EVALUATION OF THE "TRIM" ECG DATA COMPRESSOR

George B. Moody, Roger G. Mark, and Ary L. Goldberger

Massachusetts Institute of Technology, Cambridge, MA, and
Beth Israel Hospital, Boston, MA, USA

Summary

Data compression for storage of long-term ECGs is increasingly common, due to the recent introduction of devices which perform the functions of Holter ECG recorders using solid-state memory rather than tape. Often, the quality of compressed ECGs obtained from these devices is poor. It is essential to evaluate ECG data compressors for clinical acceptability. We present a method for performing such evaluations, and a case study in which we evaluate the TRIM compressor. We conclude that TRIM, when adjusted to produce a mean output rate of 200 bits per channel per second, does not measurably influence the accuracy of clinicians' diagnoses with respect to cardiac rhythm or conduction. Our evaluation methodology is general, and suitable for standardized testing of ECG data compressors.

The need to evaluate ECG data compressors

Most of the early work on ECG data compression was motivated by an interest in real-time transmission of ECGs by telephone[1], or by the possibility of reducing the computational requirements of an ECG analysis program by preprocessing its input to reduce the data rate[2]. Current interest in the subject has been stimulated by the prospect of long-term digital ECG recording using solid-state memory rather than magnetic tape.

Power and size constraints limit the memory capacity of a tapeless ambulatory monitor to about 4 megabytes. If such a device is to perform the function of a typical Holter tape recorder, which can record two channels of ECG for 24 hours, memory must be filled at a rate of 200 bits per channel per second (bpcs) or less. By contrast, rates of 1200 bpcs and more can be used for telephone transmission of ECGs. New compression methods have been developed to achieve the significantly lower data rates needed for tapeless ambulatory monitors. One such method is the TRIM compressor we have described previously[3].

TRIM, and other compression methods for this application which have been described or put into practice, do not represent the original digitized input signal exactly. Rather, they provide approximations to the original input, with quality varying from adequate to poor. Since compressors modify the characteristics of the ECG which is to be interpreted, it is essential to establish the extent to which compression influences the accuracy of clinicians' diagnoses. Clearly, any compression method which adversely affects diagnostic accuracy is of questionable value at best.

It is not enough to assert that any less than perfect representation is unacceptable, however, because *any* ECG recording is less than perfect. Neither is it defensible to assert that a given compression method is adequate because its root-mean-square errors are small. As we have observed[3], large amplitude errors during rapidly-changing portions of the QRS complex are more tolerable than much smaller errors in the baseline which may conceal or mimic P-waves. For this reason, simple criteria based on measuring the residual errors of ECG data compressors fail to distinguish acceptable from unacceptable compression.

We believe that the best way to assess the clinical acceptability of a compressor is to compare diagnoses made using compressed and uncompressed versions of the same data. In the present work, we describe a method for making such an assessment, and present the results of an evaluation of the TRIM compressor.

A database for evaluating ECG data compressors

We have compiled a database of 168 two-channel ECG strips from 38 carefully selected Holter recordings for use in evaluating ECG data compressors. Each strip is 20 seconds long, and was digitized in accordance with the specifications for the AHA database[4] (250 samples per second per channel, with 12-bit resolution over a ±10 mV range, bandpass-filtered from 0.1 to 100 Hz to limit analog-to-digital converter saturation and for anti-aliasing). Strips were chosen to include a wide variety of complex atrial, AV junctional, and ventricular arrhythmias, conduction disturbances, and noise, in order to challenge the ability of a compressor to retain the subtle features of the ECG needed for accurate diagnosis.

Digital ECG recording offers potentially higher quality than can be obtained from Holter tape recording, since the former method is not subject to wow, flutter, and poor signal-to-noise ratio and low-frequency response. Nevertheless, we selected excerpts from Holter recordings for use in evaluating compressors, principally because the availability of over 14,000 two-channel Holter recordings in the library of the Beth Israel Hospital Arrhythmia Laboratory made it possible to obtain a wide variety of unusual but clinically important waveforms which could not easily have been obtained

0276-6574/89/0000/0167$01.00 © 1989 IEEE

prospectively using digital recording. It may be argued that using Holter recordings as references introduces a bias against compressors, since the cumulative degradation of signal quality due to the analog recording and playback process and the compressor may be unacceptable even if either process by itself would be acceptable. If such a bias exists, it may be considered to provide a margin of safety in the evaluation, making the test somewhat more stringent than strictly necessary.

Methods

If diagnoses were not subject to inter- and intra-observer variability, one might readily measure the effect of compression. It would suffice to present the database in compressed form to observer A, in uncompressed form to observer B, and then to compare their diagnoses. Any discrepancies could be attributed to compression. If observers A and B are actually copies of the same ECG analysis program, this assumption is reasonable, and such an experiment is trivial to conduct. The more clinically relevant question, however, is to measure the effect of compression on diagnoses made by clinicians.

Inter-observer variability in clinicians' diagnoses complicates the experimental design, since a discrepancy may be due to a genuine difference of opinion between observers. Intra-observer variability introduces further complications; a discrepancy may result from a simple error of omission, for example. Inter-observer variability may be estimated by presenting both observers with recordings in the same format (either compressed or uncompressed). Similarly, one may estimate intra-observer variability by presenting each observer with the same recordings more than once. Based on multiple readings, it may be possible to arrive at a set of reference diagnoses which can be accepted as correct. Inter- and intra-observer variability can then be measured directly by comparing individual diagnoses against the reference set, and counting false positive and false negative diagnoses. These experimental refinements present further problems of their own: they lengthen the experiment, and they introduce the possibility that successive diagnoses will be made with the help of information recalled from a previous reading.

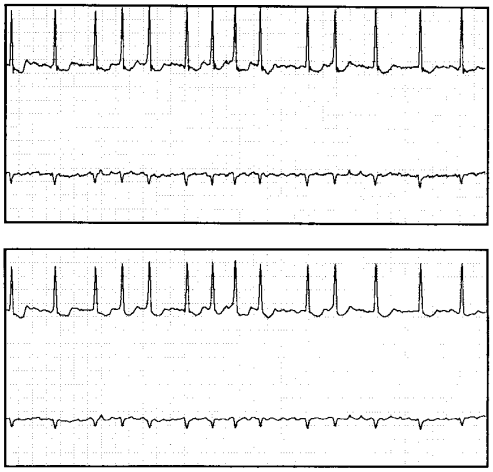

Figure 1. Examples of uncompressed (above) and compressed (below) ECG strips. The strips used in the test were printed at standard ECG scales (25 mm/sec, 10 mm/mV) by a laser printer; a diagnostic checklist (see table 1) was printed on the reverse side of each sheet.

Two of the authors (RGM and ALG) served as the expert observers for the present study. The experiment consisted of six tests (three for each observer). The database of 168 strips was arranged in random order and divided into 24 sets of seven strips each. For each of the six tests, fourteen of these sets (i.e., 98 strips) were selected. Seven sets of strips were printed in uncompressed form, and the other seven sets were TRIM-compressed to an average of 200 bpcs and printed in the same format as the uncompressed strips (figure 1). A checklist of diagnoses (table 1) was printed on the back of each sheet. The 98 selected strips were rearranged in random order, numbered serially for reference, and collected in a notebook.

The expert observers did not participate in the preparation of the data as discussed above. Each received one notebook at a time, which he annotated and returned. The observers were informed only that they would find a given strip no more than twice over the entire experiment, and not more than once in any given notebook; that each notebook would contain 49 compressed and 49 uncompressed strips in random order; and that each notebook would contain a significant number of strips which would not appear in the other notebooks they would be given. The intent was to instill in them a reasonable doubt about having seen a strip before, in order to limit any tendency to rely on memory. Over a period of five months, each observer annotated three notebooks, each with a different selection of strips. The observers reported that they spent on the order of ten hours annotating the 98 strips in each notebook. This effort was generally distributed over a week or more.

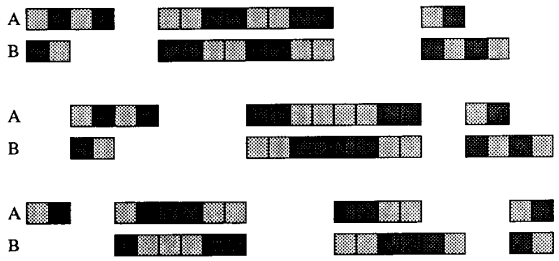

Figure 2. Allocation of sets of strips during the six tests. Each row of boxes represents one test; the observer (A or B) for each test is noted at left. Each column of boxes represents a unique set of seven strips. Dark boxes designate sets of compressed strips, light boxes represent sets of uncompressed strips, and missing boxes indicate sets of strips which were not included in a given test. Strips were rearranged in random order after being allocated according to this scheme. The observers were asked not to compare notes; strips which were in circulation simultaneously (indicated by closely spaced rows above) were never identical. Five months elapsed between the first pair of tests, at the top of the figure, and the last, at the bottom.

	Observer A					Observer B				
Diagnostic category	*Se(U)*	*N*	*Se(C)*	*N*	*p*	*Se(U)*	*N*	*Se(C)*	*N*	*p*
All diagnoses	67.70	613	65.53	618	0.515	70.66	610	67.33	603	0.116
Sinus rhythm	87.93	116	83.76	117	0.408	93.91	115	92.92	113	0.928
Sinus arrhythmia	22.00	50	19.05	42	0.906	56.10	41	61.22	49	0.813
Sinus bradycardia (HR < 60 bpm)	75.00	28	85.19	27	0.386	21.43	28	40.00	25	0.035
Sinus tachycardia (HR > 100 bpm)	75.00	8	87.50	8	0.718	0.00	7	0.00	7	1.000
Sinus pause or arrest	76.92	13	55.56	9	0.297	55.56	9	63.64	11	0.904
Ectopic atrial or junctional rhythm	82.35	17	70.00	10	0.604	0.00	9	0.00	15	1.000
Wandering atrial pacemaker	70.00	10	44.44	9	0.240	70.00	10	33.33	12	0.006
Atrial premature complexes, normally conducted	69.23	65	54.10	61	0.003	82.14	56	75.36	69	0.407
Atrial or junctional premature complexes, non-conducted	25.00	8	50.00	8	0.334	100.00	7	60.00	10	0.001
APCs with aberrant intraventricular conduction	55.00	20	57.14	21	0.986	73.68	19	69.57	23	0.935
Atrial tachycardia (regular, sustained, 1:1)	20.83	24	5.88	17	0.084	64.71	17	50.00	26	0.385
Atrial tachycardia, multifocal	50.00	10	62.50	8	0.803	14.29	7	12.50	8	0.993
Supraventricular tachycardia	77.08	48	80.65	31	0.890	87.10	31	73.47	49	0.038
Atrial flutter	50.00	8	33.33	9	0.665	87.50	8	66.67	6	0.435
Atrial fibrillation	68.18	22	84.21	19	0.169	61.90	21	83.33	24	0.011
AV junctional escape complexes or escape rhythm, slow (HR < 60 bpm)	77.78	27	60.71	28	0.067	80.00	30	70.83	24	0.553
AV junctional rhythm, accelerated (60 bpm < HR < 110 bpm)	50.00	4	0.00	2	0.253	50.00	2	66.67	3	0.937
Ventricular premature complexes, uniform	90.48	21	96.15	26	0.547	76.00	25	81.82	22	0.821
Ventricular premature complexes, multiform	68.42	19	69.57	23	0.995	66.67	24	73.68	19	0.814
Ventricular premature complexes, in pairs (2 consecutive)	0.00	7	0.00	8	1.000	100.00	8	100.00	6	1.000
Ventricular tachycardia	100.00	2	100.00	7	1.000	100.00	6	100.00	2	1.000
Ventricular tachycardia, polymorphic	42.86	7	70.00	10	0.251	87.50	8	62.50	8	0.224
Accelerated idioventricular rhythm (60 bpm < HR < 100 bpm)	100.00	1	100.00	1	1.000	0.00	1	0.00	1	1.000
Ventricular escape complexes or rhythm (HR < 60 bpm)	100.00	1	0.00	1	-	0.00	1	100.00	1	-
Ventricular fibrillation	50.00	2	100.00	2	0.423	100.00	2	100.00	2	1.000
AV block, 1st degree	84.00	25	79.41	34	0.847	63.64	33	86.36	22	0.001
AV block, 2nd degree, Mobitz type I (Wenkebach)	50.00	2	71.43	7	0.803	88.89	9	100.00	2	0.842
AV block, 2:1, 3:1, 4:1	66.67	3	66.67	3	1.000	0.00	2	50.00	2	0.423
AV block, 3rd degree	-	0	50.00	2	-	100.00	4	-	0	-
AV block, varying	-	0	0.00	1	-	50.00	2	-	0	-
Short P-R interval (with sinus rhythm and normal QRS duration)	100.00	1	66.67	3	0.826	50.00	2	0.00	1	0.795
Pre-excitation (Wolff-Parkinson-White syndrome)	100.00	1	100.00	4	1.000	75.00	4	0.00	1	0.170
AV dissociation	100.00	2	75.00	4	0.680	33.33	3	0.00	2	0.591
RBBB, fixed	100.00	3	40.00	5	0.015	33.33	3	0.00	2	0.591
RBBB, intermittent	-	0	0.00	1	-	100.00	2	-	0	-
LBBB, fixed	33.33	3	40.00	5	0.979	28.57	7	33.33	3	0.986
LBBB, intermittent	25.00	4	0.00	4	0.356	100.00	4	50.00	4	0.024
Intraventricular conduction disturbance (IVCD)	85.00	20	86.21	29	0.989	77.42	31	47.37	19	<0.001
Aberrant intraventricular conduction with supraventricular arrhythmia	63.64	11	50.00	12	0.690	8.33	12	10.00	10	0.987

Table 1. The 39 diagnostic categories in the first column were used to describe the strips in the study. They are a subset of a checklist of 60 possible diagnoses printed on the back of each strip; the other 21 were not used. For each observer, and each diagnostic category, the table shows sensitivity for uncompressed and compressed data, $Se(U)$ and $Se(C)$, in percentage units (see text). The p-values indicate the significance of the difference between $Se(U)$ and $Se(C)$ for each observer.

Figure 2 summarizes the allocation of sets of strips between the six tests. Ninety-eight of the 168 strips were annotated four times, once by each observer in uncompressed and compressed forms. Of these, the compressed form was presented first in half of the cases. The remainder of the strips were annotated three times (twice by one observer and once by the other); in half of these cases the compressed form was annotated twice, and in the others, only once. Although no markings distinguished the compressed strips from the others, it was not difficult to identify which strips had been compressed. The observers shared a subjective impression that the compression did not hinder their analysis. We did not consider it necessary to disguise the compressed strips in order to avoid bias, although this can be accomplished quite convincingly by adding a small amount of noise to the compressed ECG.

The annotations were compared to obtain reference annotations. Obvious errors of omission were corrected at this stage (for example, if a diagnosis was written on the face of the strip but not ticked on the checklist). All discrepancies relating to ventricular ectopy were resolved by consensus of the authors. The reference annotations include 806 positive diagnoses, and 115 unresolved discrepancies which arise from conflicting annotations of the uncompressed strips. The remaining discrepancies appear to be genuine differences of opinion between the observers. More than half of these relate to atrial ectopy. Both observers remarked on the difficulty of reading many of these strips in either compressed or uncompressed form, and in some cases, it was impossible to reach a conclusion. The statistical analysis of the results in the following section excludes diagnoses related to these unresolved discrepancies.

Results and discussion

There were essentially no false positive diagnoses made on either the compressed or the uncompressed strips, so it can be concluded that the compressor which we evaluated has no measurable influence on diagnostic specificity. The remainder of the discussion will address the distribution of false negative diagnoses.

Certain diagnoses were inferred from context: for example, it was assumed that sinus rhythm would have been diagnosed if any of the related diagnoses (sinus arrhythmia, sinus bradycardia, or sinus tachycardia) had been checked. These and similar cases were not counted as false negatives.

We determined sensitivity for each of 39 diagnostic categories, for each observer, for compressed and uncompressed data (thereby obtaining $39 \times 2 \times 2 = 156$ sensitivity measurements). Sensitivity for a given diagnostic category was defined as the ratio of the number of true positive diagnoses to the total number of positive diagnoses (i.e., the sum of true positives and false negatives), expressed as a percentage. As noted above, cases in which a positive diagnosis was not established as correct were excluded from these calculations. These measurements ranged from 0% to 100%, for both compressed and uncompressed data, for both observers. Significant intra-observer variability with respect to the use of the diagnostic checklist made it undesirable to pool detection statistics from the two observers. One observer, for example, neglected to note the presence of ventricular couplets on any of the strips he annotated; the other observer annotated all of them. Sensitivity for all diagnoses together showed no significant differences between observers or between compressed and uncompressed data ($p > 0.1$).

For human experts analyzing uncompressed ECGs, these statistics may appear surprisingly poor. They reflect the very considerable difficulty of the database, which was selected to contain many subtle rhythm and conduction abnormalities. In a number of cases, it appears that less clinically important features were overlooked in the context of correctly diagnosed, more serious or complex arrhythmias. Furthermore, the test requires diagnoses to be made on isolated 20-second rhythm strips. In clinical practice, such constraints are rarely rigid; ambiguous cases can usually be clarified by obtaining more data. In this sense, the test imposes a somewhat unrealistic demand on the clinician.

The TRIM compressor has no obvious influence on diagnostic sensitivity. To compensate to some extent for inter-observer variability, sensitivity measurements on compressed and uncompressed data were paired for each observer. We determined the significance of the difference in each pair of sensitivity measurements using a *t*-test. Table 1 shows these results.

Given that there were 78 such pairs, one may expect to find one or two differences below the $p = 0.02$ level due to statistical fluctuation alone. In fact, we found seven such cases. In two of these, sensitivity was higher using compressed data than uncompressed data, for atrial fibrillation and for 1° AV block. In the remaining five cases, sensitivity was higher using uncompressed data, for wandering atrial pacemaker, normally conducted APCs, non-conducted APCs, fixed right bundle branch block, and intra-ventricular conduction defects. In each case, there is no significant difference in the sensitivity of the other observer ($p > 0.2$).

Given that none of these differences are common to both observers, it is reasonable to suggest that they are dependent on random variations in the difficulty of the particular strips which were annotated in each case. There is considerable overlap between the data sets analyzed by the two observers, and between the sets of compressed and uncompressed strips analyzed by each observer, but the sets are not identical. It seems unlikely that the compressor is responsible for the differences, since one should expect that it would have a similar influence on both observers if this were the case. In fact, the results are consistent with the hypothesis that the differences are independent of the compressor, since (as this hypothesis would predict), in three cases the differences for the other observer favor the compressed data, and in four cases, the uncompressed data.

It is nevertheless possible that one or more of these results reflects an influence of the compressor on the diagnoses. In particular, those with relatively large n are of interest. It should be noted with respect to IVCD that observer B had a higher sensitivity than did observer A, and that the best results (by an insignificant margin) were actually obtained using compressed data. More specific diagnoses (e.g., right bundle branch block) appear at such low frequencies that it is difficult to draw a conclusion, but (given the design of TRIM) it is possible that this method of compression may affect one's ability to distinguish between types of IVCD on short strips, although not in the long run.

The observations with respect to 1° AV block and atrial fibrillation, if they are not due to chance fluctuation, may be related to the low-pass filtering performed by TRIM, which tends to produce relatively clean baselines with easily recognizable P-waves. In contrast to many of the noisy uncompressed strips, the TRIM-compressed strips may be somewhat more readable. This experience is unlikely to apply if, as is the more usual case, one avoids noisy ECGs rather than seeking them out.

References

1. 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)).
2. 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)).
3. G B Moody, K Soroushian, and R G Mark, "ECG data compression for tapeless ambulatory monitors," *Computers in Cardiology* **14**, pp.467-470 (1987).
4. R E Hermes and G C Oliver, "Use of the American Heart Association database," pp. 165-181 in *Ambulatory Electrocardiographic Recording*, ed. N K Wenger, M B Mock, and I Ringqvist, Yearbook Med Pub, Chicago (1980).