НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

UDC 681.327.12

ANALYSIS OF THE AUTOMATED SPEAKER RECOGNITION SYSTEM OF CRITICAL USE OPERATION RESULTS

Bisikalo O. V. – Dr. Sc., Professor, Dean of Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine.

Kovtun V. V. – PhD, Associate Professor, Computer Control Systems Department, Vinnitsia National Technical University, Vinnitsa, Ukraine.

Yukhimchuk M. S. – PhD, Associate Professor, Computer Control Systems Department, Vinnitsia National Technical University, Vinnitsa, Ukraine.

Voytyuk I. F. – PhD, Associate Professor, Computer Science Department, Ternopil National Economic University, Ternopil, Ukraine.

ABSTRACT

Context. The article summarizes the statistical learning theory to evaluate the long-term operation results of the automated speaker recognition system of critical use (ASRSCU) taking into account the features of the system's operation object and the structural specificity of such a class of recognition systems.

Objective. The goal of the represented work is the development of a complex set of methods for the ASRSCU's quality parameters stabilization during its long-term operation.

Method. The article formulated set of methods for the ASRSCU's operational risks estimation of its long-term operation. In particular, the dependence of the risk of an incorrect speaker recognition on the features space dimension is described. Based on the formulated measure of informativity, obtained a set of methods to analyze the training sample to identify examples that lead to increased risk. The influence of the phenomenon of the drift of the speech signal parameters on the quality indicators of the ASRSCU is described analytically. An estimation of the operation duration of the ASRSCU, during which it is impractical to re-train its the classifier, is carried out. Recommendations for choosing an optimal ASRSCU's classifier are formulated from the position of its complexity minimization, taking into account the risks of the ASRSCU's long-term operation and the possibility of re-training.

Results. Represented in the article theoretical results are verified by the DET-curves experiments data, which summarize the information from long-term experiments with the ASRSCU, in which, during the features space configuration were taken into account the features based on the power normalized cepstral coefficients based and the features based on the spectral-temporal receptive fields theory. Within the framework of the created theoretical concept, an estimation of the influence of the features space configuration and the type and complexity of the classifier on the stability of the ASRSCU's quality parameters during its long-term operation has been carried out.

Conclusions. For the first time the theoretically analyzed the problem of average risk minimization by empirical operation results of a ASRSCU, where, unlike existing approaches, non-stationary input data with the drift of individual speech signals features and the characteristic parameters of the recognition system classifier were taken into account, which allowed to estimate the risk's confidence interval for conditions for re-training sessions.

KEYWORDS: automated speaker recognition system of critical use, experiment planning theory, factor analysis, statistical learning theory.

ABBREVIATIONS

ASRSCU is automated speaker recognition system of critical use;

CNN is a convolution neural network;

STRF is a spectral-temporal receptive fields;

VAD is a voice activity detection;

NOMENCLATURE

Δ is a impulsive variable;

Ω is a parameter of temporal impulse filter responses;

α is a set of parameters of the classifier;

ε is a accuracy of classifier training;

DOI 10.15588/1607-3274-2018-4-7

θ is a phase of spectral impulse filter feedback;

κ_1 is a weight constant;

κ_2 is a time constant;

κ_3 is a constant, which set the CNN parameters;

κ_4 is a constant, which set the CNN parameters;

κ_5 is a constant, which set the CNN parameters;

κ_6 is a constant, which set the CNN parameters;

κ_7 is a training rank;

ξ_n is a set of n independent, equally distributed random variables;

ρ is a probability that, at least in one of the N functions $Q(z,\alpha_k)$, $k=1,\ldots,N$, the upper limit of the risk $R_{\max}$ exceeds R_m;

τ is a mathematical expectation of the interval between sequences of re-training;

φ is a phase of spectral impulse filter feedback;

$\phi(\ldots)$ is a function of estimation of the drift degree of the input data;

ω is a density parameter of the spectral impulse response filters;

$\{(x_i, y_i)\}$ is a the element of the training sample, X and Y are the set of empirical input and output data of the system;

$1-\delta$ is a reliability of the classifier's training;

b_n^u is a offset for n-i source map,

d is a degree of complexity of the training algorithm;

$F_1(t,\omega)$ is a SNRF-individual feature;

$F_2(t,\omega)$ is a SNRF-individual feature;

$F_3(t,k)$ is a SNRF-individual feature;

f is a frequency;

$G(l)$ is a measure of the learning process complexity;

$g(x)$ is a classification function;

h is a measure of Vapnik-Chervonenkis;

$\hat{h}$ is a Gilbert transformation;

h_S is a spectral impulse response;

h_{sc} is a spectral scale factor;

h_T temporal impulse response;

h_{rt} is a temporal scale factor;

$h(t,f)$ is a impulse response of each filter in bank;

H is a class of hypotheses of indicator functions g;

i is a iterator;

$I(\ldots)$ is a informative measure of rejection of empirical data from educational;

j is a iterator;

$L_q(h)$ is a loss function, which describes the average difference between a random variable Y and $h(X)$;

$\hat{L}_q(h)$ is a empirical risk;

l is a basic length of the training sample;

l_m is a expected average durability of the recognition system exploitation term without re-training;

m is a iterator;

M^{u-1} is a number of entrance maps;

n is a iterator;

N_ω is a number of elements h_{sc};

N_k dimension of the feature $F_3(t,k)$;

k is a iterator;

O is a degree of sufficiency of the training sample;

P_α is a probability of the first kind errors of ASRSCU;

P_β is a probability of the second kind errors of ASRSCU;

P is a unknown probabilities distribution;

P_1 is a empirical risks by sampling S_1 subset A;

P_2 is a empirical risk by sampling S_2 subset A;

$P^l(\cdots)$ is a function of estimating the marginal size of the initial sample;

P_r is a probability of a situation, when testing the system on a plurality m of elements, the empirical risk R_e will exceed the threshold value of risk R_{tr};

$Q(\ldots)$ is a set of corresponding indicator functions;

R assessment of empirical risk R_e;

$R(\alpha)$ is a functional risk;

$R_e(\alpha)$ is a functional of empirical risk;

R_g is a risk of incorrect classification when training a classifier on a sample from which elements were removed at the initial sample length, estimated by the error rate;

R_m is a threshold value of the risk, determined by the testing results;

$R_{\max}$ is a upper limit of risk;

$STRF(t,f,\Omega,\omega,\varphi,\theta)$ is a operation of allocation of SNRF features;

t is a time;

$w_{n,m}^u$ is a convolution core between m-input and n-output source maps;

x_m^u is a input feature map m of a layer u;

$y(t,f)$ is a SNRF spectrogram;

$y_A(t,f)$ is a model of hair cells work;

$y_C(t,f)$ is a affine wavelet transform of the speech signal frame $s(t)$;

$y_{LIN}(t,f)$ is a model of the lateral inhibitory network;

y_n^u is a output feature map n of a layer u;

$Z=\{z_1=(\bar{x}_i,y_i),z_2,\ldots,z_l\}$ is a set of empirical data.

DOI 10.15588/1607-3274-2018-4-7

INTRODUCTION

The automated speaker recognition system of critical use [1], as well as all speaker recognition systems, performs the speaker's person recognition by analyzing the individual attributes isolated from the phonogram with the recording of the speech signal. Of course, the speaker is characterized by the pronunciation variability, due both to internal and external factors. To the internal speech variability factors, we will relate the style, tempo and volume of speech. External speech variability factors are characterized by the type and level of noise in the acoustic and hardware channels of the speech signal propagation, as well as distorted perception of the speech signal due to the reverberation of the speaker's spatial surroundings. Also highlighted such high-level individual characteristics of speech as dialect and speech style, which manifests itself in the acoustic characteristics of the speech signal and the tempo of speech. To establish in ASRSCU the potential for distinguishing internal variability factors, taking into account the high-level individual characteristics and resistance to external variability factors, can be used in the systematic approach to forming the features space, the selection and parameterization of classifier, the formation of a training sample and the regulation of the training process. There are other "extreme" volatility sources of the amplitude-frequency characteristics of the speech signal due to the state of the speaker's health, the acoustic parameters and the geometry of the room where the system is operate, the parameters and the location of the microphones. However, these types of variability are so significantly distorting the meaning of informative speech recognition features that they are reasonably easily identifiable and taken into account when deciding on the result of a speech recognition session, taking into account the degree of distortion and the scope of use and the type of the recognition system.

However, the study [2] showed that during prolonged use of the speaker recognition system, the speech signal parameters drift is due to simple normal physiological processes in the articulatory apparatus of the human, as a result of which the time difference between the training session of the ASRSCU classifier and the recognition session can significantly affect the quality system performance. Consequently, the possible critical use of the speaker recognition system necessarily requires the study of the influence of the operating time on the qualitative performance indicators of the recognition system in order to stabilize them.

The object of study is the individual features of the process of human speech activity and the process of hearing perception of speech signals by a human being and their analysis by the auditory cerebral cortex.

The subject of study is the methods of the pattern recognition theory for the modeling of the recognition system, the methods of the statistical training theory for the analysis of the risks arising from the long-term use of the recognition system, the methods of the neural networks theory for the implementation of the optimal classifier for the recognition system and methods of spectral-temporal receptive fields to describe the process of perception speech signals to the acoustic cerebral cortex of a human.

The purpose of the work is to estimate the risks of long-term operation of the ASRSCU and to propose measures to reduce them.

1 PROBLEM STATEMENT

Let there exist an abstract teacher who offers the ASRSCU's classifier a finite set of examples of an unknown indicator function g over the domain X. On the basis of a sequence of examples $\{(x_i, y_i)\}$, $i = 1,\ldots,l$, $x_i \in X$, $y_i \in Y$, where $y_i = g(x_i)$, $1 \le i \le m$, and $X \subset R$, $Y = \{0,1\}$, randomly selected in accordance with an unknown distribution of probabilities above X, it is necessary to train a classifier with given accuracy $\varepsilon > 0$ and reliability $1-\delta$. Suppose the existence of an unknown probabilities distribution P over $X \times Y$, a classifying function $g(x) = E(Y|X = x)$ and a limited trainning sample with examples $\{(x_i, y_i)\}$, $i = 1,\ldots,l$, where (x_i, y_i) are taken independently, respectively P over $X \times Y$. The classifying function belongs to the hypothesis class H, which does not necessarily contain g. Determine the loss function $L_q(h)$, which describes the average difference between the random variable Y and $h(X)$: $L_q(h) = \left(E|Y - h(X)|^q\right)^{q^{-1}}$, and the empirical risk $\hat{L}_q(h)$: $\hat{L}_q(h) = l^{-1}\sum_{i=1}^{l}\left(y_i - h(x_i)^q\right)^{q^{-1}}$, where the $q \ge 1$ averaging takes place in accordance with P. If we take into account $Y = g(X)$, then the loss function will take the form $L_q(h) = \|g - h\|_{L_q(P)}$, that is L_q the norm over P. Suppose the existence of a classifier training algorithm, a coincidence $l(\varepsilon, \delta)$ for examples of a training sample, for any objective function g and any probability distribution P on X, if the hypothesis is $\hat{h} \in H$ fulfilled $L_q(\hat{h}) \le L^*_{g,q} + \varepsilon$ with probability greater than $1-\delta$, where $L^*_{g,q} = \inf_{h \in H}\|g - h\|_{L_q(P)}$ – the loss of the optimal hypothesis H. This assumption suggests that it is possible to minimize the empirical risk by using the classifier training algorithm, which $\hat{h}$ is the result of a choice h with H a minimum value $\hat{L}_q(h)$, and formulate the purpose of the training procedure as the choice of an element from H, which minimizes the generalization error $R = \int H(z)dP(z)$ based on empirical data $Z = \{z_1 = (\bar{x}_i, y_i), z_2, \ldots, z_l\}$. Next in the article, we will develop the concept formulated above in direction of

DOI 10.15588/1607-3274-2018-4-7

identifying the relationship between the performance indicators of the ASRSCU and the training process parameters and the recognition system classifier parameters for its long-term operation.

2 REVIEW OF THE LITERATURE

In the theory of pattern recognition, one of the applied applications of which is the speaker recognition systems, one of the central problems is to minimize the average risk based on the analysis of empirical data, which developed into the theory of statistical learning [3]. In this theory, many complex problems are investigated, in particular, the restoration of dependencies and distributions density (pattern recognition) and the interpretation of the indirect experiments results. As already noted, the speech signal parameters are inherent in drift, which over time leads to a decrease in the qualitative characteristics of the ASRSCU. In studies [4.5] a hypothesis is formulated on the insignificant effect of the drift of the speech signal parameters on the quality of speaker recognition and is proposed to be compensated by introducing a number of corrective coefficients. So in [6] on the basis of the assumption of constant in time, but a small absolute value of the drift of the speech signal characteristic parameters, its probabilistic estimation is based on the study of the sequence of recognition sessions results and the upper limit of the degree of drift with the given error probability is estimated and taking into account the recognition system classifier training algorithm, but the question of the influence of the number of evaluated data on the reliability of the estimates isn't investigated. The paper [7] describes a method for determining the maximum drift rate allowed for a corresponding recognition system classifier, among which, however, there are no neural networks. In papers [8, 9], the phenomenon of drift recognizes and formulates the requirements for optimizing the parameters of the speaker recognition system classifier re-training process stating, in particular, the requirements regarding the phonetic composition of language materials for re-training, thereby reducing the total amount of study sample. In work [10] the influence of "extreme" variations of speech signals on the quality of the speaker recognition system is estimated, and the permissible limits of variation of spectral individual parameters are estimated. However, in all the aforementioned works, a priori assumptions are made about the nature and parameters of drift in speech signals, therefore, an urgent task is the generalization of the theory of statistical learning to the problem of a speech signal parameters drift in the long-term operation of the speaker recognition system.

3 MATERIALS AND METHODS

In an unknown distribution $P(x,y)$ you can only estimate the empirical risk $R_e = l^{-1}\sum_{i=1}^{l} Q(z_i)$. This goal can be achieved by adjusting the parameters of the classifier α according to the condition $\alpha^* = \arg\min_{\alpha} R_e(\alpha, l)$, evaluating the complexity of the training process as $G(l)\begin{cases} = 2^l, \ if \ l \le d, \\ < l^d (d!)^{-1}, \ if \ l > d, \end{cases}$, where $d+1$ – the smallest size of the set, under which the condition $G(l) = 2^l$ is violated.

On the basis of the foregoing, we consider the problem of minimizing the risk functional

$$R(\alpha) = \int Q(z,\alpha) P(z), \ \ \alpha \in \Lambda, \tag{1}$$

as a task of minimizing the functional of empirical risk

$$R_e(\alpha) = l^{-1}\sum_{i=1}^{l} Q(z_i,\alpha), \ \alpha \in \Lambda \tag{2}$$

over a set of indicator functions $Q(z,\alpha) = \{0,1\}$. In this case (1) is characterized by the probability of incorrect classification $A_\alpha = \{Q(z,\alpha) = 1\}$, and (2) – by the frequency of occurrence of such an event. If all empirical data z are taken from the same distribution, then with probability $1-\eta$ simultaneously for all N functions from a set $Q(z,\alpha_k)$, $k = 1,2,\ldots,N$, inequality is performed

$$R(\alpha_k) < R_{\max}(\alpha_k) = R_e(\alpha_k) + l^{-1}(\ln N - \ln \rho) \cdot \left(1 + \sqrt{1 + 2R_e(\alpha_k) / (\ln N - \ln \rho)^{-1}}\right). \tag{3}$$

The equation (3) allows us to describe the dependence of the incorrect classification risk on the factor space dimension, which can be reduced by applying, in particular, the principal component analysis [11], thus reducing the computational complexity of the recognition task. However, in the context of the critical use of the recognition system, the increase in the wrong classification risk is unacceptable, which can be prevented by removing examples that increase risk from the training sample. This operation is proposed to be carried out on the basis of Shannon's informativeness [12]:

$$R_g = R(s) + \kappa_1 \left(d\left(\ln\left(2d^{-1}(p-s)\right)+1\right)s\left(\ln\left(ps^{-1}\right)+1\right)\right) \cdot (p-s)^{-1}. \tag{4}$$

If the parameters of the initiating probability distribution are unknown, then it is suggested to use the test to identify the distribution point in the context of the speaker's identity, which will be recognized by the ASRSCU [13]:

$$\phi(P_1, P_2) = \sup_A \left(|P_1 - P_2| \left(\min(0.5(P_1+P_2), 1-0.5(P_1+P_2))\right)^{-0.5} \right), \tag{5}$$

where $\sup|S_1 - S_2| = 0.5 dist(S_1, S_2)$ – the empirical measure of the distance between the samples S_1 and S_2. This test also allows us to determine the marginal sample size $P^l(\phi(S,P) > \varepsilon) \le (2l)^d e^{\cdot 0.25 l \varepsilon^2}$ and the boundary of the second kind of errors probability

$$P^{2l}(\phi(S_1, S_2) > \varepsilon) \le (2l)^d e^{\cdot 0.25 l \varepsilon^2}. \tag{6}$$

DOI 10.15588/1607-3274-2018-4-7

In addition to the parameters of the training procedure on the quality parameters of the ASRSCU, the drift of the speech signal parameters is also influenced by the physiological changes in the speech apparatus of the human. If ASRSCU will operate for a long time, the quality of recognition will decrease, as the initiating distribution of input data will change. Next, we call this phenomenon a drift of a compatible distribution $P(\bar{x}, y)$. The theory of machine learning regulates the definition of the adequacy of the amount of training sample O for "drifting" data in the form $O = \varepsilon^{-2}\left(d + \log\delta^{-1}\right)$ [13]. However, the relevance of the relationship of drift with the empirical and true risk is relevant. Assume that the recognition is performed by the Bayesian classifier in terms of deterministic connection $P(\bar{x}, y) \in \{0,1\}$. If the example provided for classification x_i is close to the the training sample data $\min\|\bar{x} + x_i\| < \varepsilon$, then the classification is carried out in accordance with a reliable estimate of conditional probability, and the error probability is $\min\{P(\bar{x}, A), P(\bar{x}, B)\} \approx 0$, and informative x_i – $I(k) \approx 1$, $I(k) = -\log P_k\left(\hat{y}_k = y_k \middle| z_1, z_2, \ldots, z_{k-1}, \bar{x}_k\right)$ which defines the deviation degree of the input data from the data of the training sample. That is, if x_i is mach different from the data of the training sample, then the error probability can be estimated as $2P(A)P(B) \approx 0.5$, and informative $I(k) \approx 2$. In the field of drift, these indicators will generally take a form $\max\{P(\bar{x}, A), P(\bar{x}, B)\} \approx 1$, $I(k) \to \infty$. Consequently, there is a link between the need for re-training of the classifier and the value of empirical risk, embodied in the value of informativity. When creating critical systems, risk management is necessary, so we will combine the re-training operation with the situation of exceeding the value of the empirical risk of some threshold. That is, we will carry out repeated training if with a probability ρ in at least one of an N functions $Q(z, \alpha_k)$, $k = 1, \ldots, N$ the upper limit of risk exceeds the thresholds by the testing results on the sample no less than from the m elements: $R_{\max}(\alpha_k) > R_m$, or by revealing this relation: $R_e(\alpha_k) + K_l\left(1 + \sqrt{1 + 2R_e(\alpha_k)K_l^{-1}}\right) > R_m$, where $K_l = (\ln N - \ln\rho)l^{-1}$, $l \geq m$. Given the previous transformations we obtain $R_e(\alpha_k) > R_m - \sqrt{2R_m K_l}$, or

$$R_e(\alpha_k) > R_m - \sqrt{2R_m(\ln N - \ln\rho)l^{-1}}. \quad (7)$$

Inequality (7) describes the ratio of empirical risk to threshold values for any one $l \geq m$. If the recognition system is used, then based on the generalization of the results of its work for a certain time you can calculate the empirical risks $R_e(\alpha_k)$ for the various classes of system parameters, for example, the length or content of the passphrase, the number of microphones for its recording, acoustic space parameters where the system is operating, etc. If for some class the empirical risk exceeds the threshold $R_e(\alpha_k) \geq R_{tr}$, then for the data of this class, you need re-training.

Let's describe the probability of a situation when after testing a system on a set of elements the empirical risk will exceed the threshold R_{tr}:

$$P_r = \sup_{1 \leq k \leq N}\left(\sum_{l=\lceil R_{tr} m \rceil}^{m}\binom{l}{\lfloor R_{tr} l \rfloor} R(\alpha_k)^{\lceil R_{tr} l \rceil}\left(1 - R(\alpha_k)\right)^{\lfloor l(1-R_{tr}) \rfloor}\right). \quad (8)$$

where unknown true risks $R(\alpha_k)$ are used, the limit values of which can be obtained by analyzing the empirical risks, but taking into account the limited sample size, these estimates will be understated. However, it is possible to obtain a reliable lower boundary r providing for a monotonous increase in the likelihood of re-training P_r with growth $R(\alpha_k)$. Let $(\xi_n, n \geq 1)$ be a sequence of intervals between the re-training procedures of the recognition system, measured in the number of recognition sessions performed. Assume that ξ_n – independent randomly distributed probability variables, then the probability that re-training will occur through sessions will describe by $p\{\xi_n = t\} = q^{t-1}P_r$, where $q = 1 - P_r$, and the probability that re-training will occur no more than through the sessions we describe like $p\{\xi_n \leq t\} = 1 - q^t$. The mathematical expectation of the interval between sequences of re-training will describe by

$$\tau = \mu(\xi_n) = \sum_{t=1}^{\infty} t q^{t-1} P_r = \sum_{t=1}^{\infty}\left(\sum_{x=t}^{\infty} q^{x-1}\right) P_r = \sum_{t=1}^{\infty}\left(q^{t-1}\right)(1-q)^{-1} P_r =$$

$$= \sum_{t=1}^{\infty} q^{t-1} = P_r^{-1},$$ from where the expected average operation duration of the recognition system without re-training is $l_m = m\tau = mP_r^{-1}$. Also, we obtain the mathematical expectation of the number of re-training procedures for sessions: $H(t) = \sum_{n=1}^{\infty} p(s_n \leq t)$, $\lim_{t \to \infty} H(t)t^{-1} = \tau^{-1} = P_r$.

Determine the effect of re-training on some limited positive rational function $G > g(t) > 0$ taking into account the corrective operator Ψ. For a given P_r correction, the evaluation (8) is based on the decomposition of a complex phenomenon on a complete set of incompatible events $P(A) = \sum_B P(B)P(A|B)$:

$$P\left\{\sup_{1 \leq k \leq N} R(\alpha_k) - R_e(\alpha_k)\left(R(\alpha_k)\right)^{-0.5} > \varepsilon\right\}. \quad (9)$$
$$< NP_r a(1-a)^{-1} + N\left(1 - P_r(1-a)^{-1}\right)a^t,$$

DOI 10.15588/1607-3274-2018-4-7

where $a = qe^{-0.5\varepsilon^2 m}$. Using inequality (9) we establish a connection between ε and reliability taking into account that

$$\rho_0 > NP_r a(1-a)^{-1} \Rightarrow a < \rho_0(\rho_0 + NP_r)^{-1} \Rightarrow qe^{-0.5\varepsilon^2 m} < \rho_0(\rho_0 + NP_r)^{-1}:$$

$$\varepsilon > \varepsilon_0 = \sqrt{-2m^{-1}\ln\left(\rho_0((\rho_0 + NP_r)(1-P_r))^{-1}\right)}. \quad (10)$$

We simplify inequality (9) taking into account the fact that the values a and $(1-a)^{-1}$ decreases with growth ε determine what value ensures reliability $1-\rho$, $\rho > \rho_0$, taking into account that $\rho > \rho_0 + Nqa^t \Rightarrow q^t e^{-0.5\varepsilon^2 mt} < (\rho - \upsilon_0)(Nq)^{-1} \Rightarrow -0.5\varepsilon^2 m < \ln\left((\rho-\rho_0)\left(N^{-1}q^{-(t+1)}\right)\right)$:

$$\varepsilon > \varepsilon_1 = \sqrt{-2m^{-1}\left(\ln\left((\rho-\rho_0)N^{-1}\right) - (t+1)\ln(1-P_r)\right)}. \quad (11)$$

Using the above considerations, we obtain on the basis of (9) taking into account (11) the expression to determine the true risk:

$$R(\alpha_k) < R_e(\alpha_k) + \varepsilon_1\left(1 + \sqrt{1 + 2R_e(\alpha_k)\varepsilon_1^{-1}}\right) = $$
$$= R_e(\alpha_k) + \sqrt{-2m^{-1}\ln\left((\rho-\rho_0)N^{-1}q^{-(t+1)}\right)} \times \quad (12)$$
$$\times\left(1 + \sqrt{1 + 2R_e(\alpha_k)\left(-2m^{-1}\ln\left((\rho-\rho_0)N^{-1}q^{-(t+1)}\right)\right)^{-0.5}}\right).$$

and generalizing (10) and (11) we obtain the constraints on the choice ρ_0 for (12):

$$\rho > \rho_0 > 0.5N\left(P_r + q^t\right) + 0.5\rho\left(\sqrt{4NP_r + \left(N\rho^{-1}\left(P_r + q^t\right) - 1\right)^2} - 1\right). \quad (13)$$

We formulate measures on the practical use of the aforementioned theory of the risk assessment of ASRSCU taking into account the procedure of the classifier re-training as a result of the drift of the speech signal parameters. In the context of the foregoing, ASRSCU requires a classifier designed to take into account the balance between the reduction of the empirical risk and the increase in the difference between the empirical and true risk with increasing complexity of the classifier, by which we mean the capacity of the set of input data that the classifier is capable of recognizing. The indicated balance is proposed to be ensured by minimizing the upper limit of true risk for the specified values of reliability and duration of the training sample. Based on [14] we formulate a kind of indicator function that minimizes empirical risk with probability $1-\rho$:

$$R_e - \sqrt{0.5\varepsilon(l)} \le R \le R_e + 0.5\varepsilon(l)\left(1 + \sqrt{1 + 4R_e(\varepsilon(l))^{-1}}\right), \quad (14)$$

where $\varepsilon(l) = 4hl^{-1}\left(\ln\left(2lh^{-1}\right) + 1\right) - 4l^{-1}\ln(0.24\rho)$.

If the re-training procedure is implemented, it's expedient to minimize the true risk, and estimate the limit of the empirical risk for the specified re-training risk. It has been previously grounded that re-training of the classifier with reliability $1-\rho$ willn't occur, if $q^t > 1-\rho$, that allows (14) to obtain an analytical expression for calculating the boundary that describes the effect of re-training on the choice of a classifier:

$$R_r \le R_e + 0.5\varepsilon(m)\left(1 + \sqrt{1 + 4R_e(\varepsilon(m))^{-1}}\right) + \quad (15)$$
$$+ \sqrt{0.5\varepsilon\left(m\left\lfloor\ln(1-\rho)(\ln(1-P_r))^{-1}\right\rfloor\right)}.$$

Consequently, the authors proposed a set of measures for assessing the operational risks of long-term use of ASRSCU. In particular, using (3) describes the dependence of the risk of incorrect classification from the dimension of the factor space. Based on the formulated measure of informativity with the help of (4) it is possible to analyze the study sample on the presence of examples that lead to increased risk. With the help of (8) we describe the influence of the phenomenon of drift of input parameters on the qualitative performance indicators of the ASRSCU, and with the help of (13) an estimation of the operation duration of the ASRSCU is performed, during which it is impractical to re-train the classifier. Also (15), it is possible to choose the optimal classifier on the position of minimizing its complexity, taking into account the risks of long-term use of the ASRSCU and the possibilities of re-training. In general, the above-mentioned material for the first time comprehensively describes the problem of minimizing the average operation risk of the ASRSCU under empirical data, generalized taking into account nonstationary input data with drift patterns and parameters of the recognition system classifier. The limits of confidence intervals of risk are calculated taking into account the procedures of classifier re-training.

4 EXPERIMENTS

The statistical data for the empirical assessment of the adequacy of the above theoretical concepts for the operational risks analysis of the ASRSCU is obtained on the basis of the analysis of the results of long-term use of ASRSCU at the Department of Computer Control Systems of Vinnytsia National Technical University. The mentioned ASRSCU has a classical architecture, which includes a block of preliminary speech signal processing, a block of informative features allocation and a classification block.

In the pre-processing block, the detection of speech activity intervals in phonograms was performed using a two-channel VAD algorithm [16]. Intervals of linguistic activity lasting 3 seconds were segmented into frames of duration 30 ms with 15 ms shift. To compensate for the Gibbs effect, the signal was weighed by the Hemming window. Effects of channel distortions at the factor level were offset by the calculation of the cepstral mean subtraction and, taking into account the sufficient duration of the analysis frameworks, the implementation of the feature warping [17].

DOI 10.15588/1607-3274-2018-4-7

In the block of informative features extraction from each of the received from the block of preliminary processing frames extracted 19 normalized by the power cepstral coefficients [18], their energy and their first and second derivatives. Also, for the presentation of speech signals, the position of the theory of spectral-temporal receptive fields was used, which describes the work of the human auditory system with the involvement of the results of psychoacoustic and neuropsychological studies of the peripheral and central auditory system of mammals in the spectral and temporal spaces [19, 20]. The STRF-description of the speech signal included two stages. At the first stage, the auditory spectrum was obtained as a result of the simulation of the peripheral auditory system. At the second stage on the basis of the first stage results the high-level representations of linguistic representations as the results of simulation of the auditory cortex of the central nervous system of human ware synthesized.

For the implementation of the first stage, an affine wavelet transform $y_C(t,f)$ of the speech signal frame $s(t)$, was initially carried out, which was passed through a bank of cochlear filters: $y_C(t,f) = s(t) *_t h(t,f)$, where $*_t$ – is a convolution operation in the time space. Further, the work of hair cells $y_A(t,f)$ was modeled, which was consistently performed: the operation of high-frequency filtration to emulate the process of converting sound pressure into the speed of hairs; nonlinear compression operation $g(u)$; low frequency filtering operation $w(t)$ to emulate phase blocking of the auditory nerve: $y_A(t,f) = g(\partial_t y_C(t,f)) *_t w(t)$. Next, the work of the lateral inhibitory network of the cochlear nucleus $y_{LIN}(t,f)$ was modeled in the form of a frequency selection operation, for which the partial derivative of the $y_A(t,f)$ frequency was passed through a half-period rectifier: $y_{LIN}(t,f) = \max(\partial_f y_A(t,f), 0)$. And the first stage was ended by receiving the auditory spectrogram $y(t,f)$ by convolution $y_{LIN}(t,f)$ in the time space with a short-term window function $\mu(t,\kappa_2)$: $y(t,f) = y_{LIN}(t,f) *_t \mu(t,\kappa_2)$, where $\mu(t,\tau) = e^{-t\kappa_2^{-1}}$. In Fig. 1 we can visually compare examples of Fourier and STRF-spectrograms.

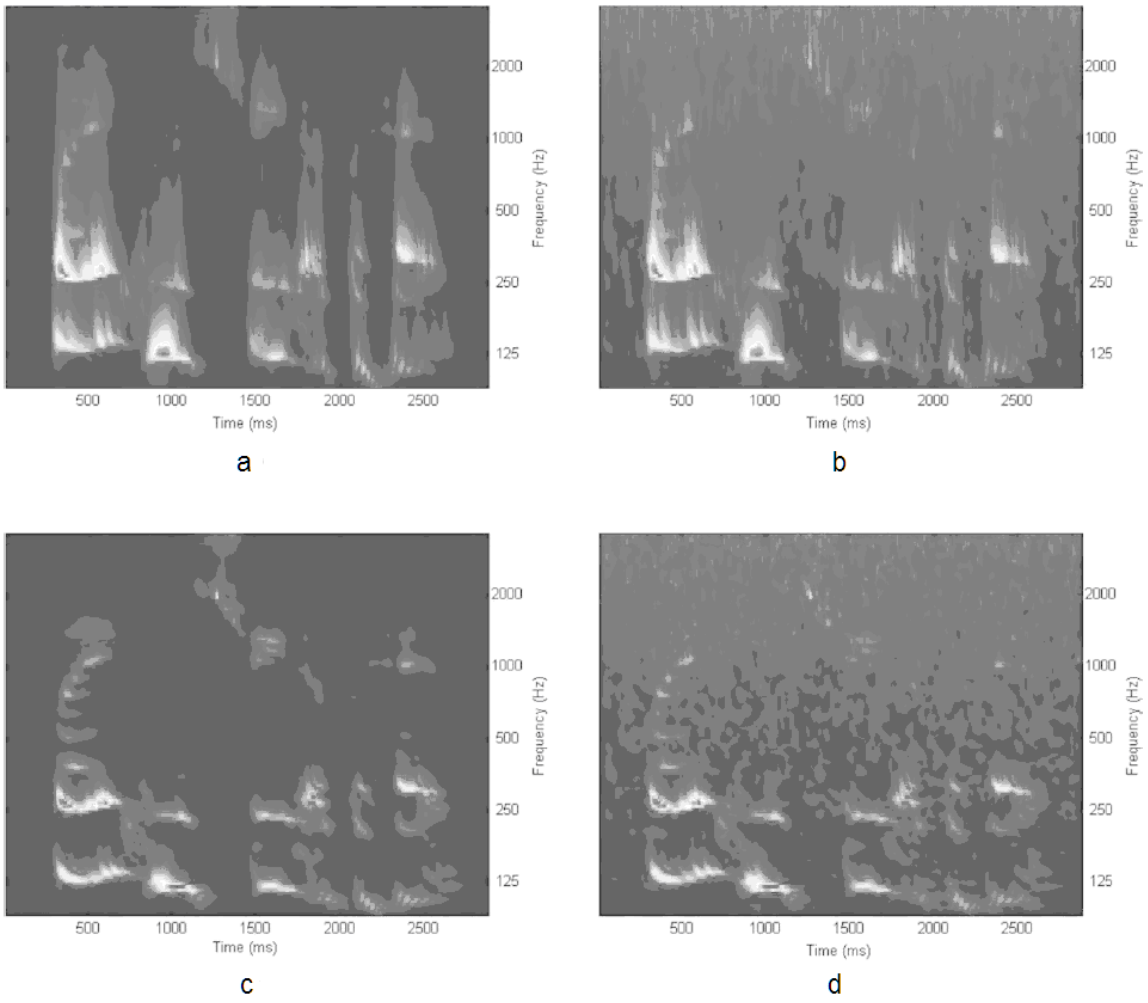

Figure 1 – Visual representation of a speech signal: a, b – Fourier spectrographs of the speech signal without / with noise respectively; c, d – STRF spectrograms of the speech signal without / with noise respectively

DOI 10.15588/1607-3274-2018-4-7

The ongoing second phase was based on obtaining the STRF-function as a result of the combination h_S and h_T: $STRF = h_S \cdot h_T$, where $h_S(f,\omega,\theta) = h_{sc}(f,\omega)\cos\theta + \hat{h}_{sc}(f,\omega)\sin\theta$, $h_T(t,\Omega,\varphi) = h_{rt}(t,\Omega)\cos\varphi + \hat{h}_{rt}(t,\Omega)\sin\varphi$. Operation $STRF(t,f,\Omega,\omega,\varphi,\theta)$ on a spectrograph $y(t,f)$ describes like $STRF(t,f,\Omega,\omega,\varphi,\theta) = y(t,f) *_{tf} [h_S(f,\omega,\theta)\cdot h_T(t,\Omega,\varphi)]$, where $*_{tf}$ is the convolution operation in both time and frequency spaces. Fig. 2 shows a scalable STRF representation of one of the speech signal frames from Fig. 1 in h_{sc}/h_{rt} space and MFCC- representation of the same frame.

In the investigated ASRSCU from the frames of the speech signal, according to the results of the STRF analysis, three informative features were distinguished. The first feature $F_1(t,\omega)$ was obtained by summing the values of all elements of the STRF representation in h_{sc}/h_{rt} spaces:

$$F_1(t,\omega) = \sum_f \sum_\Omega |STRF(t,f,\Omega,\omega,0,0)|, \qquad (16)$$

where $\omega = 1,2,\ldots,N_\omega$, N_ω – is the number h_{sc} elements, and the values of the phase parameters φ and θ, given their small informativity for a speaker recognition task [19], was considered equal to 0 for a simplification of calculations. The second feature was obtained by logarithm $F_1(t,\omega)$:

$$F_2(t,\omega) = \log(F_1(t,\omega)). \qquad (17)$$

The third STRF feature was obtained using the discrete cosine transform (DCT) [20] to $F_2(t,\omega)$:

$$F_3(t,k) = \sum_{\omega=1}^{N_\omega} F_2(t,\omega)\cos\left(2\pi\omega k N_\omega^{-1}\right), \qquad (18)$$

where $k = 1,2,\ldots,N_k$, $N_k \le N_\omega$.

Thus, the vector of informational attributes for one frame of the input speech signal after its processing consisted of 79 elements that are visually represented in the form of a spectrogram-like structure, where the axis of the ordinates is postponed by the number of frames along the abscissa, the values of the ordinate axes correspond to the numbers of informative features, and the intensity of the color shows the value of the corresponding features within the frame, multiplied by the corresponding weighting factor. Such a way of presenting informative features is due to the type of ASRSCU classifier.

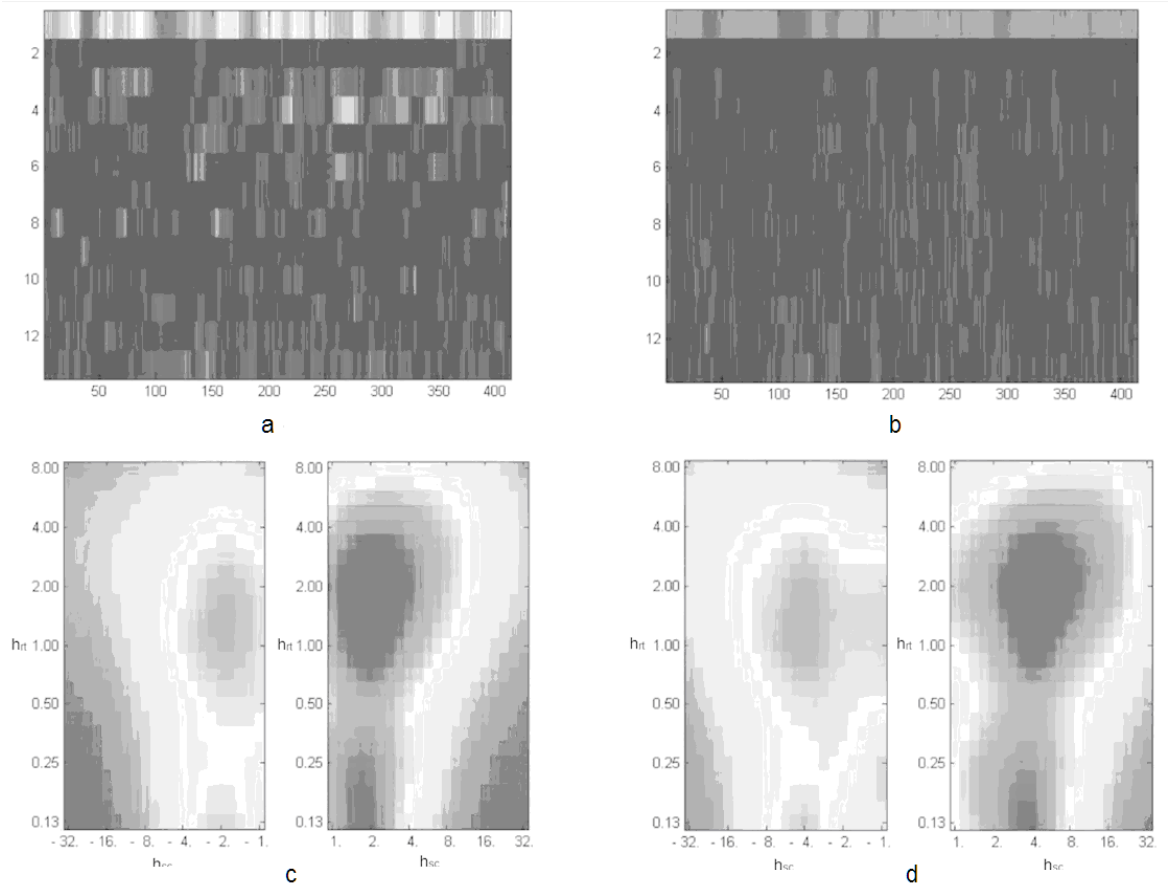

Figure 2 – Visual representation of a speech signal: a, b – MFCC-presentation of speech recording without noise / with noise; c, d – STRF-representation of spectrographs of speech recording without noise / with noise in space h_{sc}/h_{rt}

DOI 10.15588/1607-3274-2018-4-7

In the classification block of the ASRSCU, a convolutional neural network [21] was implemented. Its architecture is designed taking into account the recommendation (15) regarding the complexity of the recognition system classifier, taking into account the features space parameters, operating conditions and the purpose of the ASRSCU. The structure of network (see Fig. 3) includes two convolution layers for features extraction, two sub-sampling layers to reduce features dimension, two local normalization layers, three full-connected layers and finalized by an output SOFTMAX layer.

The convolutional layer of the neural network performs a two-dimensional convolution operation of the fragment of the input image and a filter to extract the height-level features based on the activation functions of the Rectified Linear Units: $y_n^u = \max\left(0, \sum_m^{M^{u-1}} w_{n,m}^u * x_m^u + b_n^u\right)$. To reduce the data dimension, the sub-sampling MAX-pooling type layers are used, at the outputs of which the maximum values of square pieces are obtained in size 3×3, on which the input card is broken off without overlapping. In order to prevent a decrease in the network training process speed the Local Response Normalization procedure implemented, in which the normalized response $B_n^u(x,y)$ at the output of $y_n^u(x,y)$ MAX-pooling sub-sampling layer at the position (x,y) is obtained as $B_n^u(x,y) =$

$$= H_n^u(x,y)\left(\kappa_3 + \kappa_5 \sum_{n=\max(0,n-0.5\kappa_4)}^{\min(M^u,n+0.5\kappa_4)} H_n^u(x,y)^2\right)^{-\kappa_6},$$

where M^u – is a total number of cores in the layer u, and the values $\kappa_3 = 2$, $\kappa_4 = 5$, $\kappa_5 = 10^{-3}$, $\kappa_6 = 0.75$ defined empirically. Dropout technology is implemented to prevent overhaul on the full-connected network layer. SOFTMAX classifier on the output layer of the network determines the probability distribution y_m of membership of the central pixel of the input fragment x_m to C speakers classes like $y_m = e^{\lambda_n}\left(\sum_{n=1}^{C} e^{\lambda_n}\right)^{-1}$, where $\lambda_n = \sum_{m=1}^{M}\left(w_{n,m} * x_m + b_n\right)$, $M = 100$. For network training a stochastic gradient descent method with step 128 was used. The rule for updating weight w_k on k iteration looks like $w_{k+1} = \Delta_{k+1} + w_k$, where $\Delta_{k+1} = 0.9\Delta_k - 0.004\kappa_7 w_k - \kappa_7 \partial L/\partial w_k$ and $\partial L/\partial w_k$ is a derivative. The initial values of the neuron weights on each layer were set using the zero mean Gaussian distribution with a standard deviation of 1. The training error was 0.0002.

5 RESULTS

The main purpose of the experiments carried out with the above-described ASRSCU was to assess the impact of the operation duration on the recognition system quality performance with the generalization of data on the informativity of the attributes space elements. For this purpose, the ASRSCU software was installed on three computers at the Department of Computer Control Systems of the Vinnytsia National Technical University, which operated for two years. Experiments were attended by 6 speakers (4 male and 2 female), each of whom conducted regular recognition sessions at least once every five days (total of over 2000 recognition sessions per speaker per study period) with fixation of results . The possible result of the recognition session was the correct speaker's recognition, the speaker's confusion (the first kind error, Miss) or denial access (second kind error, False Alarm). The results of experiments are presented in the form of detection error trade-off curves, which show the dependence of the likelihood of the first kind errors P_α occurrence from the second kind errors P_β occurrence

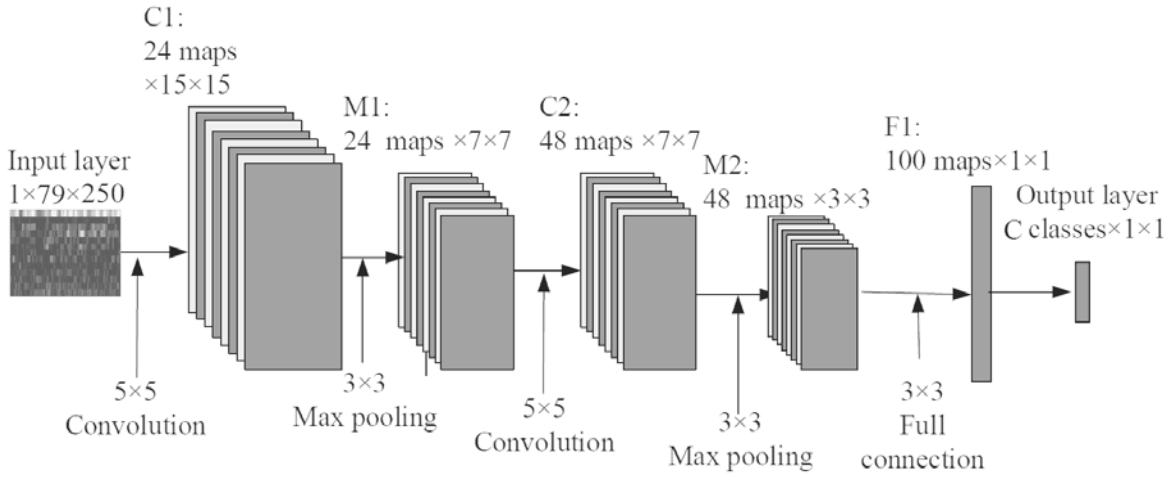

Figure 3 – Architecture of the ASRSCU's convolutional neural network classifier

DOI 10.15588/1607-3274-2018-4-7

probability, with the same threshold decision making recognition system's classifier. In particular, Fig. 4 shows the DET curves depending P_α / P_β on the operation duration of the ASRSCU without the re-training of the classifier, to evaluate the drift of the individual features that characterized the speakers in the recognition process.

Fig. 5 shows the DET curves for P_α / P_β compliance with the recommendations for the re-training frequency and the length of the training sample, taking into account the drift of individual speech parameters. Parameters of the periodicity of re-training and the length of the training sample were determined for the ASRSCU by the formulas (8) and (13) respectively.

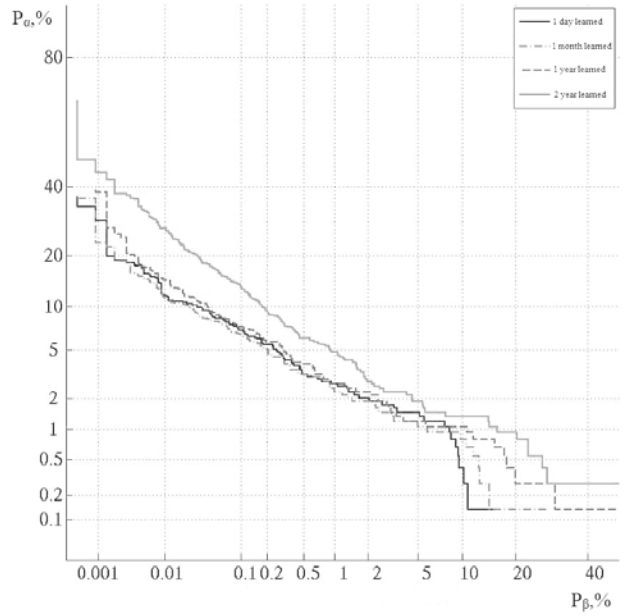

Figure 4 – DET curves P_α / P_β of ASRSCU depending on the duration of operation without re-training of the classifier

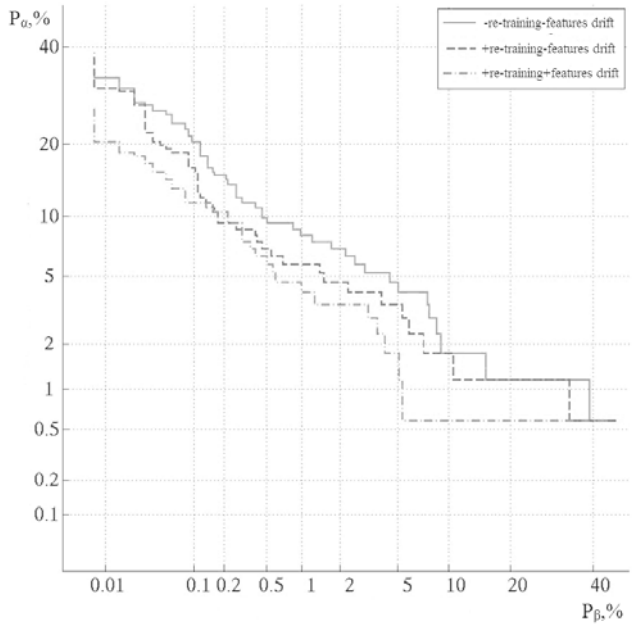

Figure 5 – DET curves P_α / P_β of the ASRSCU, depending on the observance of the recommendations for the re-training frequency and the length of the training sample, taking into account the drift of individual speech parameters

Fig. 6 shows the DET curves P_α / P_β depending on the configuration of the ASRSCU features space, which regularly passed re-training procedures with the parameters of the training sample, regulated by the above theoretical results.

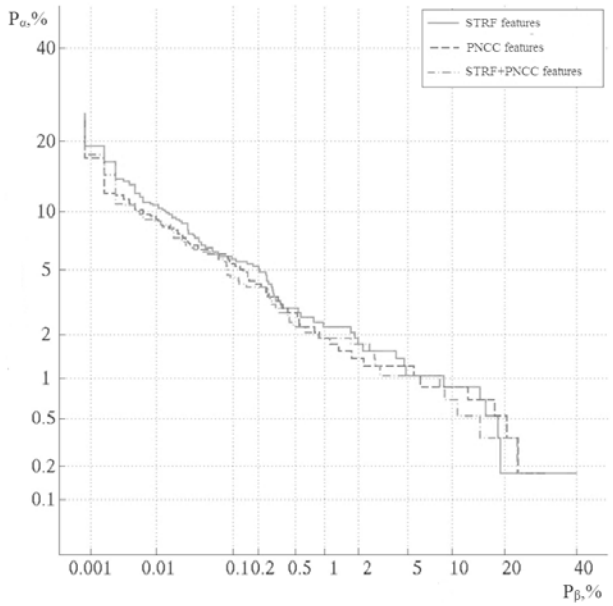

Figure 6 – DET curves P_α / P_β of the ASRSCU, depending on the feature space configuration

The obtained DET curves confirm the theoretical assessments adequacy of the sufficiency of the classifier's complexity to make decisions on the speaker personality of the ASRSCU, confirming the expediency of the re-training procedure of the ASRSCU classifier and correctness determined on the basis of theoretical estimates of this procedure parameters.

6 DISCUSSION

The results of the experiments on a Fig. 4 show that the quality indicators of the ASRSCU during a long-term exploitation process are reduced stochastically without the possibility of identifying an adequate tendency that to some extent allows the use of the speaker recognition system for its intended purpose, but makes its critical application impossible, one of the conditions of which is predictability of the work results.

The results shown in Fig. 5 clearly confirm the expediency of the re-training procedure of the classifier, whose parameters are regulated by the theoretical results obtained in part 4 of the article. It should be noted that, in addition to observing the periodicity of re-training, the obtained results reveal the relationship between the ASRSCU's first and second kind errors probabilities and the composition and the size of the training samples used

DOI 10.15588/1607-3274-2018-4-7

for re-training. On the basis of the results analysis of a long-term exploitation of the ASRSCU, the effect the informative features drift on the quality of the system's operation was found which provides objective material for optimization of the ASRSCU factor space by reevaluating the weight of the informative features, which are subsequently visualized before using the convolutional neural network classifier.

The results shown in Fig.6 on the one hand show a greater informativeness of the features based on the power-normalized cepstral speech signals analysis. However, the features that result from the practical application of the theory of spectral-temporal receptive fields make up only about 4% of the features space, but not only can significantly increase the quality of the ASRSCU, but also make the DET curves more linear, that is, in general, stabilize the decision-making process by the system critical use.

CONCLUSIONS

In the article a theoretical analysis of the long-term operation process of the ASRSCU was conducted, on the basis of which the practical recommendations for the stabilization of the quality indicators of the recognition system are formulated.

The scientific novelty of the obtained results can be attributed to the fact that for the first time a theoretical analysis of the problem of an average risk minimization has been made on the empirical operation results of the speaker recognition system for critical use, in which, unlike the existing approaches, non-stationary input data with drift patterns and characteristic features of the recognition system classifier are taken into account, which allowed to estimate the limits of the risk confidence intervals, provided that the re-training sessions were carried out. The practical consequence of the theoretical analysis is the formulated set of measures for assessing the operational risks of long-term use of the ASRSCU. In particular, using (3) the dependence of the wrong classification risk on the dimension of the factor space is described. Based on the formulated measure of informativity (4), an analysis of the training sample on the identification of elements that lead to increased risk was made. Using (8), we describe the influence of the phenomenon of drift of the speech signals parameters on the qualitative performance indicators of the ASRSCU. With the help of (13), an estimation of the operation duration of the ASRSCU was carried out, during which it was impractical to re-train the classifier. Applying (15) the optimal classifier was chosen from the position of minimization of its complexity, taking into account the risks of long-term use of ASRSCU and the possibility of re-training. In particular, the resulting ASRSCU-based convolutional neural network classifier has a compact structure and confirmed the predicted efficiency. The formulated recommendations correctness is confirmed by empirical results presented in the form of DET curves.

Subsequent studies are planned to devote to the detection of the final potential of the spectral-temporal receptive fields theory in the context of the informative features for speaker recognition synthesis. As the results of experiments have shown, their use not only significantly increases the qualitative performance of the ASRSCU, but also make the DET curves more linear, that's, in general stabilizes the decision-making process by a system of critical use. It is planned to investigate the potential of introducing into the list of information features used in the ASRSCU the human speech source parameters and to make the final factor space optimization.

ACKNOWLEDGEMENTS

The work was carried out within the framework of the cathedral scientific research work number 46K4 "Methods of modeling and optimizing complex systems on the basis of intellectual technologies" at the Department of Computer Control Systems of the Vinnytsia National Technical University with the support of the department staff and the staff of related Department of Automation and Information Measuring Technologies.

REFERENCES

1. Kovtun V. V., M. M. Bykov Ocinjuvannja nadijnosti avtomatyzovanyh system rozpiznavannja movciv krytychnogo zastosuvannja, *Visnyk Vinnyc'kogo politehnichnogo instytutu,* 2017, No. 2, pp. 70–76.
2. Speaker verification over the telephone [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/cad0/bfdec3f4fb1198f63c959580d7217d541a0f.pdf
3. Introduction to Statistical Learning Theory [Electronic resource]. Access mode: http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/pdfs/pdf2819.pdf
4. Learning deep architectures for AI [Electronic resource]. - Access mode: https://www.iro.umontreal.ca/~bengioy/papers/ftml_book.pdf
5. Scaling learning algorithms towards AI [Electronic resource]. Access mode: http://yann.lecun.com/exdb/publis/pdf/bengio-lecun-07.pdf
6. Learning a similarity metric discriminatively, with application to face verification [Electronic resource]. Access mode: http://yann.lecun.com/exdb/publis/pdf/chopra-05.pdf
7. Jang G., Lee T., Oh Y. Learning statistically efficient feature for speaker recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7–11 May 2001: proceedings. Salt Lake City, UT, USA: IEEE, 2002,* pp. 4117–4120. DOI: 10.1109/ICASSP.2001.940861.
8. Unsupervised feature learning for audio classification using convolutional deep belief networks [Electronic resource]. Access mode: http://www.robotics.stanford.edu/~ang/papers/nips09-AudioConvolutionalDBN.pdf
9. Learning methods for generic object recognition with invariance to pose and lighting [Electronic resource]. Access mode: http://yann.lecun.com/exdb/publis/pdf/lecun-04.pdf
10. Learning a nonlinear embedding by preserving class neighbourhood structure [Electronic resource]. Access mode: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.8635&rep=rep1&type=pdf
11. A tutorial on Principal Components Analysis [Electronic resource]. Access mode: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
12. Gustafson J. L., Montry G. R., Benner R. E. Development of parallel methods for a 1024-processor hypercube, *SIAM Journal on Scientific and Statistical Computing,* 1988, Vol. 9, No. 4, pp. 609–638.

DOI 10.15588/1607-3274-2018-4-7

13. Bell J., Casasent D., Bell C. G. An Investigation of Alternative Cache Organizations, *IEEE Transactions on Computers,* 1974, Vol. C-23, No. 4, pp. 346–351.
14. Sergienko I. V. Topical directions of informatics. In memory of V. M. Glushkov. New York, Heidelberg, Dordrecht, London: Springer, 2014, 286 p.
15. Sampling – 50 years after Shannon [Electronic resource]. Access mode: http://bigwww.epfl.ch/publications/unser0001.pdf
16. Mak M. W., Yu H. B. A study of voice activity detection techniques for NIST speaker recognition evaluations, *Computer, Speech and Language*, 2014, Vol. 28, No. 1, pp. 295–313. DOI: 10.1016/j.csl.2013.07.003.
17. Front-end factor analysis for speaker verification [Electronic resource]. Access mode: http://habla.dc.uba.ar/gravano/ith-2014/presentaciones/Dehak_et_al_2010.pdf
18. Power-normalized cepstral coefficients (PNCC) for robust speech recognitions [Electronic resource]. Access mode: http://www.cs.cmu.edu/~robust/Papers/OnlinePNCC_V25.pdf
19. Speech Processing with a Cortical Representation of Audio [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/f1d8/f93cdb64390b3a65f930cee4346c30bd86e4.pdf
20. Using spectro-temporal features to improve AFE feature extraction for automatic speech recognition [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/c7c5/04087f2107f0ea9a3cedeeaf5cc0c48c0c92.pdf
21. Kovtun V. V., Bykov M. M. Doslidzhennja efektyvnosti oznak rozpiznavannja movciv pry vykorystanni zgortal'nyh nejromerezh, *Optyko-elektronni informacijno-energetychni tehnologii*', 2016, No. 2(32), pp. 22–28.

Received 03.06.2018.
Accepted 25.07.2018.

УДК 681.327.12

АНАЛІЗ РЕЗУЛЬТАТІВ ЕКСПЛУАТАЦІЇ ВТОМАТИЗОВАНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МОВЦЯ КРИТИЧНОГО ЗАСТОСУВАННЯ

Бісікало О. В. – д-р техн. наук, професор, декан факультету комп'ютерних систем і автоматики Вінницького національного технічного університету, Вінниця, Україна.
Ковтун В. В. – канд. техн. наук, доцент, доцент кафедри комп'ютерних систем управління Вінницького національного технічного університету, Вінниця, Україна.
Юхимчук М. С. – канд. техн. наук, доцент, доцент кафедри комп'ютерних систем управління Вінницького національного технічного університету, Вінниця, Україна.
Войтюк І. Ф. – канд. техн. наук, доцент кафедри комп'ютерних наук Тернопільського національного економічного університету, Тернопіль, Україна.

АНОТАЦІЯ

Актуальність. У статті узагальнюється теорія статистичного навчання для оцінювання результатів довготривалої експлуатації автоматизованої системи розпізнавання мовця критичного застосування (АСРМКЗ) із урахуванням особливостей об'єкту, із яким працює система, та структурної специфіки такого класу систем розпізнавання.
Мета роботи. Розроблення цілісного комплексу заходів для стабілізації якісних параметрів АСРМКЗ при її довготривалій експлуатації.
Метод. У роботі сформульовано комплекс заходів для оцінювання експлуатаційних ризиків тривалого використання АСРМКЗ. Зокрема, описано залежність ризику неправильної класифікації від розмірності факторного простору. Базуючись на сформульованій мірі інформативності, проаналізовано заходи щодо аналізу навчальної вибірки для виявлення прикладів, які призводять до зростання ризику. Аналітично описано вплив явища дрейфу параметрів мовних сигналів на якісні показники ефективності АСРМКЗ. Здійснено оцінювання тривалості експлуатації АСРМКЗ, на протязі якої здійснювати повторне навчання класифікатора недоцільно. Сформульовано рекомендації щодо вибору оптимального класифікатора АСРМКЗ з позиції мінімізації його складності із урахуванням ризиків тривалої експлуатації АСРМКЗ та можливістю процедури повторного навчання.
Результати. Підтверджено адекватність отриманих у роботі теоретичних результатів представленими у вигляді DET-кривих даними, які узагальнюють інформацію від довготривалих експериментів із АСРМКЗ, у якій при формуванні конфігурації простору ознак враховувалися нормовані за потужністю кепстральні коефіцієнти та похідні від них характеристики і ознаки, отримані на основі теорії спектрально-темпоральних рецептивних полів. В рамках створеної теоретичної концепції проведено оцінювання впливу конфігурації простору ознак та виду і складності класифікатора на стабільність якісних параметрів АСРМКЗ при її довготривалій експлуатації.
Висновки. Вперше теоретично проаналізовано проблему мінімізації середнього ризику по емпіричним результатам експлуатації системи розпізнавання мовця критичного застосування, де, на відміну від існуючих підходів, враховано нестаціонарність вхідних даних із дрейфом індивідуальних параметрів мовних сигналів та характеристичні параметри класифікатора системи розпізнавання, що дозволило оцінити межі довірчих інтервалів ризику за умови здійснення сеансів повторного навчання.
КЛЮЧОВІ СЛОВА: автоматизована система розпізнавання мовців критичного застосування, планування експерименту, факторний аналіз, теорія статистичного навчання.

УДК 681.327.12

АНАЛИЗ РЕЗУЛЬТАТОВ ЭКСПЛУАТАЦИИ АВТОМАТИЗОВАННОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ ДИКТОРА КРИТИЧЕСКОГО ПРИМЕНЕНИЯ

Бисикало О. В. – д-р техн. наук, профессор, декан факультета компьютерных систем и автоматики Винницкого национального технического университета, Винница, Украина.
Ковтун В. В. – канд. техн. наук, доцент, доцент кафедры компьютерных систем управления Винницкого национального технического университета, Винница, Украина.
Юхимчук М. С. – канд. техн. наук, доцент, доцент кафедры компьютерных систем управления Винницкого национального технического университета, Винница, Украина.
Войтюк И. Ф. – канд. техн. наук, доцент кафедры компьютерных наук Тернопольского национального экономического университета, Тернополь, Украина.

DOI 10.15588/1607-3274-2018-4-7

АННОТАЦИЯ

Актуальность. В статье обобщается теория статистического обучения для оценки результатов длительной эксплуатации автоматизированной системы распознавания диктора критического применения (АСРДКП) с учетом особенностей объекта, с которым работает система, и структурной специфики такого класса систем распознавания.

Цель работы. Разработка целостного комплекса мер по стабилизации качественных параметров АСРДКП при ее длительной эксплуатации.

Метод. В работе сформулирован комплекс мер для оценки эксплуатационных рисков длительного использования АСРДКП. В частности, описано зависимость риска неправильной классификации от размерности факторного пространства. Основываясь на сформулированной степени информативности, сформулированы меры по анализу обучающей выборки для выявления примеров, которые приводят к росту риска. Аналитически описано влияние явления дрейфа параметров речевых сигналов на качественные показатели эффективности АСРДКП. Осуществлена оценка зависимости длительности эксплуатации АСРДКП, в течение которой осуществлять повторное обучение классификатора нецелесообразно. Сформулированы рекомендации по выбору оптимального классификатора АСРДКП с позиции минимизации его сложности с учетом рисков длительной эксплуатации АСРДКП и возможности процедуры повторного обучения.

Результаты. Подтверждена адекватность полученных в работе теоретических результатов представленными в виде DET-кривых данными, которые обобщают информацию от длительных экспериментов с АСРДКП, в которой при формировании конфигурации пространства признаков учитывались нормированные по мощности кепстральные коэффициенты и производные от них характеристики и признаки, полученные на основе теории спектрально-темпоральных рецептивных полей. В рамках созданной теоретической концепции проведена оценка воздействия конфигурации пространства признаков и вида и сложности классификатора на стабильность качественных параметров АСРДКП при ее длительной эксплуатации.

Выводы. Впервые теоретически проанализирована проблема минимизации среднего риска по эмпирическим результатам эксплуатации системы распознавания диктора критического применения, где, в отличие от существующих подходов, учтена нестационарность входных данных с дрейфом индивидуальных параметров речевых сигналов и характерные параметры классификатора системы распознавания, что позволило оценить пределы доверительных интервалов риска с условием осуществления сеансов повторного обучения.

КЛЮЧЕВЫЕ СЛОВА: автоматизированная система распознавания диктора критического применения, планирование эксперимента, факторный анализ, теория статистического обучения.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Ковтун В. В. Оцінювання надійності автоматизованих систем розпізнавання мовців критичного застосування / М. М. Биков, В. В. Ковтун // Вісник Вінницького політехнічного інституту. – 2017. – № 2. – С. 70–76.
2. Speaker verification over the telephone [Electronic resource]. – Access mode: https://pdfs.semanticscholar.org/cad0/bfdec3f4fb1198f63c959580d7217d541a0f.pdf
3. Introduction to Statistical Learning Theory [Electronic resource]. – Access mode: http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/pdfs/pdf2819.pdf
4. Learning deep architectures for AI [Electronic resource]. – Access mode: https://www.iro.umontreal.ca/~bengioy/papers/ftml_book.pdf
5. Scaling learning algorithms towards AI [Electronic resource]. – Access mode: http://yann.lecun.com/exdb/publis/pdf/bengio-lecun-07.pdf
6. Learning a similarity metric discriminatively, with application to face verification [Electronic resource]. – Access mode: http://yann.lecun.com/exdb/publis/pdf/chopra-05.pdf
7. Jang G. Learning statistically efficient feature for speaker recognition / G. Jang, T. Lee, Y. Oh // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7–11 May 2001: proceedings. – Salt Lake City, UT, USA: IEEE, 2002. – P. 4117–4120. DOI: 10.1109/ICASSP.2001.940861.
8. Unsupervised feature learning for audio classification using convolutional deep belief networks [Electronic resource]. – Access mode: http://www.robotics.stanford.edu/~ang/papers/nips09-AudioConvolutionalDBN.pdf
9. Learning methods for generic object recognition with invariance to pose and lighting [Electronic resource]. – Access mode: http://yann.lecun.com/exdb/publis/pdf/lecun-04.pdf
10. Learning a nonlinear embedding by preserving class neighbourhood structure [Electronic resource]. – Access mode: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.8635&rep=rep1&type=pdf
11. A tutorial on Principal Components Analysis [Electronic resource]. – Access mode: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
12. Gustafson J. L. Development of parallel methods for a 1024-processor hypercube / J. L. Gustafson, G. R. Montry, R. E. Benner // SIAM Journal on Scientific and Statistical Computing. – 1988. – Vol. 9, № 4. – P. 609–638.
13. Bell J. An Investigation of Alternative Cache Organizations / J. Bell, D. Casasent, C. G. Bell // IEEE Transactions on Computers. – 1974. – Vol. C-23, № 4. – P. 346–351.
14. Sergienko I. V. Topical directions of informatics. In memory of V. M. Glushkov / I. V. Sergienko. – New York, Heidelberg, Dordrecht, London : Springer, 2014. – 286 p.
15. Sampling – 50 years after Shannon [Electronic resource]. – Access mode: http://bigwww.epfl.ch/publications/unser0001.pdf
16. Mak M. W. A study of voice activity detection techniques for NIST speaker recognition evaluations / M. W. Mak, H. B. Yu // Computer, Speech and Language. – 2014. – Vol. 28, № 1. – P. 295–313. DOI: 10.1016/j.csl.2013.07.003.
17. Front-end factor analysis for speaker verification [Electronic resource]. – Access mode: http://habla.dc.uba.ar/gravano/ith-2014/presentaciones/Dehak_et_al_2010.pdf
18. Power-normalized cepstral coefficients (PNCC) for robust speech recognitions [Electronic resource]. – Access mode: http://www.cs.cmu.edu/~robust/Papers/OnlinePNCC_V25.pdf
19. Speech Processing with a Cortical Representation of Audio [Electronic resource]. – Access mode: https://pdfs.semanticscholar.org/f1d8/f93cdb64390b3a65f930cee4346c30bd86e4.pdf
20. Using spectro-temporal features to improve AFE feature extraction for automatic speech recognition [Electronic resource]. – Access mode: https://pdfs.semanticscholar.org/c7c5/04087f2107f0ea9a3cedeeaf5cc0c48c0c92.pdf
21. Ковтун В. В. Дослідження ефективності ознак розпізнавання мовців при використанні згортальних нейромереж / М. М. Биков, В. В. Ковтун // Оптико-електронні інформаційно-енергетичні технології. – 2016. – № 2 (32). – С. 22–28.

DOI 10.15588/1607-3274-2018-4-7