The Possibility of the Classification of Infrasonic Signals Using Methods for Checking Statistical Hypotheses

A. I. Chulichkov\textsuperscript{a}, N. D. Tsybulskaya\textsuperscript{a}, and S. N. Kulichkov\textsuperscript{b}

\textsuperscript{a} Faculty of Physics, Moscow State University, Moscow, 119991 Russia
\textsuperscript{b} Obukhov Institute for the Physics of the Atmosphere, Russian Academy of Sciences, Moscow, 119017 Russia

e-mail: \textsuperscript{a} achulichkov@gmail.com, \textsuperscript{b} sandratsy@list.ru

Received December 9, 2011

Abstract—This work is devoted to the possibility of the classification of signals using methods for checking statistical hypotheses. On the basis of analysis of the typical peculiarities of signals that belong to the same class we performed an empirical derivation of the shape of the class. The mechanism of the definition of the separability of signals of each class, as well as the level of criterion for defining critical domain were suggested. A classification algorithm was obtained on the basis of the research. The efficiency of the suggested methodology was tested on the problem of separability of infrasonic signals recorded in the atmosphere.

Keywords: classification of signals, checking statistic hypotheses, morphologic analysis, empiric derivation of shape.

DOI: 10.3103/S0027134912020087

INTRODUCTION

The present work considers a variant of the combining of two approaches to treating signals in application to the classification problem. The first approach, the morphologic analysis of images, allows one to distinguish typical peculiarities of signals and to derive the shape of each class empirically \cite{1, 2}. The shape in methods of morphologic analysis provides information that is common for elements of a given class and does not depend on the recording conditions. For example, when the signal amplification coefficient is unknown, the shape must be invariant against change in the signal amplitude.

The second approach is connected with methods for checking statistical hypotheses \cite{3}. It defines the possibility of splitting classes and is the basis of the classification algorithm.

The classical approach to problems of checking statistical hypotheses is connected with the names K. Pirsov and Yu. Neiman \cite{3} and is concluded in setting a decision rule that allows one to accept or reject a hypothesis according to observation \( \xi \). A decision rule is selected in such a way as to make as few errors as possible when accepting (false) hypothesis; at the same time it is accepted that in a certain percentage of cases one will be wrong by rejecting (true) hypothesis.

THE CLASSIFICATION PROBLEM IN TERMS OF ANALYSIS OF STATISTICAL HYPOTHESES

Let us consider the splitting of vectors into two classes. It is assumed that the elements of each class are arbitrary vectors from Euclidian space \( R_n \) with zero mathematical expectation and correlation matrix \( V \) for the first class and \( W \) for the second class. In order to solve the problem we will use a non-randomized criterion that splits the space \( R_n \) into two domains, viz., the domain for accepting hypothesis \( S \) and complement to it called as critical domain \( \bar{S} \). If the realization of an arbitrary vector belongs to the domain, then it is referred to the first one, otherwise, to the second.

Domain \( S \) will be based on the following concepts. We assume that a hypothesis is true. Let us consider basis \( \{ \mathbf{e}_j, j = 1, ..., n \} \) of Karunen-Loev consisting of eigen vectors of matrix \( V \), which correspond to eigen values \( \sigma_j^2, j = 1, ..., n \), ordered in such a way that \( \sigma_1^2 \geq \sigma_2^2 \geq ... \geq \sigma_n^2 \); then an arbitrary vector \( \xi \in R_n \) with zero mathematical expectation and covariance matrix \( V \) will be denoted as \( \xi = \sum_{j=1}^{n} \alpha_j \mathbf{e}_j \), where expansion coefficients \( \alpha_j \) are uncorrelated arbitrary quantities with zero mathematical expectation and dispersion \( \sigma_j^2, j = 1, ..., n \) \cite{2}. After transformation using matrix \( V^{-1/2} \) we obtain vector \( V^{-1/2} \xi = \sum_{j=1}^{n} \alpha_j \mathbf{e}_j \), whose expansion coefficients have unit dispersion and whose squared norm \( t(\xi) = \| V^{-1/2} \xi \|^2 = (\xi, V^{-1} \xi) \) has a mathematical expectation equal to dimension \( n \) of space \( R_n \). Then,
according to the Chebyshev inequality for any number \( \varepsilon > 0 \), one can denote \( P(\alpha(\xi) \geq \varepsilon) \leq \frac{n}{\varepsilon} \).

The last correlation is used for the agreement characteristic of the realization \( x \) of an arbitrary vector \( \xi \in R^n \) with the hypothesis. Substituting \( \varepsilon = \alpha(x) \) we have
\[
P(\alpha(\xi) \geq \alpha(x)) \leq \frac{n}{\alpha(x)} = \frac{n}{\alpha(x)},
\]
which can be interpreted as follows: the greater the value of \( \alpha(x) \), the lower is the probability that for a true hypothesis there will be a value \( \alpha(\xi) \) that exceeds \( \alpha(x) \). The value \( \alpha(x) = \frac{n}{\alpha(x)} \) is the upper boundary of the probability of obtaining realization \( \xi \) that agrees with the hypothesis no better than \( x \). The arbitrary quantity \( \alpha(x) \) is called the reliability of the hypothesis and is used as an agreement characteristic of realization \( x \) with the hypothesis [4].

Similarly we obtain that the agreement of the realization of vector \( x \) of an arbitrary vector \( \xi \in R^n \) with the alternative is defined by the quantity \( \alpha_u(x) = \frac{n}{\alpha(x)} \).

Since errors of the first and second types lead to different losses, we will assume that vector \( \xi \) by realization \( x \) refers to the hypothesis on the difference \( \alpha(x) - \alpha_u(x) \geq c \), where the threshold value is a problem parameter that regulates correlation between errors of the first and second kinds. After performing the corresponding transformations, we obtain that domain \( S \) for accepting the hypothesis is defined by the following expression:
\[
S = \{ x \in R^n : (x, V^{-1}x) \leq \frac{n}{\alpha(x)} \}. 
\]

**EMPIRICAL DERIVATION OF THE SHAPES OF CLASSES**

In order to derive a shape, all the signals belonging to a current class are split into regions using the “caterpillar” method [5]. The obtained vectors were considered as realizations of the arbitrary vectors of dimension \( n \). The mathematical expectations of arbitrary vectors were assumed equal to zero and selective vectors were normalized.

The thus-obtained selection of vectors of the first class was used for building selective covariance matrix \( V \). The number of selective vectors is \( (N - n)kL \) (\( L \) is the number of signals of selected class, \( k \) is the number of sensors registering the signal, and \( N \) is the number of measuring references).

Vectors that were similarly obtained for the second class were considered as the selective values of an arbitrary vector that was distributed according to the alternative and they were used for building the covariance matrix \( W \).

**CHECKING THE SEPARABILITY OF CLASSES AND THE DEFINITION OF CRITICAL LEVELS**

In order to check the separability for each signal of the \( i \)th class we used function \( d_i(c_i) = (x, V^{-1}x) - (x, W^{-1}x) \). After this, the estimations of the correct acceptance of the hypothesis \( P_1(c_i) = d_i(c_i)/N_i \) (\( N_i \) is the number of vectors of the first class) and estimation of the incorrect acceptance of the alternative \( P_2(c_i) = d_i(c_i)/N_j \) (\( N_j \) is the number of vectors of the second class) were defined.

The obtained data allowed us to define the possibility of separating the classes, as well as the threshold \( c_\alpha \) setting criterion (1).

**CLASSIFICATION ALGORITHM**

Final classification at the selected thresholds was performed according to the algorithm below:

1. \( (N - n)kL \) selective vectors were built for a classified signal using the “caterpillar” method.
2. Each selective vector was classified on the basis of criterion (1).
3. It was assumed that the signal can be reliably referred to the class with number \( i \) if the sum of the number of regions of the signal referred to this class divided by the number of vectors of the class exceeded some threshold value \( h \).

**EMPIRICAL DERIVATION OF THE MODEL OF SIGNAL CLASSES**

The efficiency of the method was tested on the problem of the classification of infrasonic signals [4]. The SigLib library that contains these signals is composed of 57 signals split into 5 classes: explosion (class No. 1, ExplosionTest), mountain creep (class No. 2,