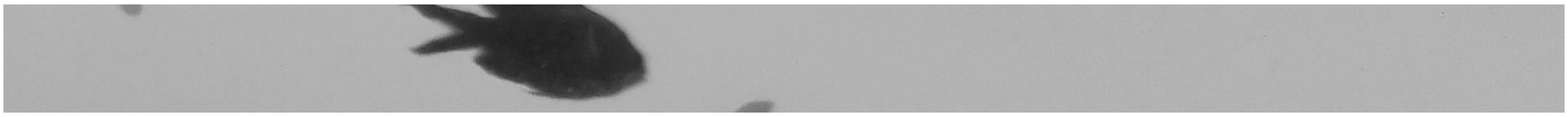

ID17- REAL-TIME LOSSLESS COMPRESSION OF MULTIBEAM ECHOSOUNDER WATER COLUMN DATA

DAVID AMBLAS[183], JORDI PORTELL[40], XAVIER RAYO[72], ALBERTO G. VILLAFRANCA[1], ENRIQUE GARCÍA-BERRO[48], MIQUEL CANALS[71]

Abstract - Multibeam echosounders can generate vast amounts of data when recording the complete water column, which poses logistic, economic and technical challenges. Lossy data compression can reduce data size up to one or two orders of magnitude, but often at the expense of significant image distortion. Lossless compression ratios tend to be modest and at a high computing cost. In this work we test a high-performance data compression algorithm, FAPEC, initially developed for Space data communications with low computing requirements. FAPEC provides good compression ratios and supports tailored pre-processing stages. Here we show its advantages over standard and high-end lossless compression solutions currently available, both in terms of ratios and speed.

Keywords - Water column, multibeam, lossless, compression, FAPEC.

I. INTRODUCTION

In the last years remarkable advances in sonar technology, spatial positioning and computation power have led to significant improvements in mapping, imaging and monitoring of the oceans. Multibeam echosounders are now capable of collecting backscatter data for the whole water column, in addition to the traditional measures of bathymetry and seafloor reflectivity. These new data sets open up a new range of applications for multibeam sonars, including direct imaging of fish and marine mammals, mapping of suspensate plumes, location of mid-water targets, proper determination of sunken structures such as shipwrecks and investigating a wide range of physical oceanographic processes [1].

Despite the potential value of water column reflectivity measurements, the enormous increase in data collection often makes the media storage requirements prohibitive, which forces many users to opt for not recording systematically water column information. The volume of data generated in multibeam water column surveys can easily be one order of magnitude larger than in conventional bottom detection assessments, especially in shallow water where the higher ping rates lead to data rates of several gigabytes per hour [2]. This complicates the efficient browsing, querying, sharing and transfer of data. It also limits the design of Autonomous Underwater Vehicles (AUVs) equipped with multibeam echosounders and remote assistance during sea works by technical support teams, which often rely on expensive satellite links.

Data compression provides a potential solution to this challenging issue. However, few published studies face this question, and most of them use lossy methods involving a certain degree of signal distortion and water column imagery degradation [2,3,4]. Even fewer sonar-dedicated lossless compression algorithms have been proposed [5,6], and commonly used lossless techniques, such as ZIP, RAR or 7z, yield only modest compression rates at very slow speeds. Here we evaluate these data compression tools and their performance on multibeam water column data. We also provide the results obtained with a lossless compression algorithm called FAPEC (Fully Adapted Prediction Error Coder) [7], initially designed to meet the tight requirements of satellite payloads and deep space communications. More precisely, it is inherited from a Technology Research Programme of the European Space Agency for Gaia, an ambitious space astrometry mission which aims at measuring the position, distances and motions of more than one billion stars with unprecedented accuracy [8]. The large amount of data and its complex data model required a tailored and extremely optimized solution. In [9], an initial solution based on stream partitioning and standard compressors was proposed, but finally a solution based on Rice-Golomb codes [10] was adopted and improved with an outlier-resilient algorithm, which is at the core of FAPEC [11]. In this work we take advantage of the lessons learned in Space research to adapt the FAPEC algorithm to multibeam water column data.

II. WATER COLUMN DATA

Multibeam echosounder raw records are usually logged as binary files using signed or unsigned integer numbers. Each sonar manufacturer has a specific file format, which in turn can vary depending on the particular sonar model, the survey purpose, its configuration and the external sensors included such as CTD probe, GPS, Compass and Gyro. Each file usually contains time-stamped information about beam geometry, sonar configuration, navigation, attitude, sound speed, bathymetry and water column backscatter measurements. Most of the data volume comes from backscatter raw samples, composing the image as a two-dimensional array (number of beams times the number of samples for that beam). Each element is typically an 8, 16 or 32 bit integer value. In our case we have signed 8-bit samples. Note that array dimensions are not necessarily uniform throughout the data file, which complicates its handling by image compression algorithms.

In this study we analyse water column data acquired with a Kongsberg EM710 multibeam echosounder. The EM710 operates at sonar frequencies in the 70-100 kHz range with a maximum ping rate of 30 Hz, up to 400 soundings per swath and a coverage that can reach 140º.

III. THE FAPEC DATA COMPRESSOR

FAPEC is a highly-optimized entropy coding algorithm which offers outstanding resiliency in front of outliers in the data [7,11]. That is, FAPEC reaches good compression ratios even on data severely contaminated by noise and values outside the typical statistics. In general, its compression efficiency is typically above 90%, which means that it barely introduces any redundancy with respect to the actual data entropy.

This highly configurable system follows a typical two-stage approach. The first stage, or decorrelator, reduces the original entropy of the data by applying a reversible (lossless) or partially-reversible (lossy) algorithm. It can be as simple as a delta stage, outputting differences between consecutive samples. Linear filters can also be used, as well as interleaving for samples following some given patterns. Elaborated stages are also available, such as pattern recognition, multiband prediction, or image compression algorithms. Some of these stages support lossy compression, but in this work we focus on the lossless case. If needed, a tailored pre-processing stage can be implemented and easily integrated into FAPEC.

The output of the first stage is a sequence of signed integers (the prediction errors). The second stage (the entropy coder in itself) generates short binary codes for more frequent values, and slightly longer codes for less frequent values. It includes mechanisms for the efficient compression of sequences with repeated values, thus going beyond the simple entropy coding.

FAPEC, implemented in highly optimized ANSI C, is available as an executable program for Linux, Windows or Mac, and as a dynamic library with a simple Application Process Interface (API) for better integration with data handling systems. A hardware prototype in VHDL language is also available for programmable electronic chips (FPGAs). FAPEC can natively handle sample sizes of 8 to 24 bits, and arbitrarily large samples by means of interleaving. Its compression performance is excellent especially on samples of 16, 32 or 64 bits. FAPEC also enforces data integrity, minimizing data loss in case of file or data transfer corruption.

IV. TEST SETUP

The two data sets used to test the lossless compression algorithms were acquired with a Kongsberg EM710 multibeam survey in the outer continental shelf southeast of Spain. Both datasets correspond to a relatively smooth and homogeneous bathymetry, the second one including a shipwreck (Fig. 1). Table 1 summarizes the data files used for these tests and their main characteristics.

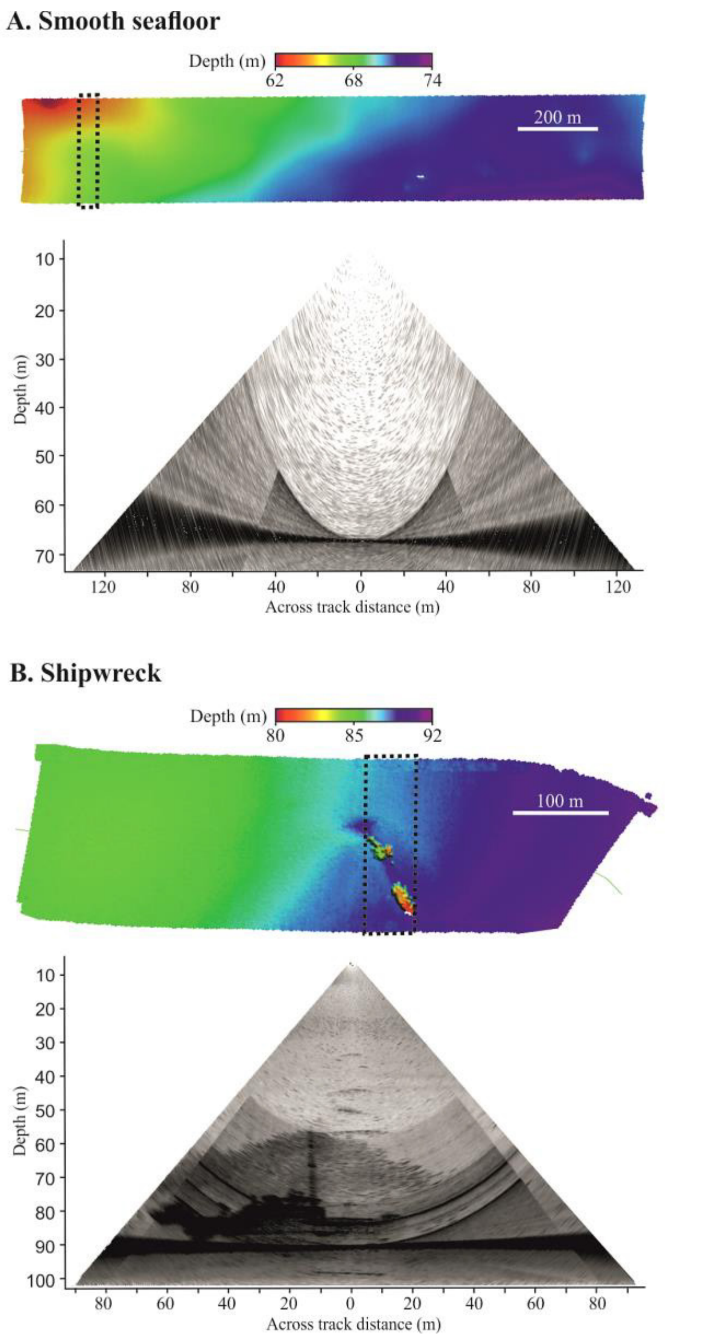

Fig. 1. Swath bathymetry and water column data of the two test sites. The black dashed frame shows the water column corridor displayed.

File	0001_20160519_ 193149_AA.wcd	0044_20140814_ 213726_AA.wcd
Size	339.5MB	149.3MB
Acquisition time	5 min	2 min 38 s
Acquisition rate	1.33MB/s	0.95 MB/s
Case	Smooth seafloor	Shipwreck
Depth range (m)	62-74	80 92

Table 1: Water column data files used in the tests.

For the data compression tests we selected the well-known gzip, bzip2, rar and 7z programs, testing the fast and best options for each of them. We forced a single-thread operation in rar to obtain better comparable results. In the case of FAPEC, we used release 2016.0 Core with anautomatic configuration, leading to an 8-bit delta pre-processing, as otherwise expected for our case. We have forced a single-thread operation as well, as FAPEC also supports multi-threaded compression. All tests were run on a laptop with Intel Core i7 2640M 2.8GHz processor running 64-bit Gentoo Linux. Only the User time (that is, the CPU time) has been taken into account to ignore the effect of disk I/O, and to account the actual time accumulated by all threads in the 7z case.

Tests were done directly on the Kongsberg wcd file (water column datagram). In order to evaluate a potential improvement, we did a test with FAPEC using a prototype pre-processing stage tailored for this kind of files. Specifically, we "transposed" most of the raw samples in order to enhance their correlation. Also, samples were compressed separately from ancillary information and attributes (such as time stamping and beam information). This approach is very similar to the one initially proposed for the Gaia space mission [9].

V. TEST RESULTS

Table 2 shows the lossless compression ratios obtained with these tests (that is, the original file size divided by the compressed file size). As predicted by the Information Theory, the shipwreck file is more difficult to compress because it contains more information, that is, the intrinsic entropy of its raw samples is higher. The best "standard" compressor regarding ratios is 7z with its ultra configuration, although bzip2 offers similar results. FAPEC, in its standard and automatic configuration, offers modest ratios of 1.43 and 1.30 for the two tested files, which is anyway better than what gzip can offer, and very similar to what rar can achieve. However, when combined with a simple pre-processing stage tailored for this kind of files, the overall compression ratios boost to 1.8 and 1.65 respectively (12% and 17% better than 7z in its ultra configuration). Specifically, 7z reduces the file sizes down to 248MB and 106MB respectively, whereas our tailored FAPEC reduces them to 221MB and 90MB. It is worth noting that the FAPEC tailoring is just a prototype. Further improvements should lead to even better ratios.

Regarding compression times, Table 3 shows the throughputs (or speeds) achieved during these tests, calculated as the original file size divided by the CPU time, shown in megabytes per second. Depending on the disk access times, the actual speeds may decrease very slightly. Here differences between the tested compressors become evident. Both gzip and rar, in their fast configuration, provide reasonably good speeds around 30MB/s, but performance drastically decreases when setting them to best. Bzip2 offers a quite uniform and reasonable performance. When considering both the ratio and speed, bzip2 appears as an interesting solution. On the other hand, 7z is quite slow even in its fast configuration, but its ultra option is even slower than the actual data acquisition rate, which is about 1.0MB/s to 1.3MB/s (Table 1). Finally, FAPEC clearly provides the fastest solution, with the tailored version slightly faster than the standard one. This is one of the features of FAPEC, namely, its compression speed is better when data is more compressible. To better illustrate overall performances, Figs. 2 and 3 plot the ratios and throughputs, respectively.

File		Smooth seafloor	Shipwreck
Gzip	--fast	1.37	1.24
	--best	1.39	1.25
Bzip2	--fast	1.55	1.38
	--best	1.58	1.40
Rar	-m1	1.38	1.25
	-m5	1.44	1.35
7z	-mx1	1.52	1.34
	-mx9	1.61	1.41
FAPEC	Standard	1.43	1.30
	Tailored	1.80	1.65

Table 2: Compression ratios obtained on the test files

File		Smooth seafloor	Shipwreck
Gzip	--fast	34.7	31.1
	--best	5.9	16.8
Bzip2	--fast	11.9	10.7
	--best	11.1	9.9
Rar	-m1	29.6	26.2
	-m5	5.4	10.6
7z	-mx1	9.7	9.0
	-mx9	1.1	1.2
FAPEC	Standard	71.3	67.9
	Tailored	79.9	74.6

Table 3: Compression throughputs (MB/s) obtained on the test files

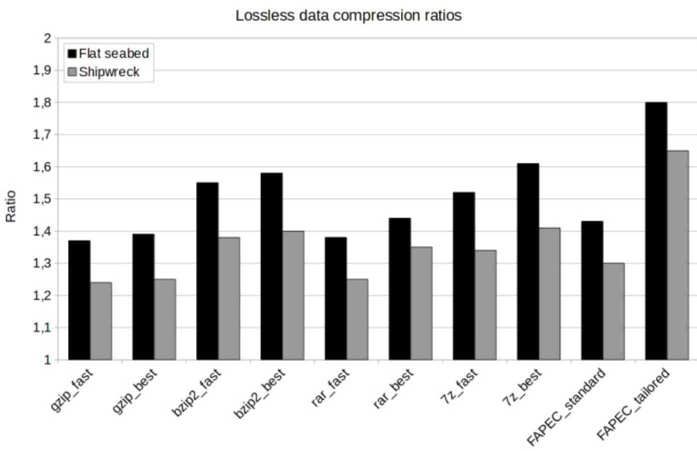

Fig. 2. Lossless compression ratios obtained in the tests.

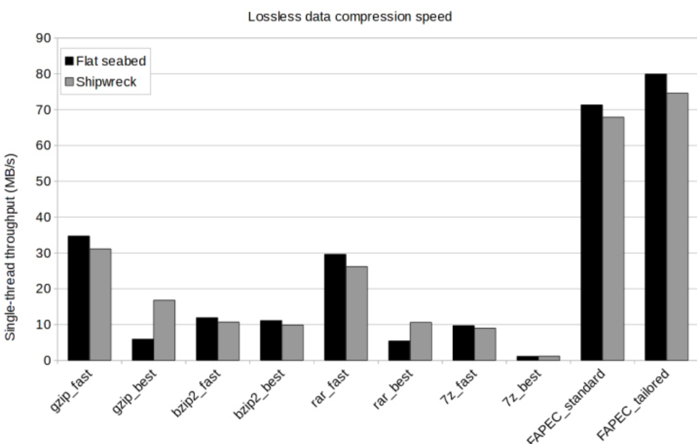

Fig. 3. Single-thread data compression speeds achieved during the tests.

Besides compression speeds, the memory requirements of the several compressors are also relevant. Both gzip and bzip2 have very modest requirements (4MB to 11MB). FAPEC needs also are about 8MB of RAM. Nevertheless, rar raises this requirement to 68MB or 97MB, depending on its configuration. The worst case is 7z, which uses just 27MB in its fast configuration, but its ultra configuration raises this requirement up to 700MB.

VI. CONCLUSIONS AND FORTHCOMING WORK

In this preliminary study we have evaluated the lossless data compression performance of a variety of algorithms on multibeam echosounder water column data. One of these solutions, FAPEC, is a high-performance algorithm initially devised for satellite payloads, which makes it very lightweight and therefore applicable to compression of large datasets or massive data handling systems. Lossless compression tests on two water column data files have assessed the very high speed of FAPEC, making it suitable for real-time compression and decompression. For example, it could be applied as a transparent compression and decompression system, allowing directly processing and handling compressed datasets on-the-fly instead of having to decompress them beforehand.

Regarding compression ratios, the default (differential) pre-processing stage of FAPEC is able to provide competitive results, and when tailoring it with an adequate pre-processing algorithm the ratios become better than any other solution tested here. Therefore, this is a lossless data compressor with the highest ratios and speeds on water column data. This is an especially remarkable result considering that compression is done on relatively small chunks of data. Specifically, each datagram is handled independently, and compression in itself is done in chunks of just 1MB. Therefore, in case of file corruption, most of the original file can still be recovered. This error resilience is one of the features inherited from the Space data communications field.

While excellent by themselves, these FAPEC results should be considered preliminary. We are working on further improvements of the tailored pre-processing stage to take better advantage of the intrinsic data redundancy, not only within a beam or between adjacent beams in a datagram, but also between adjacent swaths. It should also be noted that the ratios obtained here are relatively low because of the 8-bit sample format. FAPEC performs even better (both in ratios and speeds) when dealing with samples of 16 bits or more. If necessary, FAPEC allows introducing some level of losses to reduce further the compressed files, which may be interesting in some scenarios. Finally, further tests will be done to include analysis in different depth settings and seafloor characteristics, and using different echosounder models.

ACKNOWLEDGEMENTS

R+D work on FAPEC is supported by the ESA Business Incubation Programme through Barcelona Activa, by the MINECO (Spanish Ministry of Economy) – FEDER through grants ESP2014-55996-C2-1-R, AYA2014-59084-P and MDM-2014-0369 of ICCUB (Unidad de Excelencia 'María de Maeztu'), and by the AGAUR. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 658358 (D. Amblas). The authors acknowledge funding received from the Spanish RTD grant NUREIEV (CTM2013-44598-R) and from EC contract MIDAS (GA-603418). GRC Geociencies Marines is recognized by Generalitat de Catalunya as an excellence research group (ref. 2014 SGR 1068). We especially thank M. Morales (Simrad Spain SLU) for valuable discussions during the progress of this study. We also thank the NUREIEV cruise shipboard party and the RV Angeles Alvariño crew, as well as the colleagues of the R3.

REFERENCES

[1] Colbo, K., Ross, T., Brown, C., Weber, T., A review of oceanographic applications of water column data from multibeam echosounders. Estuarine, Coastal and Shelf Science 145, 41-56, 2014. [2] Beaudoin, J., Application of JPEG 2000 Wavelet Compression to Multibeam Echosounder Mid-water Acoustic Refectivity Measurements. Center for Coastal and Ocean Mapping, paper 779, 2010. [3] Chybicki, A., Moszyński, M., Poćwiardowski, P., Applications of compression techniques for reducing the size of multibeam sonar records. Proceedings of the 2008 1st International Conference on Information Technology, pp. 1-4, 2008. [4] Chybicki, A., Łubniewski, Z., Moszyński, M., Using wavelet techniques for multibeam sonar bathymetry data compression. Hydroacoustics 13, 31-38, 2010. [5] Wu, L., Zielinski, A., Bird, J., Lossless Compression of Hydroacoustic Image Data. IEEE Journal of Oceanic Engineering 22 (1), 93-101, 1997. [6] Moszynski, M., Chybicki, A., Kulawiak, M., Lubniewski, Z., A novel method for archiving multibeam sonar data with emphasis on efficient record size reduction and storage. Polish Maritime Research 20(1), 77-86, 2013. [7] Portell, J., Villafranca, A. G., and García-Berro, E., Quick outlier-resilient entropy coder for space missions. Journal of Applied Remote Sensing 4, 339–363, 2010. [8] Perryman, M. A. C., de Boer, K. S., et al., A&A 369, 339, 2001. [9] Portell, J. García-Berro, E., Luri, X., Villafranca, A.G., Tailored data compression using stream partitioning and prediction: application to Gaia. Experimental Astronomy 21, 125–149, 2006. [10] R. F. Rice, Some practical universal noiseless coding techniques. JPL Technical Report 763 79–22, 1979. [11] Portell, J., Villafranca, A.G., García-Berro, E., Outlier-resilient entropy coding, in Satellite data compression, Springer, ed.: B. Huang, ISBN 978-1-4614-1183-3, 2011.