. Such error control codes defined over the real or complex field have been recognized as advantageous [1–5]. Some of these works were stimulated by Blahut's [6], who casted several error control codes in the terminology of the Fourier transform over a finite Galois field, and made it more accessible to signal processing engineeers, well acquainted with spectral techniques. Sampling with unknown locations [7] is an interesting related problem.

As opposed to conventional error-correcting codes defined over the binary or other finite fields, which are usually implemented using special hardware, the codes over the real or complex fields, to which we will collectively refer as "real number codes", lend themselves to implementation with standard digital signal processors.

The error correction procedures based upon the discrete Fourier transform (DFT) lead to codes that allow efficient techniques for combined source and channel coding using standard digital signal processing techniques. They require only standard arithmetic units, and can be conveniently implemented in software, using any standard high-level programming language. They allow complete freedom in the size of encoder input and output block lengths, usually denoted by K and N, respectively. Maximum distance separable real number (N,K) codes exist for all nontrivial N and K [2], whereas in finite field codes the choice of these parameters is severely restricted.

Partially supported by the FCT.

But real number codes also have disadvantages. The estimation procedure will be perturbed by the presence of background and round-off noise, and there is evidence of noise sensitivity in this context [8]. Burst errors are notoriously difficult to handle, and serious problems may occur for block sizes as small as 32 or 64 [9]. For the analysis of low-pass DFT codes under bursty erasures see [10]. Bounds for the condition number or eigenvalues of the linear operators involved in the erasure decoding process are of great interest to understand their numerical stability (see [11, 12], for example).

The numerical stability of the error detection and correction problems depends critically on the distribution and magnitude of the errors. In this paper we show that the condition number of the problem ranges over an extremely wide range and critically depends on the error pattern (although it is invariant under cyclic shifts). We examine the effect of the amplitudes and distribution of the errors on the stability. Theoretical results are given that effectively decouple these two factors, and show how the relative magnitude of the errors and their position influences the conditioning.

2. LOCATING THE ERRORS

Consider a signal x with N samples and let $y = x + e$, where e (the error signal) is nonzero at the the locations

$$U := \{i_0, i_1, \ldots i_{m-1}\}.$$

The set U will also be called the error pattern. Thus, $e(k) = 0$ or $y(k) = x(k)$ for all $k \notin U$. Typically, the cardinal of U is much less than N, that is, the density m/N of the incorrect samples is small.

Let F be a general linear transform and assume that some components of the transformed vector Fx are zero, that is, $(Fe)(k) = 0$ when $k \in S$, S being a certain set of integers. Then, because $(Fe)(k) = (Fy)(k)$, the restriction of Fe to S depends only on the observed data. Under certain conditions, this knowledge about the transform of the error signal can be used to estimate the error signal itself. Subtracting e from the observed data yields the original, errorless data vector x.

We will take the general linear transform F to be the DFT from now on (thus, one can think of F as the Fourier matrix). A well-known technique similar to Prony's method [13] can then be used to approach the error-correction problem. One introduces the error-locating polynomial

$$P(z) := \sum_{k=0}^{m} h_k z^k,$$

which is determined by the conditions $h_m = 1$ and

$$P\left(e^{-\mathrm{i}\frac{2\pi}{N} i_p}\right) = 0, \quad (0 \le p < m). \tag{1}$$

The error pattern can be found from the zeros of the polynomial $P(z)$, since its coefficients can be expressed as functions of the observed data. The key equation is

$$Th = b.$$

Assuming for simplicity that N is even, and setting $r = N/2$, the elements of the matrix T are

$$T_{ab} := \hat{e}(r+b-a) \quad (0 \le a < m, \quad 0 \le b < m). \tag{2}$$

where $\hat{e} = Fe$ is the discrete Fourier transform of e. Noting that

$$\sum_{k=0}^{m-1} h_k \hat{e}(r+k-\ell) = -\hat{e}(r+m-\ell), \quad (0 \le \ell < m),$$

one sees that $2m$ samples of $\hat{e}$ need to be known,

$$\hat{e}(r-m+1), \hat{e}(r-m+2), \ldots, \hat{e}(r+m).$$

In the case considered, the matrix T is square, $m \times m$, and Toeplitz. A number of methods can be used to solve the linear equations, including singular value decomposition, or Levinson's iteration. In practice, m is unknown. However, if only $k < m$ errors ocurred, the principal submatrices of T of order greater than k will be singular. This can be detected, say, by the basic Levinson recursion, and the algorithm stopped.

To locate the errors one must find the zeros of the polynomial $P(z)$, the coefficients of which are $\{h_k\}_{k=0}^{m-1}$. Equation (1) shows that the zeros can be found by padding

$$h_0, h_1, \ldots, h_{m-1}, 1$$

with $N-(m+1)$ zeros to form a new Nth-dimensional vector

$$g := [h_0, \ldots, h_{m-1}, 1, \underbrace{0, \ldots, 0}_{N-(m+1)}]^T \tag{3}$$

and evaluating the FFT of the N data obtained:

$$\hat{g}(q) := \sum_{p=0}^{N-1} g(p) e^{-\mathrm{i}\frac{2\pi}{N}pq}.$$

The kth sample of the observed signal $y(k)$ is considered in error if and only if the kth coefficient of this FFT is zero, that is, if $\hat{g}(k) = 0$.

Ideally, one expects that $\hat{g}(k) = 0$ if and only if $k \in U$, where $U = \{i_0, i_1, \ldots, i_{m-1}\}$. But in practice the finite precision of the computations and the round-off errors lead to possibly small, but nonzero, $\hat{g}(i_p)$, and the importance of the numerical stability issues is once again brought to light.

3. STABILITY ANALYSIS

The stability of the problem depends on the Toeplitz matrix T, defined by (2). Its elements can be written as

$$T_{ab} = \frac{1}{\sqrt{N}} \sum_{p=0}^{m-1} e(i_p) (-1)^{i_p} e^{-\mathrm{i}\frac{2\pi}{N} i_p (a-b)}. \tag{4}$$

Recall that for simplicity we have assumed that N is even, and to simplify the notation we wrote $r = N/2$.

Proposition 1 *Let N be even and $r = N/2$. If the error signal is real, T and any principal submatrix of T will be Hermitian.*

This follows from (4), or from the properties of the DFT: if e is real,

$$\hat{e}(r+i) = \hat{e}^*(-r-i) = \hat{e}^*(N-r-i) = \hat{e}^*(r-i),$$

and (2) shows that $T_{ab} = T_{ba}^*$.

The following is a simple but useful fact: if the positions of the errors (the error pattern) are cyclically shifted, the conditioning of T remains unchanged. Here we tackle the general case: the number of errors m is not necessarily equal to the maximum number of errors, n, and T is $n \times m$.

Proposition 2 *The condition number of the $n \times m$ matrix T is invariant under cyclic shifts of $U = \{i_0, i_1, \ldots, i_{m-1}\}$.*

To simplify the notation, let $S := T^*T$. We write $S(U)$, to stress the dependence of S on the set U. Thus,

$$S_{ab}(U) = \sum_{p=0}^{n-1} \hat{e}^*(r+a-p)\hat{e}(r+b-p).$$

A cyclic shift of the error pattern U changes i_p into $(i_p + c)$ mod N. The DFT of the error signal changes from $\hat{e}(k)$ to

$$\hat{e}(k) e^{-\mathrm{i}\frac{2\pi}{N}kc}.$$

Let $U + c$ denote the set U cyclically shifted by c. Then,

$$S_{ab}(U+c) = e^{\mathrm{i}\frac{2\pi}{N}c(a-b)} \sum_{p=0}^{n-1} \hat{e}^*(r+a-p)\hat{e}(r+b-p).$$

The quadratic form associated with $S(U+c)$ can be written as

$$\begin{aligned} v^*S(U+c)v = \\ &= \sum_{a,b=0}^{m-1} v(a)v^*(b) e^{\mathrm{i}\frac{2\pi}{N}c(a-b)} \sum_{p=0}^{n-1} \hat{e}^*(r+a-p)\hat{e}(r+b-p) \\ &= \sum_{a,b=0}^{m-1} u(a)u^*(b) \sum_{p=0}^{n-1} \hat{e}^*(r+a-p)\hat{e}(r+b-p), \end{aligned}$$

where

$$u(a) := v(a) e^{\mathrm{i}\frac{2\pi}{N}ca}.$$

The Euclidean norms of u and v are clearly identical, and so one can take the maximum or minimum of the quadratic form with respect to u or v, subject to $\|v\| = \|u\| = 1$, without affecting the result. But these maximum and minimum values of $v^*T^*Tv = \|Tv\|^2$ are the extreme singular values of T, which are therefore unchanged by a cyclic shift of U, completing the proof.

It is convenient to have an idea of the extent to which the position of the errors $U = \{i_0, i_1, \ldots i_{m-1}\}$ determines the condition number $\kappa(T)$ of the $m \times m$ matrix T. This condition number is plotted in Fig. 1 as a function of the error pattern U. We considered $m = 5$ errors of unitary magnitude, and a block size of $N = 20$. The number of possible error patterns satisfying $i_0 = 0$ is 3876, and the patterns were sorted by lexicographic order. Despite the unrealistic small N and

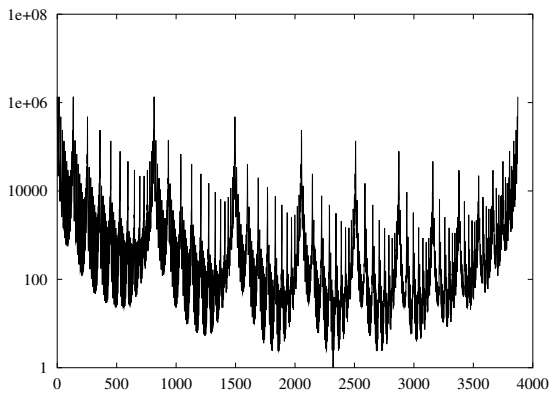

Figure 1: The condition number $\kappa(T)$ as a function of the distribution of $m = 5$ errors of unitary magnitude, subject to $i_0 = 0$, for a blocksize $N = 20$. The 3876 possible error patterns are sorted by lexicographic order.

m, the condition number already ranges over several orders of magnitude.

The constraint $i_0 = 0$ reduces the number of error patterns that have to be tested by a factor of m/N (from C_m^N down to C_{m-1}^{N-1}, where C_j^i is the binomial coefficient). This entails no loss of generality, since, by proposition 2, the conditioning of T is invariant under circular shifts of U.

For larger values of N and m, the condition number varies over an extremely wide range, and the numerical difficulty of the problem will range from "very easy" to "extremely difficult", depending on the relative error amplitudes and pattern. An in-depth understanding of the numerical stability of the problem and conditioning of T is crucial for practical applications.

An interesting fact concerning T is that it can be positive definite, negative definite, or indefinite, depending on the sign of the $e(i_k)$ and the parity of i_k.

Proposition 3 *The $m \times m$ matrix T is (a) positive definite if and only if $e(i_k)(-1)^{i_k} > 0$ for all $0 \le k < m$, (b) negative definite if and only if $e(i_k)(-1)^{i_k} < 0$ for all $0 \le k < m$ (c) otherwise, it is indefinite.*

The proof depends on (4), which can be used to show that

$$v^*Tv = \frac{1}{\sqrt{N}} \sum_{p=0}^{m-1} e(i_p)\,(-1)^{i_p} \left| \sum_{a=0}^{m-1} v(a) e^{\mathrm{i}\frac{2\pi}{N} i_p a} \right|^2,$$

leading to the proposition.

We need to introduce two more matrices. The first is the diagonal $m \times m$ matrix W, whose main diagonal elements are

$$W_{kk} := \sqrt{N}\, e(i_k)(-1)^{i_k}. \tag{5}$$

Assume, without loss of generality, that W is nonsingular (if $e(i_k) = 0$ then i_k can be deleted from U). We also need the following $m \times m$ matrix:

$$B_{pq} := \frac{1}{\sqrt{N}} e^{\mathrm{i}\frac{2\pi}{N} i_p q}, \quad (0 \le p, q < m). \tag{6}$$

This is a nonsingular Vandermonde matrix. It turns out that W and T are star-congruent (see [14] for a discussion of congruence).

Proposition 4 *The $m \times m$ matrix T is star-congruent to the diagonal matrix W, that is,*

$$T = B^*WB,$$

where B and W are given by (6) and (5).

The proof is routine: compute B^*WB and verify the equality. We now have the following proposition.

Proposition 5 *The eigenvalues of the $m \times m$ matrix T and those of the diagonal matrix W satisfy*

$$\lambda_{\min}(B^*B)\lambda_k(W) \le \lambda_k(T) \le \lambda_{\max}(B^*B)\lambda_k(W).$$

The eigenvalues are labeled according to nondecreasing size.

The proof depends on a theorem of Ostrowski [14], that asserts that for each $k = 1, 2, \ldots, m$ there exists a positive real number θ_k such that

$$\lambda_1(B^*B) \le \theta_k \le \lambda_m(B^*B)$$

and

$$\lambda_k(B^*WB) = \theta_k \lambda_k(W).$$

It can be shown that this and $T = B^*WB$ yields the result.

The condition number is submultiplicative, and so $T = B^*WB$ implies $\kappa(T) \le \kappa(B^*)\kappa(W)\kappa(B) = \kappa(B)^2\kappa(W)$. However, a stronger result holds, namely, $\kappa(T) \le \kappa(B^*B)\kappa(W)$ (stronger because $\kappa(B^*B) \le \kappa(B)^2$).

Proposition 6 *The condition number of T satisfies*

$$\kappa(T) \le \kappa(W)\kappa(B^*B) = \frac{e_{\max}}{e_{\min}} \kappa(B^*B),$$

where

$$e_{\max} = \max_k |e(i_k)|, \quad e_{\min} = \min_k |e(i_k)|$$

are the largest and smallest errors in absolute value.

The proof is omitted, due to space constraints. Note, however, the following closing remarks:

1. If all the errors have the same absolute value a, $\kappa(T)$ will be equal to $\kappa(B^*B)$ provided that $e(i_p) = a(-1)^{i_p}$. Equation (5) shows that $W = aI$, and so $\kappa(W) = 1$. Thus, the upper bound in proposition 6 can be met.
2. Note how the dynamic range of the errors enters the problem. For finite precision representations of the real numbers (fixed point or floating point) $e_{\max}/e_{\min}$ can be a very large number, increasing the chances of ill-conditioning.
3. We have reduced the study of the stability of T to two separate problems: the conditioning of W, which is essentially determined by the magnitude of the errors, but is independent of their position; and the conditioning of B^*B, which depends entirely on the error positions, but not on their magnitude.
4. The (Hermitian) matrices $P = B^*B$ and $S = BB^*$ have the same (real) eigenvalues. They are related to the matrices that occur in the error correction or interpolation problem (in which the positions $U = \{i_0, \ldots i_{m-1}\}$ of the unknown samples are known; this is also known as "erasure correction").
5. The matrix P occurs in the frequency-domain formulation of the error correction / interpolation problem, whereas S occurs in the time-domain formulation (compare with [15]).

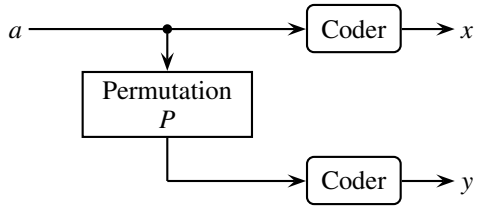

Figure 2: Parallel concatenation of two DFT codes with an interleaver.

4. CONCATENATION AND OPEN PROBLEMS

A structure less sensitive to variations in the error pattern cannot be obtained simply by interleaving the coded data. The interleaver is a one-to-one mapping. Therefore, there will always exist error patterns that it will map into contiguous patterns (consider the inverse image under the interleaver of a contiguous pattern).

Two data paths, one in which the data are coded normally, and another in which they are interleaved, as in Fig. 2, might lead to more uniform performance with respect to error patterns. In fact, the concatenated DFT code depicted in Fig. 2 was discussed recently in [11], and appears to have a good potential for handling error bursts without numerical problems.

One of the open issues regarding it concerns the role of the permutation. Fig. 3 shows the condition number of the linear operator that arises when decoding contiguous erasures using the two-channel structure, for 100 randomly selected permutations. The same figure also shows the condition number for the one-channel DFT code of the same rate. The concatenated code clearly outperforms the one-channel code of the same redundancy. We do not know how to characterize the "best" permutations and the "worst" nontrival ones (those that lead to condition numbers close to the one-channel case). As far as we know, the average performance (with respect to the set of possible permutations) is also unknown.

There are several other open problems regarding the two-channel structure. For a fixed permutation P, the numerical performance when decoding bursty erasures or errors seems to be much better than for the one-channel DFT code of the same rate. In fact, the condition numbers can differ by 10 orders of magnitude or more (see the results in [11]). But are there specific non-bursty error patterns that lead to poor condition numbers? How can such error patterns be characterized? Are they rare? How do they depend on the permutation?

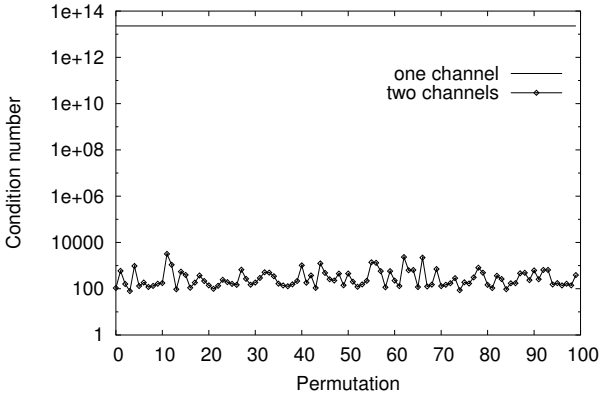

Figure 3: The figure shows the condition number of the two-channel problem (for 100 random permutations) and that of the equivalent one-channel problem.

REFERENCES

[1] J. K. Wolf. Redundancy, the discrete Fourier transform, and impulse noise cancellation. *IEEE Trans. Commun.*, 31(3):458–461, Mar. 1983.

[2] T. G. Marshall, Jr. Coding of real-number sequences for error correction: A digital signal processing problem. *IEEE J. Select. Areas Commun.*, 2(2):381–391, Mar. 1984.

[3] T. G. Marshall, Jr. Codes for error correction based upon interpolation of real-number sequences. In *Proceedings of the 19th Asilomar Conf. Circuits Syst.*, pages 202–206, Pacific Grove, CA, Nov. 1986.

[4] F. A. Marvasti and M. Nafie. Sampling theorem: A unified outlook on information theory, block and convolutional codes. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, E76–A(9):1383–1391, Sep. 1993.

[5] J.-L. Wu and J. Shiu. Discrete cosine transform in error control coding. *IEEE Trans. Commun.*, 43(5):1857–1861, May 1995.

[6] R. E. Blahut. *Theory and Practice of Error Control Codes*. Addison-Wesley, Reading, MA, 1983.

[7] P. Marziliano and M. Vetterli. Reconstruction of irregularly sampled discrete-time bandlimited signals with unknown sampling locations. *IEEE Trans. Signal Processing*, 48(12):3462–3471, Dec. 2000.

[8] C. K. W. Wong, F. Marvasti, and W. G. Chambers. Implementation of recovery of speech with impulsive noise on a DSP chip. *Electron. Letters*, 31(17):1412–1413, Aug. 1995.

[9] F. Marvasti, M. Hasan, M. Echhart, and S. Talebi. Efficient algorithms for burst error recovery using FFT and other transform kernels. *IEEE Trans. Signal Processing*, 47(4):1065–1075, Apr. 1999.

[10] G. Rath and C. Guillemot. Performance analysis and recursive syndrome decoding of DFT codes for bursty erasure recovery. *IEEE Trans. Signal Processing*, 51(5):1335–1350, May 2003.

[11] P. J. S. G. Ferreira and J. M. N. Vieira. Stable DFT codes and frames. *IEEE Sig. Proc. Letters*, 10(2):50–53, Feb. 2003.

[12] P. J. S. G. Ferreira. Mathematics for multimedia signal processing II — discrete finite frames and signal reconstruction. In J. S. Byrnes, editor, *Signal Processing for Multimedia*, pages 35–54. IOS Press, 1999.

[13] S. M. Kay and S. L. Marple. Spectrum analysis — a modern perspective. *Proc. IEEE*, 69(11):1380–1419, Nov. 1981.

[14] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1990.

[15] P. J. S. G. Ferreira. Interpolation in the time and frequency domains. *IEEE Sig. Proc. Letters*, 3(6):176–178, June 1996.