DEVELOPMENT AND EVALUATION OF A 2-LEAD ECG ANALYSIS PROGRAM

George B. Moody and Roger G. Mark

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

Summary

Simultaneous analysis of two ECG leads offers the possibility of reducing the rate of PVC classification error below that of single-channel analysis. We have developed a two-channel analysis program which on the MIT/BIH ECG database demonstrates a 42% reduction in the number of PVCs misclassified as non-PVCs, and significant improvement in detection of ventricular couplets and runs, when compared to a single-channel analysis. The technique is a straightforward extension of the feature extraction and clustering approach used in the single-channel program. Features from both channels are used in cluster definition, a technique which avoids having to arbitrate between decisions which have been made for each channel independently. A linear discriminant function of noise level and signal-to-noise ratio is used to recognize periods during which signal quality in one channel is too poor for reliable two-channel analysis; during these periods, single-channel analysis is performed on the better channel.

In human analysis of long-term ECG recordings, it has been accepted that a second lead is highly useful. The redundancy thereby obtained permits analysis to continue when, as occasionally happens, one electrode set fails. Motion artifact which mimics PVCs can often be recognized by its absence on a second lead. PVCs which closely resemble normal beats in one lead are often distinctly different from normals in another lead. Beats which are isoelectric in one lead are usually more prominent in another lead, particularly if the second lead is nearly orthogonal to the first.

One should therefore expect a well-designed computer program for two-channel analysis to exhibit better noise immunity, more reliable QRS detection, and greater accuracy in PVC identification than an analogous single-channel program. In the analysis of the ECG in figure 1, for example, the slow, bidirectional ventricular tachycardia which begins with the second beat from the left might easily be misdiagnosed as bigeminy by a program analyzing only one channel. (This has been observed to be the case for our single-channel program.) A two-channel program, however, should have little difficulty with this example.

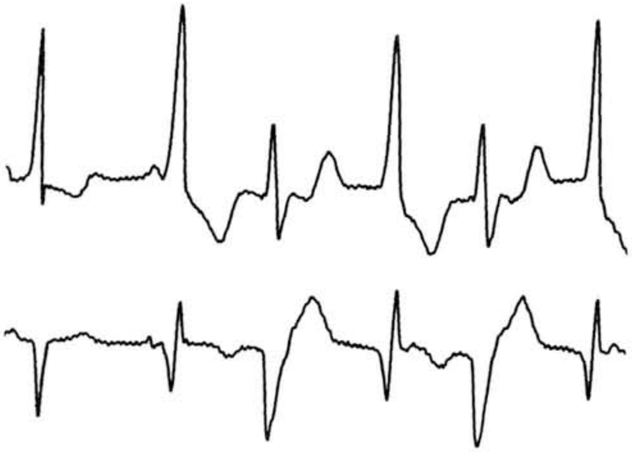

Figure 1. An excerpt from MIT/BIH ECG database tape 223 (17:21 from the beginning). The first beat is typical of the normals, the remainder are PVCs at a rate of 100-105 bpm. Note the similarity in the upper channel between the normal beat and the first, third, and fifth PVCs; and in the lower channel between the normal beat and the second and fourth PVCs. Although the ST segments and T-waves are quite different from the normals, the QRS complexes are similar enough to be confused, especially in the context of moderate noise, ST segment changes, frequent fusion PVCs, and axis shifts.

Since noise, when it occurs, is frequently confined to a single channel, a two-channel analysis should be able to use the alternate channel to continue analysis without interruption. For this technique to be successful, however, a reliable method of assessing signal quality is required; otherwise, the likely result will be worse performance compared to a single-channel analysis. Because of the statistical independence of noise, a greater fraction of the monitoring period may be expected to be noise corrupted in at least one channel than in any single channel. For example, if a two-channel analysis of the ECG in figure 2 failed to recognize the deterioration in signal quality on the lower channel which occurs near the third beat, erroneous beat labels might easily be generated as a result of reliance on poor data. An analysis which ignored the lower channel after the third beat would be unlikely to err.

0276-6574/83/0000/0039$01.00 © 1982 IEEE

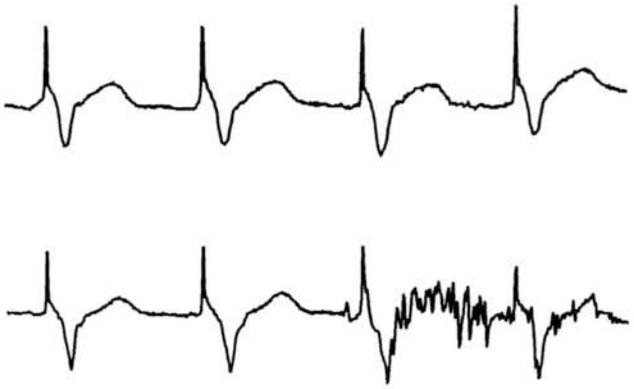

Figure 2. An excerpt from MIT/BIH database tape 104 (8:22 from the beginning). The rhythm is paced at 72 bpm. Using the signal quality LDF described in the text, analysis mode is switched from two-channel to analysis of the upper channel only after the third beat.

The single-lead ECG analysis program selected as the basis of this study was developed during the past five years in our laboratory[1].

The QRS detector uses a matched filter[2], dynamic threshold adjustment, and a look-back procedure to reduce false negatives. It operates without feedback from later processing stages, and is considered to be of sufficient accuracy that no further artifact rejection logic is required. Table 1 summarizes the results of an evaluation of the QRS detector operating on channel 1 of the MIT/BIH ECG database and on the 25 AHA database tapes which were available to us. QRS sensitivity, as used in the table, is defined as the fraction of QRS complexes which are detected; QRS positive predictivity, or positive predictive accuracy, is defined as the fraction of total detections which are genuine QRS complexes.

A simple estimate of noise level, N, is derived from a segment of the ECG, v(t), selected to begin at a time t_0 midway between R-wave peaks:

$$N = \frac{1}{n}\sum_{i=1}^{n} |v(t_0+i\triangle t) - v_{pred}(i)|$$

where $\triangle t$ is the sampling interval, n is a constant chosen so that the estimate will be sensitive to noise close in frequency to the QRS complex, and

$$v_{pred}(i) = v(t_0)+\frac{i}{n}(v(t_0+n\triangle t) - v(t_0))$$

N is thus the mean absolute error of a linear interpolation between the endpoints of the segment; the RMS amplitude of the signal over the same interval is a lower bound on N.

Table 1: Single-channel QRS detector evaluation

Database	Gross			Average	
	QRS Se	QRS +P		QRS Se	QRS +P
MIT/BIH(1)	99.77%	99.80%		99.77%	99.81%
AHA(2)	99.77%	99.87%		99.82%	99.90%
	Missed beats			False detections	
	N	V	F		
MIT/BIH(1)	163	73	5	212	
AHA(2)	134	5	0	81	

(1) Excluding ventricular flutter and 50-beat learning periods at the beginning of each tape.

(2) 25 tapes: 1001-1009, 2001-2004, 2006-2010, 3001-3007.

Column headings: Se = sensitivity, +P = positive predictivity (defined in text). "Gross" statistics are derived from totals of all beats in the data set; "average" statistics are the means of tape-by-tape statistics. Under "missed beats", "V" and "F" are PVCs and fusion PVCs respectively; "N" includes all other beats.

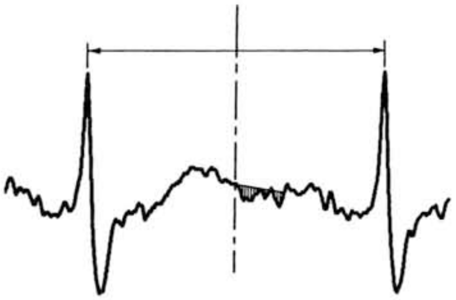

Figure 3. The noise level measurement is the mean length of the vertical lines in the figure.

QRS delineation is accomplished by a three-pass procedure. The baseline amplitude is estimated by searching backwards from the R-wave peak for a segment (of the same length as that of the segment used for noise estimation) in which the range of variation of v(t) is less than twice the estimated noise level. The QRS onset and termination are found in the second and third passes, which use the same criterion as for the baseline search, but relax the segment length requirement if v(t) approaches the previously-measured baseline.

Each QRS is characterized by seven features, which include offset, amplitude, and absolute area as used in ARGUS[3]; a T-wave measure, as described by Lovelace et al.[4]; signed area; an estimate of width derived from the absolute area and the amplitude; and a weighted mean, T_c, of the onset-to-peak and onset-to-nadir intervals. All features other than the T-wave measure are computed over the interval between QRS onset and termination, as determined by the QRS delineator. A clustering algorithm which uses a Mahalanobis distance approximation (ignoring off-diagonal elements of the covariance matrix) constructs clusters of similar QRS complexes in the seven-dimensional feature space.

Clusters are labelled supraventricular or ventricular by an algorithm which incorporates heuristics based on frequency of occurrence, mean prematurity of beats, and similarity to previously recognized clusters. Each beat is given a label according to its prematurity and the label of its cluster. The possible labels are normal, SVPB, PVC, ventricular escape, and unknown. Finally, arrhythmia detection algorithms use the stream of labelled beats and R-R intervals to identify couplets, runs of ventricular tachycardia, idioventricular rhythm, bigeminy, trigeminy, atrial couplets, SVTA, and sudden rate changes.

As shown in table 2 (under the heading "Channel 1 only"), the single-channel analysis achieves generally satisfactory performance on the MIT/BIH database and on the AHA database subset. The present single-channel technique achieves significantly higher PVC sensitivity than has been previously reported in other evaluations using the MIT/BIH database[5,6]. We therefore consider it to be a suitable testbed for examination of two-channel analysis strategies.

Given the modular structure of the single-channel program, it is straightforward to extend it to dual-channel analysis in several stages. Since the single-channel QRS detector performed adequately, and is computationally the most expensive step in the analysis, we did not use a two-channel QRS detector in these experiments, but focussed instead on improving the accuracy of QRS labelling.

Several investigators have reported on approaches to multiple-channel analysis in which the outputs of parallel single-channel analyses are merged with the application of arbitration logic to settle disagreements[7-9]. We have taken a fundamentally different approach, in which the parallelism extends only to the level of feature extraction, and clustering is performed using an ensemble of features selected from measurements taken on both channels. This "blended analysis" approach should have two significant advantages over the "parallel analysis" approaches:

1. Each channel makes a probabilistic contribution to the outcome of the labelling process. The weight of evidence in both channels put together may suffice to make a decision possible in a context in which neither channel alone carries information sufficient for a confident decision.
2. The "blended analysis" approach uses significantly less computation and memory than comparable "parallel analysis" approaches.

For the "blended analysis" approach to work advantageously, the feature set selected for use by the clustering algorithm must possess a discriminant power at least as great as that of the single-channel program. The ability to make useful distinctions between beat types on the basis of a given feature set is weakened when some or all of the features used are corrupted by noise. If a feature set for two-channel analysis contains features from two channels, the likelihood of at least some of those features being noise-corrupted is higher than it would be were all features chosen from a single channel, for reasons discussed above. Clearly, in the extreme case of a second channel containing only white noise, one should expect reduced discriminant power from a feature set which includes features from both channels. One way to address this problem is to perform the blended analysis only when signal quality is roughly equal on both channels, switching to single-channel analysis mode when signal quality deteriorates on one channel.

An effective technique for choosing the analysis mode is by application of a linear discriminant function (LDF) of the form

$$D = \triangle Q + \triangle N$$

where

$$\triangle Q = \begin{cases} \frac{Q_1}{Q_2} - 1, & Q_1 \geq Q_2 \\ 1 - \frac{Q_1}{Q_2}, & Q_1 < Q_2 \end{cases}$$

$$Q_i = \frac{\bar{P}_i}{N_i}$$

where the subscripts refer to the channel number, $\bar{P}_i$ is a moving average of the peak-to-peak amplitude on channel i, and N_i is the channel i noise estimate. Q_i is constrained so that $1 \leq Q_i \leq 20$. Finally,

$$\triangle N = k(N_1 - N_0)$$

where k is a weighting constant which will depend on the gain of the signal.

The value of the signal-quality LDF becomes large in magnitude when a significant disparity in estimated noise level or signal-to-noise ratio exists between the two channels. In selecting decision boundaries based on the LDF, it is desirable to incorporate some hysteresis to minimize instability in the context of frequent episodic noise on one channel. If settings are chosen so that the analysis mode is switched no more often than once in five beats or so, results (as summarized below) appear quite acceptable. Figure 4 illustrates the range of variation of the LDF com-

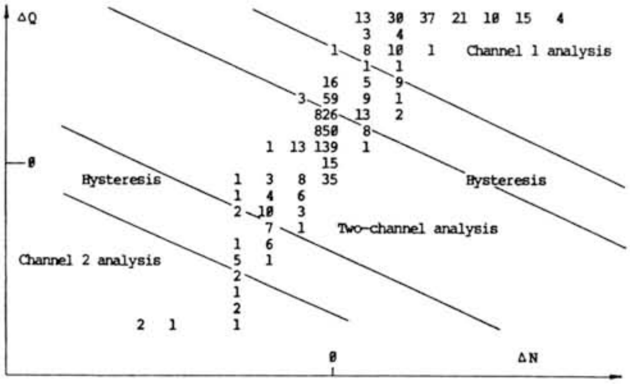

Figure 4. Signal-to-noise ratio vs. noise plot for MIT/BIH database tape 104. In this scatter histogram, each count represents one detected QRS complex. Moving to the right-hand side of the figure, channel 1 becomes relatively less noisy than channel 2; moving upward, the signal-to-noise ratio becomes relatively higher on channel 1 than on channel 2. The diagonals represent the decision boundaries for the LDF described in the text. The regions marked "hysteresis" represent beats to be analyzed in either two-channel or one-channel mode, depending on which of the adjacent regions was most recently visited. The tape shown here has short episodes of severe muscle noise on both channels, usually not simultaneously on both (see figure 2).

ponents on a moderately noisy tape from the MIT/BIH database.

Using the MIT/BIH ECG database as input, the LDF selected two-channel analysis 60% of the time, channel 1 analysis 36% of the time, and channel 2 analysis 4% of the time. On this database, channel 2 is sometimes totally unreadable; on many other occasions, QRS amplitude is quite low on channel 2 although the signal may be relatively clean. On the 25 AHA database tapes, the LDF selected two-channel operation 94% of the time, channel 1 analysis 5% of the time, and channel 2 analysis less than 1% of the time. On these tapes, signal quality is generally good on both channels; almost all of the channel 1 analysis occurs on tapes 1009 (very low amplitude in channel 2) and 3003 (noise on channel 2).

Given a suitable method for recognizing unreadable data, the remaining problem is to select a suitable feature set for blended analysis. In view of the reduced effectiveness of a two-channel feature set in the presence of noise, we decided to use three different feature sets, defining one for each analysis mode: all seven channel 1 features for "channel 1 only" analysis, all seven channel 2 features for channel 2 analysis, and seven features chosen from among the total of fourteen for two-channel analysis.

The choice of which seven features to use for two-channel analysis was based on a series of experiments using a subset of the MIT/BIH database which included several tapes with poor second channels. The experiments produced rather surprising results: when the fourteen features were ranked by discriminant power, six channel 1 features ranked highest, followed by channel 2 offset, channel 2 signed area, and the seventh channel 1 feature (width estimate). It is conjectured that the marked asymmetry of this result reflects that of the MIT/BIH database itself; generally, channel 1 (usually modified lead 2) is roughly parallel to the heart axis and has larger normal QRS complexes with less noise than in channel 2 (usually V1), which is nearly orthogonal to channel 1. The seven features which ranked highest on the basis of these experiments were used for two-channel mode in the "blended analysis" program.

Five parallel analysis algorithms were evaluated for comparison with the blended analysis and single-channel analysis techniques. Table 2 presents the results for all algorithms tested. The simplest parallel analysis approach used "best channel" arbitration: disagreements were resolved in favor of the channel with highest signal quality, as measured by the LDF described above. This approach appeared superior to other parallel analysis techniques when measured in terms of total errors.

Two other approaches did not use signal quality measures to resolve disagreements. The "AND" algorithm required PVCs to be identified as such on both channels; this technique had the lowest PVC gross sensitivity (76.12)%, but achieved the highest PVC gross positive predictivity (93.22%) of any algorithm tested. The "OR" algorithm required PVCs to be identified as such on at least one channel; this technique achieved the highest PVC gross sensitivity (94.05%) and the lowest PVC gross positive predictivity (78.91%) of the parallel analysis algorithms.

The last two approaches combined the "best channel" approach (when signal quality was considered adequate in only one channel) with the "AND" and "OR" criteria when signal quality appeared roughly equal in both channels. The results obtained using these techniques were intermediate between the "best channel" and the "AND"/"OR" approaches.

Blended analysis produced better results in general than any of the parallel analyses. The "AND" algorithms generated fewer false positives, but at a cost of many more false negatives. The "OR" algorithms had good PVC average sensitivity at a cost of many more false positives.

In comparison to single-channel analysis, only blended analysis appeared clearly superior among the algorithms tested. If QRS detection errors are ignored (73 PVCs were missed out of 7114), there were 451 false negatives due to misclassification using analysis of channel 1 only, while only 261

Table 2: Comparison of blended, parallel, and single-channel analyses

Experiment	Gross		Average		Estimated(1)	
	PVC Se	PVC +P	PVC Se	PVC +P	PVC Se	alpha
MIT/BIH database(2)						
Blended	95.31%	90.65%	87.46%	60.49%	90.06%	9.06
Best channel	89.79%	88.75%	83.48%	55.94%	83.80%	5.17
"AND"	76.12%	93.22%	75.19%	62.72%	71.54%	2.51
"OR"	94.05%	78.91%	88.25%	49.28%	90.52%	9.55
Best/"AND"	82.65%	91.32%	78.69%	60.17%	74.77%	2.96
Best/"OR"	93.10%	81.83%	87.62%	52.33%	89.74%	8.75
Channel 1 only	92.63%	90.21%	85.21%	58.32%	86.85%	6.61
Channel 2 only	80.37%	78.87%	78.61%	48.86%	75.44%	3.07
AHA database(3)						
Blended	97.57%	78.63%	94.55%	56.42%	95.64%	21.92
Channel 1 only	97.63%	81.62%	93.32%	59.18%	94.76%	18.07

(1) Estimated PVC sensitivity is calculated in accordance with the method of Hermes and Cox[10].

(2) MIT/BIH database, excluding ventricular flutter and 50-beat learning periods at the beginning of each tape. Fusion PVCs ignored.

(3) AHA database, 25 tapes. Fusion beats ignored.

were observed using blended analysis, a 42% decrease.

Furthermore, detection of clinically important ventricular couplets and runs exhibited significant improvement as well. As suggested by Schluter et al.[5], the probability of correct detection of a sequence of n PVCs, given a gross PVC sensitivity of P, is well predicted by P^n, implying that each PVC detection is an independent event statistically. We have observed that this prediction breaks down for n greater than about 4 or 5, probably a consequence of the interdependence of detection probabilities in runs with many similar events. If a PVC of a given morphology is correctly labelled, the probability that the next occurrence will be correctly labelled is higher than the gross PVC sensitivity would indicate. Nevertheless, the P^n behavior for small n implies that small improvements in gross PVC sensitivity are likely to be accompanied by larger improvements in couplet and run sensitivity, as is observed in table 3.

In table 3, couplet sensitivity is the fraction of couplets correctly identified as couplets; couplet positive predictivity is the fraction of events called couplets by the program which were genuine couplets. Couplets are defined for this purpose as two consecutive PVCs preceded and followed by non-PVCs. For "all runs" statistics, the program must correctly identify at least three consecutive PVCs for a correct detection. For ">5 beat runs" statistics, the program must correctly identify at least six consecutive PVCs for a correct detection.

Table 3. Ventricular couplet and run detection

Experiment	Couplets	
	Se	+P
Blended analysis	87.96%	79.61%
Channel 1 only	84.18%	78.65%
	All runs	
	Se	+P
Blended analysis	84.85%	54.90%
Channel 1 only	75.00%	49.48%
	>5 beat runs	
	Se	+P
Blended analysis	94.12%	80.00%
Channel 1 only	87.50%	77.78%

MIT/BIH database, excluding 50-beat learning periods at the beginning of each tape. Fusion PVCs ignored.

Conclusions

Based on experiments with the MIT/BIH ECG database, blended analysis permits significantly higher sensitivity for PVCs, couplets and runs than single-channel or any of the five varieties of parallel analysis which were tested. Only marginal improvements in PVC positive predictivity were demonstrated for blended analysis.

The asymmetry of the feature set which was chosen for two-channel analysis remains a topic for further investigation. It appears clear that poor signal quality throughout much of channel 2 in the MIT/BIH database is responsible to at least some degree for the anomaly.

Single-channel analysis on the 25 available AHA database tapes left little room for improvement, and none was noted.

References

1. Mark, R.G., Moody, G.B., Olson, W.H., Peterson, S.K., Schluter, P.S., and Walters, J.B., Jr. Real-time ambulatory arrhythmia analysis with a microcomputer. Computers in Cardiology 6:57-62. Long Beach, California: IEEE Computer Society. 1979.

2. Arnold, J. Time-domain filtering of electrocardiograms. S.B. thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering, Cambridge, Mass., 1972.

3. Nolle, F.M. ARGUS, a clinical computer system for monitoring electrocardiographic rhythms. D.Sc. dissertation, Washington University, Sever Institute of Technology, St. Louis, Missouri, 1972.

4. Lovelace, D.E., Knoebel, S.B., and Zipes, D.P. Recognition of ventricular extrasystoles in sedentary vs. ambulatory populations. Computers in Cardiology 3:9-12. 1976.

5. Schluter, P.S., Mark, R.G., Moody, G.B., Olson, W.H., and Peterson, S.K. Performance measures for arrhythmia detectors. Computers in Cardiology 7:267-270. 1980.

6. Zeelenberg, C., and Meij, S.H. Evaluation and optimisation of an existing arrhythmia detection system by using an annotated ECG database. Computers in Cardiology 8:103-108. 1981.

7. Clark, K.W., Hitchens, R.E., Ritter, J.A., Rankin, S.L, Oliver, G.C., and Thomas, L.J., Jr. ARGUS/2H: a dual-channel Holter-tape analysis system. Computers in Cardiology 4:191-198. 1977.

8. Rosenberg, N.W., and Tartakovsky, M.B. The TELAVIV system -- three-channel evaluation of long-term ECG records for atrial and ventricular identification and verification of arrhythmia. Computers in Cardiology 6:29-32. 1979.

9. Bragg-Remschel, D.A., and Harrison, D.C. A computerized two-channel ambulatory arrhythmia analysis system. Computers in Cardiology 7:197-200. 1980.

10. Hermes, R.E., and Cox, J.R., Jr. A methodology for performance evaluation of ventricular arrhythmia detectors. Computers in Cardiology 7:3-8. 1980.