Methods Assessment the Probability Density of Discrete Signals in Telecommunications

Yuriy Kropotov, Aleksey Belov

Murom Institute (branch) "Vladimir State University named after Alexander and Nicholay Stoletovs", Murom, Russia

kaf-eivt@yandex.ru

Abstract. This paper is devoted to investigation of problems and methods of acoustic signals modeling in the information and control systems for audio exchange communications. The problems of estimation and approximation of probable density functions, which may assist in distinction of acoustic speech signals and external acoustic noise. We consider the direct and indirect methods, techniques histogram evaluation, ways to overcome incorrect problems.

Keywords: Probability density, discrete signals, telecommunication systems, distribution function.

1   Introduction

Evaluation distributions speech signals and noise, as well as any nature of data, based on empirical derived from experimental results of measurements [1]. There are many methods of preparing such estimates, divided into many parametric and nonparametric, direct and indirect methods.

Under the parametric or understood by classical methods and the methods in which the probability density is known to an accuracy of parameters, it has the form $f(x, \theta) \equiv f_\theta(x)$, where $x \in R^n$ and $\theta \in R^m$ are respectively the vectors of random variables and unknown parameters [2].

The job distribution is characterized, for example, for the detection of problems and evaluation of signals.

The detection tasks assumed that the observed data belongs to one of two or more classes, each of which is characterized by its known a priori probability density $f_k(x)$, or in particular, its own set of parameters θ_k. The density $f_k(x) = f(x, \theta_k)$ and the problem is correlating the observed data to one of the known distributions.

On the contrary, problems of estimation parameter vector θ is considered to be unknown, though the function $f(x, \theta)$ itself may be a known probability density.

If the function $f(x, \theta)$ is not a probability density, the parameter vector estimation methods θ are considered to be non-parametric. In this case - it is a task of approximation or approximation of the observed data. The resulting approximation function $f(x, \theta)$ must satisfy the constraints [1, 3]

$$f(x, \theta) \geq 0 \text{ and } \int_{-\infty}^{\infty} f(x, \theta)dx = 1. \tag{1}$$

A clear distinction between parametric and non-parametric methods is not always possible. Thus, the problem of data closer mixture of known distributions represented density functions $\psi_k(x, \theta_k)$, $f(x, \theta) = \sum_k a_k \psi_k(x, \theta_k)$, $\sum_k a_k = 1$,, more appropriately be classified as non-parametric tasks. However if the coefficients $a_k \geq 0$ are known, the task can be seen as a parametric. For nonparametric problems are the problems of least squares or linear and nonlinear regression. Methods for solving such problems is also called projection methods. It should be noted that the definition of non-parametric methods above only used in mathematical statistics. In the field of systems theory, optimization, approximation and approach them, on the contrary, it is called parametric [4, 7], based on the meaning of the tasks is to find a finite number of unknown parameters.

2   Direct and indirect methods of estimating the probability density

A number of studies estimating the probability density methods are divided into direct and indirect methods. This hallmark of the direct methods is to use a direct link with the required density of empirical data. For example, to direct methods include methods based on the solution of the integral equation relating the probability density of the empirical distribution function

$$\int_{-\infty}^{\infty} I(x - v)f(v)dv = F_n^*(x), \tag{2}$$

where $F_n^*(x)$ is the empirical distribution function of the stepped type. The solution of equation (2) gives the desired estimate of the probability density. The empirical distribution function is given by

$$F_n^*(x) = \frac{1}{N} \sum_{l=1}^{N} I_{(-\infty, x]}(x_l) \tag{3}$$

where $I_{(-\infty, x]}(x_l) -$ the indicator of the set $(-\infty, x]$,

$$I_{(-\infty, x]}(x_l) = \begin{cases} 1, & x_l \in (-\infty, x] \\ 0, & x_l \notin (-\infty, x] \end{cases} \text{ and } N - \text{the sample size.}$$

Problem solving equation (2) with the function (3), as already indicated, it relates to a class of incorrect and requires the use of special techniques. Especially the incorrectness is shown with a small sample size [5]. Thus the need for recovery of the probability density limited amount of data arises frequently, for example, in connection with the analysis and segmentation unsteady, particularly speech signals, the statistical characteristics can only be considered as constant intervals of similar sounds.

Unlike the direct, indirect methods are based on the average risk minimization functional described by expressions of the form

$$R(\theta) = \int Q(x, \theta) dF(x) \cong R_n(\theta) = \frac{1}{N} \sum_{l=1}^{n} Q(x_l, \theta).$$

or their corresponding empirical functionals

$$R_n(\theta) = \frac{1}{N} \sum_{l=1}^{n} Q(x_l, \theta).$$

According to this criterion to indirect methods include, such as the maximum likelihood method [6].

Direct, in principle, other methods, such as histogram techniques and methods based on approximation $\delta -$ functions of a regular feature in the in the expression

$$f(x) = \int_{-\infty}^{\infty} \delta(x - v) f(v) dv. \tag{4}$$

However, a clear distinction between direct and indirect methods, in general, is not always possible. And due to the fact that in both cases, the problem of finding the density estimates may result in one way or another, to the problem of minimizing a functional of the empirical data, in particular, from the empirical distribution function.

3   On nuclear and projection estimates the probability density

The nuclear method for obtaining estimates of the density based on the approximated $\delta -$ function under the integral sign in (4) is a function $K(x)$ defined on some interval of the argument. This function must satisfy the condition

$$\lim_{h \to 0} \frac{1}{h} K\left(\frac{x}{h}\right) = \delta(x) .$$

As a function $K(x)$ frequently used expressions

$$K(x) = \begin{cases} 1/2, & |x| \le 1 \\ 0, & |x| > 1 \end{cases}, \quad K(x) = \frac{1}{\pi(x^2 + 1)}, \quad K(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}, \quad K(x) = \frac{1}{\pi} \frac{\sin x}{x}.$$

The right-hand side of equation (4) after such a substitution is a function of expectation $\dfrac{1}{h}K\left(\dfrac{x}{h}\right)$, which can be replaced by the empirical mean value

If we consider that option is chosen on the basis of the sample size, the probability density estimate in accordance with equation (4) can be written as

$$\hat{f}(x) = \frac{1}{nh(n)}\sum_{l=1}^{n}K\left(\frac{x-x_l}{h(n)}\right) \quad . \tag{5}$$

The convergence of this expression to the desired density estimation provided by the conditions: 1) $h(n) \to 0$ if $n \to \infty$ and 2) $a > 0$ for any number of inequality

$$\sum_{l=1}^{\infty}e^{-\sigma nh(n)} < \infty.$$

The definition of function $K(x)$ can be seen that with the decrease of the parameter h is an increase in the accuracy of the approximation functions δ, but at the same time, increasing the chances of erroneous classification evaluation to class multimodal densities. Conversely, increasing this setting may lead to an erroneous assessment of the assignment to the unimodal density. The problem of choosing a parameter h that arises in this regard stems from the incorrect density estimation problem and for this reason has no unique solution. We can only assert that in assessing unimodal distributions require higher values h than in the case of multimodal. Equation (4), and you can use when assessing the probability density projection method. In this method, an unknown probability density is represented by a polynomial system of normalized orthogonal functions $\left\{\phi_k(x)\right\}_1^m$, while assessment [1]

$$\hat{f}(x) = \sum_{k=1}^{m}a_k\phi_k(x). \tag{6}$$

Substitution of this polynomial in (4) gives the equation

$$f(x) \cong \sum_{k=1}^{m}a_k\phi_k(x) \quad . \text{ and } a_k \cong \int_c^d \phi_k(x)f(x)dx \approx \frac{1}{n}\sum_{l=1}^{n}\phi_k(x_l)$$

Substituting this expression in (6) leads to the formula $\hat{f}(x) = \dfrac{1}{n}\sum_{l=1}^{n}\sum_{k=1}^{m}\phi_k(x)\phi_k(x_l).$

Finally, if you enter the kernel function $K(x, x_l) = \sum_{k=1}^{m}\phi_k(x)\phi_k(x_l)$ the estimate of the density takes the form similar to (5)

$$\hat{f}(x) = \frac{1}{n}\sum_{l=1}^{n}K\left(x, x_l\right).$$

The use of projection methods in which the score is represented by formula (6), is not limited to the case considered. There are tasks that are equally based on a projec-

tion methods, and the integral equation (2). In one of the approaches are evaluated on the smoothed data that is provided by a non-degenerate linear operator of the form

$$B \circ g(x) = \int_c^d K(x, v) g(v) dv \; .\backslash$$

The action of the operator on (2) leads to the equation

$$G \circ f(x) = Q_n^*(x) \tag{7}$$

where

$$G \circ f(x) = \int_c^d K(x, z) \int_c^d I(z - v) f(v) dv dz,$$

$$Q_n^*(x) = B \circ F_n^*(x) = \frac{1}{n} \sum_{l=1}^n \int_{x_l}^d K(x, v) dv,$$

and (6) is an expansion with respect to functions $\{\phi_k(x)\}_1^m$ of operator $G \circ G^H$.

The solution of equation (7) because of its incorrectness reduced to the problem of minimizing the functional

$$J(\hat{f}) = \int_c^d \left(G \circ \hat{f}(x) - Q_n^*(x) \right)^2 dx + \alpha_j \int_c^d \hat{f}^2(x) dx . \tag{8}$$

It is shown that this functional reaches a minimum at values of the coefficients of the polynomial-patients (6)

$$a_k = \frac{\lambda_k b_k}{\lambda_k^2 + \alpha_j},$$

where $b_k = \int_c^d Q_n^*(x) \psi_k(x) dx$ and $\psi_k(x)$, λ_k — its own functions and values of the operator $G^H \circ G$.

In the particular case when the core $K(x, v) = K(x - v)$ of the operator B the operator to convolution $B \circ f(x) = \int_c^d K(x - v) f(v) dv$.

This allows for the minimization of the functional (8) to take advantage of the Fourier transform [1, 6, 7]. Using this, evaluation of density

$$\hat{f}(x) = \frac{1}{n \sqrt{\alpha_j}} \sum_{l=1}^n g\left((x - x_l) / \sqrt{\alpha_j} \right) \; .$$

Here the function $g(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{g}(\omega) e^{j\omega u} d\omega$. It is the inverse Fourier transform

$$\tilde{g}(\omega) = \frac{\tilde{K}(\omega) \tilde{K}(-\omega)}{\tilde{K}(\omega) \tilde{K}(-\omega) + \alpha_j \omega^2} \quad \text{and} \quad K(\omega) = \int_{-\infty}^{\infty} K(u) e^{-j\omega u} du .$$

4   The histogram assessment of the probability density

The histogram is called a bar chart of the distribution of the random variable. The height of each column represents the number of values of the random value falling within the appropriate interval, generally different widths (see Fig. 1.). The ratio of the random variable values n_l from the interval $(x_{l-1}, x_l]$ to the total number of values N is the empirical probability of the event $x \in (x_{l-1}, x_l]$.
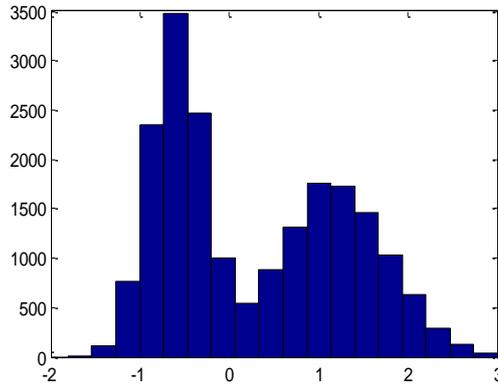

Fig. 1. Histogram of a mixture of two normal distributions

The theoretical value of this probability is written at the same time through a probability density $P\big(x \in (x_{l-1}, x_l]\big) = \int\limits_{x_{l-1}}^{x_l} f(x)dx$.

If we equate the theoretical and empirical density and assume that within each interval change in the probability density can be neglected, the density estimation can be written as

$$f_l = \frac{n_l}{\Delta x_l N}, \; l = 1, \ldots, q, \tag{9}$$

where $\Delta x_l = x_l - x_{l-1} -$ the lenght of the $l -$ interval. When splitting field $(c, d]$ of the random variable values q at equal intervals of length $\Delta x_l = (d - c)/q$ and formula (9) can be written as

$$f_l = \frac{n_l q}{(d - c)N}, \; l = 1, \ldots, q, \tag{10}$$

Count value obtained by the formula (9) or (10), etc. may be used for approximation of the probability density. Units corresponding to estimates, in the first approximation can be found from the expressions

$$\overline{x}_l = \frac{1}{n_l} \sum_{x_i \in (x_{l-1}, x_l]} x_i \,, \ l = 1, \ldots, q \,.$$

The graph obtained by approximating the probability density based on points $(\overline{x}_l, f_l)$, $l = 1, \ldots, q$, in the coordinates x, f .. Thus, depending on the amount of data it can be used as interpolation numerical techniques and approximations of functions. In both cases the problem is to construct the system polynomial

$$P(x, a) \quad \text{by} \quad \text{the} \quad \text{functions} \, \phi_i(x), \quad i = 1, \ldots, m \ P(x, a) = \sum_{i=1}^{m} a_i \phi_i(x) = a^T \phi(x) \,.$$

where $a = \left(a_1, \cdots, a_m \right)^T$ and $\phi(x) = \left(\phi_1(x), \cdots, \phi_m(x) \right)^T$.

In the case of vector interpolation a is the system of equations

$$a^T \phi(x_l) = f_l \,, \ l = 1, \ldots, q = m \,.$$

In matrix form, this system takes the form

$$\Phi^T a = f \,, \tag{11}$$

where $f = \left(f_1, \cdots, f_m \right)^T$ and $\Phi = \left(\phi(x_1), \phi(x_2), \cdots, \phi(x_m) \right)$.

In evaluating you can also take advantage of the generalized method of local interpolation. In this method, a sequence of the form of formula (11) as defined for the corresponding sequence of interpolation intervals. At the same time these formulas are supplemented by restrictions, providing the necessary conditions of conjugation of local solutions, and the order of the polynomial is not required to match the number of points $(\overline{x}_l, f_l)$, that is $q \geq m$.

Approximation of probability density smoothing means is the task of the least squares. The challenge here is to minimize the residual sum of squares polynomial smoothing and density f_l ratings. Functional to be minimized is recorded at the same time as

$$J(a) = \sum_{l=1}^{n} \left(a^T \phi(x_l) - f_l \right)^2 \tag{12}$$

In order to smooth the data, as in the interpolation, you can use the methods of the local approximation, generalizing them in relation to the desired, in particular, a smooth interface polynomials defined on a sequence of intervals and delivering the minimum values of functionals of the form (12) under the constraints set by the terms of pairing.

Histogram methods [1, 6] of estimation of the probability density, especially by interpolation, the problem inherent in the partition of the set of values of the random variable into intervals for small sample sizes. Fig. 2. a, b shows two histo-

grams mixture of normal distributions, the same as in Fig. 1. for a sample of 100 samples.
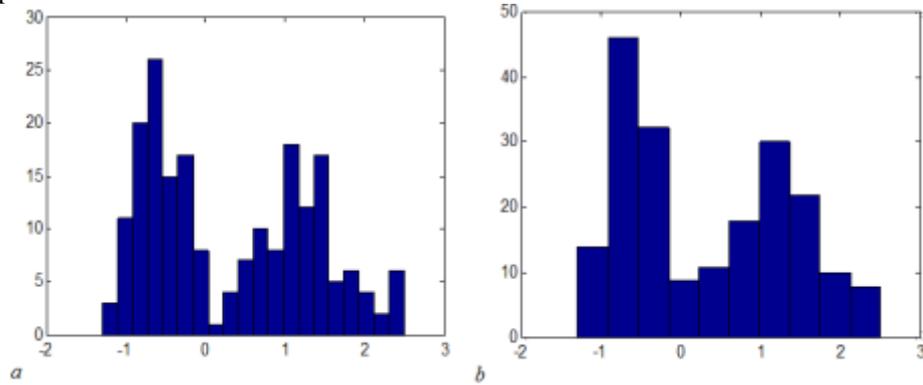

Fig. 2. Histograms mixture of normal distributions by partitioning the range of values in (a) and (b) intervals

This figure shows that the partition of the set of values of the random variable by 20 intervals (Fig. 2 a) interpolation approach does not restore the true form of distribution and draw the right conclusions. The situation is improved by splitting the plurality of slots 10 (Fig. 2 b). In this case, a graph similar in shape to the true probability density bimodal. The solution to this problem, in principle, feasible in the framework of the adaptive partition of the set of values of the random variable in the interval, not necessarily of the same length. Optimal partition is in this case, by varying the lengths and intervals of the centers and of the results of comparison, possibly followed by averaging them.

At the local, including generalized local approximation, partition problem is less acute, and is connected, on the contrary, ensuring sufficient to smooth the number of intervals. However, there is a new question - the question selection algorithm that ensures optimal degree of smoothing empirical estimates (9). Resolution of this issue in principle, feasible methods based on the variation of the free parameters of the algorithm and then selecting the best according to some evaluation criteria.

Another problem for interpolation and approximation of methods for smoothing is a problem of assessment $\hat{f}(x)$ belonging to the class of probability density functions. These conditions within the local approximation can be taken into account by introducing into the problem of minimizing the functional (12), corresponding limitations and within the interpolation approach - by varying the lengths and intervals of the partition centers.

Finding the coefficients of the polynomial (6) optimization methods is the task of the linear regression. In practice, however, these polynomials are often built on the systems standard probability densities nonlinearly depend on a certain set of parameters. In this case, if the input vector of parameters $\theta = \left(\theta_1, \cdots, \theta_r \right)^T$, it is possible to

determine   the   polynomial $P(x, a, \theta) = \sum_{k=1}^{m} a_k \phi_k(x, \theta) = a^T \phi(x, \theta)$,   where

$\phi(x, \theta) = \left(\phi_1(x, \theta), \cdots, \phi_m(x, \theta) \right)^T$.

Accordingly, the estimate of density can be written as

$$\hat{f}(x) = \sum_{k=1}^{m} \hat{a}_k \phi_k(x, \hat{\theta}) = \hat{a}^T \phi(x, \hat{\theta})$$

where the evaluation parameters are the solution to the minimization problem

$$\left\{ \hat{a}, \hat{\theta} \right\} = \arg \min_{a, \theta} \sum_{l=1}^{n} \left(a^T \phi(x_l, \theta) - f_l \right)^2 \ .$$

Finding the vector of parameters a and θ and in this case refers to a class of non-linear problems, which are usually solved by constrained optimization methods.

5   Conclusion

This paper is a study of direct and indirect estimating the density methods of acoustic signals and the probability of interference occurring in the information and control telecommunications systems. Investigated models of nuclear projection probability density estimate that are based on probability density signals approximation in the case of unimodal and multimodal distributions. Applying method of histogram mixture normal distributions estimation shows that the true form of distributions in the partition of values set of a random variable on a different number of slots is not always possible to restore. This solution is provided by an adaptive optimal partition by varying the lengths and intervals of partition centers.

References

1. Kropotov Y.A., Paramonov A.A. Methods of designing information processing telecommunications systems sharing audio algorithms: monograph.-Moscow-Berlin: Direct Media, 2015. 226 p (in Russian).
2. Kropotov Y.A. The time interval determine the probability distribution of the amplitude of the speech signal law Radiotekhnika, 2006. № 6. pp. 97-98 (in Russian).
3. Ermolaev V.A., Kropotov Y.A. About correlation estimating model parameters of acoustic echo. Questions electronics, Vol. 1. №1. pp. 46-50 (in Russian).
4. Kropotov Y. A., Bykov A.A. Algorithm acoustic noise suppression and interference with concentrated formant distribution rejection bands. Questions electronics. 2010. Vol. 1. № 1. pp. 60-65 (in Russian).
5. Kropotov Y.A., Bykov A.A. Approximation of law probability distribution of acoustic noise signal samples. Radio engineering and telecommunication systems. 2011. № 2. pp. 61-67 (in Russian).
6. Ermolaev V.A., Eremenko V.T., Karasev O.E., Kropotov Y.A. Identification of model of discrete linear systems with variable, slowly varying parameters. Radio Engineering and Electronics, 2010. Vol. 55. №1. pp. 57-62 (in Russian).

7. Ermolaev V. A., Karasev O.E., Kropotov Y.A. Interpolation method filtration in problems of speech signal processing in the time domain// Journal of Computer and Information Technology, 2008.- №7.- pp. 12-17 (in Russian).