NEWS
OF THENATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 1, Number 335 (2021), 32 – 38 **https://doi.org/10.32014/2021.2518-1726.5**

UDC 004.89
IRSTI 28.23.37

O. Mamyrbayev[2], A. Karelova[1]

[1] Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[2]Institute of Information and Computing Technologies, Almaty, Kazakhstan.
E-mail: morkenj@mail.ru, karelovaayagul@gmail.com

AN OVERVIEW OF THE RECOGNITION ALGORITHM OF A HUMAN VOICE

Abstract. Speech recognition has various applications, including human-machine interaction, sorting phone calls by gender classification, categorizing videos with tags, and so on. Currently, machine learning is a popular field that is widely used in various fields and applications, taking advantage of the latest developments in digital technologies and the advantages of data storage capabilities from electronic media. In this article, we will focus on voice gender recognition for a class of text-dependent systems using the Dynamic time distortion (DTW) algorithm and for a class of text-independent systems, the Gaussian mixture model. With this method, it is possible to distinguish a person's voice with the highest accuracy, since the components of Gaussian mixtures can simulate the personality of the voice. The article presents the results of testing the algorithm, and concludes that the Gaussian mixture model is applicable to solving the problem of identifying a person by voice.

Keywords: algorithm; Gaussian mixture; identification; recognition; classification.

1. Introduction.

Speech is one of the most popular and significant means of communicating people, expressing their emotions, cognitive states and intentions with each other. Speech is produced by humans through a natural biological mechanism in which the lungs release air and convert it into speech, passing through the vocal cords and organs, including the tongue, teeth, lips, etc. As a rule, you can use a speech and voice recognition system to identify your gender. The natural voice recognition system is the human ear. The human ear has an excellent mechanism that can effectively distinguish gender by voice and speech based on attributes such as frequency and volume. Similarly, a machine can be taught to do the same by selecting and including the correct features from the voice data in the machine learning algorithm.

Due to the ease of use, there is currently a growing interest in biometric technologies and the method of biometric identification. Identification is a comparison of features of an object. The advantage of voice identification is convenience and affordable price, and the disadvantage is low reliability[[1]] .

The fundamental disadvantage of all biometrics methods, except speech, is the constancy of the biometric code used, since fingerprints or palms, the pattern of the iris and facial features are unchanged for the individual. This disadvantage prevents the use of these methods in cases requiring particularly high reliability of identity identification, since the immutable biometric code can be read by malicious intrusion into the recognition program[[2]].

Unlike fixed-parameter biometrics, voice verification has virtually unlimited potential to reduce error by using increasingly long speech messages. Voice verification can be used in the dark, at a distance, in particular, over a standard telephone channel, in conditions where it is impossible to get a face image.

In modern voice identification systems, text-dependent identification is used to increase reliability, for example, the utterance of a passphrase, which is randomly generated each time. The use of individual characteristics and matching of the generated and the detected passphrases increases the reliability. Text-independent identification implies the use of only individual features[[3]].

An important characteristic of the voice identification system is the speed (speed) of identification. Performance is especially important for applications that process large databases of voice data and work in

real time. Performance improvement can be achieved through the use of new fast algorithms of data processing. Thus, voice identification of a person, despite the shortcomings indicated in this work, under certain conditions has significant advantages that need to be developed. In the problem of voice identification, various mathematical, algorithmic, and technical methods are used, starting with the stage of voice recording and ending with the stage of classification. Virtually every identification system has four main stages: signal acquisition, signal preprocessing, feature extraction, and feature classification.

2. The stage of receiving the signal.

The method of receiving or recording a voice signal, in most cases, is to record the signal using a microphone and present the signal digitally using an analog-to-digital Converter. As an analog-to-digital Converter, a personal computer sound card or a digital voice recorder is usually used. However, it is still necessary to ensure minimal computing costs while maintaining accuracy, noise immunity to various types of interference, and sufficient reliability with common hardware[[4]-[5]].

3. Pre-processing stage.

The received digital signals, as well as analog ones, contain a certain amount of distortion and interference. Distortions are understood as distortions of the speech-forming tract (for example, throat disease) and the speech-transmitting channel (for example, distortions of the telephone channel).

Pre-processing stage. The received digital signals, as well as analog ones, contain a certain amount of distortion and interference. Distortions are understood as distortions of the speech-forming tract (for example, throat disease) and the speech-transmitting channel (for example, distortions of the telephone channel).

4. The stage of feature extraction.

Feature extraction usually takes place using Fourier transform, wavelet transform, linear prediction, and others. The transformation coefficients are used as features. Currently, the voice characteristics that can uniquely identify a person's identity are not precisely defined. The choice of features also affects the reliability of identification. There are methods that describe the integral characteristics of the human voice and are used to extract tones, speech dynamics, and prosodic characteristics. Such methods are the Fourier transform (amplitude-frequency distribution), the cepstral transform (amplitude-time distribution), and the linear prediction transform (amplitude-frequency distribution). There are also formant methods and phoneme extraction methods.

5. Stage of classification of features.

This stage includes the application of mathematical classification methods, which are used to make decisions, as well as the calculation of classification errors.

Speech recognition systems are based on the principles of recognition of recognition forms. The methods and algorithms that have been used so far can be divided into the following large classes [[6]-[7]]:

Classification of speech recognition methods based on comparison with the standard.

- Dynamic programming — time dynamic algorithms (Dynamic Time Warping).
- Context-sensitive classification. When it is implemented, separate lexical elements are distinguished from the speech stream-phonemes and allophones, which are then combined into syllables and morphemes.
- Methods of discriminant analysis based on Bayesian discrimination (Bayesian discrimination);
- Hidden Markov models (Hidden Markov Model);

Neural networks (Neural networks).

Dynamic Time Warping (DTW) is a dynamic time scale transformation algorithm, a dynamic programming method that allows you to find the distance between two time series. As a rule, such sequences have different lengths, so you have to make measurements at different speeds. The main advantage of this algorithm is the ease of implementation [[8]-[9]].

Gaussian mixture models can be applied not only to model the characteristics of the speaker's voice, but also to record the voice signal and the environment. Each of the components of the model reflects some common features of the voice, but individual when they are reproduced by each speaker. Gaussian mixture models have proven to be effective because they have high recognition accuracy. That is why this approach can be successfully used to solve the problem of identifying a text-independent speaker [[10]].

The weighted sum of M components representing the Gaussian mixture model is calculated using the formula [[11]]

$$P(\bar{x}|\lambda) = \sum_{i=1}^{M} p_i \, b_i(\bar{x}) \quad (1)$$

where $\bar{x}$ is a d-dimensional vector of random variables, pi, 1≤ i ≤ M – weights of the model components, $b_i(\bar{x})$, 1≤ i≤ M – density functions of the distribution of the model components:

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\sum i|1/2} exp\left\{-\frac{1}{2}(\bar{x} - \bar{\mu_i})^T \sum_i^{-1} \bar{x} - \bar{\mu_i}\right\} \quad (2)$$

where $\bar{\mu_i}$ is the expectation vector and $|\sum i|$ is the covariance matrix. The weights of the mixture must satisfy the condition:

$$\sum_{i=1}^{M} \bar{p_i} = 1 \quad (3)$$

The entire Gaussian mixture model is defined using expectation vectors, covariance matrices, and mixture weights for each of the model components:

$$\lambda = \{p_i, \bar{\mu_i}, \sum i\}, i = 1, \dots M \quad (4)$$

When using the method, each speaking person can be represented by their own Gaussian mixture model.

To build a system of automatic identification of a person by voice using Gaussian mixtures, it is necessary to solve the following subtasks:

Extract and process the features of the input speech signal;

Develop an algorithm for initializing and evaluating model parameters;

Determine the number of components of the Gaussian mixture model.

First, an analog-to-digital conversion of the audio signal is performed. During sampling, the signal is divided into separate values of the quantized amplitude at certain time intervals.

The entire signal recording is viewed by Windows of pre-set duration that overlap. It is recommended to choose the duration of the time window within 20-30 MS. In this paper, to simplify the calculations, the duration of each window was chosen to be 25 MS.

Then the digitized signal is viewed in small fragments (frames) that are characteristic of individual vocal components of the speech signal and for which it is assumed that the signal retains its properties constant for a given period of time. Next, the window function is selected. The time window function must take a non-zero value inside a certain time interval, and it must be zero outside of it. Then the window function is sequentially superimposed on the signal frames, and information is extracted from the speech frame. This information is extracted by multiplying the value of the signal x[t] taken at time t with the value of the window function w[t] taken at time t:

$$y[t] = w[t]x[t] \quad (5)$$

The characteristics of the window function are the following parameters: width (in milliseconds), offset (the number of milliseconds between the borders of consecutive Windows), and shape. In this paper, a Hamming window with a width of L = 30ms and an offset of 10 MS is used.:

$$w(t) = \begin{cases} 0{,}54 - 0{,}46\cos\left(\frac{2\pi t}{L}\right), 0 \le t \le L - 1 \\ 0, \text{иначе} \end{cases} \quad (6)$$

After filtering each segment, we get a complete signal, which is free of noise, interference and other distortions that can interfere with the correct recognition of the speaker.

Next, it is necessary to extract information about the spectral components from the signal obtained at the previous stages of the algorithm, for which a discrete Fourier transform is used. A signal divided into frames is fed to the input of the computer, and at the output of the computer for each of The t frequency ranges, we get a complex number X[k], which is the amplitude and phase of the original signal. X[k] is calculated by the formula:

$$x_k = \sum_{n=0}^{N-1} x_n \, exp\left(-\frac{2\pi i}{N} kn\right) \quad (7)$$

where k = 0,..., N-1.

Then you need to go from the value of the sound frequency f to the value of the height (Mel). First, you need to place the resulting spectrum on the chalk scale. This operation is carried out according to the formula

$$B(f_{Hz}) = 1127{,}01048 * \ln\left(1 + \left(\frac{f_{Hz}}{700}\right)\right) \quad (8)$$

This operation is necessary to simulate the fact that human hearing has different sensitivity in different frequency ranges.

Then it is necessary to form triangular filters that serve to accumulate the energy value in each of the frequency ranges (10 filters are distributed linearly below 1000Hz, and the rest are logarithmically above 1000Hz) and take the logarithm of each obtained chalk value. The use of the logarithm is necessary so that differences in the input signal delivery methods have less impact on the assessment of individual speech characteristics.

Next, we translate the obtained values into a scale with frequencies. At the next step of the algorithm, the signal kepstr is calculated. This transformation allows you to separate the source of the sound wave from the filter, whose properties allow you to generate the corresponding sound when a wave with the frequency of the main tone of speech passes through the voice channel. At the same time, the filter contains most of the useful information.

Each signal segment can be described using 12 Mel-frequency cepstral coefficients. To find them, use the formula

$$c(n) = \sum_{m=0}^{M-1} S(n) \cos\left(\frac{\pi n\left(m+\frac{1}{2}\right)}{M}\right) \quad (9)$$

where 0≤ n <M

Figure 1 shows a graph of the dependence of the Mel-frequency cepstral coefficients on time for two frames of the speech signal of two different speakers who uttered the same speech phrase. On the graph, you can see that the recording coefficients differ for different speakers. The dependence of the Mel-frequency cepstral coefficients on time for two different recordings of the same speaker's speech is shown in figure 2. From the graph of figure 2, you can see a small difference between the Mel-frequency cepstral coefficients.

After all the coefficients are calculated, the recording signal must be compared with the reference signal stored in the database. The criterion for matching these signals will be the Euclidean distance measure.

Figure 3 shows a complete block diagram of the algorithm, on the basis of which a program for identifying a person by his vocal data is developed.

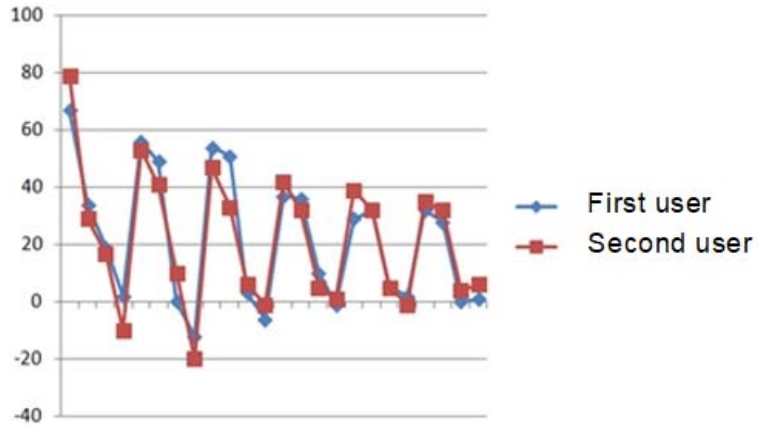

Figure 1 - Dependence of Mel-frequency cepstral coefficients of speech recordings of two different speakers on time in the first two frames of the speech signal

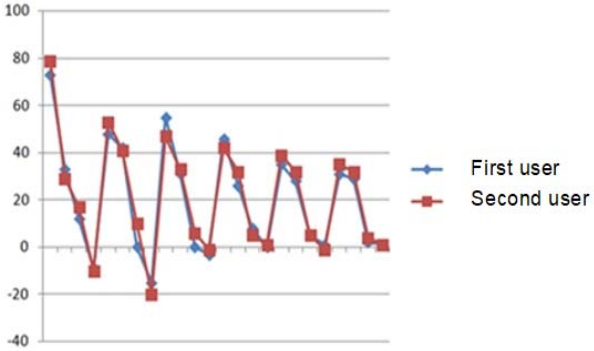

Figure 2 - Dependence of Mel-frequency cepstral coefficients of speech recordings of the same person on time in the first two frames of the speech signal

To initialize the initial parameters of the model, in this paper we used the algorithm of cluster analysis for vectors of speech signal features. The K-means++ algorithm was chosen as the clustering algorithm, which uses Euclidean distance as a distortion measure [[12]].

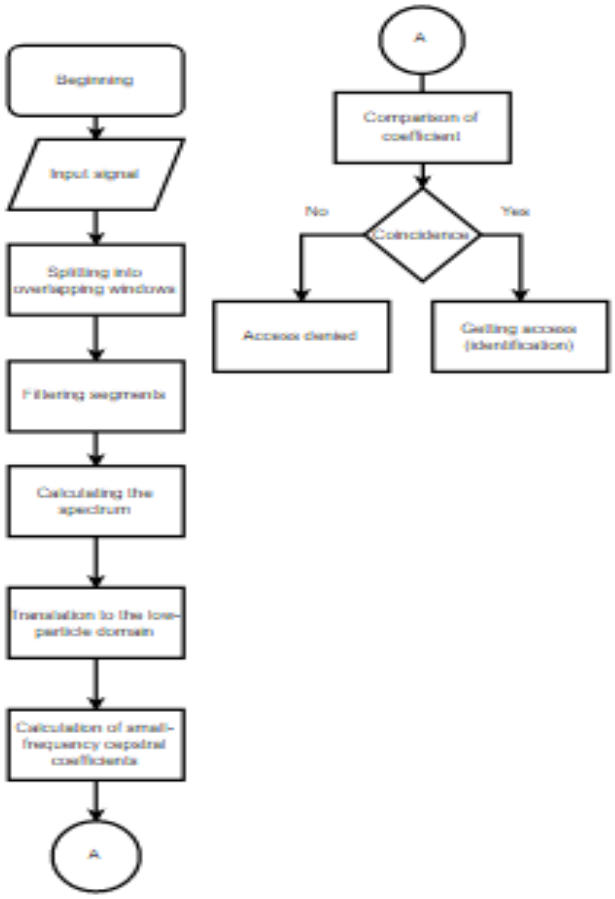

Figure 3 - Block diagram of the algorithm for automating the speaker identification process by voice

the K-means++ Algorithm is a modification of the K-means algorithm. In this algorithm, the center of the first cluster is randomly selected, and then each subsequent center can be selected from the remaining data points with a probability proportional to the square of the distance to the nearest existing cluster center. After that, the standard K-means algorithm is executed. The advantage of this approach is a large reduction in the error of the final result.

To test the developed algorithm, a software tool in the C++language was developed. The voice signals of twenty people were selected. Speech recordings were made in mono mode using a microphone

built into the computer, which has a sampling rate of 16 kHz and an ADC bit rate of 16 bits. The duration of the speech signal was 50 seconds, and the duration of the test signal was 15 seconds. The algorithm was tested for a different number of components of the Gaussian mixture model. Figure 4 shows the dependence of the number of correctly identified speakers (in %) on the number of components of the Gaussian mixture model.

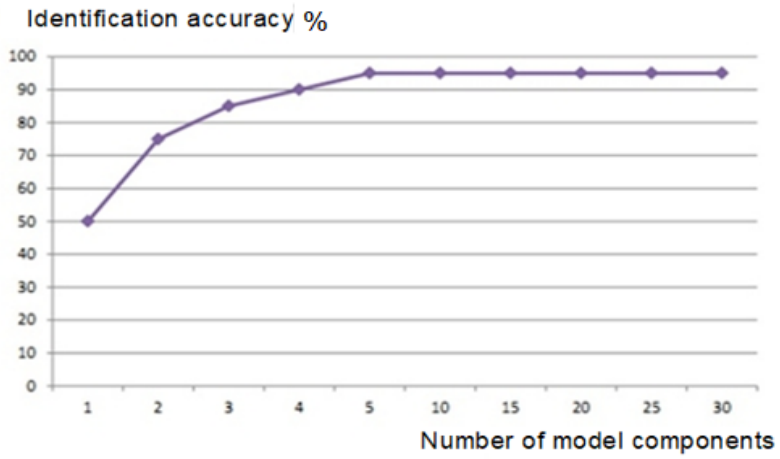

Figure 4 - Dependence of the number of correctly identified speakers (in %) on the number of components of the Gaussian mixture model

6. Conclusion

The results of the development of an algorithm for automatic identification of a person by voice for authorizing access to information obtained in this paper allow us to draw the following conclusions:

For modeling individual voice characteristics, the components of Gaussian mixtures are best suited, since they allow you to recognize speakers with high accuracy.

Determining the initial parameters of the model using the K-means++ algorithm can significantly increase the learning rate and improve the accuracy of identification.

The number of components that is optimal for the effective operation of the system is five. With this number of components, the speaker identification accuracy is 96%, which indicates that the implemented algorithm can be successfully used to authorize access to information by the user's voice.

А. Г. Карелова[1], О. Ж. Мамырбаев[2]

[1]Казахский национальный университет им. аль-Фараби, Алматы;
[2] Институт информационных и вычислительных технологий КН МОН РК, Алматы, Казахстан

ОБЗОР АЛГОРИТМА РАСПОЗНАВАНИЯ ЧЕЛОВЕКА ПО ГОЛОСУ

Аннотация. Распознавание речи имеет различные приложения, включая взаимодействие человека и машины, сортировку телефонных звонков по гендерной категоризации, категоризацию видео с тегированием и т.д. В настоящее время машинное обучение является популярным направлением, которое широко используется в различных областях и приложениях, используя последние разработки в области цифровых технологий и преимущества возможностей хранения данных с электронных носителей.

В статье раскрывается четверной этап классификации признаков. Сосредоточимся на распознавания по голосу, используя Dynamic Time Warping (DTW) – алгоритм. Для характеристики голоса говорящего и для записи сигнала голоса применяется модель гауссовской смеси. С помощью этого метода можно отличить голос человека с высочайшей точностью, поскольку компоненты гауссовых смесей могут моделировать индивидуальности голоса. В статье показаны результаты тестирования алгоритма, делается итог о применимости модели гауссовых смесей для решения задачи идентификации личности по голосу.

Ключевые слова: алгоритм; гауссовская смесь; идентификация; распознавание; классификация.

А. Ғ. Карелова[1], О. Ж. Мамырбаев[2]

[1]Әл-Фараби атындағы ҚазҰУ, Алматы, Қазақстан;
[2]ҚР БҒМ БК Ақпараттық және есептеуіш технологиялар институты, Алматы, Қазақстан

АДАМДЫ ДАУЫС АРҚЫЛЫ ТАНУ АЛГОРИТМІНЕ ШОЛУ

Аннотация. Сөйлеуді танудың түрлі қосымшасы бар, соның ішінде адам мен машинаның өзара әрекеті, телефон қоңырауын гендерлік категория бойынша сұрыптау, тегтеу бейнелерін санаттау және т.б. Қазіргі уақытта машиналық оқыту – сандық технологияның соңғы әзірлемелерін және электронды медиадан деректерді сақтаудың артықшылықтарын қолдану арқылы түрлі сала және қосымшада кеңінен қолданылатын бағыт.

Мақалада белгілерді жіктеудің төртінші кезеңі көрсетілген. Dynamic Time Warping (DTW) алгоритмін қолдану негізінде дауыс тануға назар аударды қ. Мәселені шешудің қолданыстағы әдістеріне шолу жасаймыз. Әдіс үшін Гаусс қоспасының моделін қолданамыз. Сөйлеушінің дауысын сипаттау және дауыс сигналын жазу үшін Гаусс қоспасының моделі қолданылады. Бұл әдіс арқылы адам дауысын жоғары дәлдікпен ажыратуға болады, өйткені Гаусс қоспаларының компоненттері дауыстың даралығын модельдей алады. Мақалада алгоритмді тестілеу нәтижелері көрсетілген, дауыстың жеке басын анықтау мәселесін шешу үшін Гаусс қоспаларының моделін қолдану туралы қорытынды жасалады.

Түйін сөздер: алгоритм, Гаусс қоспасы, сәйкестендіру, тану, жіктеу.

Information about authors:

Mamyrbayev O.Zh., PhD, Associate Professor, head of the Laboratory of computer engineering of intelligent systems at the Institute of Information and Computational Technologies, Almaty, Kazakhstan; morkenj@mail.ru; https://orcid.org/0000-0001-8318-3794;

Karelova A. G., 2nd year master's degree in “Computer engineering", Institute of Information and Computing Technologies, Almaty, Kazakhstan; karelovaayagul@gmail.com; https://orcid.org/0000-0003-3960-4226

REFERENCES

[1] Rybin S. V. Sintez rechi. Uchebnoye posobiye po distsipline “Sintez rechi” [Synthesis of speech. Textbook on the discipline “Synthesis of speech.”] / S. V. Rybin. SPb: Universitet ITMO, 2014. 92p. [in Russian].

[2] Sorokin V. N. Verifikatsiya diktora po spektral’no-vremennym parametram rechevogo signala [Speaker verification using the spectral-temporal parameters of a speech signal] / V. N. Sorokin, A. I. Tsyplikhin // Informatsionnyye protsessy. [Informational processes]. 2010. T.10. № 2. P. 87–104 [in Russian].

[3] Akhmad Khassan Mukhammad: Issledovaniye i razrabotka algoritmov parametrizatsii rechevykh signalov v sisteme raspoznavaniya diktora [Research and development of algorithms for the parameterization of speech signals in the speaker recognition system]: dis…. PhD in Engineering: 05.13.01: defense of the thesis 26.11.08: approved 12.06.09/ Akhmad Khassan Mukhammad. Vladimir, 2008. 157 p. [in Russian].

[4] Pervushin Ye. A. Obzor osnovnykh metodov raspoznavaniya diktorov [Review of the main speaker recognition methods] / Ye. A. Pervushin // Matematicheskiye struktury i modelirovaniye.[Mathematical Structures and Modeling]. 2011. Vyp. 24. P. 41-54 [in Russian].

[5] Campbell J. P., Speaker Recognition: A Tutorial / J. P. // Proceedings of the IEEE. 1997. V. 85, N 9. P. 1437-1462.

[6] Martin A., Przybocki M. The NIST 1999 Speaker Recognition Evaluation – An Overview // Digital Signal Processing. 2000. V. 10.

[7] Kim S. H. Pattern Matching Trading System Based on the Dynamic Time Warping Algorithm. Sustainability / S. H. Kim, H. S. Lee, H .J. Ko and others.2018, 10, 4641.

[8] Thi-Thu-Hong Phan Dynamic time warpingbased imputation for univariate time series data. Pattern Recognition Letters / Phan Thi-Thu-Hong, Emilie Poisson Caillault, Alain Lefebvre, André Bigand., Elsevier, 2017, <10.1016/j.patrec.2017.08.019>. <hal-01609256>

[9] Bayev N. O. Ispol’zovaniye metoda opornykh vektorov v zadachakh klassifikatsii [Using the support vector method in classification problems]/ N. O. Bayev // Mezhdunarodnyy zhurnal informatsionnykh tekhnologiy i energoeffektivnosti. [International Journal of Information Technology and Energy Efficiency]. 2017. T.2 №2(4). P. 17-21 [in Russian].

[10]Chow D. Speaker Identification Based on Perceptual Log Area Ratio and Gaussian Mixture Models / D. Chow, H. Waleed, A. Robust. Auckland, New Zealand: 2002. 65 p.

[11]Sadykhov R. KH. Modeli gaussovykh smesey dlya verifikatsii diktora po proizvol’noy rechi [Models of Gaussian Mixtures for Speaker Verification by Arbitrary Speech] / R. KH. Sadykhov, V. V. Rakush // Doklady BGUIR.[Reports of BSUIR]. 2003. №4. P.98–103 [in Russian].

[12]Shokina M. O. Primeneniye algoritma k-means++ dlya klasterizatsii posledovatel’nostey s neizvestnym kolichestvom klasterov [The use of the k-means ++ algorithm for clustering sequences with an unknown number of clusters][Electronic resource] / M. O.Shokina // Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh.[New information technologies in automated systems]. 2017. № 20. URL:https://cyberleninka.ru/article/n/primenenie-algoritma-k-means-dlya-klasterizatsii-posledovatelnostey-s-neizvestnym-kolichestvom-klasterov (accessed: 15.01.2019). [in Russian].