5

AUTOMATED SYSTEM FOR THE DETECTION OF ERRONEOUS DATA

Nopporn Khaewpong, Chi Associates
Jeff Marcus, NHTSA

QEI SYSTEM

Introduction :

Since 1979 the national highway traffic safety administration has contracted with several organizations around the country and abroad to perform experimental work on human surrogates under various accident conditions. The results of these tests are sent to NHTSA in the form of magnetic tape and stored in the NHTSA data base. Up to the present time there are about 2000 tests stored in this data base. On average there are about 35 channels per test. In addition there are about 12 new submitted tests per month. Before these data can be analyzed they must go through data quality control procedures. These procedures are extremely tedious and time consuming when carried out manually. Because the need for reliable valid data is so great, and because checking the data is a tedious procedure, it was the aim of this project to develope an automated method to detect error and data in need for correction.

the proposed QEI system is part of this work.

Objective :

The main objective of QEI system is to automate checking of data quality. The system is not designed as a tool to label the data as good or bad but rather to identified questionable data so that engineers can look at them more closely before they are used in analysis.

Types of Data Quality Checks :

From the experience it is found that data quality checking can be divided into two groups according to the type of errors to be checked. They are named here as "general checking" and "specific checking"

General Checkings :

Errors that can be detected under general checking are those that are commonly found in most digitized data. Their characteristics do not depend on test configuration or test speed. The type of errors that are included in this category are as follow.

1. Lead-in and Lead-out.
2. Bias.
3. Empty Channel.
4. Clipping.
5. Channel Failure.
6. Noise.

Specific Checkings :

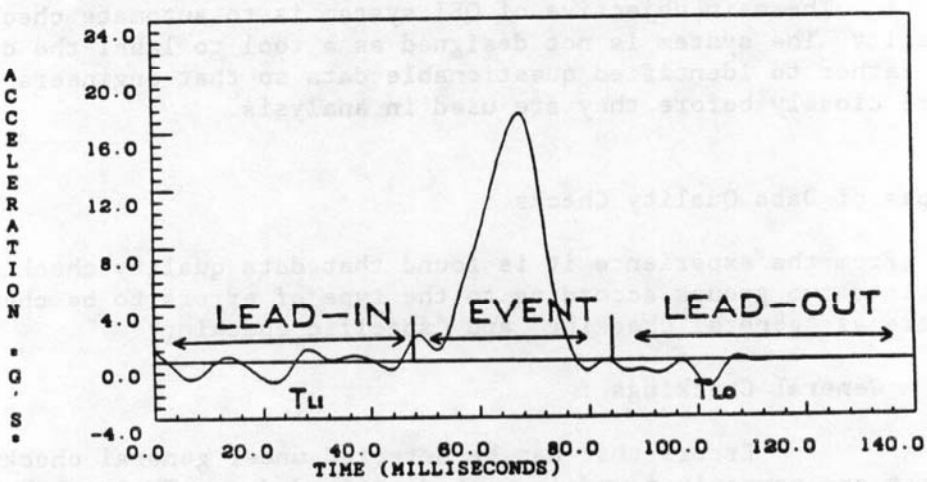
As the name "specific checkings" implies, the errors that can be checked here are those whose characteristics are considered to be more particular and specific. As a result their characteristics seem to depend very much on the test set up and the test speed. The errors that are included here are ;

1. Magnitude.
2. Polarity.
3. Final Velocity.
4. Phasing.
5. Axis.

Check Lead-in and Lead-out.

Lead-in and lead-out are defined as the quiescent portions of the signal just before and after the actual event data, as shown in figure 1. For the purpose of establishing the level of bias in the signal it is required that the length of these portions should be at least 30 millisecond .

CHECKING LEAD-IN, LEAD-OUT



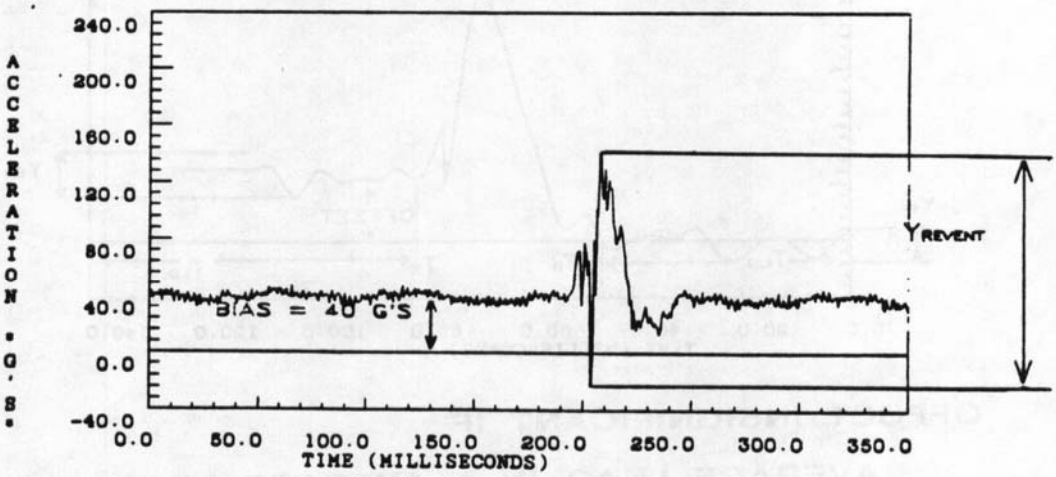
CRITERION: T_{LI} OR T_{LO} > 30 MILLISECONDS

FIGURE 1.

Check for Bias

If the average magnitude of the lead-in and lead-out portions are significant compared with the range of magnitude of the signal,as shown in figure 2., it is said that the signal has bias. When bias is present, the magnitude of the signal will be larger than its true value. The bias can be removed by subtracting the bias value from the curve.

CHECKING AND REMOVING BIAS



BIAS = AVERAGE (Y (FIRST 30 MSECS))

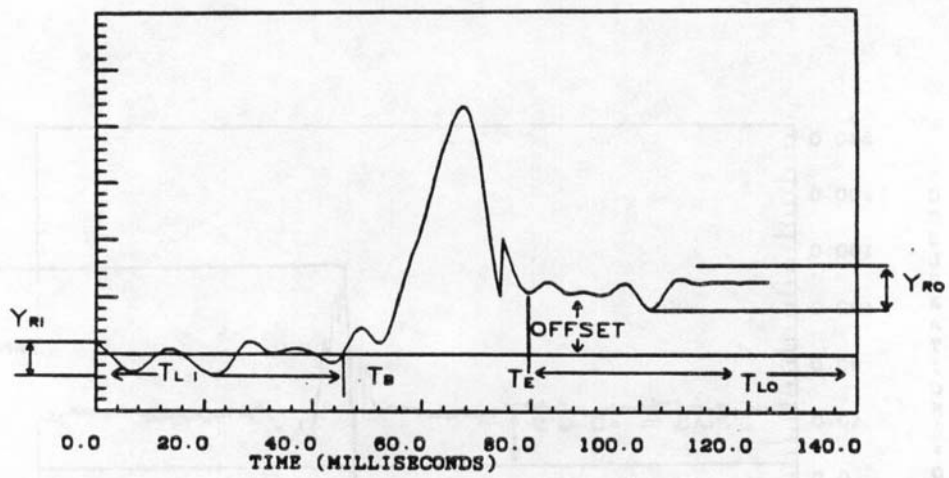
BIAS INSIGNIFICANT IF BIAS < .01 • Y_{REVENT}

BIAS REMOVED BY SUBTRACTING BIAS VALUE FROM CURVE

FIGURE 2.

If the level of bias in lead-in and lead-out are substantially different it then faces another type of error that is call "offset". The example of this type of error is shown in figure 3. When offset is present it indicates that the measurement device was damaged during the test. Therefore the signal is not usable.

OFFSET



OFFSET INSIGNIFICANT IF

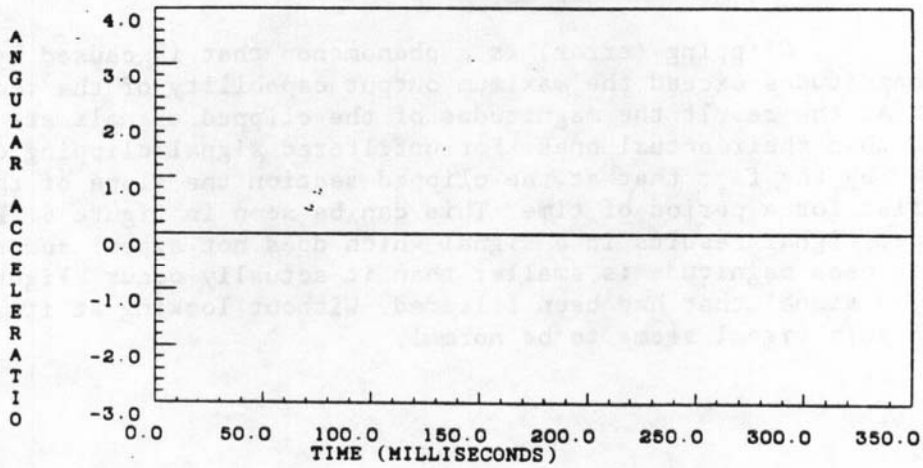
AVERAGE LEAD-IN ≈ AVERAGE LEAD-OUT

Figure 3.

Check for Empty Channel.

Data result from empty channel could either be a blank channel i.e. all data points are zero, as shown in figure 4 or the values are just noise as shown in figure 5. These data are flagged in the database as meaningless.

CHECKING FOR EMPTY CHANNEL

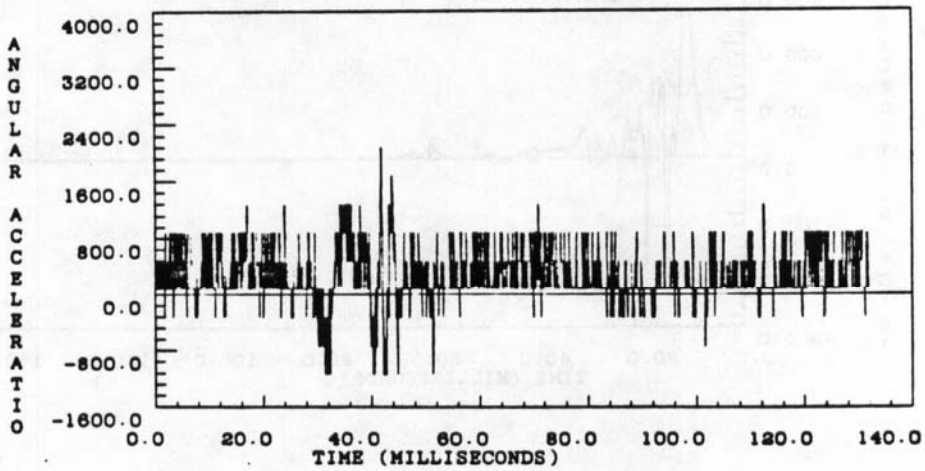


FAILS LEAD-IN, LEAD-OUT TEST

ALL POINTS = 0

FIGURE 4.

EMPTY CHANNEL WITH NOISE



FAILS LEAD-IN, LEAD-OUT TEST

BIAS REMOVED, AVERAGE OF POINTS ≈ 0

FIGURE 5.

Check for clipping.

Clipping (error) is a phenomenon that is caused by the signals whose magnitudes exceed the maximum output capability of the recording system. As the result the magnitudes of the clipped signals are smaller than their actual ones. For unfiltered signal clipping can be detected by the fact that at the clipped section the slope of the signal stays flat for a period of time. This can be seen in figure 6. Filtering a clipped signal results in a signal which does not appear abnormal except that its peak magnitude is smaller than it actually occur. Figure 7 shows a clipped signal that had been filtered. Without looking at its unfiltered version this signal seems to be normal.

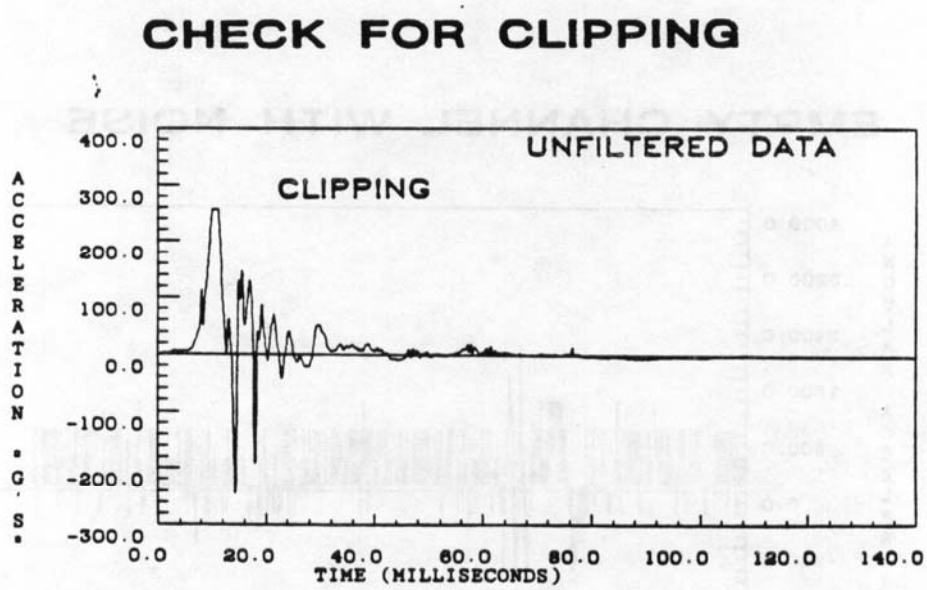


FIGURE 6.

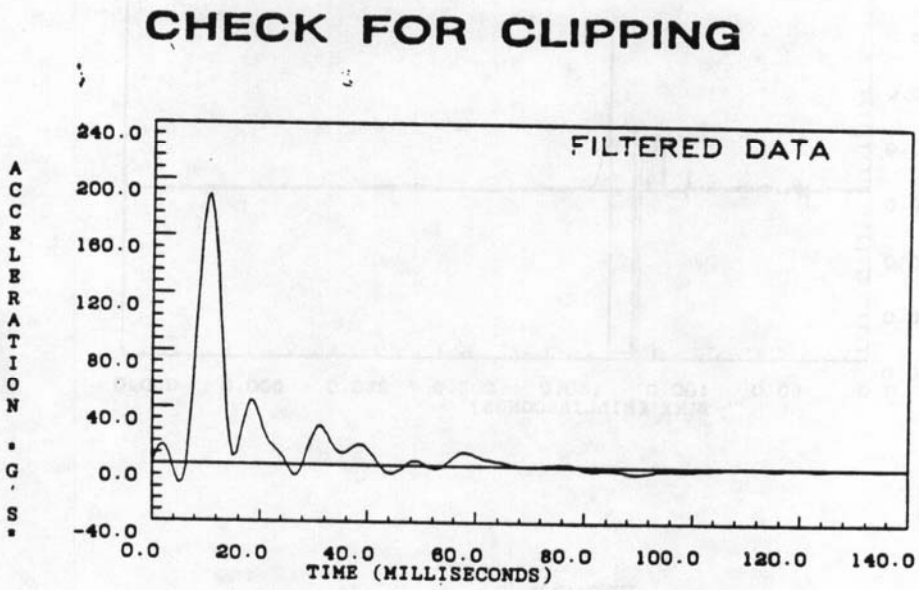


FIGURE 7.

Check for Channel Failure

Channel failure error is typified by a dramatic increase in slope that becomes flat immediately for a certain period of time. As the result, the shape of the signal looks more like a step-type function. This makes them easy to identify since all of the signals that we measure are transient continuous function. Channel failure is caused by either a mechanical or electrical interruption on a sensing bridge of the measurement device. This has been a particular problem with certain types of older accelerometers. The interruption could be either transient or permanent. a typical error signal caused by transient interruption is shown in figure 8. In this figure the signal is interrupted briefly, as shown by the flat portion, and returns to normal at the end of the event. In contrast to the transient interruption case, the flat portion of of the signal that is caused by permanent interruption would stay flat permanently as shown in figure 9. The signal is considered to be bad in either case.

CHECKING FOR CHANNEL FAILURE

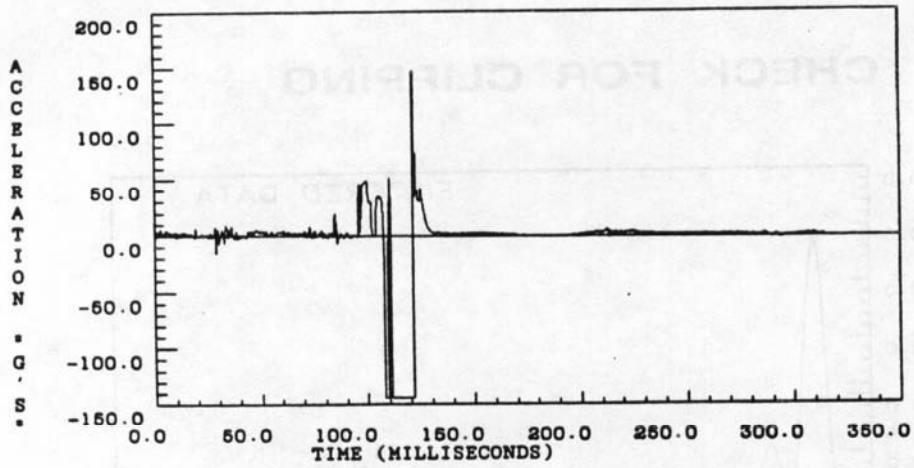


FIGURE 8.

CHECKING FOR CHANNEL FAILURE

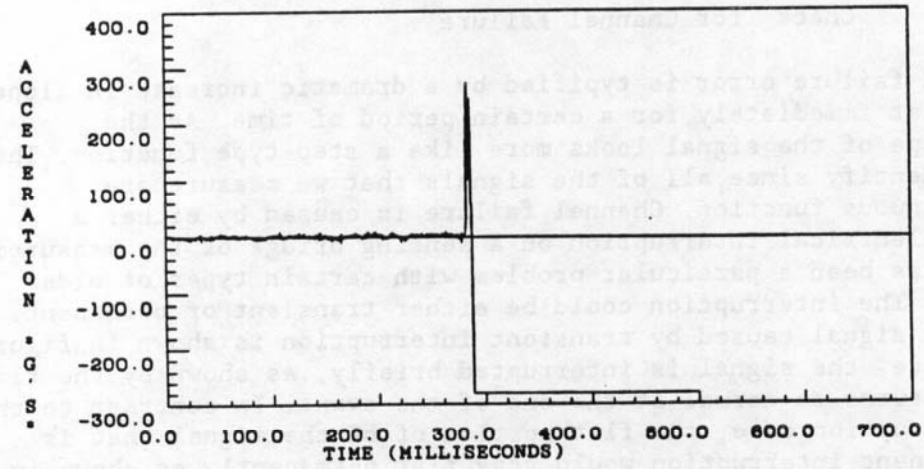


FIGURE 9.

Check for Noise

All the data that are measured in the experiment are considered to be transient phenomenon. By this it means that a normal signal should consist of quiescent portions plus the actual event data itself. The data is said to be very noisy if there are significant signal over entire time interval include lead-in and lead-out. A signal is considered to be too noisy to be used if it has no lead-in and lead-out and more than 50% of the pulse peaks are greater than the average pulse peak. Figure 10 shows a typical signal that is considered to be very noisy.

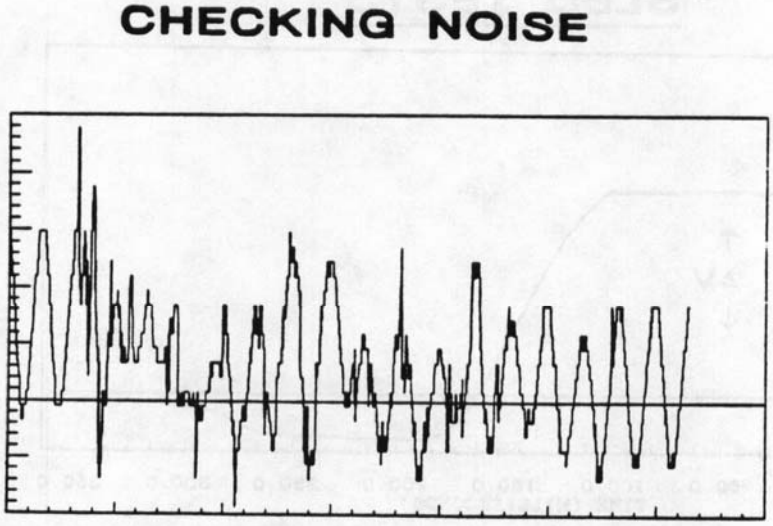


FIGURE 10.

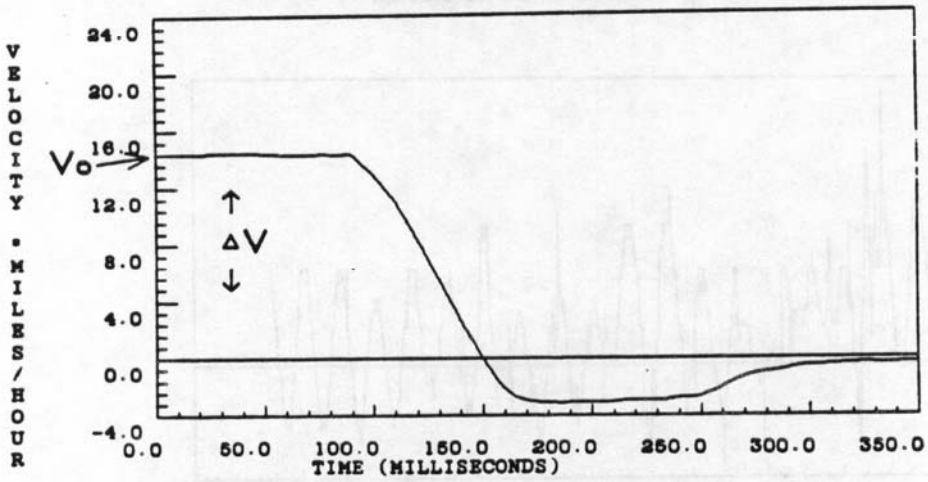
SPECIFIC CHECKINGS.

Check Magnitude

To check the magnitude of a signal it is necessary to know first the type of test set up. Is it a sled test or pendulum test. The characteristics of these two types are completely different. Figure 11 shows the typical velocity curve from a sled test . From the test set up one knows that the sled is brought up to the level of the test speed and coasts at this speed before it strikes a barrier,stops and bounces back. Therefore through the change of its corresponding velocity, delta v, one can check if the signal possesses a reasonable magnitude or not.

CHECKING MAGNITUDE

SLED TESTS



ΔV SHOULD APPROXIMATE Vo (CLSSPD)

FIGURE 11.

Check Polarity

The purpose here is to check if the sign of the signal is correct or not. Knowing the test set up and the sign convention for the global coordinate polarity of the signal can be checked. The fact that in sled tests a signal that has the correct polarity its corresponding velocity curve must possess a rebound point. Figure 12 shows two velocity curves that demonstrate this. In this figure the lower spine which has the wrong polarity has no rebound point.

POLARITY CHECKING

SLED TEST

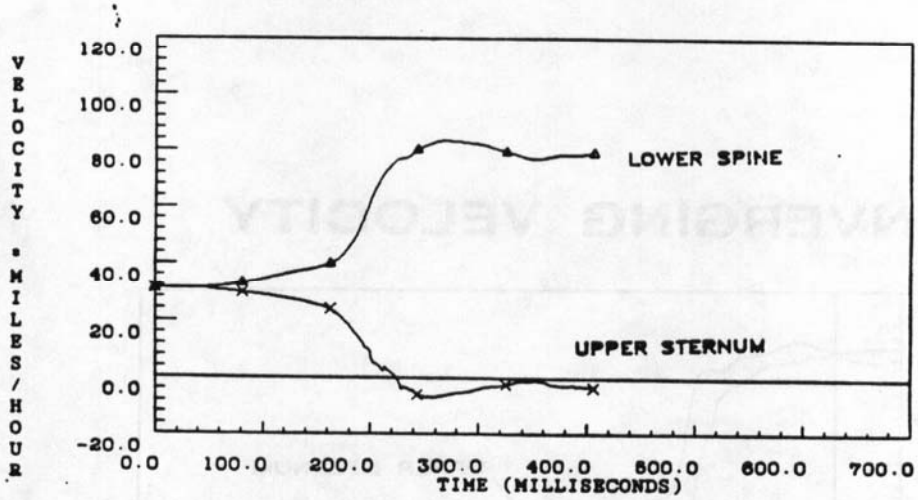


FIGURE 12

Check Converging Velocity

The main purpose of this check is to look for some abnormality that took place during the testing. In normal circumstances the test surrogate (either dummy or cadaver) will come to rest. Therefore velocity curves of all signals should approach a common velocity at the end of the event. For example curves no. 1 and 2. ,shown in figure 13, represent velocity curves of upper sternum and lower spine respectively. Here curve no. 2 shows some sign of abnormality since it diverges from the common velocity at the end of the event. It implies that instead of coming to rest, the upper spine starts to accelerate at the end of the event.

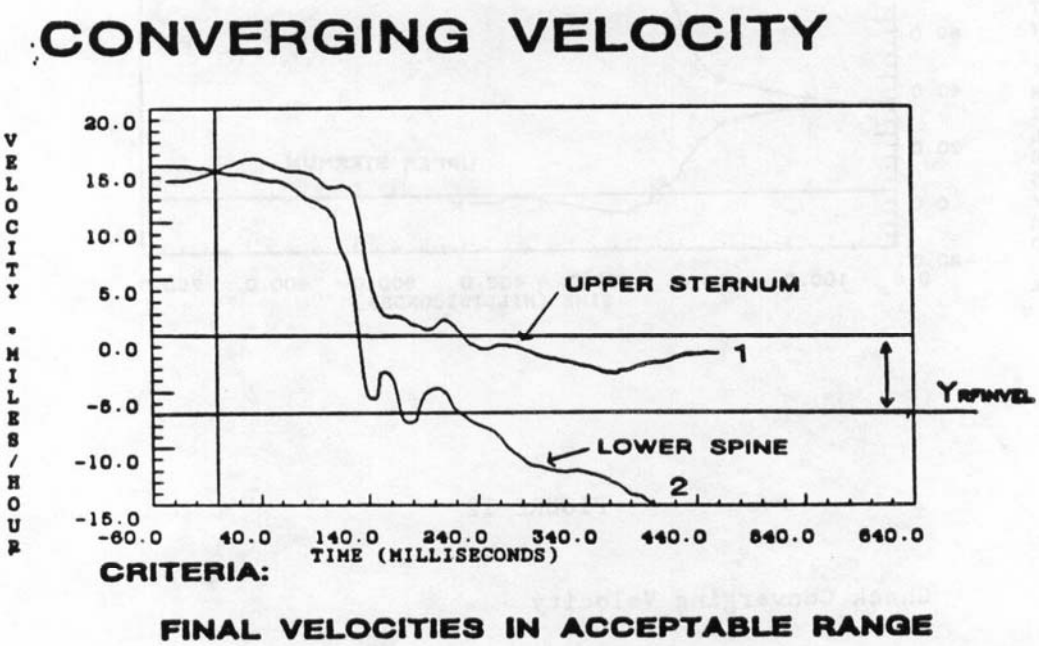
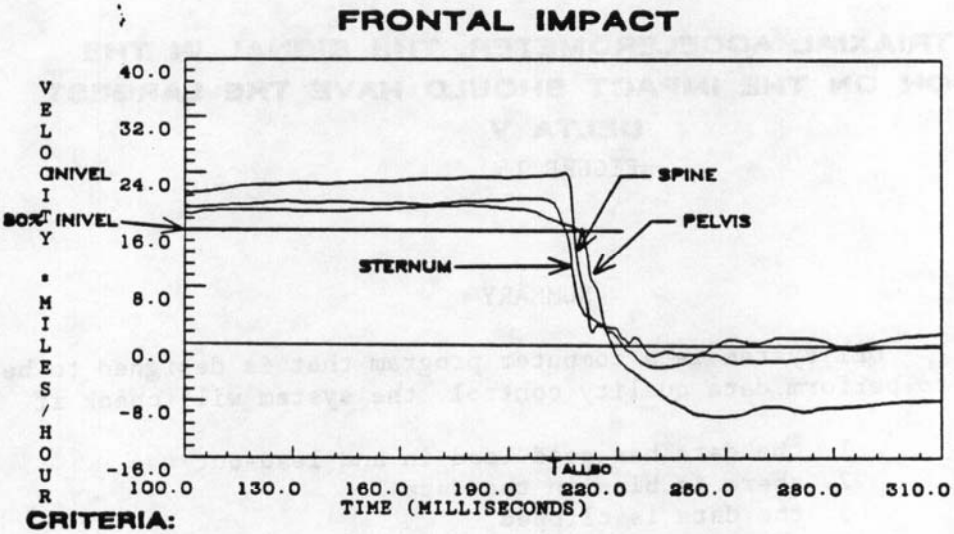


FIGURE 13.

Check Phasing

Here the purpose is to detect if some abnormality has taken place during the event. The aim here is to check if the test surrogate has been stimulated in a manner consistent with the test set up. For example, in a frontal pendulum impact the sternum should be stimulated before the pelvis. The order of stimulation is determined by the order in which the various responses fall below 80% of the initial velocity. As shown in figure 14.

PHASING



CRITERIA:

– SIGNALS REACH 80% INITIAL VELOCITY IN PROPER ORDER

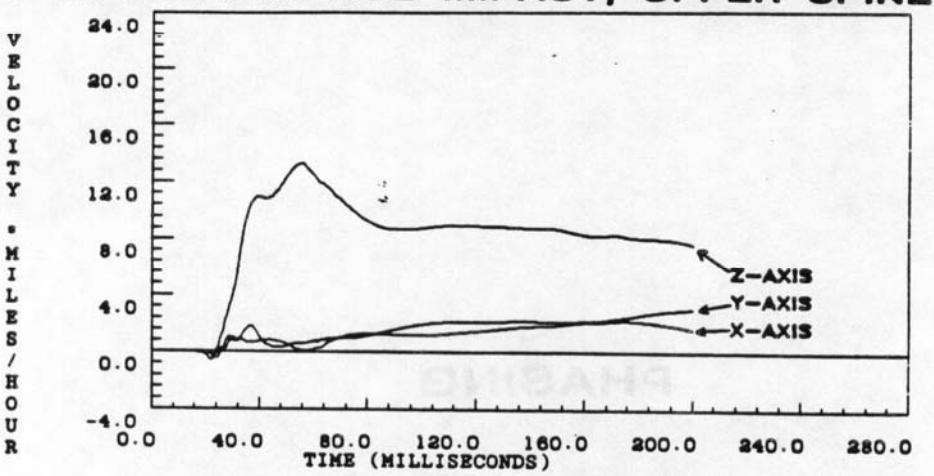
FIGURE 14.

Check Axis

From the experience it is quite common that some data have their axes mislabeled. This type of error can be checked quite easily with triaxial accelerometers through the magnitude of the signal, particularly through the magnitude of its corresponding velocity curve. Figure 15 shows the velocity curves of the upper spine in side impact pendulum test. For this particular test the global Y axis coincides with impact direction. Therefore the magnitude of velocity in this Y direction must be the largest. But it has been mislabeled as the z axis.

CHECKING AXIS

20 MPH PEND. SIDE IMPACT, UPPER SPINE



IN A TRIAXIAL ACCELEROMETER, THE SIGNAL IN THE DIRECTION ON THE IMPACT SHOULD HAVE THE LARGEST DELTA V

FIGURE 15.

SUMMARY

QEI system is a computer program that is designed to be used as a tool to perform data quality control. the system will check if

1. the data has good lead-in and lead-out
2. there is bias in the data
3. the data is clipped.
4. the data is very noisy.
5. the data is meaningless due to instrumentation failure.
6. the data has correct magnitude.
7. the data has the correct polarity.
8. the data behaves consistently with the test set up in terms of phasing and final velocity.
9. the data has the correct axis.

The main objective of the system is to identify questionable data so that engineers can look at them more closely before they are used in any analysis. The system is not designed as a tool to label the data as good or bad.

DISCUSSION

PAPER: Automated System for the Detection of Erroneous Data

SPEAKER: Nopporn Khaewpong, Chi Associates

Q. Sam Shaibani, Insurance Institute for Highway Safety
I have a misgiving about the selection of the value of 80 percent to determine the phase of the measurements. What is the physical justification for that particular value? It seems from that slide if you move that line up or down you can determine what phase you want almost arbitrarily.

A. Yes, it is almost arbitrary, This is the one I showed in particular for the sled test. We know that the initial velocity of all signals should be equal to the initial velocity so that sometimes the signal has some type of bias even though it's not great enough to justify being removed. So we pick the point where the signal falls below 80.

Q. But wouldn't you accept from the basis of the figure that you showed that if you made it at 90 or 70 you would have gotten a different result?

A. The reason we checked 80 is because we didn't want to pick something to close to the beginning of the event as we are not sure what's happening at that time, particularly if bias occurs. We would prefer to move to either 80 or lower because you can be sure that the signal is okay.

Q. John Melvin, General Motors Research Labs
I'd like to know how this is intended to be used. Maybe Jeff can answer that better. Is this to be applied to the data that has been sent to NHTSA for qualification on the data base or will it be supplied to the generators of the data to check their data before it is sent to NHTSA?

A. No, we want to make sure the data is correct before we use it and before NHTSA uses it in their analysis. We don't want to have questions later. Basically, NHTSA wants to use it to screen the data that they receive from the various contractors.

Q. I think the best people to make judgments on whether there is a problem with the data are the contractors themselves. For instance, you mentioned clipping, Guy Nusholtz told us about a mechanical phenomenon where the pressure he measured in an animal would clip because it was a cavitation phenomenon. If you assume that it's an electronic problem instead of an actual phenomenon you've just thrown the data away. That's why I think that the contractors are the best people to judge whether the data is okay. This could really help them find some problems when they're turning out 30, 40, channels every time.

A. Well, from my experience with NHTSA, when they believe something is wrong with the data they will send a letter to the contractor asking him to explain why the data behaves as it does.

A. Jeff Marcus, NHTSA

I would like to comment a little bit more on what John said. He is absolutely right that the best person to know good data from bad is the person who generated it in the first place. The problem that I think the people who generate our data have is the same problem that we have. That is, if you have 10 tests, each of which has 35 channels on it, it is so voluminous and tedious to go through that it would be difficult to find someone who was really smart but willing to do the really tedious work. In the real world you can't find anybody like that. So what we hope to do is have a system which will not go through and say, "This data is bad", but which will point out data where it is worth our while to get back to the contractor and say, "Something looks suspicious; can you explain it?" Our ultimate plan is to use the data that we already have to test the system until we feel that it works. Then we would like to do two things. One, is to put this into our standard process of reading the data so that as soon as it comes in we would know that it bears examination. We also intend to supply it to the different laboratories generating data for us so that they would be able to check it before they send it to us.

Q: Anil Khadilkar, Mobility Systems

I think I'm just elaborating a little bit on what has been said. Do you intend in this program to flag in the biodata base that there is a questionable channel? Similar, for example, to the feedbacks that you have when you're calculating a HIC number and it tells you that you are not using a resultant. Would this program then put on the tape when a channel was questionable?

A: Jeff Marcus

We maintain a field in the data base for each individual channel. The field has several codes available to it. The one we most often like to see is "as measured", which is an indication that it's been checked and that it's good. We also have a field that says this is "meaningless data", which means that we still have it available if it later turns out to be good data and we inadvertently labeled it as meaningless. We also have several other codes for things like channels which are good up to a point and beyond that point they're not usable. We have various ways of labeling these things and our analysis programs look at that field and if they find anything other than an "as measured" signal, it will query the person who wants to use that signal.

Q: Pat Kaiker, UMTRI

What kind of computer is this program on now? Is it on a NOVA-size thing or are you running it on an Amdahl or is it PC adaptable?

A: Nopporn Khaewpong
No. It's all under VAX right now. We develop on FORTRAN on a mainframe, a VAX mainframe.

A. Pat Kaiker
Thank you.