Lossless Compression Method for Acoustic Waveform Data Based on Linear Prediction and Bit-recombination Mark Coding

Ming Cai, Wenxiao Qiao, Xiaodong Ju, and Xiaohua Che

***Abstract*— In acoustic logging, large amounts of data put forward severe challenges for data transmission, storage and processing. Data compression techniques, to some extent, may relieve this problem. In this paper, we first review the basic principle of the linear prediction, and appropriately improve common optimal linear prediction method to get a new optimal linear prediction method that maps integers to integers. Then, we explore an appropriate bit-recombination mark coding approach according to the characteristics of prediction errors sequence. Finally, we propose a new lossless compression method for acoustic waveform data based on linear prediction and bit-recombination mark coding. The compression and decompression programs are developed according to the proposed method. Field acoustic logging waveform data are then applied to compression and decompression tests and the compression performances of our method and several other lossless compression methods are compared and analyzed. Test results validate the correctness of our method and demonstrate its advantages. The new method is potentially applicable to acoustic waveform data compression.**

***Index Terms*— acoustic, linear prediction, logging, lossless compression, mark coding**

I. Introduction

IN well logging, large amounts of data need to be sent from downhole to the surface by means of a very band-limited cable telemetry system. The limited bandwidth usually results in prolonging of expensive rig-time and/or the sacrificing of borehole information [1][2]. In logging while drilling (LWD), the data transmission rate of telemetry system becomes more evidently incapable of satisfying field requirements [3][4][5]. The amount of acoustic logging data is particularly large, and the data transmission efficiency is limited by the low data transmission rate more severely. Therefore, it is necessary and important to introduce data compression [1][6] techniques to acoustic logging.

Acoustic waveform data compression technology has a wide application prospect in the acoustic logging business. In recent years, scholars all over the world have conducted numerous studies on the acoustic waveform data compression. Robinson used a simple predictive model of waveform and Huffman coding in the compression of audio data [7]. Wu applied linear prediction and Huffman coding to the compression of acoustic exploration signals [8]. Qiang et al. studied a lossy wavelet transform compression method for acoustic logging waveform data [9]. Bernasconi et al. proposed a lossy data compression algorithm based on the wavelet transform for LWD [10][11]. This method might achieve a large compression factor (the inverse of compression ratio, defined as the ratio of the size of input stream to that of output stream [6]) on the condition that the quality of signals did not degrade a lot. Wu et al. [12], Li et al. [13] and Zhang et al. [14] studied lossy compression methods for downhole waveform signals. Liu et al. presented a lossless compression scheme for acoustic LWD data [15]. Jia et al. tested the compression effects of several compression algorithms, including lossless and lossy algorithms, on the field acoustic logging waveform data [2]. Moreover, Davis [16], Mandyam et al. [17], Ergas et al. [18] and Stromberg et al. [19] studied compression methods for seismic waveform data.

Above review shows that common acoustic waveform data compression methods are mostly lossy, and it is difficult to find a lossless compression method with large compression factor for acoustic logging waveform data. This condition has motivated the present work, which presents a new lossless compression method for acoustic waveform data. This approach is lossless and promising, with the advantages of a simple algorithm, large compression factor and stable compression capability. Field acoustic logging waveform data are applied to test our method and several other lossless compression methods. Test results show that our method can compress and decompress the acoustic waveform data losslessly. Moreover, the compression performance (the compression factor is used to evaluate the compression performance of a compression method) of our method is superior to that of the other tested methods.

This paper is organized as follows. Section II gives a review of the basic theory of linear prediction, and introduces the integer-to-integer optimal linear prediction; Section III describes the bit-recombination mark coding method; Section IV delineates the compression and decompression methods; Section V shows compression and decompression test results, and compares and analyzes the compression performances of all tested methods; Finally, the discussion and conclusions are

Manuscript received June 21, 2013; revised July 11, 2013. This work was supported in part by National Natural Science Foundation of China (61102102, 11134011, 11204380), National Science and Technology Major Project (2011ZX05020-009), Science Foundation of China University of Petroleum, Beijing (KYJJ2012-05-13) and China National Petroleum Corporation Project (2008A-2702, 2011A-3903).

Ming Cai, Wenxiao Qiao, Xiaodong Ju, and Xiaohua Che Authors are with the State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum (Beijing). E-mail: caiming2006@163.com; qiaowx@cup.edu.cn; juxdong@cup.edu.cn; aclab@cup.edu.cn. Phone: +8615117974961 (Cai).

presented in Section VI.

II. Linear Prediction

Linear prediction is a prediction method that predicts the current sample of a signal by a linear combination of the N immediately-preceding samples [17][20][21]. In a stream of correlated acoustic waveform samples $s(t)(t=1,2,3,\cdots)$, almost every sample $s(t)$ is similar to its predecessor $s(t-1)$ and its successor $s(t+1)$. Thus, a simple subtraction $s(t)-s(t-1)$ normally produces a small difference. Consecutive acoustic waveform samples may become larger and larger and be followed by smaller and smaller samples. It therefore makes sense to assume that an acoustic waveform sample is related in a simple way to several of its immediate predecessors and several of its successors. This assumption is the basis of the technique of linear prediction. A predicted value $\hat{s}(t)$ for the current sample $s(t)$ is computed from the N immediately-preceding samples by a linear combination

$$\hat{s}(t)=\sum_{i=1}^{N} a_i s(t-i), \tag{1}$$

where N is the order of the predictor, $a_i(i=1,2,\cdots N)$ are predictor coefficients. Then the residuals (also termed prediction errors) can be obtained by

$$e(t)=s(t)-\hat{s}(t), \tag{2}$$

When the predictor order N is a constant, the predictor coefficients are vital to the performance of the predictor. Generally, the predictor coefficients determined by the optimization method may produce a good prediction, and the corresponding prediction method is called as the optimal linear prediction [20].

Notice that the predictor coefficients determined by the optimization method are generally not integers, which may result in the residuals are not integers even if the original signal consists of integer samples. Yet, for lossless compression it would be of interest to be able to characterize the residuals completely again with integers. This aim can be achieved by improving common optimal linear prediction. By analyzing (1) and (2), it is easy to find that what results in the residuals are not integers is that the predicted values are not integers. Therefore, we can obtain the integer residuals by truncating the predicted values. According to this idea, we can derive the integer-to-integer optimal linear prediction formula from (1) and (2), and it can be written as

$$e(t)=s(t)-\left\lfloor \sum_{i=1}^{N} a_i s(t-i)+0.5 \right\rfloor, \tag{3}$$

where $\lfloor x \rfloor$ stands for the largest integer not exceeding x, the 0.5 added at the end is to reduce the error caused by truncating, and the coefficients a_i are determined by using the optimization method. According to (3), we can easily calculate the original samples from the residuals by

$$s(t)=e(t)+\left\lfloor \sum_{i=1}^{N} a_i s(t-i)+0.5 \right\rfloor, \tag{4}$$

By comparing (3) with (4), it is easy to find that the predictor and the truncating processing in (4) are all the same as those in (3). Therefore, we can reconstruct the original signal losslessly from the residuals by (4).

III. Bit-recombination Mark Coding

The optimal linear prediction is used to remove correlation, which reduces the redundancy of the original signal. However, the residuals still need to be transmitted or stored in a storage device finally. To compress the data file (to reduce the size of the data file) as much as possible, the selection of an appropriate coding means is vitally important. In this paper, we explore a novel effective bit-recombination mark coding means according to the characteristics of residuals. We apply the arithmetic coding, LZW (a dictionary compression method that developed by Lemple, Ziv, and Welch, and named by the acronym of their names [20]) and our coding means to compress the residuals of acoustic logging waveform data. The results show that our coding means is superior to arithmetic coding and LZW apparently.

A. Mark Coding

The idea of mark coding is inspired by the Index-Data record format [13]. However, the Index-Data record format is not advanced enough and is effective only when the probability that 0 occurs is significantly larger than that of other data. In this paper, we explore a more effective mark coding means that can encode the integer sequence adaptively according to its characteristics.

The mark coding is to encode the data stream by a special means. That is, we use a shorter code word (substitution mark) that occupies as less bits as possible to replace the datum that has the largest occurrence probability while without changing other data (each datum is regarded as a nonreplaced code word) to record the input data stream. Moreover, a non-substitution mark is given before each nonreplaced code word. Except for the substitution marks, the other nonreplaced code words have the same length. The given non-substitution marks are to distinguish the replaced code words from the nonreplaced code words during decoding and to guarantee the accuracy of the decoding.

Two types of marks are applied to our mark coding means that can adaptively select the correct record format to encode the input data stream according to the probability distribution characteristics of local data. Our mark coding means can be implemented by three procedures. The first procedure is to divide the input data stream into several data blocks, in which the fore data blocks have the same size while the last one may be smaller. The second procedure is to count the occurrence probability of all data in the selected block and then to record the largest probability value and the corresponding datum. The third procedure is to select the mark record format or the normal record format to encode the selected data block according to the largest probability value recorded in the second procedure. The selection rule of the record format is that when the largest probability value is larger than the given appropriate threshold, the mark record format is selected; otherwise, the normal record format is selected. The third procedure is the most important but the most difficult part. In practical coding, each data block of the input data stream is encoded successively until the last data block is encoded. First, a special bit mark that indicates which record format is selected for the encoded data block is outputted, which is the first type of mark. The data block is then encoded according to the selected record format. If the mark record format is selected, the datum that has the largest occurrence probability is outputted first and each datum of the data block is then encoded successively. When the input datum is the datum that has the largest occurrence

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

probability, the substitution mark is outputted to replace it. Otherwise, the input datum is outputted without discarding or changing any bit and a non-substitution mark is outputted before it. The substitution or non-substitution mark is the second type of mark. If the normal record format is selected, each datum of the data block is outputted successively without discarding or changing any bit. The whole input data stream can be encoded successfully according to this rule. The processes of the mark record format and the complete mark coding method are depicted in Fig. 1 and Fig. 2, respectively.

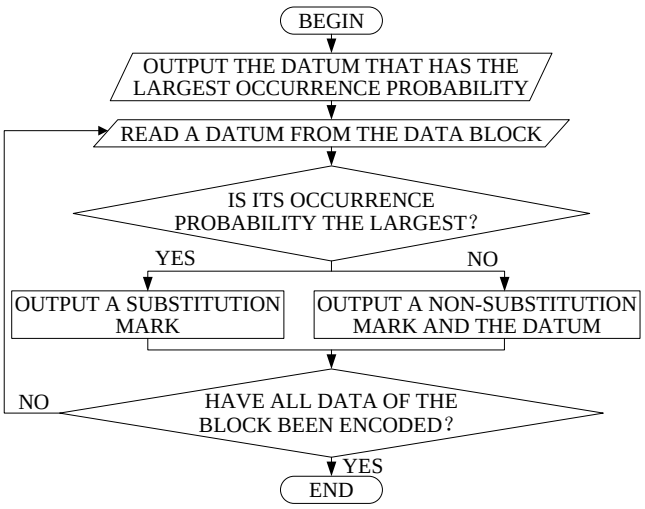

Fig. 1. Process of the mark record format.

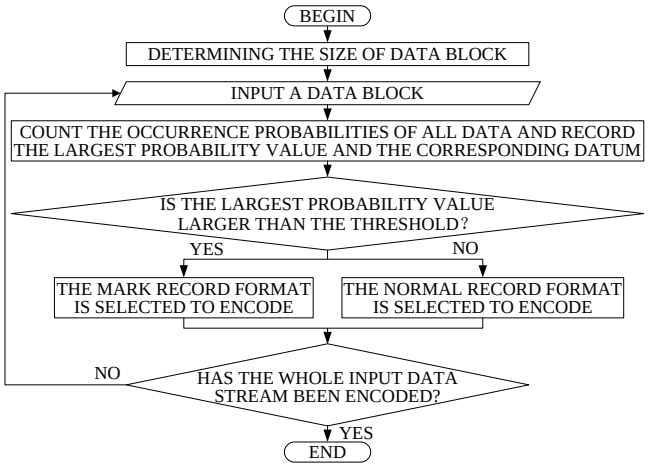

Fig. 2. Process of the mark coding method.

The input data stream is divided into several data blocks for encoding. The advantage is that the correct coding format can be selected adaptively to encode the input data stream according to the probability distribution characteristics of local data. However, the common mark coding format does not have this adaptability. When the input datum is not the datum that has the largest occurrence probability and if the common mark coding format is applied to encode, it is necessary to output the input datum without discarding or changing any bit and a non-substitution mark. However, an excessive number of non-substitution marks may also occupy a large number of bytes, which is not beneficial for improving the compression performance. Therefore, if the common mark coding format is still applied to encode when there is no datum that has a significantly larger occurrence probability than others, numerous non-substitution marks have to be outputted. This may result in the size of compressed data file is much larger than that of the original data file. In this case, the bad result can be avoided by selecting the normal record format to encode.

Our mark coding method can adaptively select the correct record format to encode according to the probability distribution characteristics of the data in each data block. When the largest counting probability value is larger than the given threshold, the mark record format is selected to encode, which results in the bytes occupied by the compressed data block are fewer or much fewer than that occupied by the original data block. Otherwise, the normal record format is selected to encode, which results in the bytes occupied by the compressed data block and the original data block are the same. Thus, the bytes occupied by each compressed data block are fewer than or equal to that occupied by the corresponding original data block, and the whole input data stream is compressed successfully. Therefore, our mark coding method, to some extent, may always have compression ability, and the size of the compressed data file will never be much larger than that of the original data file.

B. Bit-recombination

The function of bit-recombination that serves for the mark coding is to increase the occurrence probability of a datum, and in most cases, the occurrence probability of zero is increased. The idea of bit-recombination is inspired by the shuffle principle [22]. The bit-recombination process that is applied to our compression method can be described as follows. First, the most significant bit (or sign bit) of each datum of the input data stream is extracted successively, followed by the succeeding bit, and so on until the least significant bit of each datum is extracted successively. Then, all the extracted bits are combined sequentially to yield a new data sequence.

The mean squared value of the residuals is significantly smaller than that of the original signal samples. Therefore, the high order bits of the residuals primarily consist of 0, which is suitable for bit-recombination. Theoretically, the bit-recombination may significantly increase the occurrence probability of zero for this type of data. The bit-recombination is applied to the residuals of acoustic logging waveform signals. Results indicate that the occurrence probabilities of some data, particularly the occurrence probability of zero, are increased, which is beneficial for improving the compression performance.

IV. Compression and Decompression Methods

The compression method mainly consists of four parts. First, the integer-to-integer optimal linear prediction is applied to the original waveform signal. Second, the data type of the residuals is converted, which converts the signed integers into unsigned integers. Third, the bit-recombination is applied to the unsigned integer residuals. Forth, the mark coding is performed. The compression process is depicted in the bottom part of Fig. 3.

The principles and implementation approaches of the integer-to-integer optimal linear prediction, bit-recombination and mark coding have already been discussed in detail. Data type conversion in the compression

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

is to convert signed integers into unsigned integers, which has two advantages. First, there is no need to process the sign bit of each datum specially in the compression and decompression. Second, it may improve the performance of bit-recombination mark coding and thus may improve the compression performance as well.

Data type conversion can be implemented by an overlap and interleave scheme, in which all values are re-assigned to some positive number in a unique and reversible way [23]. According to this scheme, the n^{th} negative value (i.e., $-n$) is mapped to the n^{th} odd number $(2n-1)$, and the m^{th} positive value is mapped to the m^{th} even number $(2m)$. This may be expressed mathematically as

$$y = \begin{cases} 2x, & (x \geq 0) \\ -2x-1, & (x < 0) \end{cases}, \tag{5}$$

where x stands for the signed integer while y for the unsigned integer.

Data decompression is the reverse process of data compression and mainly consists of four parts. First, mark decoding, which is the reverse process of mark coding, is applied to the compressed acoustic waveform data. Second, bit-recovery, which is the reverse process of bit-recombination, is applied to the data. In this process, each bit of all data is returned to its original position before bit-recombination. Third, data type conversion is performed. This process is to convert the data into signed integers from unsigned integers, and the conversion formula can be easily obtained according to (5). Fourth, signal reconstruction is implemented according to (4). The decompression process is depicted in the upper part of Fig. 3.

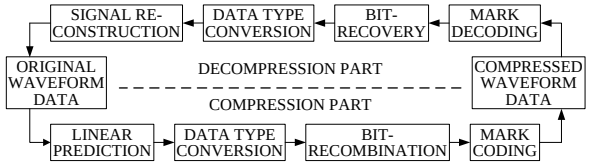

Fig. 3. Compression and decompression processes.

The implementation method of each part of decompression can be easily obtained according to the principle and implementation method of the corresponding part of compression. Since all parts of compression are reversible, the entire compression method is also reversible. That is, the original waveform signal can be exactly reconstructed from the compressed acoustic waveform data by the decompression method. This is proven by compression and decompression tests of field acoustic logging waveform signals.

V. Application Effect and Analysis

Compression and decompression programs are developed according to the corresponding compression and decompression methods discussed above. 12 acoustic logging waveform traces are selected arbitrarily from the field data acquired by a Multi-Pole Acoustic Logging (MPAL) tool and are applied to compression and decompression tests. Each waveform trace consists of 4 near monopole waveforms, 8 far monopole waveforms and 4×8 dipole waveforms.

An arbitrary waveform trace is first applied to the compression and decompression tests of our method. The decompressed waveform is then compared with the original waveform, as shown in Fig. 4 (only a part of the waveform is shown for ease of observation). From the Fig. 4, it is easy to find that the decompressed waveform completely coincides with the original waveform and that the error values (the differences between the decompressed waveform samples and the corresponding original waveform samples) are all zero. This proves that our compression method is lossless.

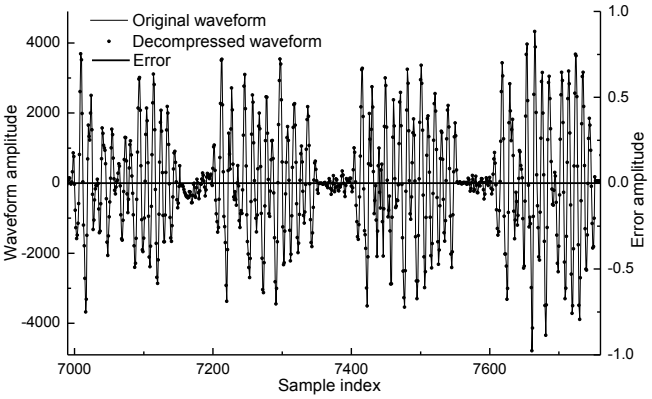

Fig. 4. Comparison between decompressed waveform and original waveform.

In addition, all the 12 acoustic logging waveform traces are applied to compression and decompression tests. The tested compression methods include classical arithmetic coding, universal WinRAR software, Free Lossless Audio Codec (FLAC) software, Monkey's Audio software and our method. FLAC and Monkey's Audio are both professional audio data compression software [24][25].

Test results show that all tested compression methods are lossless. Table I shows the compression performances of all tested compression methods for acoustic waveform data. Where A, B, C, D and E refer to the arithmetic method, WinRAR, FLAC, Monkey's Audio and our method, respectively; the ACF refer to the average compression factor; the size of each original trace data file is 21968 bytes; and an eighth-order predictor is used in our method.

TABLE I. Compression Performances of All Tested Methods

Trace NO.	Compression factor				
	A	B	C	D	E
1	1.15	1.45	0.94	1.55	1.65
2	1.17	1.46	0.95	1.57	1.66
3	1.13	1.44	0.93	1.51	1.62
4	1.15	1.46	0.94	1.54	1.65
5	1.16	1.47	0.94	1.55	1.68
6	1.15	1.47	0.94	1.56	1.66
7	1.15	1.49	0.96	1.60	1.71
8	1.15	1.50	0.96	1.60	1.73
9	1.16	1.48	0.96	1.58	1.71
10	1.14	1.43	0.93	1.51	1.56
11	1.14	1.45	0.93	1.54	1.66
12	1.15	1.46	0.95	1.55	1.67
ACF	1.15	1.46	0.94	1.55	1.66

Table I shows that the size of data file compressed by

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

FLAC exceeds that of the corresponding original waveform data file, that is, FLAC has no compression ability. The arithmetic coding, to some extent, has some compression ability, but the performance is not good enough. By contrast, WinRAR, Monkey's Audio and our method all have relatively satisfactory compression performances. In addition, according to the compression factors of a trace or the average compression factors, it is easy to find that the compression performance of our method is superior to that of other tested methods.

VI. Discussion and Conclusions

In this paper, we propose a novel lossless compression method for acoustic waveform data. Compression tests show that our method is potentially applicable to acoustic waveform data compression.

About the predictor order N, generally, when the N is small, the mean squared value of residuals sequence decreases with the increase of the N; and when the N is large enough, the increase of the N will not result in a diminution of the mean squared value of residuals sequence. When the N is small, the compression performance of our method may become better with the increase of the N, but the time complexity and space complexity of the algorithm will increase as well. Therefore, it is necessary to reach a compromise between the complexity of the algorithm and the compression performance when selecting the predictor order. For the same type of data sequences, such as music data with the similar music style and the same type of acoustic logging waveform data from a well, the optimal predictor orders are similar and may be determined by experience.

From this study, we mainly draw four conclusions as follows. (1) Our method with stable compression capability, can compress and decompress the waveform data losslessly; (2) The compression factor of our method is large and its compression performance is superior to that of the classical arithmetic coding, universal WinRAR software, FLAC and Monkey's Audio; (3) Bit-recombination mark coding is a quite promising coding approach for the integer sequence with a small mean squared value; and a smaller mean squared value of the integer sequence generally results in a better compression performance; and (4) Our method may be also suitable for compressing seismic waveform data, natural gamma ray logging curve data, spontaneous potential logging curve data, resistivity logging curve data, acoustic velocity logging curve data, and so on; and in most cases, the amplitude variation ranges and frequencies of these waveforms or curves are significantly lower than that of the acoustic logging waveform, which is theoretically beneficial for achieving a better compression performance.

References

[1] X. G. Li, "Compression of well-logging data in wavelet space," in *1996 Annual Meeting, SEG*, Denver, Colorado, pp. 1615-1618.

[2] A. X. Jia, W. X. Qiao, X. D. Ju, X. H. Che, R. Lu, and R. J. Wang, "Effect test on compression algorithms of acoustic logging downhole data," *Well Logging Technology*, vol. 35, no. 3, pp. 288-291, Jun. 2011.

[3] W. R. Gardner, and W. C. Sanstrom, "Real-time compression of logging data," in *Society of Petroleum Engineers*, Cannes, France, 1992, pp. 557-566.

[4] J. Aron, S. K. Chang, R. Dworak, et al., "Sonic compressional measurements while drilling," in *35th Annual Logging Symposium, SPWLA*, 1994, pp. 1-17.

[5] X. Y. Zhang, Y. J. Guo, and J. N. Wang, "The Logging While Drilling: past, present and future," *Well Logging Technology*, vol. 30, no. 6, pp. 487-492, Dec. 2006.

[6] D. Salomon, Data compression: the complete reference (Fourth Edition). Springer-Verlag, 2007.

[7] T. Robinson, "Shorten: Simple lossless and near-lossless waveform compression," Technical Report, Cambridge University Engineering Department, 1994, pp. 1-16.

[8] L. N. Wu, "Prediction coding of acoustic exploration signals," *Oil Geophysical Prospecting*, vol. 30, no.4, pp. 505-508, Aug. 1995.

[9] L. Qiang, and G. Z. Liu, "Use of wavelet transform for compressing acoustic wave data from acoustic logging," *Journal of Xi'an Jiaotong University*, vol. 33, no. 3, pp. 35-38, Mar. 1999.

[10] G. Bernasconi, V. Rampa, F. Abramo, and L. Bertelli, "Compression of Downhole Data," in *1999 Society of Petroleum Engineers Drilling Conference/IADC*, Amsterdam, Netherlands, pp. 1-9.

[11] G. Bernasconi, and V. Rampa, "High-quality compression of MWD signals," *Geophysics*, vol. 69, no. 3, pp. 849-858, May 2004.

[12] P. T. Wu, P. H. Campanac, A. Sinha, and J. G. L. Thompson, "Methods and systems for compressing sonic log data," U. S. Patent 0062081 A1, Mar. 23, 2006.

[13] C. W. Li, D. J. Mu, A. D. Li, and G. H. Yao, "A real-time data compression algorithm for acoustic wave logging while drilling," *Journal of Southwest Petroleum University (Science & Technology Edition)*, vol. 30, no. 5, pp. 81-84, Oct. 2008.

[14] Y. Zhang, K. Xiong, Z. D. Qiu, S. H. Wang, and D. M. Sun, "A new method for real-time LWD data compression," in *2009 International Symposium on Information Processing*, Huangshan, China, pp. 163-166.

[15] Y. J. Liu, Y. K. Zhou, and H. B. Xiao, "A lossless acoustic logging while drill compression scheme based on differential prediction," in *2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, Harbin, Heilongjiang, China, pp. 1478-1481.

[16] A. J. Davis, "Linear prediction coding for compressing of seismic data," U. S. Patent 4 509 150, Apr. 2, 1985.

[17] G. Mandyam, N. Magotra, and W. McCoy, "Lossless seismic data compression using adaptive linear prediction," in *Geoscience and Remote Sensing Symposium, 1996. IGARSS'96. 'Remote Sensing for a Sustainable Future.', International. IEEE*, Lincoln, NE, pp. 1029-1031.

[18] R. A. Ergas, P. L. Donoho, and J. Villasenor, "Method for reducing data storage and transmission requirements for seismic data," U. S. Patent 5 745 392, Apr. 28, 1998.

[19] J. Stromberg, A. Averbuch, F. G. Meyer, and A. A. Vaddiliou, "Fast compression and transmission of seismic data," U. S. Patent 6 594 394 B1, Jul. 15, 2003.

[20] D. Salomon written, L. N. Wu et al. translated, The principle and application of data compression (Second Edition). Publishing House of Electronics Industry, 2003.

[21] D. Salomon, A concise introduction to data compression. Springer-Verlag, 2008.

[22] G. Q. Wu, and H. Chen, "A lossless compression scheme for scientific data from simulation," *Computer Engineering and Application*, no. 5, pp. 172-175, May 2006.

[23] R. F. Rice, "Some practical universal noiseless coding techniques," Part Ⅲ, Module PSI14, K+: JPL Publication 91-3, Jet Propulsion Laboratory, 1991.

[24] J. Coalson. 2000, FLAC. Available: http://flac.sourceforge.net/

[25] M. Ashland. 2000, Monkey's Audio. Available: http://www.monkeysaudio.com/

ISBN: 978-988-19252-3-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)