RUNS OF SIGNIFICANT SAMPLES FOR PROCESSES WITH SHARP NON-STATIONARITIES: APPLICATION TO SEISMOGRAM COMPRESSION

V.F. BABKIN, N.E. RYBEVA, YU. M. SHTARKOV
Institute for Space Research, USSR

Summary. An algorithm for threshold compression of the processes with the sharp variations of a level is considered. In broadening the concept on an significant sample we are successful in transmitting completely almost all the samples appropriate to the phenomenon studied; at the same time the compression ratio at the quiet parts is kept at the acceptable level. The experimental results of seismogram compression are given.

Introduction. For the global seismic investigations of the Earth to be performed a great number of the automatic research stations can be located on the planetary surface including the almost inaccessible regions. It is convenient to carry out the acquisition of the data obtained by means of an artificial Earth satellite. Here each station transmits the data to the satellite; these data are stared (from all the stations) on board the AES and then are transmitted to the Earth during sessions.

Due to the limitation of the AES memory volume and very short time of sessions the necessity arises for the economical description (compression) of the data recorded. The similar problem suggests itself in studying seismometrically the Moon or the other planets. Finally, compression of the seismometrical data can be useful for the stationary seismometrical stations on the Earth also.

Below a rather simple technique, of compression is discussed that can be also used either on board the space vehicles or on the surface of celestial bodies studied. The choice of the compression device location depends on the problem under solution and the considerations on the technical expediency.

Any compression algorithm can be, naturally, correlated to some extent with the character of a signal recorded. Hence let us consider first of all the features of seismograms. Fig. 1 gives the very typical seismogram obtained from the “Moscow” seismometrical station. The characteristic features of this seismogram. (fairly typical) are not difficult to outline.

a) For a considerable interval of observation time the signal shows fluctuations of noise-type,relative to zero level. It is “quiet” parts almost not including the data on the seismic phenomena and hence are of little interest for the experimentators.

b) Sharp oscillatory bursts of the high level occupy a short time interval in comparison with the total duration of recording. They characterize the events that allow to define the earthquake properties. It should be noted that of interest can be the parts directly adjacent with a burst (e.g. for studying the earthquake heralds).

The techniques of seismogram. compression. According to the description given above it is not difficult to formulate one of the main principles of compression of the seismometrical data. This principle means that the data associated with the quiet part are not recorded (as far as it can be realized). The active parts are completely reproduced without any attempts to reduce redundancy.

Now let us describe the specific algorithm that allows to realize the chosen approach to the data compression.

1) For each sample of the process y_i the comparison is performed

$$|y_i| \geq d \tag{1}$$

where $|y_i|$ is an absolute value of the sample y_i and d is a preset threshold which generally can vary.

2) If the condition (1) is satisfied the appropriate sample is believed to be “significant” and has to be recorded. Otherwise the sample is not registered. It is clear that this algorithm permits to realize the principle formulated above (with the sufficiently high degree of approximation). The threshold d should be chosen as adequately great so that fluctuations rarely exceeded d at the quiet parts. At the same time the threshold should not be very great since otherwise the serious distortions can occur in recording the bursts.

The compression algorithm modification. The significant (recorded) samples appear irregularly. Therefore not only their value but also the time of their appearance have to be recorded. Using the well-known terminology the number determining the time of registration will be called as “address” of a sample. Two techniques of the formation of addresses will be considered below.

(1) Address - sample. Let us denote the number of binary symbols as s and t that are used for the description of a value of sample and its address, respectively. The initial sequence of samples (up to compression) should be divided into the groups of equal

lengths $L \leq 2^t$. Then t-gigit address corresponds simply to the item number of a sample in the group (telemetry frame).

If the number of samples (significant) in "adaptive" frame (after omission of insignificant samples) is equal to ℓ their description requires

$$K_1 = \frac{sL}{(s+t)\ell} \tag{2}$$

times less binary symbols than for the initial presentation (in (2) the number of symbols used for synchronization are not taken into account that for the initial and adaptive presentations are sufficiently less than the numerator and the denominator, respectively). Thus, K_1 is a compression ratio with the use of "individual addresses" of samples.

(2) Address-run of sufficient samples. When an event occurs the sufficient samples are grouped into runs and this can be used for economical description of their addresses. Actually, in this case it is sufficient to indicate only the address of the first sample and to record subsequent samples (in the run) without addresses. Here for one-valued decoding it is necessary to introduce the additional markers for address and the values of sample (see e.g. [1]). Then the compression ratio is equal to

$$K_2 = \frac{sL}{(t+1)m + (s+1)(\ell_1 + \ell_2 + \ldots + \ell_m)} \tag{3}$$

where m is a number of runs per a frame; ℓ_i is a length of i-th ran (i = 1, 2....., m). It is expected that, due to specific character of a signal studied, K_2 should be more than K_1 . In conclusion it should be noted that it is necessary to choose s = t; otherwise bit errors can result in the synchronization errors at the prolonged parts of data[i].

(3) Recording of the generalized runs-1. As it has been briefly mentioned the recording of some samples turns out to be rather desirable or necessary before the beginning and (or) after the end of the run. Let us give several substantiations of the expediency of such an additional registration.

- Recording a of the insignificant samples before the beginning of bursts can be of interest in studying the development of a burst or investigating the earthquake heralds.

- Oscillatory character of seismodetector output signal means that its amplitude takes relatively often the values close to zero. Hence using the algorithm formulated above the burst will be described by the set of single runs of the samples the total number of which can be sufficiently great. But the latter means that some samples "inside" the burst will not be recorded that is evidently undesirable.

It is impossible to make allowance for these notes at the cost of decreasing the level of d-threshold since the location of the insignificant samples which are desirable to be recorded is rather rigidly bound with that of the significant samples. The decrease of d-threshold will lead to the considerable increase of the number of samples recorded during the quiet parts and, as a consequence, to the essential diminution of the compression ratio.

Taking into account the last note it is necessary to modify the algorithm formulated above. The following modification is of interest. All the runs of samples have to be recorded that satisfy the following conditions:

a) The run begins precisely from a insignificant samples (i.e. (a + 1)-th sample is significant).

b) The run ends exactly with insignificant samples (i.e. (b + 1)-th sample from the end is significant).

c) Between the first and last significant samples of the run "subruns" of insignificant samples may occur which lengths do not exceed $\alpha = a + b$. The latter condition follows directly from the former two. As an example, Fig. 2 gives the subdivision of samples into generalized runs with some values of a and b.

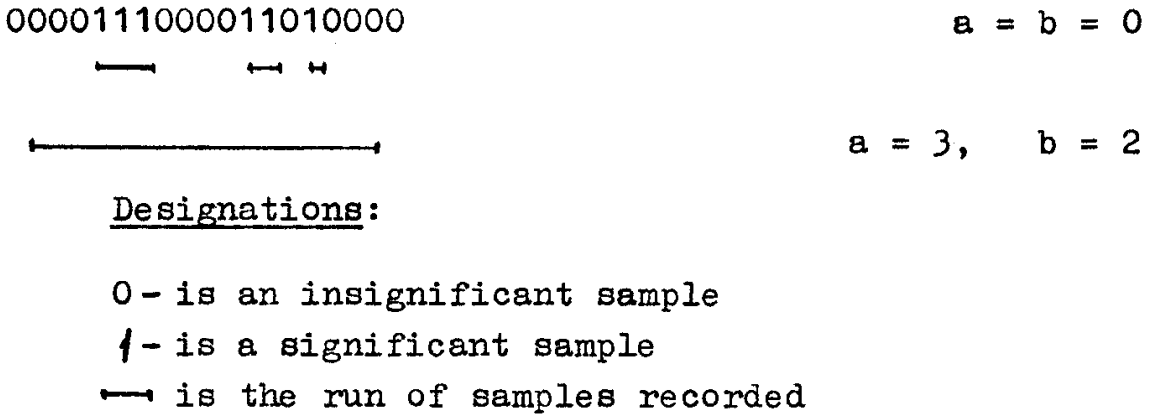

Fig. 2

If the problem of technical realization and effectiveness of compression is not dealt with the chosen value a is determined by the burst "prehistory" duration, the experimentator is interested in, but the value b is defined by the duration of decay of seismograph pickup oscillation. Various combinations of a and b are possible. For example, there is either a = 0 or b = 0 in the simplest case. Further we shall not discuss various combinations of parameters as they are determined only by the experimental objectives instrument performance characteristics and technical requirements.

As a result of modification each run contains not less than a+b+1 samples. Many runs are combined as compared with the initial algorithm according to a = b = 0 (Fig. 2 illustrates the latter). It is natural that in this case it is expedient to use only one method of address formation, the signel address for the run. Therefore in calculating the compression ratio the expression (3) should be used with changes that take into account decrease of m and increase of run lengths.

(4) Recording of generalized runs-2

With increase of a and b the compression ratio K_2 can decrease rather essentially. Therefore the previous method can be slightly changed.

Let us consider the runs of samples to be registered that have been determined in the previous item. Sometimes it seems possible not to record the first values of a and (or) the last b of insignificant samples. As a result the slightly changed form of generalized run is obtained which can be defined as follows:

a) The run begins with a significant sample which is preceded by not less than a insignificant samples.

b) The run ends by a significant sample which is followed by not less than b insignificant samples.

c) Just the same as in the previous item.

And as before the expediency of description of run addresses rather than individual samples is obvious. In calculating the compression ratio Eq. (3) should be also used taking into account the change of m and ℓ_i , i = 1, 2, ... m.

(5) The elimination of "isolated" significant samples- The duration of the investigated phenomena (earth-quake, etc.) allows to confidently affirm that during the burst a great number of significant samples is recorded. But the latter means that "isolated" rare significant samples can be caused by fluctuations at the quiet part and they can be omitted.

We can regard a significant sample as isolated (not to be recorded) if it is preceded and followed by not less than C insignificant samples.

The experimental results. For an experimental estimation of the efficiency of the algorithms suggested the real seismogram was used a fragment of which is given in Fig. 1. Three-component seismograph was used for recording the total duration of which corresponded to 12 hours of observation. A single recording component was processed. A

number of quantization levels on the digitized seismogram was $2^s = 128$; the flame length was chosen to be equal to 128 samples (it corresponds to s = t). The whole seismogram used was divided into 424 frames, the event being recorded over the interval equal to 20 frames, the rest of 404 frames can be used for the quiet parts.

The analysis of the recording at the quiet parts showed that zero bursts (or microseisms) rarely exceed the value of three quantization levels, hence during the experimental study d = 3 was chosen. The analysis of frames containing the event registered showed that according to the first two compression technique modifications only 80% of samples at the parts being of interest for investigator were transmitted. To improve this value other algorythms of compression were also considered and choosing a = b = 3 or 5 in frames containing the event almost all the samples are transmitted.

The calculation results are summarized in Table I. Here (i) i = 1, 2, 3, 4, 5 denotes the number of the compression techniques used according to items of the previous section.

Table I

techniques used	(1)	(2)	(3)+(2) a=b=3	(4)+(2) a=b=3	(4)+(2) a=b=5	(5)+(3)+(2) a=b=c=3
compression coefficient of the quiet parts	6.6	7.5	4.8	6.4	4.8	9.8

On the complexity of technical realization of the methods described To realize the methods using the concept of generalized runs given above the internal memory is needed. It is due to the necessity to record sometimes insignificant samples in a combination with significant ones. In particular, for the realization of (3) and (4) methods while dividing sequences of samples into runs it is necessary to store (a + b - 1) successive insignificant samples that requires $(a + b - 1) \log_2 d$ bit storage cells. Using the most effective method (5) together with (3) and (4) the device that singles out generalized runs should be preceded by another device for recognition and omission of the isolated significant samples.

REFERENCES

1. V.F. Babkin, et al., “A Study of Simple Chromatogram-Compression Algorithm Efficiency”. The National Telecommunications Conference, 1972, Houston, Texas.

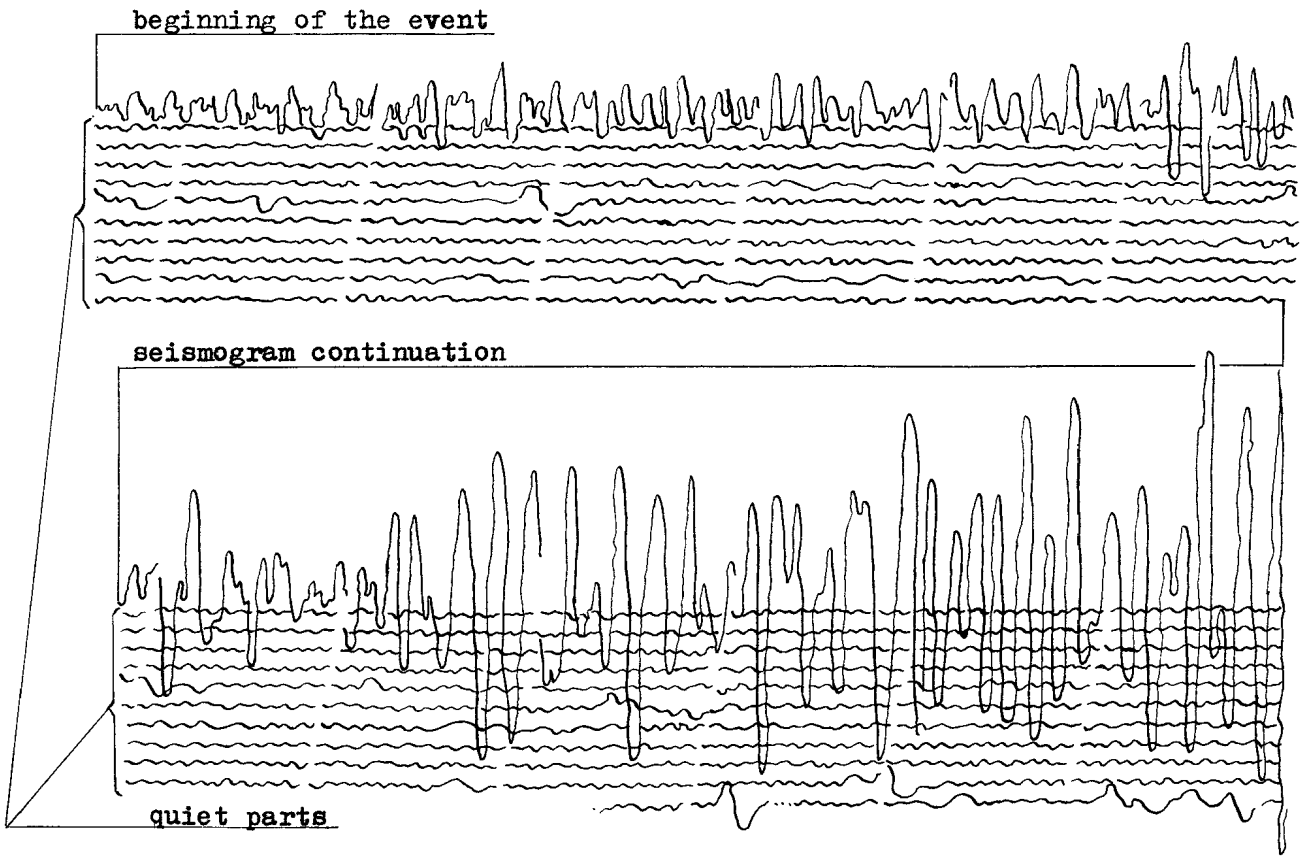
beginning of the event
seismogram continuation
quiet parts

Fig. 1