УДК 681.3.01

СИБИРСКИЙ СОЛНЕЧНЫЙ РАДИОТЕЛЕСКОП: ФОРМАТ РСА КОМПРЕССОРА ДАННЫХ ССРТ

А. В. Джурик[1], **П. Г. Жилин**[1]

Дан краткий обзор современного состояния технологии сжатия информации без потерь. Приведен формат записи данных наблюдений Сибирского Солнечного Радиотелескопа (ССРТ), анализ характеристик избыточности исходных данных и основные из опробованных в процессе разработки компрессора методов моделирования. Описана архитектура разработанного компрессора, математический аппарат прогнозирования сигнала на стадии моделирования. Рассмотрены некоторые результаты практического применения методики.

Ключевые слова: компрессия, сжатие, моделирование, прогнозирование.

1. Введение. ССРТ — специализированный солнечный радиотелескоп, один из крупнейших астрономических инструментов, предназначенный для изучения солнечной активности в микроволновом диапазоне (5,7 ГГц). Телескоп расположен в долине, разделяющей два горных хребта Восточных Саян (220 км от Иркутска). Регулярные наблюдения Солнца на ССРТ позволяют иметь ежедневные данные в радиодиапазоне. Объем данных за день наблюдения, в зависимости от времени года, порядка 300 Мб. На настоящее время архив данных на CD-носителях составляет около 500 Гб.

Для оперативного доступа к данным ССРТ при поддержке РФФИ (грант № 99–07–90038) разработана система общего пользования, базирующаяся на использовании информационного сервиса сети Интернет (`http://ssrt.iszf.irk.ru`). К сожалению, до настоящего момента не было возможности предоставления в сети Интернет доступа к большей части архива данных наблюдений ССРТ. Высокая стоимость цифровых накопителей и низкая пропускная способность каналов передачи данных обусловили необходимость разработки программного обеспечения, предполагающего эффективное сжатие без потерь представляемой информации.

С учетом характерных особенностей данных наблюдений ССРТ был разработан простой, свободный от потерь компрессор на основе адаптивных предсказывающих фильтров, показавший такие же или лучшие результаты, чем большинство современных кодер-декодеров без потерь.

2. Область применения компрессора. Общеизвестно, что нет алгоритма, способного сжать без потерь входные данные любого типа. Высокая степень компрессии поступающих с радиотелескопа данных может быть достигнута только с помощью алгоритма, учитывающего характерные особенности данных ССРТ. Разработанный компрессор рассчитан на работу с данными рутинных наблюдений, полученных при помощи ССРТ (формат PC1, разрядность данных — 16 бит, число каналов — $4 \cdot 500$). Это, однако, не исключает возможности адаптации реализованного алгоритма для архивирования данных, которые по своим параметрам (частота дискретизации, степень корреляции между соседними каналами и степень корреляции между соседними выборками) достаточно близки к рассматриваемым.

3. Современное состояние технологии сжатия информации без потерь. Сжатие данных без потерь не является новой технологией. Данный принцип используется известными утилитами сжатия типа PkZip, Compress или Gzip для компрессии текста и двоичных файлов. Это сжатие без потерь, потому что файлы после декомпрессии остаются идентичны оригиналам.

Согласно принятой терминологии, данные, получаемые с ССРТ, являются мультимедиа-данными, т.е. данными, включающими в себя различные формы естественной информации. К сожалению, указанные выше программы, обычно использующие варианты алгоритмов Лемпела–Зива [2], не дают хороших результатов при сжатии мультимедиа данных. В то время как большинство текстовых файлов может быть сжато с коэффициентами, большими чем 2 : 1, размер мультимедиа файлов практически не меняется.

Утилиты сжатия, основанные на алгоритме Лемпела–Зива, заменяют группу символов указателем на тот фрагмент текста, где они встречаются ранее. На примере мультимедиа файлов символам соответствуют выборки сигнала. К сожалению, последовательности повторений тех же самых выборок сигнала

[1] Институт солнечно-земной физики СО РАН, ул. Лермонтова, 126, 664033, г. Иркутск-33, а/я 4026; e-mail: sasha@iszf.irk.ru

очень редки, так что коэффициент сжатия получается весьма низким. Это справедливо и для данных ССРТ.

Кроме того, эти утилиты сжатия не используют важную особенность мультимедиа файлов — существенную зависимость между соседними выборками данных и между соседними каналами в случае многоканальных данных. Именно поэтому все современные алгоритмы сжатия мультимедиа информации включают стадию декорреляции данных в целях снижения такой статистической зависимости.

В современных компрессорах обычно присутствуют как минимум две стадии декорреляции: межканальная декорреляция и прогнозирующее моделирование. Остаточный декоррелированный сигнал обычно сжимается без потерь одним из известных методов кодирования энтропии. Фактически, все современные компрессоры используют по крайней мере один из следующих методов: кодирование Хаффмана, кодирование повторов (RLE) или кодирование Райса [3].

4. Формат исходных данных ССРТ. Прежде чем перейти к описанию формата файла исходных данных ССРТ, необходимо кратко описать работу приемника и структуру поступающей на него информации [1]. В силу использующегося на ССРТ метода частотного сканирования, приемник является фактически анализатором спектра в диапазоне частот, составляющем 2 % от центральной частоты ССРТ. Центральная частота 5730 ГГц, диапазон частот 120 МГц. Таким образом, задача, решаемая приемником, — регистрация спектра мощности радиосигнала. Это подразумевает многоканальность приемника.

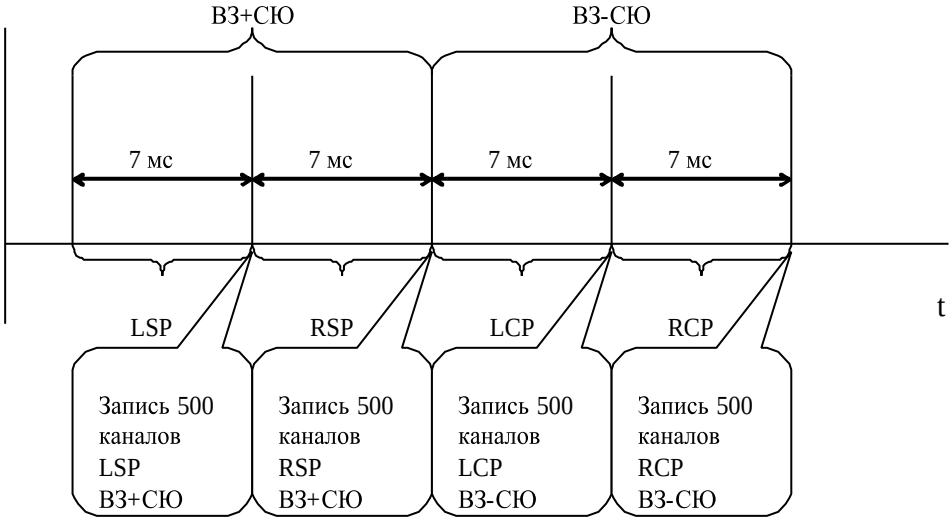

Диаграмма регистрации одного отсчета исходных данных ССРТ

В настоящий момент для рутинного режима наблюдений используется 500 каналов. ССРТ измеряет два вида поляризации радиосигналов: левую круговую (LP) и правую круговую (RP) поляризации. Для этого антенная система переключается каждые 7 мс на прием LP или RP. Для получения двумерного изображения на ССРТ используется метод фазно-противофазной модуляции. Для реализации этого метода сигналы, соответствующие линейкам антенн “Восток–Запад” (ВЗ) и “Север–Юг” (СЮ), складываются синфазно (ВЗ+СЮ) или противофазно (ВЗ-СЮ). Переключение между суммой и разностью на ССРТ осуществляется каждые 14 мс. Таким образом, приемник регистрирует 500 каналов спектра мощности каждые 7 мс. Если по окончании первых 7 мс накопления сигнала регистрируется LP-компонента суммарного сигнала (LSP), тогда после следующих 7 мс будет зарегистрирована RP-компонента суммарного сигнала (RSP). Затем (см. рисунок) то же самое повторяется для разностного сигнала (LCP и RCP).

Основу файла исходных данных ССРТ составляет блок последовательности спектров мощности суммарного и разностного сигналов левой и правой поляризаций. Каждый блок содержит кроме данных о спектрах некий признак и информацию о состоянии приемника на момент регистрации. Последовательность блоков предваряется заголовком, содержащим информацию, общую для всех блоков.

5. Анализ характеристик избыточности исходных данных ССРТ. В целях оптимального выбора метода компрессии исходных данных ССРТ был проведен выборочный анализ нескольких файлов формата PC1. Усредненные результаты проведенного анализа подтвердили предположение о сильной зависимости между соседними выборками и соседними каналами в данных рутинных наблюдений радиотелескопа. Так, средний коэффициент корреляции между соседними выборками данных оказался равен 0.998, а средний коэффициент корреляции между соседними каналами данных в файле формата PC1 — 0.997.

Как можно заметить, данные наблюдений ССРТ обладают высокой избыточностью и, следовательно, могут быть эффективно сжаты за счет устранения такой избыточности применением вышеописанных методов.

6. Архитектура компрессора. Обычно современные многоканальные компрессоры мультимедиа данных без потерь строятся по общей схеме, включающей в себя четыре основных этапа:

— разбиение сигнала на блоки;

— межканальная декорреляция;

— моделирование сигнала (прогнозирование);

— кодирование остатка.

Разбитые на блоки данные подвергаются межканальной декорреляции и передаются далее на стадию прогнозирования. На этой стадии моделирование входного сигнала осуществляется с помощью двухступенчатого адаптивного фильтра. На кодировщик энтропии передается разность между оригинальным сигналом и спрогнозированным. Остаточный сигнал сжимается с использованием кодов Райса. Рассмотрим архитектуру программы подробнее.

Выбор оптимального размера блока очень существенно от выбранного алгоритма компрессии на стадии моделирования. В общем случае для обеспечения возможности восстановления поврежденных файлов размер блока должен быть как можно меньшим. Уменьшение размера блока ведет к увеличению количества заголовков фреймов, что снижает степень сжатия файла. Увеличение размера блока затрудняет редактирование сжатого цифрового потока. В случае статического моделирования размер блока также не должен быть большим, так как в этом случае невозможно эффективное моделирование сигнала. Разрабатываемый компрессор использует динамическое (адаптивное) моделирование сигнала. Применение алгоритмов этого класса эффективно лишь на больших фреймах. Для сжатия файлов исходных данных ССРТ формата PC1 наиболее эффективным оказалось кодирование каждого из 2000 $(4 \cdot 500)$ каналов исходного файла в отдельном фрейме.

Поступающие на вход компрессора многоканальные данные могут быть подвергнуты межканальной декорреляции. В случае двухканальных данных два исходных канала преобразуются к среднему и разностному по формулам

$$\text{средний} = \frac{\text{первый} + \text{второй}}{2}, \quad \text{разностный} = \text{первый} - \text{второй}\,.$$

В целях исключения потери данных описанная формула приводится к виду:

$$\text{разностный} = \text{первый} - \text{второй}, \quad \text{средний} = \text{первый} - \frac{\text{разностный}}{2}\,.$$

Для многоканальных данных с хорошей корреляцией соседних каналов это обычно приводит к значительному увеличению степени сжатия.

На этапе моделирования компрессор пытается аппроксимировать сигнал такой функцией, чтобы полученный после ее вычитания из оригинала результат (называемый разностью, остатком, ошибкой) был минимальным. Если в процессе разработки существующего алгоритма компрессора другие стадии (межканальная декорреляция и сжатие остатков) остаются практически неизменными, то стадия моделирования претерпевает ряд значительных изменений. Ниже приведены основные из опробованных в процессе разработки компрессора методов:

— аппроксимация сигнала с помощью множества постоянных предикторов [4];

— моделирование сигнала методом линейного предсказания (LPC) [5];

— моделирование сигнала с помощью адаптивных фильтров [6 – 8].

Из всех рассмотренных методов наилучшие результаты показал метод моделирования сигнала с помощью адаптивных фильтров.

В этом методе используются IIR-фильтры (Infinite Impulse Response model), параметры которых изменяются адаптивно в процессе работы. Базовым элементом системы является p-мерный нерекурсивный фильтр, в общем случае описываемый следующими выражениями:

$$x'(n) = \sum_{k=1}^{p} v(n,k) \cdot x(n-r-k)\,, \tag{1}$$

где $x'(n)$ — предсказанное значение новой выборки $x(n)$, $v(n,k)$ — очередное значение весового коэффициента фильтра и r — задержка сигнала. Весовые коэффициенты фильтра определяются по формуле

$$v(n+1,k) = v(n,k) + m \cdot \operatorname{sign}\big(e(n)\big) \cdot x(n-r-k), \tag{2}$$

где m — коэффициент, определяющий скорость адаптации и $e(n)$ — очередной отсчет выходного сигнала (сигнала ошибки).

Минимизация остатка $e(n) = x(n) - x'(n)$ может быть реализована с помощью различных алгоритмов, такими как алгоритм минимума среднего квадрата ошибки Уидроу–Хоффа (LMS), основанного на статистическом подходе, или же рекурсивный алгоритм наименьших квадратов (RLS). Второй из них, хоть и имеет более высокую скорость сходимости, использует значительно больше ресурсов процессора. Поэтому в разрабатываемом компрессоре был использован LMS-алгоритм. Сходимость алгоритма осуществляется по методу наискорейшего спуска, причем для упрощения вычислений применяется стохастическая аппроксимация градиента по Уидроу–Хоффу. Для ускорения сходимости в качестве критерия оптимальности используется минимум модуля ошибки фильтра. Независимо от начального вида матрицы коэффициентов фильтра $\mathbf{v}$, который может быть произвольным, алгоритм сходится в среднем и остается устойчивым до тех пор, пока параметр m удовлетворяет условию $1/\lambda_{\max} > m > 0$, где $\lambda_{\max}$ — максимум собственного значения автокорреляционной матрицы входных сигналов. Выходной остаточный сигнал фильтра определяется как разность между реальным сигналом и его предсказанным значением, которое вычисляется как свертка реального сигнала с весовыми коэффициентами трансверсального фильтра в соответствии с (1). Импульсная характеристика этого фильтра (или вектор весовых коэффициентов размерностью p) обновляется на каждом дискретном моменте времени n в соответствии с (2).

Примененный в компрессоре алгоритм незначительно отличается от описанного выше. Входной сигнал проходит двухступенчатую фильтрацию. На первой стадии в качестве фильтра выступает предиктор нулевого порядка. Это означает, что сигнал ошибки определяется формулой $e(n) = x(n) - kx(n-1)$, где k достаточно близко к 1. На следующей, последней стадии фильтрации используется аналогичный описанному выше широкополосный фильтр 32-го порядка.

Описанный метод является наиболее эффективным из рассмотренных методов моделирования сигнала как по точности предсказания, так и по скорости работы. К недостаткам метода можно отнести неэффективность метода на блоках малого размера и, как следствие, сложность редактирования сжатого потока.

Когда модель подобрана, кодировщик вычитает приближение из оригинала. Если модель не описывает сигнал точно, разность между оригинальным сигналом и спрогнозированным представляет собой остаточный сигнал, который затем кодируется без потерь. При кодировке учитывается то обстоятельство, что разностный сигнал обычно имеет распределение Лапласа. При этом применяется набор специальных кодов Хаффмана [9] (эти коды, называемые кодами Райса [10], позволяют эффективно и быстро кодировать сигналы без использования словаря). Кодирование Райса состоит из нахождения одного параметра, отвечающего распределению сигнала, с последующим использованием его для составления кодов. При изменении распределения меняется и оптимальный параметр, поэтому имеется метод, позволяющий пересчитывать его в случае необходимости. Если предсказание сделано удачно, остаточный сигнал будет занимать меньше бит на выборку, чем оригинальный сигнал. Обычно остаток разбивается на подблоки, у каждого из которых будет свой параметр Райса. Выбор размера подблока также влияет на эффективность кодирования. Разработанный компрессор для кодирования остатка использует адаптивный метод кодирования с динамическим определением эффективного параметра Райса. При этом разбиение остаточного сигнала на подблоки не производится.

7. Формат архива PCA. Файл архива состоит из пяти частей: заголовка; поля, содержащего количество записей; массива состояний приемников; битового массива данных; цифровой подписи. Заголовок архива совпадает с заголовком исходного файла данных PC1, за исключением уникального идентификатора формата. Такой подход дает возможность получать общую информацию о данных без распаковки архива. За заголовком следует четырехбайтное поле, содержащее количество отсчетов в неупакованном файле, затем массив состояний приемников и битовый массив упакованных данных. Упакованные данные записываются блоками поканально. Завершает файл 16-байтная цифровая подпись по алгоритму MD5 [11], предназначенная для проверки целостности архивного файла.

8. Заключение. Разработанный потоковый компрессор рассчитан на быстрое кодирование-декодирование данных Сибирского солнечного радиотелескопа, предусматривает возможность легкой аппаратной реализации и работы в реальном режиме времени. По результатам тестирования разработанный компрессор показал хорошую скорость работы и степень сжатия (в среднем 0.41) для файлов формата PC1.

Аналогов разработанному компрессору, применимых для сжатия цифровых данных наблюдений ССРТ, не существует. Благодаря разработанному компрессору стала возможной организация интерактивного доступа к многолетнему архиву данных наблюдений ССРТ, что является одной из задач проекта создания

“Информационной системы комплекса гелио-геофизических инструментов Института солнечно-земной физики СО РАН”. Проект поддерживается РФФИ (грант № 03–07–90054в).

Для демонстрации эффективности примененных в настоящем проекте алгоритмов разработан свободный от потерь аудиокомпрессор, предназначенный для сжатия 8- и 16-разрядных аудиофайлов формата Wav.

Тестовая версия компрессора доступна на сайте проекта по адресу `http://tta.iszf.irk.ru`.

СПИСОК ЛИТЕРАТУРЫ

1. *Лесовой С.В.* Кандидатская диссертация. Иркутск. 1999.
2. *Bell T.C., Cleary J.G., Witten I.H.* Text compression. Englewood Cliffs: Prentice-Hall, 1990.
3. *Hans M., Schafer R.* Lossless audio coding. Technical Report. CSIP TR–97–07. Atlanta, 1997.
4. *Robinson T.* SHORTEN: Simple lossless and near-lossless waveform compression. Technical Report. Cambridge University Engineerig Department. Cambridge, 1994.
5. *Bruekers A.A.M.L., Oomen A.W.J., van der Vleuten R.J.* Lossless coding for DVD audio. 101st AES Convention. Los Angeles, 1996.
6. *Craven P.C., Law M.J., Stuart J.R.* Lossless compression using IIR prediction filters. 102nd AES Convention. Munich, 1997.
7. *Gabor D., Wilby W.R., Woodcock R.A.* A universal nonlinear filter, predictor and simulator which optimizes itself by a learning process // Proc. Inst. Electr. Engrs. 1961. **108**, part B, N 40. 85–98.
8. *Колмогоров А.Н.* Проблема синтеза оптимального предсказывающего фильтра // Изв. АН СССР. Сер. матем. и естеств. наук. 1941. № 5. 112–129.
9. *Knuth D.E.* Dynamic Huffman coding // J. Algorithms. 1985. **6**, N 2. 163–180.
10. *Rice R.F.* Some practical universal noiseless coding techniques. Technical Report. JPL-79-22. Jet Propulsion Laboratory. Pasadena, 1979.
11. *Madson C., Glenn R.* The use of HMAC-MD5-96 within ESP and AH, RFC 2403. Internet Draft. 1998 (www.faqs.org/rfcs/rfc2403.html).

Поступила в редакцию
21.09.2003