31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Multiscale Electrophysiology Format: An Open-source Electrophysiology Format Using Data Compression, Encryption, and Cyclic Redundancy Check

Benjamin H. Brinkmann, Ph.D., Mark R. Bower, Ph.D., Keith A. Stengel, Gregory A. Worrell, M.D., Ph.D., Matt Stead, M.D., Ph.D

Abstract— **Continuous, long-term (up to 10 days) electrophysiological monitoring using hybrid intracranial electrodes is an emerging tool for presurgical epilepsy evaluation and fundamental investigations of seizure generation. Detection of high-frequency oscillations and microseizures could provide valuable insights into causes and therapies for the treatment of epilepsy, but requires high spatial and temporal resolution. Our group is currently using hybrid arrays composed of up to 320 micro- and clinical macroelectrode arrays sampled at 32 kHz per channel with 18-bits of A/D resolution. Such recordings produce approximately 3 terabytes of data per day. Existing file formats have limited data compression capabilities, and do not offer mechanisms for protecting patient identifying information or detecting data corruption during transmission or storage. We present a novel file format that employs range encoding to provide a high degree of data compression, a three-tiered 128-bit encryption system for patient information and data security, and a 32-bit cyclic redundancy check to verify the integrity of compressed data blocks. Open-source software to read, write, and process these files are provided.**

I. Introduction

THERE is accumulating evidence that high frequency oscillations measured via intracranial electroencephalography (iEEG) can localize the epileptogenic zone in focal epilepsy [1]-[3]. Large-scale high frequency recording is challenging, however, due to the need for specialized recording equipment and the practical challenges presented by the large amounts of recorded data generated by this approach. Commonly the data management challenges have been addressed by imposing *a priori* limitations on data acquisition, reducing the duration, number of channels, sampling rate, dynamic range, or resolution of recordings in order to limit acquired data to manageable quantities. At our institution, large-scale, high-frequency recordings are routinely performed for presurgical evaluation of epilepsy patients using hybrid intracranial microwire and macroelectrode arrays with up to 320 channels [3]-[5]. Continuous recordings are performed for up to 10 days at a sampling frequency of 32 kHz. Eighteen bits of information are needed per sample in order to resolve the full range of electrophysiological activity, from high amplitude field oscillations to microvoltage single-neuron action potentials. Using conventional data formats this approach has the potential to generate (at 4 bytes/sample) 140 gigabytes per hour, or up to 23 terabytes for a full session recording.

As the size of recorded data files increases, so does the probability of data errors within the file, either through network transmission errors or disk storage errors. Keeping on-demand backup copies of these files is impractical, and data backup is necessarily limited to a single tape copy of each file. As the iEEG data will likely be used in planning surgical resection of the epileptic focus, detection and isolation of any corruption that may occur in the data is crucial. In addition, collaboration between institutions either for clinical care or basic research involving this data requires addressing the issue of patient confidentiality. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) requires any patient protected health information transmitted over a public network (such as the internet, or an unsecured institutional intranet) to be encrypted with a minimum 112-bit symmetric encryption [6]. Removal of patient information prior to transmission is possible, but this method requires subsequent secured transmission of the relevant protected information in some other way. This raises the possibility of mismatching the recorded electrophysiological data and the clinical patient information, in addition to the institutional cost of the required personnel time to complete the process. Finally, data analysis needs routinely require random access to limited time points within files, but this is difficult to achieve when new data are appended to an existing file. This becomes even more difficult when data are compressed in blocks of unequal size.

Current commonly-used digital EEG file formats such as the European Data Format (EDF) and its subsequent revision, EDF+ [7], [8], Extensible Biosignal Format (EBS) [9], XLTek format, and Neuralynx DMA format do not completely address these challenges. EDF, EBS, and XLTek encode data as 16-bit integers, which requires discarding two bits from our recordings. While XLTek and EBS formats employ limited data compression, no current format employs safeguards against data corruption, or protects patient

Manuscript received April 7, 2009. This work was supported by the National Institutes of Health (Grant 1R01-NS063039) and by an Epilepsy Therapy Development Project grant from the Epilepsy Foundation of America.

B. H. Brinkmann, M. R. Bower, G. A. Worrell, and M. Stead are with the Mayo Systems Electrophysiology Lab, Rochester, MN 55905 USA (507-538-9140; fax: 507-284-4795; e-mail: brinkmann.benjamin@mayo.edu).

K. A. Stengel is with Neuralynx Inc., Bozeman, MT, USA.

information with data encryption.

Our group has developed Multiscale Electrophysiology Format (MEF), a novel digital EEG data format to deal with these challenges. Data up to 24-bits in length are compressed using range-encoded differences in a series of data blocks, each of which employs a 32-bit cyclically redundant checksum (CRC) to verify data integrity. A fixed-length one kilobyte file header allows optional 128-bit AES encryption [10] of patient identifying information, as well as optional encryption of the file's technical parameters and data. Individual recording channels are stored in individual files, and a series of data block offsets is stored at the end of the file to allow direct access to individual compressed blocks within the file. Source code to store and read these files is freely available under the GNU open-source license (http://mayoresearch.mayo.edu/mayo/research/msel/).

II. Multiscale Electrophysiology Format

Multiscale Electrophysiology Format (MEF) files consist of three main parts: 1) a fixed-length 1024-byte header, containing patient information and technical information about the recording, 2) a data section, comprised of a series of encoded data blocks, and 3) a time index section, consisting of three 8-byte element blocks holding block start time, file offset, and sample index values to facilitate rapid random access to the data (Table 1). Each file's header begins with an unencrypted section of data containing technical information necessary to read the file, including the file's byte order, the encryption algorithm used, the file's version, the length of the header, and boolean values denoting what areas of the file are encrypted. The file's header uses a dual-tiered encryption scheme, with both sections being encrypted independently. A "subject" section contains all the subject-identifying data, while a "session" section contains information regarding data acquisition, such as filter settings and sampling frequency. The session encryption can optionally be applied to the leading coefficients of the statistical model in each of the data block headers, making the data impossible to decompress without the encryption key. The subject-encrypted header region contains the session password, so that if the subject password is provided, all header information is accessible. If only the session password is provided, the subject information remains inaccessible, but the technical details of the recording and data can be decrypted. Subject and session encryption use 128-bit AES encryption [10] with passwords chosen by the file's creator. Use of encryption is not required, and subject or session encryption may be omitted.

The data section of the file consists of recorded samples stored in compressed blocks, the length of which can be specified by the file's creator. Lossless data compression is accomplished via the range encoded differences (RED) algorithm [11], [12]. Range encoding is a class of integer arithmetic encoding that uses byte-wise scaling to improve encode and decode speeds. RED compression encodes data in two stages: first, differences between sequential samples in the data block are computed; second, the frequencies of difference values are computed. The frequencies of values in the statistical model are then used to encode values within the block. Differencing time series data efficiently reduces its variance, a property that range encoding benefits significantly from, i.e. as the inherent variance in a signal decreases its compression ratio increases.

A 32-bit cyclically redundant checksum (CRC) value [13] is calculated from each compressed block using the polynomial reported by Koopman [14], which has been shown to have a Hamming distance of 4 for message lengths up to 114,663 bits. The CRC is stored as the first entry in each block's header, providing the ability to detect data corruption arising from network transmission errors or disk errors during long-term storage. The block-wise compression scheme used has the advantage that each compressed data block is independent of other blocks in the file. In the event that a particular data block is corrupted beyond repair, the affected block can be removed with no effect on the remaining data. In comparison, a single corrupt value in a conventional difference-encoded file propagates the error to all remaining data in the file. Discontinuities in the recording are indicated by a flag in each block's header, and maximum and minimum recorded values are stored in each block header to facilitate processing and display. The compressed data blocks are stored with 8-byte alignment to enable direct access to header variables, and to assist file recovery if damage to the file results in alignment loss.

Following the compressed data blocks is a series of 8-byte integer triplets encoding the time stamp (in microseconds) of the start of each compressed data block, the file offset to the start of each block, and index number of the first sample in each block. These values allow data blocks within the file to be accessed directly based either on a desired time index or recording sample number. Time stamps are stored in Microsecond Coordinated Universal Time (□UTC), which is a variation of standard Unix or Posix UTC time defined by the number of microseconds since midnight January 1, 1970, GMT. Integer microseconds provide sufficient temporal resolution for high-frequency EEG recordings without requiring the use of floating point data types, which can cause errors from truncation of the least significant bits.

III. Methods and Results

The MEF format was tested on large-scale electrophysiology recordings obtained from a series of 20 patients using subdural and depth hybrid electrodes[3]. A series of randomly-selected 32 kHz macro and micro-electrode iEEG channel recordings were stored in MEF format using RED compression with varying block lengths. MEF file sizes are reported as a percentage of the theoretical size of each file if it had been stored on disk with 18 bits per recorded sample (i.e. 4 samples for every 9 bytes) with no header. For all channels, the data is compressed to less than 30% of its theoretical size with blocks 1627 samples (50 msec) or longer (Figure 1).

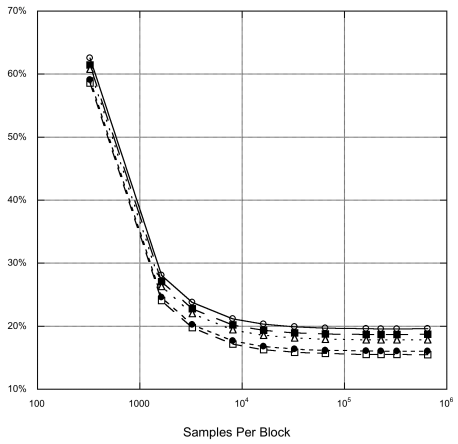

Figure 1. MEF compression improves with increasing block lengths, though good results are obtained for block sizes as short as 50 msec. The compressed file size is plotted against the log of the number of samples in each block for three clinical macroelectrode channels (black symbols) and two microwire channels (white symbols).

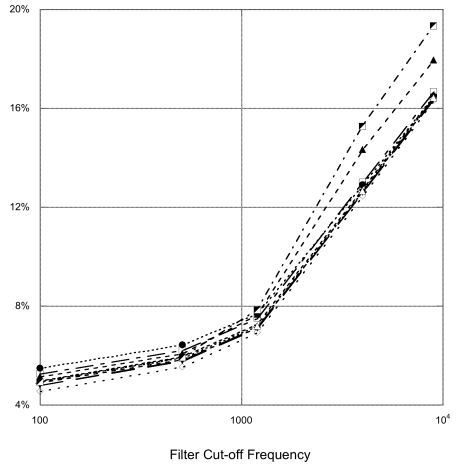

Figure 2. The RED compression algorithm improves its compression ratio as the high-frequency information in the raw data is removed. The compressed file size is plotted against the log of the low-pass filter's cut-off frequency for ten microwire channels.

Recorded data from microwire and clinical macrowire channels were low-pass filtered with varying cutoff frequencies between 100 and 9000 Hz, (constant sampling frequency) and were encoded in MEF format to illustrate the RED algorithm's ability to increase compression as the variance of the recorded data decreases. Results are shown in Figure 2, again as a percentage of a theoretical 18 bit sequential sample file.

We also compared the data compression achieved with the MEF file format to recorded data files in widely-used formats. A 32 kHz, 395.8 second iEEG recording with 40 channels in Neuralynx DMA format (7.03 Gb) was converted to a series of MEF format files with a 1.0 second block interval, compressing to 247.6 Mb, or 3.44% of the original file's size. For the XLTek file format 32 kHz data was not available, so a 500 Hz data file was used containing 76 recording channels and spanning 65267.5 seconds (4.71 Gb). Conversion into MEF format files with a 10.0 second block interval resulted in a total of 1.91 Gb, or 40.53% net compression. Conversion of a 28327.6 sec EDF file (1.72 Gb) to MEF format (150.7 Mb) resulted in 8.56% net compression. It should be noted that because of the formats' limitations, the XLTek and EDF data files contained only 16 bit sample resolution.

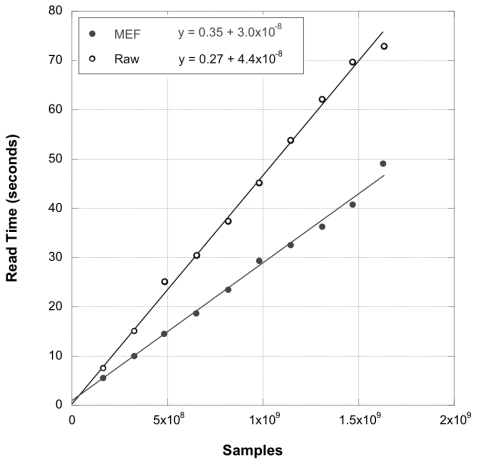

Figure 3. Reading and decompressing the MEF data is faster than reading raw 32-bit integers from disk directly. The read time in seconds is plotted against the number of samples read for MEF and 32-bit raw integer binary files.

The speed of reading and decompressing MEF data was compared to the speed of reading uncompressed raw 32-bit data from disk. Varying lengths of iEEG data were read from a MEF data file, decompressed, and stored on disk as a binary file of 32-bit integers. Custom software written in C and compiled with the Intel Compiler version 11.0 (Intel Corporation, Santa Clara, CA) was used on an Apple Macintosh computer (Apple Inc., Cupertino, CA) running Mac OS X version 10.5.5 with a 3.2 GHz Intel Xeon 8-Core processor and 32Gb of RAM to read the raw data from disk, and to read the corresponding MEF file from disk and decompress the data into 32-bit integers in memory. The MEF decompression was single-threaded, removing any potential advantage to the machine's multiple processors. As shown in Figure 3, reading and decompressing the MEF data is faster than reading raw uncompressed data. Further improvement can be achieved by multithreading the data decompression, and test results on our system suggest that MEF read times using 8 threads are less than 20% of the corresponding raw data read time. Data block header encryption was not used in these examples, but it typically adds 0.5% to the encoding time.

IV. Discussion

MEF has been designed to facilitate data storage, transmission, access and processing with high-frequency electrophysiological recordings spanning long durations in human or animal subjects. The block structure of the data and 32-bit CRC makes minor damage to the file detectable and limits damage to the affected block. The index data permits rapid access to individual data blocks, regardless of the length of the overall file. Sampling frequencies are

channel-specific, making it possible to downsample low-impedance macroelectrode signals for additional data reduction. The MEF format is suitable for any time series data that can be stored as 24-bit or smaller integers, including scalp EEG, polysomnography, electro-cardiography, and analytic transforms of recorded data. The MEF format is equally applicable to human and animal recordings, and header fields have been designed to accommodate either type of subject. The 128-bit AES encryption of patient information is compliant with HIPPA standards, allowing patient files to be transferred across networks without specialized security protocols. In addition, the MEF data format's division of channels into separate files and independent compressed data blocks facilitate parallel processing. While the MEF format is flexible enough to be used with other block-wise compression algorithms, including lossy algorithms if desired, RED encoding offers significant advantages for compression of time-series data. Principal among these advantages are the algorithm's high lossless compression rate and its computational speed. An additional advantage is the algorithm's ability to adapt to the statistical variation in the raw data, which is particularly useful in non-stationary signals such as EEG [15], resulting in improved compression ratios in filtered or slowly varying data without requiring changes to the algorithm.

The format specification, C source code, and Matlab functions to generate and read MEF files are freely available (http://mayoresearch.mayo.edu/mayo/research/msel/) under the GNU open-source software license in the interest of facilitating widespread use of this file format. In addition, Neuralynx Inc recording equipment will soon be capable of saving recordings directly in MEF format (http://www.neuralynx.com).

V. Conclusion

Systems electrophysiology recordings sampled at high frequency from a large number of electrodes over a wide dynamic range are important in clinical and research neurophysiology. This paper describes a novel file format which incorporates data compression, encryption, 32-bit CRC, and a block index structure to help address the practical challenges of managing the massive data volumes generated with high spatiotemporal electrophysiology recordings. Range encoded difference compression reduces the size of data files, while increasing the speed at which recorded data is accessed. AES encryption with a 128-bit key meets most patient information privacy restrictions. The 32-bit cyclically redundant checksum is capable of detecting data corruption, and MEF's block-wise approach limits the effects of data errors. The MEF index table provides access to any arbitrary point in the recorded data without requiring all previous blocks to be decoded.

Acknowledgment

The authors acknowledge the contributions of Lammert Bies, Arturo Campos, Andrew Gardner, and PK Niyaz.

References

[1] Bragin A, Mody I, Wilson CL, Engel J Jr., Local generation of fast ripples in epileptic brain. *J Neurosci*. 2002 Mar 1;22(5):2012-21.

[2] Urrestarazu E, Chander R, Dubeau F, Gotman J. Interictal high-frequency oscillations (100-500 Hz) in the intracerebral EEG of epileptic patients. *Brain*. 2007 Sep;130(Pt 9):2354-66.

[3] Worrell GA, Gardner AB, Stead SM, et.al. High-frequency oscillations in human temporal lobe: simultaneous microwire and clinical macroelectrode recordings. *Brain*. 2008 Apr;131(Pt 4):928-37

[4] Van Gompel JJ, Worrell GA, Bell ML, Patrick TA, Cascino GD, Raffel C, Marsh WR, Meyer FB. Intracranial electroencephalography with subdural grid electrodes: techniques, complications, and outcomes. *Neurosurgery*. 2008 Sep;63(3):498-505

[5] Van Gompel JJ, Stead SM, Giannini C, Meyer FB, Marsh WR, Fountain T, So E, Cohen-Gadol A, Lee KH, Worrell GA. Phase I trial: safety and feasibility of intracranial electroencephalography using hybrid subdural electrodes containing macro- and microelectrode arrays. *Neurosurg Focus*. 2008 Sep;25(3):E23.

[6] Health Insurance Reform: Security Standards; Final Rule. *Federal Register* 2003 Feb. 68(34):8334-81.

[7] Kemp B, Olivan J. European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data. *Clin Neurophysiol*. 2003 Sep;114(9):1755-61.

[8] Kemp B, Värri A, Rosa AC, Nielsen KD, Gade J. A simple format for exchange of digitized polygraphic recordings. *Electroencephalogr Clin Neurophysiol*. 1992 May;82(5):391-3.

[9] Hellmann G, Kuhn M, Prosch M, Spreng M. Extensible biosignal (EBS) file format: simple method for EEG data exchange. *Electroencephalogr Clin Neurophysiol*. 1996 Nov;99(5):426-31.

[10] NIST, Federal Information Processing Standards Publication 197, November 2001. Announcing the ADVANCED ENCRYPTION STANDARD (AES). Springfield, VA: NTIS.

[11] Bodden E, Clasen M, and Kneis J. Arithmetic Coding in a nutshell. In *Proseminar Datenkompression* 2001. University of Technology Aachen, 2002.

[12] Martin, GNN. Range encoding: an algorithm for removing redundancy from a digitised message. *Video & Data Recoding Conference, Southampton*, 1979.

[13] Peterson, WW and Brown, DT. Cyclic Codes for Error Detection. Proceedings of the IRE. January 1961. 49: 228.

[14] Koopman, P. 32-Bit Cyclic Redundancy Codes for Internet Applications. *The International Conference on Dependable Systems and Networks* (June 2002). 459.

[15] Cranstoun SD, Ombao HC, von Sachs R, Guo W, Litt B. Time-frequency spectral estimation of multichannel EEG using the Auto-SLEX method. *IEEE Trans Biomed Eng*. 2002 Sep;49(9):988-96.