ECG DATA COMPRESSION FOR TAPELESS AMBULATORY MONITORS

George B. Moody, Kambiz Soroushian, and Roger G. Mark

Massachusetts Institute of Technology, Cambridge, MA, USA

Summary

Recent increases in the density of solid-state memory chips make it possible to construct ambulatory monitors which can record the ECG continuously without the use of tape. To record 2 channels of ECG for 24 hours, currently feasible devices will require that the ECG be compressed to an average of 200 bits/channel/second or less. We describe an ECG data compressor, called TRIM, which meets this goal. Each ECG signal is sampled at 120 Hz with 8-bit resolution, and digitally low-pass filtered to remove components above 40 Hz. TRIM selects a subset of these samples which allows accurate reconstruction of the filtered signal by linear interpolation between them. Significant turning points are selected as the initial output set. Segments of the signal which are bounded by the selected points are partitioned at intermediate points, until the maximum error of linear interpolation between the selected points is below an adaptive threshold. The augmented output set contains about 15 points per second, which are Huffman-encoded using about 10 bits per point.

Tapeless ambulatory monitors

It is now technically feasible to build a tapeless multichannel ECG recorder, powered by batteries, somewhat smaller and lighter than conventional Holter tape recorders. By recording the ECG digitally in solid-state memory, many of the inherent problems of analog recording and playback can be avoided. Low-frequency response, important for ST segment analysis, is limited only by the preamplifier in a digital recorder. Signal-to-noise ratio in a properly designed digital system is dependent only on the input signal and on the resolution of the analog-to-digital converter. Signal quality does not deteriorate once the recording has been made. The time base of a digital recording is stable; it suffers none of the problems associated with the mechanics of tape handling (wow, flutter, sticking, slipping, stretching, and breakage). Inter-channel skew, if present, is fixed and correctable. The recorder itself, with no need for moving parts, is mechanically robust and silent, requires little maintenance, and can be manufactured inexpensively (although, at mid-1987 prices, the cost of semiconductor memories was too high to permit construction of a tapeless monitor at the price of a conventional Holter recorder).

Analysis of the recorded ECG can be performed in real time if the recorder is equipped with a suitable microprocessor. Since the device provides a continuous record of the ECG, a major obstacle to clinical acceptance of real-time monitors is thereby avoided. Alternatively, high-speed off-line ECG analysis may be performed, using any of the conventional methods for Holter tape processing.

Power and size constraints suggest that first-generation tapeless ambulatory ECG monitors will have about 4 megabytes of solid-state memory. For a 24-hour two-channel recorder, this implies that the rate at which memory is filled should be 200 bits per channel per second (bpcs) or less, on average. Typical data acquisition rates are five times higher or more. An effective method for ECG data compression is therefore a necessity for a feasible tapeless ambulatory monitor.

ECG compression

The purpose of any data compression method is to represent its input in a more compact form, by reducing the amount of redundant information. Redundancy may arise from a lack of statistical independence between symbols in the input, or from an uneven distribution of symbols. In the ECG, both sources of redundancy are present, and may be removed using different methods.

Redundancy which arises from interdependencies among the input symbols is treated by recognizing patterns and encoding them efficiently. Typically, one begins with a model of the input. When the model is fitted to the data, its parameters describe most of the significant features of the data. We chose to investigate a relatively "low-level" model of the ECG, a piecewise-linear model. Higher-level models of the ECG, such as those based on identifying waveforms and representing them using templates or coefficients of orthonormal functions, do not appear to offer superior compression, and they tend to be noise-sensitive and computationally expensive.

The problem of removing redundancy which arises from uneven distribution of symbols has a well-defined solution from information theory. Huffman encoding[1] provides a method for reducing the average number of bits per symbol to a value arbitrarily close to the first-order source entropy in bits, defined as:

$$H(I) = -\sum_i p_i \log_2 p_i \qquad (1)$$

where p_i is the probability of observing symbol i in the input, and the sum is taken over all possible symbols.

0276-6574/88/0000/0467$01.00 © 1988 IEEE

The differences between the data and the fitted model are the residuals. If the model is a good one, the residuals will be distributed over a narrow range, and they will be encodable in fewer bits than the original data. If the model parameters and the residuals can be represented in a more compact form than the input, one has achieved "lossless" compression: the input can be reconstructed exactly from the compressed data.

ECG recording, unlike text storage, does not require "lossless" compression. It is sufficient to record only the parameters of the fitted model, provided that the fit is faithful to the original ECG to the extent that any differences have no bearing on the clinical interpretation. Thus the scope of compression methods which may be useful is very wide, but it is difficult to formulate an objective criterion to determine if a particular method is adequate.

One obvious method of assessing the adequacy of a compressor is to measure its residuals. A major flaw in this method is that the significance of an error is highly dependent on its context in the ECG. During rapidly-changing segments of the QRS complex, for example, errors as large as 200 μv are tolerable, since the visually important features are the extrema rather than the connecting segments, which the eye does not readily distinguish from straight lines. Figure 1 (trace C) illustrates the output of a compression method which is quite acceptable despite the large residuals during the QRS complexes.

In the baseline, however, residuals as small as 50 μv may be intolerable, since they may conceal P-waves. Trace E in figure 1 shows an ECG compressed using a method which strictly limits the size of the residuals. Although the residuals in trace F are much smaller than those in trace D, the compressed ECG is unacceptable, since P-waves are not visible. Thus clinically acceptable quality is neither guaranteed by low residuals nor ruled out by high residuals.

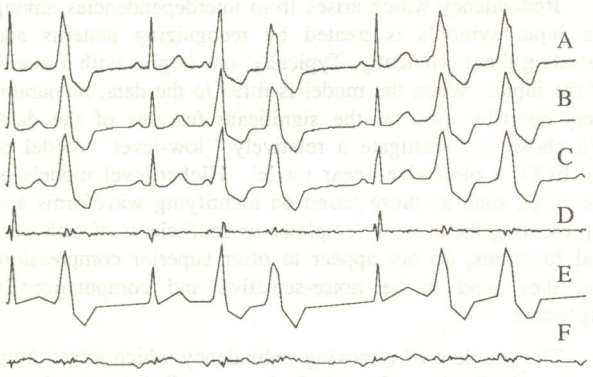

Figure 1. Reference signal, compressed ECGs, and residuals. Trace A: original ECG from the MIT-BIH database; B: low-pass filtered, 8-bit, 120 samples/sec reference signal; C: output of a clinically acceptable compressor with high residuals; D: residual error signal (C-B); E: output of a clinically unacceptable compressor with low residuals; F: residual error signal (E-B). The method used in C is turning point compression, using a fixed ΔT for reconstruction. In E, the method is recursive partitioning compression, using a fixed threshold.

Many ECG data compression schemes have been reported previously; few approach 200 bpcs. Several investigations were aimed at real-time 2- or 3-lead telephone transmission, which (using 2400 baud synchronous modems available at the time) becomes feasible at rates of 1200 or 800 bpcs respectively[2, 3]. The well-known AZTEC method[4], used in conjunction with Huffman encoding, can achieve the necessary data rates, but the quality of the reconstructed signal (given the amount of compression required) is clinically unacceptable. Our own earlier work[5] produced clinically acceptable compressed data at 400 bpcs without the use of Huffman encoding. As we began the present study, then, our goals were to achieve at least a twofold reduction with respect to the best previously reported result, and to improve the quality of the compressed signal in order to make it usable for automated analysis as well as visual review.

The divergent requirements of automated analysis and visual review can be understood in terms of the aims of each. Automated ECG analysis for identification of arrhythmias and ischemic change is dependent on details of the QRS complexes and the ST segment; the quality of the baseline is irrelevant to algorithms which do not attempt to detect P-waves. Thus AZTEC has been a useful tool for arrhythmia analysis because the details which it records are those which are needed for that purpose. For visual review, however, the smallest details of QRS morphology are rarely significant, while the appearance of the baseline is of paramount importance for the identification of P-waves, an essential part of a comprehensive visual analysis of rhythm.

The reference signal

Common practice in long-term ECG analysis has converged on use of 8-bit analog-to-digital converters, having an input range of about 10 millivolts, operated at rates of about 120 samples per second. Typically, the antialiasing filters needed at such sampling rates remove frequencies above 40 to 50 Hz from the ECG. We adopted these specifications for the reference signals which were presented as input to the compressors, since clinical experience with commercial systems has shown that ECGs obtained in this way retain the information needed for accurate identification of arrhythmias and ischemic change.

In the present study, we used reference signals derived from the MIT/BIH and AHA arrhythmia databases, as well as additional signals obtained by high-speed digitization of Holter recordings. In each case, we obtained the reference signal from a higher resolution signal by digital low-pass filtering, rescaling the data to 8 bits, and decimation (to 120 Hz for the MIT/BIH tapes, and to 125 Hz for the others). Pilot studies indicated no significant difference between results obtained using a cosine-window (Hanning) filter and a rectangular-window (moving average) filter; the rectangular filter was chosen because of its computational efficiency.

Trace B of figure 1 illustrates the appearance of the reference signals using an example from the MIT/BIH database. Low-pass filtering occasionally improves the signal quality visibly, by removing EMG noise; in most cases, however, the most obvious effect is a slight reduction in the

amplitude of the Q- and S-waves, and sometimes the R-waves as well, and loss of high-freqency detail in the QRS complex. On standard chart recordings, quantization of the signal is not visible (at 100 μv/mm, one amplitude step is about 0.4 mm, less than the width of the trace made by a thermal "pen").

Turning point compressors

Since (as noted above) the visually important features of the ECG are the extrema, compressors which accurately represent extrema are of interest. Mueller described a turning point compressor (TPC)[6] which performs 2:1 compression given any input data rate by selecting for output one sample from each pair of consecutive input samples. The second sample of each pair is chosen unless the first is a turning point (i.e., a local extremum). Mueller suggested applying the TPC to its own output to achieve 4:1 compression, and found that by doing so, acceptable results could be obtained with an output rate of 50 samples per second.

The TPC may be trivially extended to produce any desired compression ratio[5], as illustrated in traces B, C, and D of figure 2. At high compression ratios, however, waveforms may appear broadened, since the fixed output rate of the TPC does not permit retention of the higher frequency components of the signal. As a result, it is difficult to identify conduction defects in overcompressed TPC output. A more serious consequence is that waveforms which should appear nearly identical often look very different as a result of small amounts of noise. A major strength of the turning point compressor is that the appearance of the baseline is excellent, even at high compression ratios. Low-amplitude P-waves remain visible, a characteristic not shared by compressors such as AZTEC which define a lower limit of "significant" amplitude variation.

Figure 2. Comparison of compression methods. Trace A: reference signal; B, C, and D: TPC output at 40, 30, and 15 samples per second; E, F, and G: RPC output at 25, 18, and 15 samples per second (mean rate); H, I, and J: TRIM output at 20, 15, and 12 samples per second (mean rate).

Turning point compressors are computationally very simple. The output data rate is independent of the characteristics of the input signal. Since (for N:1 compression) one sample is saved from each group of N, the time of each output sample is known to within ±N/2 sample intervals. For purposes of visual analysis, then, it may not be necessary to save any explicit time information with each sample, thereby allowing considerable savings in storage requirements as compared to other compression techniques with non-uniform output data rates.

Recursive partitioning compressors

Another approach to deriving a piecewise linear approximation of the ECG is recursive partitioning compression[7] (RPC), illustrated in traces E, F, and G of figure 2. An arbitrarily chosen segment of ECG is initially represented by linear interpolation between its endpoints. At the intermediate point most distant from the interpolated line, the segment is partitioned into two shorter segments. This procedure is applied recursively to each of the shorter segments, until the maximum error of the linear interpolation lies below a threshold.

A strength of the RPC method is that the error of the representation is strictly bounded, but (as a result) the compression ratio is data-dependent. The appearance of the QRS complexes is excellent using the RPC; P-waves, however, are often concealed if they lie below the threshold. It is always possible to lower the threshold, but the compression ratio decreases as well.

By varying the threshold inversely with the segment length, considerable improvement in the baseline representation can be achieved. In this way, a reasonable QRS representation can be obtained with a modest number of short segments, while long baseline segments are partitioned until high accuracy is obtained.

The dependence of compression ratio on signal quality can be limited by varying the threshold directly with the mean output data rate. The effect of such a variation is to make the storage requirements more predictable, at a cost of providing a less accurate representation of noisy data than of clean data.

The TRIM compressor

The strengths of the turning point and recursive partitioning compressors may be combined by using a TPC-based method to determine the initial partitioning, and then by refining the representation using the RPC. The TRIM (*t*urning point/*r*ecursive *im*provement) compressor uses this strategy to achieve clinically acceptable yet highly compressed output, illustrated in traces H, I, and J of figure 2.

TRIM begins by finding all of the turning points in the signal. (This process differs from that used by the stand-alone TPC, in that there is no fixed number of input samples per output sample.) The "significance" of each turning point is calculated as:

$$S = k_1 \, |\Delta T_{\min}| + k_2 \, |\Delta V_{\min}| - 1 \qquad (2)$$

where $\Delta T_{\min}$ is the time interval to the nearer of the two

adjacent turning points, and ΔV_{min} is the voltage difference with respect to the less different of the two adjacent turning points. Suitable values for k_1 and k_2 are on the order of 10 sec^{-1} and 10^6 v^{-1} respectively. Those turning points for which $S > 0$ and $\Delta V_{min} \geq 80$ μv are deemed "significant", and form the initial output set (typically 5-10 points per second).

The next phase of TRIM uses the RPC approach to improve the representation of the segments delimited by the significant turning points. The tolerance, δV, is a function of both the length of the segment to be partitioned, ΔT, and the median output rate:

$$\delta V = k_3 \cdot \frac{\overline{\Delta T} + \hat{\Delta T}}{\Delta T \cdot \overline{\Delta T}} \qquad (3)$$

where $\overline{\Delta T}$ is the observed median segment length (the reciprocal of the median instantaneous output rate), $\hat{\Delta T}$ is the expected median segment length (typically about 30 msec), and k_3 is adjusted to produce the desired compression ratio. We obtained excellent signal quality using $k_3 = 40$ volt-seconds. The additional points selected using the RPC are merged with the significant turning points to form the final output set of selected points (typically about 15 per second).

Finally, the voltage and time intervals between the selected points are Huffman-encoded using independent code dictionaries. (Since the output data rate is not fixed, time intervals must be recorded.) The time intervals typically require 4 bits on average, while the voltage intervals need about 6 bits.

Results and discussion

Figure 3 shows the reduction in redundancy which is accomplished by TRIM. The results summarized in the figure were obtained by dividing the MIT-BIH arrhythmia database into 960 one-channel segments, each three minutes in length. The first-order source entropy was determined using equation (1) for several different representations of each of the 960 segments. These measurements were multiplied by the number of points per second to arrive at ideal data rates in bpcs. The box plots summarize the distributions of data rates for each representation.

The database in its original form (360 samples/second with 11-bit resolution) was used to obtain the top group of four distributions, which show the data rates for amplitude and first, second, and third difference representations. Correlation between successive samples tends to keep first differences small, and second differences are clustered somewhat more tightly. Third- and higher-order differences are less efficient representations.

The next group of four distributions shows the effect of low-pass filtering the data. There is essentially no effect on the data rates for amplitudes, but there are noticeable effects on those for difference representations. The third group of distributions shows the effect of lowering the resolution and the sampling rate. Although second differences are most efficient at a sampling frequency of 360 Hz, they are less clearly favored at 120 Hz, which reflects the increased likelihood of a substantial change in slope over the longer sampling interval. The worst-case data rate is about 400 bpcs for second differences, which makes them an ideal choice if "lossless" compression is required.

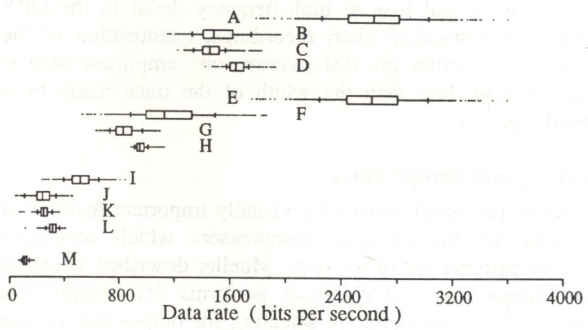

Figure 3. Distributions of data rates required for various representations. For each distribution, the box plot indicates the 10th, 25th, 50th, 75th, and 90th percentiles, and all outliers. Sets A-D: Original MIT-BIH database segments (A: 11-bit amplitudes; B, C, and D: first, second, and third differences). Sets E-H: Low-pass filtered data (E: 11-bit amplitudes; F, G, and H: first, second, and third differences). Sets I-L: Reference signals (I: 8-bit amplitudes; J, K, and L: first, second, and third differences). Set M: TRIM output.

Finally, the single distribution at the bottom of figure 3 shows the measured data rates for TRIM compression. The worst-case data rate is 182 bpcs; 90% of the segments required below 145 bpcs, and the median rate was only 109 bpcs, well below the 200 bpcs goal.

References

1. D A Huffman, "A method for construction of minimum-redundancy codes," *Proc IRE* **40**, pp.1098-1101 (1952 (Sep)).
2. M Bertrand, R Guardo, F A Roberge, and P Blondea, "Microprocessor application for numerical ECG encoding and transmission," *Proc IEEE* **65**, pp.714-722 (1977 (May)).
3. U E Ruttimann and H V Pipberger, "Compression of the ECG by prediction or interpolation and entropy encoding," *IEEE Trans Biomed Eng* **BME-26**(11), pp.613-623 (1979 (Nov)).
4. J R Cox, F M Nolle, H A Fozzard, and G C Oliver, Jr, "AZTEC, a preprocessing program for real-time ECG rhythm analysis," *IEEE Trans Biomed Eng* **BME-15**(4), pp.128-129 (1968 (Apr)).
5. R G Mark, G B Moody, W H Olson, S K Peterson, P S Schluter, and J B Walters, Jr, "Real-time ambulatory arrhythmia analysis with a microcomputer," *Computers in Cardiology* **6**, pp.57-62 (1979).
6. W C Mueller, "Arrhythmia detection program for an ambulatory ECG monitor," *Biomed Sci Instrum* **14**, pp.81-85 (1978).
7. P S Schluter, in *The Design and Evaluation of a Bedside Cardiac Arrhythmia Monitor*, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA (1981). [Ph.D. thesis]