Auto clustering of the variety of pulse signals based on their symbolic description

Yury Senkevich[1,*]

[1] Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS, Laboratory of Acoustic Research, Mirnaya str., 7, Paratunka, Kamchatskiy kray, 684034, Russia

Abstract. In a number of applied studies of geophysics, medicine, cosmophysics, atomic physics and other fields of knowledge, useful information is often hidden in the character of the behavior of a stream of frequency modulated pulses, which are represented by a large variety of forms, significantly different from each other up to several orders value of magnitude amplitude and durations. Noise is often present in the signal. Under these conditions, the problem arises of identifying both individual pulses and groups of pulses to assess the connection between their dynamic characteristics and the state of system. To solve the problem by a method is proposed that includes signal cleaning from interference, the operation of extracting and converting pulses into a code representing a sequence of invariant amplitude and time transformations of similar pulses combined by a single graphic pattern called "symbol". All symbols extracted from the signal make up the alphabet. A procedure for narrowing the dimension of the alphabet is shown, which allows you to automatically divide it into clusters according to the degree of coincidence of the code. The results of the practical application of the developed method for the selection of base classes of the geoacoustic emission (GAS) signals related to the objective data of the state of the signal-generating medium are presented. The study used data from the archives of observations IKIR FEB RAS.

1 Introduction

Studies have shown that signals of geoacoustic emission (radio) indirectly reflect the characteristics of processes occurring in the lithosphere [1]. The physical nature of radio pulses is well studied and represents processes associated with changes in external fields [2-4], first of all, tense Earth's crust under the influence of tectonic movements, which also contribute to the detection of seismic waves. The list of other acting physical fields is quite wide - these are natural physical polar fields (gravitational, electromagnetic, thermal, atmospheric pressure, radioactive, the influence of weather phenomena – rain, snow, wind and the like, as well as anthropogenic impact). Acoustic waves in the sedimentary rocks are recorded in the range from 10 Hz to 11 kHz. In the presence of sensors that allow fixing the accompanying physical fields, it is possible to take into account their influence on the radio signal. Thus, it is possible to obtain data on changes in the state

* Corresponding author: senkevich@ikir.ru

of the lithospheric-atmospheric system and collect information about the possible occurrence of seismic events.

The variety of radio pulse shapes is determined by the presence in the medium fractions, voids, leakage of gas bubbles, as well as its layered nature. The heterogeneity of the medium is the result of noticeable distortions of the shape of the radio pulses. In the signal exist multiple interferences, as well as re-reflection effects in the layers, when presenting signals from individual sources. The difference in the number of pulses can be up to three orders of magnitude in intervals of units of seconds. In addition, at the point of signal reception, radios are observed against the background of natural and artificial relatively weak acoustic noises of various physical nature. The circumstances listed above make it difficult to obtain useful information in a detailed analysis of the radio. Most often, such indicators as the integrated energy of incoming pulses are used for analysis [2] or dimension of the set of shapes of the emitted pulses [5]. Methods of statistical data processing are data obtained as a result of current or most common retrospective observations [6]. The accuracy in calculating the size and number of obvious pulses that determine the nature and natural effects is largely determined by the degree of confidence in geophysics. To date, the information stored in various amplitude-phase diagrams has not been investigated due to the lack of sufficient scientific material for constructing effective models for transmitting seismic events and radio, on the one hand, and the imperfection of established methods for processing the multiplicity of radio pulses, on the other hand. The emergence of powerful computing resources for the implementation of algorithms for assessing the state of dynamic systems has opened the possibility of revising the known results of the study of the information content of radio signals.

The search of a connection to changes in the quantitative and qualitative characteristics of radio pulses during the manifestation of changes in the state of the lithosphere according to recorded seismic events is the purpose of the presented study.

2 Description of the method and algorithms

It was assumed that revealing the signs of group behavior of a number of detected radio pulses from the point of view of information theory will allow solving the problem of further clarifying the relationship between the detected signs and the state of the geophysical system of the upper layers of the lithosphere at the moments preceding before during and after seismic events.

The search for the obtained signs of radio signals was carried out in the range from 10 Hz to 4800 Hz from depths recorded in the sensitivity radius of the installed equipment (up to 10 meters). To process the pulse stream of the received radio signal, a method for extracting information (hereinafter referred to as the structural description method) was developed, which is based on a number of principles of the theory of assessing the state of dynamic systems. One of the important provisions of this theory is the ability to characterize the state of the system through a time series of values of the measured parameter [7]. However, it is not possible to directly apply the well-known provisions of the theory of dynamical systems related to the identification of stable and chaotic states based on the calculation of the indicators of the correlation dimension or fractal dimension [8] with respect to the analysis of radio signals. The reason for the difficulty was the basic procedure for compiling the so-called z-vectors from time series. Due to the inevitable hit or loss in the z-vector of a part or a whole individual pulse and / or group of pulses, the radio will be reflected in the graphs in the form of jumps the phase space function, which will lead to incorrect calculation of the indicators of the selected dimension. In addition, the practical numerical calculation of the correlation measurement requires a large amount of data containing several thousand samples, which leads to gross errors in calculating the results of dynamic changes

associated with the duty cycle and the spread of the pulse amplitude stream to the radio signal. The practice of observations shows that when the time interval for data processing extends to units of minutes, the pulse train of the radio signal repeatedly changes the nature of its dynamic behavior. Therefore, it is necessary to find an instrumental indicator that is more suitable for reflecting the dynamic behavior of the observed system.

To highlight informative signs, the study applied the approach known as "Symbolic Dynamics" [7, 9]. Following the general principles of the approach, it is necessary to find and compare a certain set of encountered signal fragments corresponding to the same type of phase trajectories. Such signal fragments should be described by a single template and designated as a symbol. Such signal fragments should be described in one pattern and indicated as a symbol. The entire set of detected symbols highlighted in a given observation interval can be combined into an alphabet characterizing this fragment of the signal. Extracting the alphabets for each subsequent time interval and comparing then with the character set of the previously selected alphabet, we can trace the variability of the composition of characters from one alphabet to another alphabet. There can later try to relate the result to the stationary or chaotic states of the signal generation system. Thus, to implement this approach, initially in the radio signal, it is necessary to distinguish some classes of pulses with close-in-shape changes in the amplitude-phase relations (pattern) and to calculate the frequency of their appearance for a certain observation period. Each class will be represented by a specific template, which is given some symbolic designation. Replacing the original signal with a built-up sequence of characters of the selected alphabet will make up a code sequence, which will be subjected to further analysis. The developed coding method was called the structural transformation of the signal [10]. The following is a detailed description this method.

Processing begins with the procedure for extracting radio signal pulses from interference [11]. For each of the selected pulses, local extrema x_i are determined (maxima and minima of the signal amplitude), as well as the values of time intervals between local extrema τ_i, where $i = 1...M$; M is the number of selected local extrema. The resulting set of pairs of elements $\{x_i, \tau_i\}$ are compared on a "each with each" basis in accordance with the following rule (Eq. 1):

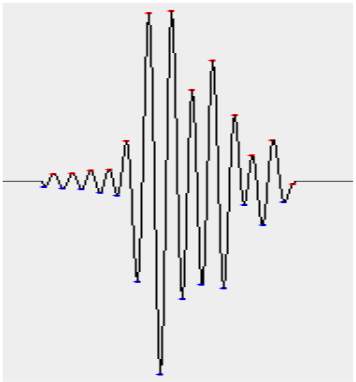
$$r_{i,i+m} = \begin{cases} 1, & x_i > x_{i+m} \\ 0, & x_i \le x_{i+m} \end{cases}, \quad \omega_{i,i+m} = \begin{cases} 1, & \tau_i > \tau_{i+m} \\ 0, & \tau_i \le \tau_{i+m} \end{cases}, \quad m = 1...M \ , \tag{1}$$

where: $r_{i,i+m}$ is the result of a logical comparison of the i-th and $i+m$-th values of the amplitudes of the extrema; $\tau_{i,i+m}$ is the result of a logical comparison of the i-th and $i+m$-th values of the intervals between extrema; M is the size of the matrix. The resulting binary matrices (Eq. 2) and (Eq. 3) have diagonal symmetry due to the algebraic property of the symmetry of the inequalities (if $a > b$, then $b < a$), and in this sense are redundant, therefore, for further operations, a combined matrix, which includes half above the main diagonal for matrix (Eq. 2) and the lower half below the main diagonal for matrix (Eq. 3).

$$\mathbf{R}_i = \begin{pmatrix} r_{i,i} & r_{i,i+1} & \cdots & \cdot(M-2) & r_{i,i+(M-1)} \\ r_{i+1,i} & r_{i+1,i+1} & & & r_{i+1,i+(M-1)} \\ \vdots & & \ddots & & \vdots \\ r_{i+(M-2),i} & & & r_{i+(M-2),i+(M-2)} & r_{i+(M-2),i+(M-1)} \\ r_{i+(M-1),i} & r_{i+(M-1),i+1} & & r_{i+(M-1),i+(M-2)} & r_{i+(M-1),i+(M-1)} \end{pmatrix}. \tag{2}$$

$$\mathbf{W}_i = \begin{pmatrix} \omega_{i,i} & \omega_{i,i+1} & \cdots & i+(M-2) & \omega_{i,i+(M-1)} \\ \omega_{i+1,i} & \omega_{i+1,i+1} & & & \omega_{i+1,i+(M-1)} \\ \vdots & & \ddots & & \vdots \\ \omega_{i+(M-3),i} & & & \omega_{i+(M-3),i+(M-2)} & \omega_{i+(M-3),i+(M-1)} \\ \omega_{i+(M-2),i} & \omega_{i+(M-2),i+1} & & \omega_{i+(M-2),i+(M-2)} & \omega_{i+(M-2),i+(M-1)} \end{pmatrix}. \quad (3)$$

The combined matrix describes the structure of the pulse and has the important property of invariance with respect to the operations of the additive phase shift of the pulse pattern, as well as the amplitude compression (stretching) of the pulse pattern due to the known properties of the inequalities: if $a > b$, then $a + d > b + d$; if $a > b$ and $c > 0$, then $ac > bc$. Matrices are constructed for each selected pulse in the radio signal. An example of radio pulse conversion is shown in Fig. 1.



```
Size = 26
10101010101010101010101010
01101010101010101010101011
11101010101010101010101010
00011010101010101010101011
11111010101010101010101010
00000110101010101010101011
11111110101010101010101010
00000001101010101010101011
11111111101010101010101010
00000000011010101010111011
11111111111010101010000000
00000000000110111111111111
11111111111111000000000000
00000000000001111111111111
11111111111110110000000000
00000000000010101101111111
11111111111110101101000000
00000000000101000111111111
11111111111110101111000000
00000000000101010101111111
11111111110101010101101000
00000000010101010101011011
11111111111010101010111000
00000000000101010101000111
11111111110101010101010110
01010101010101010101010101
```

Fig. 1. Description of the pulse pattern by the relationship matrix (explanations in the text).

In order to build the basic elements of classes (hereinafter referred to as symbols), the procedure of searching for similarly shaped patterns of selected pulses is applied using the following rule for selecting symbols:

$$\text{if}((\mathbf{R}_i \cap \quad \cup \quad \cap \quad \quad \quad (M) \text{ then Symbol},$$

where $\mathbf{R}_i$ and $\mathbf{R}_j$, $\mathbf{W}_i$ and $\mathbf{W}_j$ are the binary matrices of the compared patterns, respectively; $G(M) = 2M^2k$ is the threshold value, k is the empirical coefficient. Symbols are stored in the form of program objects that include information about its size (binary matrix dimension), repeatability statistics over the selected row, a list of sample numbers of the series where pulses of similar shape were detected, and also the length of the pulse length in the sampling frequency and the value of the sum of the squares of the local extremes for estimating the symbol power.

3 Evaluation of noise immunity of the applied algorithms for preprocessing radio-pulses signal

An important role for high-quality signal coding by symbols of the alphabet is the correct selection of radio pulses from the accompanying it noise. For this computational experiment

was conducted to evaluate the noise immunity of the present method. The artificial signal consisted of 100 Berlage pulses [5] with an equal amplitudes and an arbitrarily varying duty cycle for simulate of the real radio pulses. In each of the 50 files, Gaussian noise was added to the signal, so that the signal-to-noise ratio decreased uniformly from 32 dB to -4.6 dB from file to file. Next, the pulse extraction was carried out by a threshold detector at a level of 2 to 5 sigma. The calculation of errors of the second kind for different values of the selected threshold for the selected of pulses was 18% (the allocation of false pulses) for a series with a signal-to-noise ratio of 4 dB (Fig. 2a).

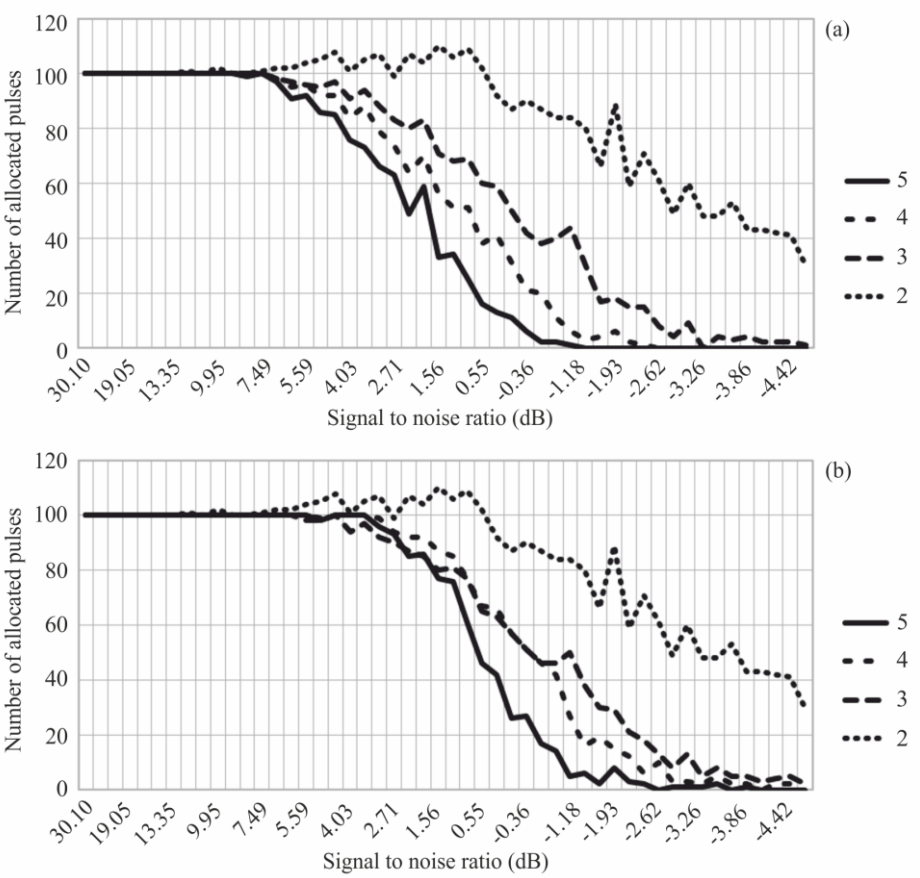


Fig. 2. Results of a computational experiment. Evaluation of the noise immunity for the structural description method of radio-pulse signals without using a logic filter (a) and with using a logic filter (b) for threshold values of 2, 3, 4 and 5 sigma.

To suppress these errors, a selection algorithm was used, which represents an empirical selection rule built on the basis of the well-known property of radio pulses, associated with the behavior of the modulating function, which, when moving between local extremes, every time crosses the center line of the signal. The essence of the rule consists in counting the number of local extrema in the selected fragments, between which there is no discontinuity in the signal function at the level of the set threshold, and comparing it with the length of this fragment. The application of the algorithm, called logical filtering, made it possible to select the first 22 of the 50 generated files for which the maximum permissible error of the first kind was less than 3% for a series with a signal-to-noise ratio of 2.7 dB (Fig. 2b). The gain in noise immunity when applying the selection algorithm, depending on the value of the selected threshold, was from 2.8 to 3.5 dB.

4 Sensitivity assessment of applied structural description algorithms to distortion of radio pulse patterns under the influence of noise

The practice of isolating the flow of radio pulses during field experiments has shown that the previously mentioned external factors, noise, and also the propagation medium noticeably distort the initial shape of the generated acoustic signals. This leads to the allocation of an indefinitely large variety of forms of the selected pulses. Thus, experiments on the identification of radio pulses using their description in various bases showed that the dimension of this diversity is extremely high and exceeds the values of several thousand elements in signals with a duration of 15 minutes. At the same time, the well-known physics of the origin of the initial forms of radio pulses in an inhomogeneous medium [1], which is the soil layer, indicates a limited number of mechanisms of radio generation. The main mechanisms of radio generation are compression and extension of soil masses and slipping of inhomogeneous soil layers under the influence of tectonic displacements, the passage of powerful seismic waves and the movement of gas bubbles. Therefore, the generated variety of radio pulses associated with these mechanisms should be limited and differ mainly in their amplitude, length, and pulse filling frequency. Will call this limited variety the base set. From here, adhering to the chosen approach to the search for the necessary information in radio pulses signals, one of the tasks of the preprocessing follows – the search for the basic set the pulse patterns in the signal. To reduce the observed diversity of the emitted pulses, an algorithm for narrowing the set was developed, reflecting the well-known idea of the formation of classes of similar objects based on the coincidence of elements of the feature vectors. By analogy, the operation of determining the degree of coincidence of the structural description matrices, which are commensurable by order values, is the basis for the mechanism of selecting elements of the basic set of radio pulses.

The matrices $\mathbf{A}_{n\times n}$ and $\mathbf{B}_{m\times m}$ $(n > m)$ will be comparable: if $g = (n - m) / n$ and $0 < g < S_0$, where S_0 is the empirically selected threshold of admissible overlap of the orders of the compared symbol matrices; m, n are the values of the orders of the matrices being compared. Similar matrices of the same order are carried out by counting the coincidences of their elements when matrices are superimposed on each other. If the values of the orders are different $(n > m)$, $n - m + 1$ comparisons are performed, and in the first comparison, a matrix of order m is embedded in a matrix of order n so that the first elements of the first rows coincide. Next, the comparison continues by shifting the smaller matrix along the diagonal of the larger matrix, that is, the smaller matrix is shifted step by step, where each step is a shift of one element to the right and down. Based on the results of $n - m + 1$ comparisons, the result with the greatest number of matches is selected. This result is compared with a threshold value calculated on the basis of an empirically composed formula: $R = gm^2$, which reflects the degree of coincidence of matrix elements. It was called the threshold for the degree of overlap of characters. If the threshold R is exceeded at several steps of comparison, the result with the greatest number of matches is selected and based on this result a decision is made on the permissible proximity of the structures of the compared pulses. In this case, the statistics of the frequency of occurrence of a pulse with a higher-order structural matrix increases by the value of the statistics of the frequency of occurrence of a lower-order matrix, and the picture with a lower-order matrix is removed from the alphabet of the analyzed signal fragment. A good example of comparing two matrices of 5 and 3 orders is shown in Fig. 3.

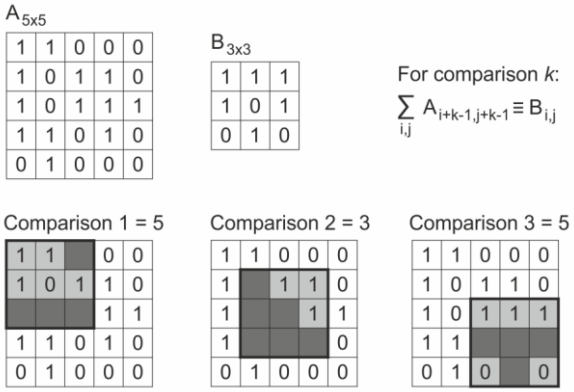


Fig. 3. Representation sequence of the operations for performing the algorithm for comparing matrices and orders 5 and 3, respectively (explanation in the text).

To evaluate the sensitivity of the developed algorithm for narrowing the set, a computational experiment was conducted. In the 21st file, which were selected as a result of the noise immunity assessment experiment, the alphabets for different values of the S_0 and R_0 coefficients were selected and the dimensions of the resulting alphabets were calculated. At the first stage, a condition was set for comparing the symbols in the alphabet with the difference in sizes (the number of extrema included in the pulse) from 0 to 30%. In this case, the fulfillment of the condition for absorption of a smaller symbol matrix by a larger symbol matrix was required. Preprocessing of prepared files was carried out with the following parameters. 22 files with a uniform decrease in the signal-to-noise ratio from a signal without noise to a value of 2.7 dB were analyzed. Low-pass filtering with a cutoff frequency of 4800 Hz (suppression of frequencies outside the pass band was -80 dB, unevenness in the pass band was 6 dB) and logical filtering with a minimum capture of 10 local extrema were used. The adaptive detection threshold was set at 3 sigma. The fixed threshold value of the degree of overlap of symbols.

The obtained results in the dependence from the alphabet dimension with an increase in the signal-to-noise ratio are presented in Fig. 4. It can be seen that the structural description of pulses strongly depends on the local effects of noise on the pulse pattern, which is explained by the "rigid" binding of the symbol matrix to the undistorted pulse pattern. This is also emphasized by the indicator of a slight discrepancy (up to 2.2 dB) the graphs of the dependence of the alphabet dimension on the signal-to-noise ratio for various values of the threshold for the allowable overlap S_0 (0.7–1.0), which is responsible for the dimension limits of the compared symbol matrices. It was experimentally confirmed that the total number of selected pulses remains constant and equal to 100, as expected.

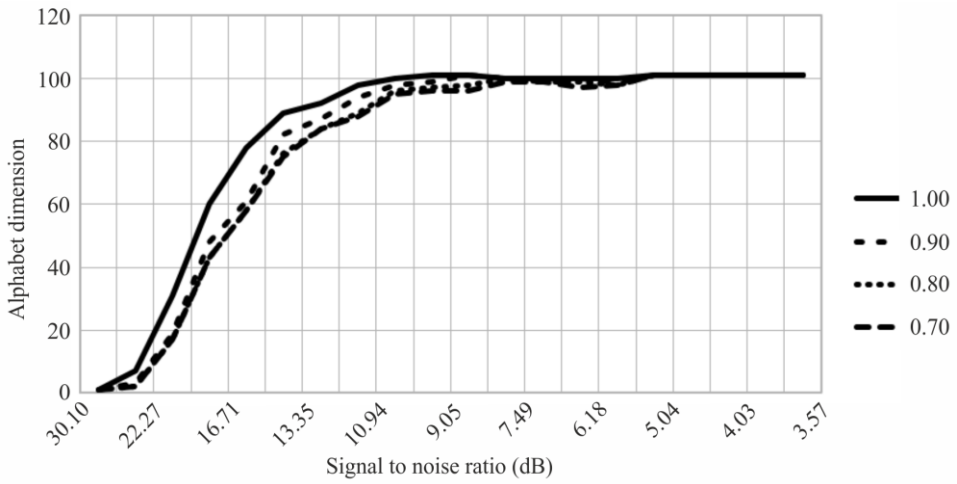


Fig. 4. Graph of the dependence of the dimension of the alphabets of the structural description of artificial radio signals with various signal-to-noise ratios from the choice of the threshold for the allowable overlap of the sizes of S_0 symbols.

The threshold change was regulated by the threshold of the degree of overlap of R_0 symbols (0.6–0.9). A series of calculations was carried out under the condition that only commensurate symbols were compared ($S_0 = 1$). The result of this series of computational experiment is shown in Fig 5.

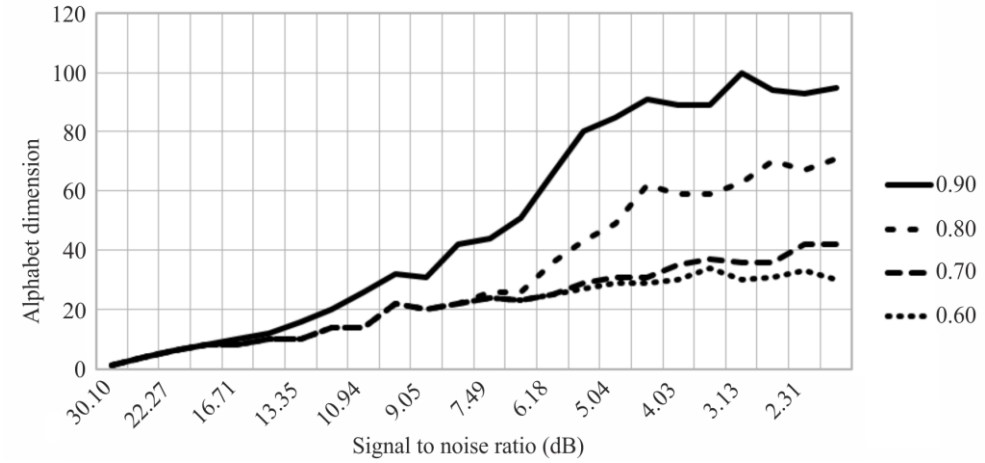


Fig. 5. Graph of the dimension of alphabets obtained during the structural description of artificial radio signals with different signal-to-noise ratios, by the choice of the threshold for the symbols degree overlap (R_0) and a fixed value of the threshold for fixed of the symbol size overlap is S_0=1.0.

The graphs show that "softening" the conditions of the degree of overlap ($R_0 < 0.9$) gives a significant effect of reducing the dimension of alphabets. Moreover, the degree of reduction in dimension of the alphabets is proportional to the change in R_0, and at a value of 0.7, the graph becomes almost linear (Fig. 6).

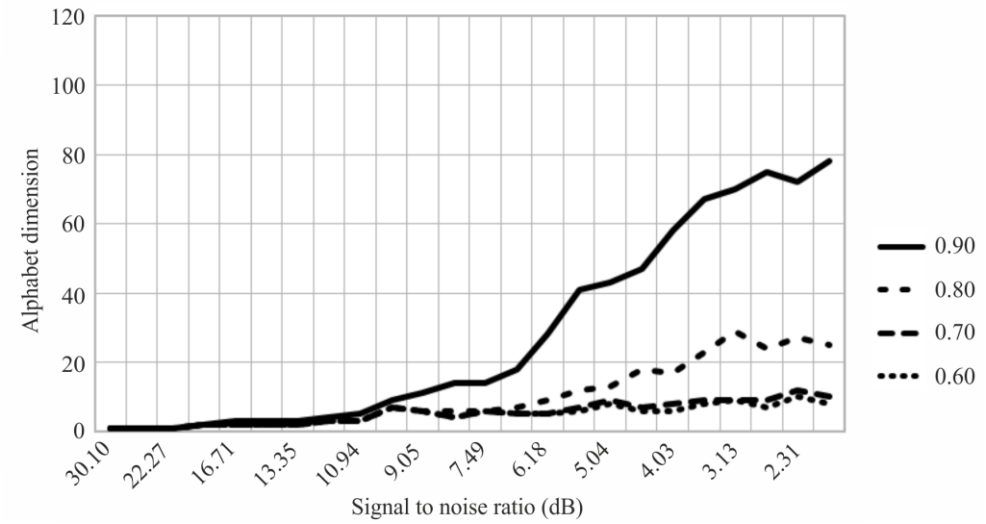


Fig. 6. Graph of the dependence of the alphabet dimension of the structural description of artificial radio signals with various signal-to-noise ratios from the choice of the threshold the degree of overlap of the symbols (R_0) for the allowable fixed threshold value overlap of the symbol sizes ($S_0 = 0.8$).

Comparison of the graphs in Fig. 5 and Fig. 6 gives a clear picture of a regular narrowing of the dimension of alphabets for the corresponding threshold values for the degree of overlap of R_0 symbols.

The final result of this stage of the computational experiment indicates the possibility of effectively lowering the dimension of the alphabet of the structural description of pulsed signals. The application of the presented algorithm for narrowing the alphabet dimension made it possible to reduce the dimensions of the resulting sets of identified pulses by an order of magnitude and more. However, it remains unclear to the end whether it is possible to distinguish some local statistical patterns occurring during a certain epoch of observation on the basis of the structural description, and also to what extent the structural description method allows one to distinguish some basic set of characters based on local statistical patterns.

The method of structural description and the algorithm for narrowing the dimension of the alphabet form the core of the mechanism of automatic clustering of pulses that are close in their structural laws. Therefore, we further solve the problem by displaying some properties of symbols in the alphabet. To solve this problem, a visual display algorithm was developed that clearly demonstrates the statistical properties (frequency of occurrence of a pulse with a specific pattern for the observation era), and sort these symbols by the size inside the alphabet array. Fig. 7 shows an example of the representation of the statistical properties of the characters of the alphabet obtained by the structural description of a simulated radio signal, which was used in a previous computational experiment. The initial signal was subjected to a structural description and selection of pulses with thresholds $S_0 = 0.8$ and $R_0 = 0.8$ in the simulated signal under the influence of additive noise with a signal-to-noise ratio of 9.9 dB. In the beginning, the number of emitted pulses of the same shape was 100. As a result of the distortion by noise, the structural description of the patterns of the multiple pulses turned out to be different. After applying the procedure clustering, the number of selected pulses in total remained equal to 100, and the size of the resulting alphabet decreased to 4 characters.

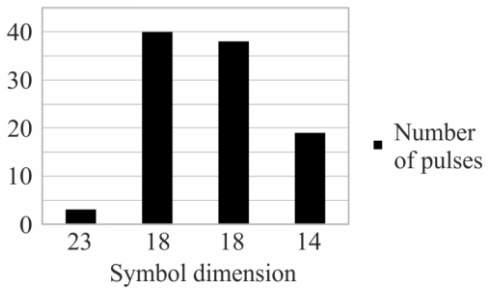


Fig. 7. The result of the structural description of the selected pulses from the radio signal. Horizontally in descending order 4 classes of a certain symbols with the different patterns and dimension are displayed from left to right. Column heights reflect statistics on the appearance this inside characters in each class.

During the repetition of the experiment, it was found that the distribution obtained is resistant to changes in the initial conditions of randomization of the start of the generator of the pseudo-noise sequence of a computer program simulating Gauss noise. This fact indicates that the induced distortions of the patterns of the emitted pulses are regular, and the resulting symbols act as clusters of a certain characteristic alphabet, which can be associated with a certain state of the system generating the signal.

The presented algorithms allow us to represent the radio signal in the form of an encoded message. This message can be used to search for hidden grammar using linguistic analysis methods, which is the subject of promising research. The next stage of research relates to the testing of the developed method of automatic classification on the results of nature observations of the GAE.

5 Results of the nature experiments and its discussion

The study used a file archive in the form of "WAV" sound file containers containing digitized GAE signal recording streams obtained by the acoustic complex recording system installed in tanks at three geophysical observation points of the IKIR FEB RAS: at the Karymshina station of the Paratunka observatory, 20 km from it. The piezoceramic sensor is 1 meter deep in the soil. The selected observation period was December 15-17, 2013, when seismic events were recorded. For processing was take few files containing a series of consecutive three-minute GAE signal recordings were collected. All the results of the preprocessing of the data series presented below were carried out with the set program parameters: $S_0 = 0.8$; $R_0 = 0.8$; $G_0 = 0.8$; the width of the window for calculating the mean square deviation value for calculating the adaptive threshold of 4096 samples at an input sampling frequency of the signal was 44100 Hz; minimum statistics of pulse selection was 24; the boundary of the minimum size of the emitted pulses was 10; logic filter window was 12 extremuma.

To assess pulse activity, the frequency of the appearance of pulses in the observation episode was calculated. Fig. 8 shows the time-ordered results of counting the number of selected pulses and the dimension of the alphabet for each of the selected files. We can assume that both series are independent characteristics, since the cross-correlation coefficient was 0.42. Noteworthy is the noticeable difference in the average deviations: 2.94 and 173.65, respectively, as well as the spread in the number of selected pulses: 692 maximum and 28 minimum value.

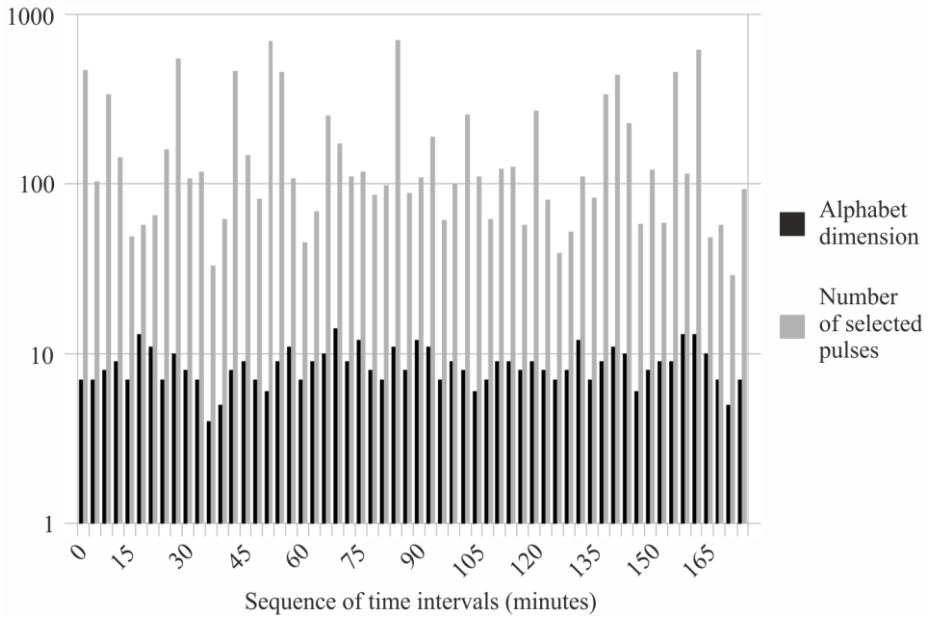


Fig. 8. Results of a full-scale experiment. The number of selected pulses and the obtained values of the dimension of the alphabet for successive intervals of the GAE observation time.

To highlight the basic characters of alphabets obtained from sound files, we used a simple technique for calculating the intersection of sets combined with the described auto-clustering technique. The logic of such selection obviously consists in the fact that in the process of finding the region of intersection of the elements of sets, only that part of them remains that is common. As an example, Fig. 9 shows screenshots of the results of processing four randomly selected files containing 3-minute GAE records received on December 15, 2013 at 20:45; December 15, 2013 at 20:48; December 16, 2013 at 18:06 and December 17, 2013 at 11:30. The upper screenshot contains the normalized ordered distribution of the total statistics of the set of intersection of alphabetic characters obtained for the first and second files. Thin lines with digital signatures indicate the size boundaries of the selected characters. Vertical columns arranged in decreasing order of sizes of characters from left to right display the normalized values of the statistics of occurrence of characters. The middle screenshot contains a similar distribution of the total statistics of the set of intersection of alphabetic characters obtained for the first and third files, and the middle screenshot contains the distribution for the first and fourth files, respectively.

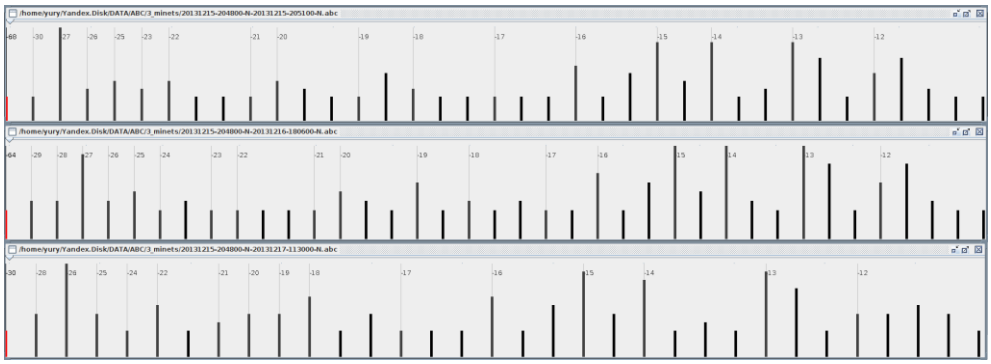


Fig. 9. The results of processing two 3-minute fragments of the GAE signal (explanations in the text).

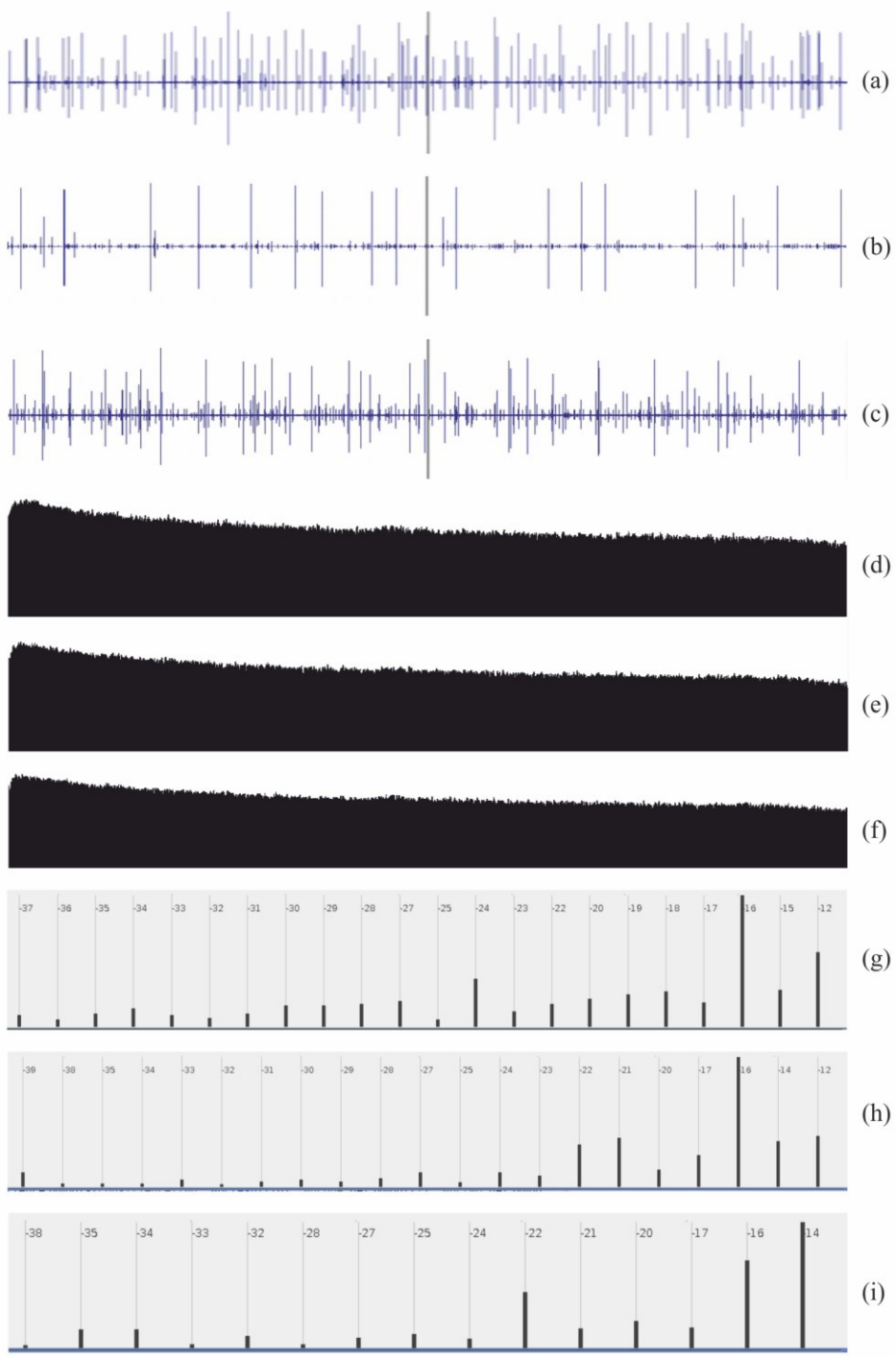


Fig. 10. Results of processing three 15-minute fragments of the GAE signal (explanations in the text).

Comparing the character-by-character patterns of the distribution of normalized statistics of intersections of alphabets for GAE signal fragments, divided by time into 3 minutes, 21 hours 54 minutes and 1 day 13 hours 30 minutes, respectively, it can be noted that their

quantitative and qualitative composition varies slightly. Considering the fact that the weather conditions during this period remained approximately the same, and no seismic events were observed at the selected measurement moments, it can be assumed that the description of the structure and the subsequent analysis of the characteristics of the auto-forming clusters in alphabetical characters, characterizes a certain stationary state of the soil. To illustrate the application of the method of symbolic description on the practice in Fig. 10 in the top are windows screenshots of the program for processing pulse activity of three 15-minute spaced apart episodes of GAE signals: January 17, 2016 at 09:00 (Fig. 10a), 22 January 22, 2016 at 23:30 (Fig. 10b) and February 29, 2016 at 12:45 (Fig. 10c) high-frequency channel in the band 6500–22000 Hz, where the visual picture of the observed changes was the most obvious. According to experts, the first episode corresponds to the background GAE signal, the second to the first manifestations of GAE anomalies, and the third to the moment of the manifestation of an intense energy anomaly, during which, according to subjective data of specialists, an anomaly was detected in the form of periodic high-amplitude bursts (gratings). This anomaly is associated with a change in the state of the signal-forming sedimentary rocks in the area of the Karymshina station. The number of highlighted pulses in each episode was 7283, 6183, and 7937. The apparent discrepancy between the number of visible pulses and the given figures is explained by the emerging change in the dynamic range of amplitudes for various states a medium, which generates the GAE pulses.

To demonstrate the lack of visibility of the application of spectral processing, the three average images (Fig. 10d,e,f) show the energy spectra for each corresponding episode of the GAE signal in the logarithmic scale of the frequency range from 100 Hz to 22050 Hz with a sampling frequency of 44100 Hz and a time window of 4096 samples. Spectra are almost indistinguishable.

The three bottom screenshots (Fig. 10g,h,i) of the working windows of the processing program present pictures of the normalized distribution of alphabet statistics for the corresponding time episodes. The alphabets corresponding to the episodes under consideration have dimensions 22, 21, and 15. There are noticeable changes in the composition of the alphabets, which can be associated with the state of the soil, witch generating GAE signal. In the latter case, a reduction in the dimension of the alphabet indicates a certain type of influence that causes the dominance of a certain mechanism of pulse generation in the soil.

At present, IKIR employees use the structural description to study the steady states of the lithosphere and detect the moments of its transitional states based on a comparison of quantitative and qualitative changes in the composition of alphabets.

6 Research findings

The symbolic description of pulse signals allows you to transfer processing and analysis from classical numerical computation to the field of processing and analysis of code sequences, making available the applicability of linguistics methods, information and meaning search present as identifying hidden rules and grammatical forms of a meta-language. The mechanism of narrowing sets and clustering, included in the developed method, based on obtaining invariant forms for describing pulse patterns, translates the representation of an indefinite variety of pulse shapes to a specific, characterizing the state of the signal source, set of patterns – the alphabet. This opens up the possibility of revealing the hidden correlation of the signal-generating system, as well as building on this basis decision rules for detecting anomalies in the signal and building state detectors on this basis.

The method of symbolic description of radio pulses as a whole expands the technical base of data preprocessing tools for analyzing the state of dynamic systems based on the results of quantitative and qualitative characteristics of signals.

As shown by the results of an experimental verification of the method for structural description of pulsed signals, for the correct operation of the built-in algorithms, it is necessary to ensure the minimum influence of noise interference. The formed basic character sets - patterns that make up the core of the clusters are sensitive to the selected thresholds of the pulse detection detector. Therefore, it is important at the first stages of processing to ensure the elimination of errors of the second kind, which is lead to the appearance of false clusters. For this purpose, empirical rules and / or techniques for isolating radio pulses should be used, based on a priori information about the possible characteristics of the pulse frequency, phase, and amplitude modulation. On the example of constructing a logical filter, the effectiveness of such an application was shown to reduce errors of the second kind.

Further development of research related to the use of the symbolic signal description method will be aimed at describing the behavior of dynamic systems.

The work was supported by Russian Science Foundation (project No. 18-11-00087).

References

1. Yu.V. Marapulets, B.M. Shevtsov, *Mesocsale acoustic emission*, 126 (Dalnauka, Vladivostok, 2012)
2. M.A. Mishchenko, Bulletin KRASEC. Phys. & Math. Sci., **1**(2), 56 (2011)
3. Yu.V. Marapulets, B.M. Shevtsov, I.A. Larionov, Russ. J. Pac. Geol., **31**(6), 59 (2012)
4. Yu. Marapulets, A. Solodchuk, A. Shcherbina, E3S Web of Conf., **11**, 00014 (2016)
5. O.O. Lukovenkova, A.B. Tristanov, V.V. Geppener, SPbGETU «LETI», **4**, 32 (2018)
6. Yu.V. Marapulets, O.O. Lukovenkova, A.B. Tristanov, A.A. Kim, *Methods for recording and for time-frequency analysis of geoacoustic emission signals*, 148 (Dalnauka, Vladivostok, 2017)
7. G. Nicolis, I. Prigogine, *Exploring Complexity* (W.H. Freeman and Company, New York, 1989)
8. V.S. Zaharov, *Analysis of the correlation dimension of seismic energy release time series* ("Dubna" Publ., Moscow, 2007)
9. V.M. Alekseev, M.V. Yakobson, *Symbolic dynamics and hyperbolic dynamical systems (*Mir, Moscow, 1979)
10. Yu. Senkevich, V. Duke, M. Mishchenko, A. Solodchuk, E3S Web of Conf., **20**, 02012 (2017)
11. Yu. Senkevich, E3S Web of Conf., **62**, 03008 (2018)