The widespread use of biometric systems entails increased interest from cybercriminals aimed at developing attacks to crack them. Thus, the development of biometric identification systems must be carried out taking into account protection against these attacks. The development of new methods and algorithms for identification based on the presentation of randomly generated key features from the biometric base of user standards will help to minimize the disadvantages of the above methods of biometric identification of users. We present an implementation of a security system based on voice identification as an access control key and a verification algorithm developed using MATLAB function blocks that can authenticate a person's identity by his or her voice. Our research has shown an accuracy of 90 % for this user identification system for individual voice characteristics. It has been experimentally proven that traditional MFCCs using DNN and i and x-vector classifiers can achieve good results. The paper considers and analyzes the most well-known approaches from the literature to the problem of user identification by voice: dynamic programming methods, vector quantization, mixtures of Gaussian processes, hidden Markov model. The developed software package for biometric identification of users by voice and the method of forming the user's voice standards implemented in the complex allow reducing the number of errors in identifying users of information systems by voice by an average of 1.5 times. Our proposed system better defines voice recognition in terms of accuracy, security and complexity. The application of the results obtained will improve the security of the identification process in information systems from various attacks

Keywords: *security system, voice identification, voice recognition, voice biometric, x-vector, i-vector*

UDC 004.932.75
DOI: 10.15587/1729-4061.2021.239186

DEVELOPMENT OF SECURITY SYSTEMS USING DNN AND I & X-VECTOR CLASSIFIERS

Orken Mamyrbayev
PhD, Associate Professor, Deputy General Director in Science
Laboratory of Computer Engineering of Intelligent Systems*
Aizat Kydyrbekova
Researcher
Department of Information Systems
Al-Farabi Kazakh National University
Al-Farabi ave., 71, Almaty, Republic of Kazakhstan, 050040
Keylan Alimhan
Doctor of Science Degree in Mathematical Sciences, Professor
Department of Mathematical and Computer Modeling
L. N. Gumilyov Eurasian National University
Satpayev str., 2, Nur-Sultan, Republic of Kazakhstan, 010008
Dina Oralbekova
Corresponding author
Researcher
Department of Cybersecurity, Information Processing and Storage
Satbayev University
Satpayev str., 22, Almaty, Republic of Kazakhstan, 050013
E-mail: dinaoral@mail.ru
Bagashar Zhumazhanov
Software Engineering*
Bulbul Nuranbayeva
Professor
Leader of "Oil and Gas Business" Programs
Caspian University
Dostyk str., 85A, Almaty, Republic of Kazakhstan, 050000
*Institute of Information and Computational Technologies
Pushkina str., 125, Almaty, Republic of Kazakhstan, 050010

Received date 08.07.2021
Accepted date 24.08.2021
Published date 31.08.2021

How to Cite*: Mamyrbayev, O., Kydyrbekova, A., Alimhan, K., Oralbekova, D., Zhumazhanov, B., Nuranbayeva, B. (2021). Development of security systems using DNN and i & x-vector classifiers. Eastern-European Journal of Enterprise Technologies, 4 (9 (112)), 32–45. doi: https://doi.org/10.15587/1729-4061.2021.239186*

1. Introduction

Nowadays, many residential areas and companies use all kinds of security systems to protect their property, for example by using a password and user ID/PIN for security. Unfortunately, all these security systems are not protected at all, because the pin code can be hacked, the ID card can be stolen and duplicated. For these reasons, a completely new security technology should emerge to increase the confidentiality of civilians in relation to the security system [1].

Biometric technology is one that uses the user function parameter as a password. The parameters of each function are unique, even if users are twins. Voice biometric systems are a set of systems based on the uniqueness of the individual biometric characteristics of a person. Thus, their areas of application are the same. A distinctive feature of voice biometric recognition systems is the complete absence of special requirements for the equipment used to obtain biometric data. In most cases, biometric voice systems can use standard microphones used in mobile and fixed phones, headsets for per-

sonal computers, laptops or tablets. Voice biometric systems for separating tasks are divided into verification systems and identification systems. The paper studies the problem of voice recognition and develops a voice recognition system for a specific spoken word [1, 2]. Any authentication system can be attacked by hackers who deliberately want to break into the security system. Such people are usually called impostors. Impostors are people who tend to impersonate real customers. A possible attack that voice authentication systems may suffer from is a replay or spoofing attack. In such an attack, during an authorization session, the impostor first presents an identity card, posing as a valid client. The impostor then plays the pre-recorded speech actually delivered by the client to the speaker verification system to gain access through the system. This attack is particularly dangerous in a voice authentication system with a fixed passphrase, where the passphrase does not change from one authentication session to the next. Thus, there is certainly a need for voice authentication systems that are resistant to such attacks. This stability can be achieved by using a verification system.

Deep neural networks (DNNs) herald a new phase in the evolution of automatic voice recognition technology, providing a powerful way to extract highly legible specific features of the voice from speech recordings. Recently, automatic voice recognition technology supports the DNN-based "*x*-vector" structure, a modern approach that uses DNN to retrieve compact representations of speakers. The performance of *x*-vectors is vastly superior to that of *i*-vectors, especially over short periods of time. DNN *x*-vectors are trained in a discriminatory manner using voice labels. A DNN *x*-vector is capable of using larger amounts of training data than an *i*-vector structure, which is saturated after a certain amount of training data.

Voice identification includes a complex of technical, algorithmic and mathematical methods covering all stages, from voice recording to the classification of voice data. The considered difficulties and disadvantages lead to the conclusion that the further development of voice identification systems urgently requires the development of new approaches aimed at processing large arrays of experimental acoustic signals, their effective analysis and reliable classification. This indicates the relevance of research on the creation of new mathematical methods for processing, analysis and classification of voice data, ensuring the reliability and validity of personal identification.

2. Literature review and problem statement

In the problem of voice identification, the same classification methods are used as in the field of pattern recognition, namely, statistical modeling methods that build certain models of vectors of acoustic features. The most common of these are Gaussian mixture models and Hidden Markov models. However, other models, such as multilayer perceptron or support vector machines, are also successfully used in this task. In addition, there has been a recent trend towards using combinations of several models.

Gaussian mixture models are often used for text-independent verification of speakers [3, 4], estimating the probability density of the variability of speech data. Due to the limited data available for training the speaker model, adaptation technologies are in demand: EM-algorithm (Expectation-Maximization) [5], maximum a posteriori probability or maximum likelihood linear regression [6]. The proposed methods provide a good privacy protection effect and restore speech quality. They are robust against the compression of speech using signal coding algorithms and useful for systems using parametric coding algorithms for compression.

The *i*-vector [7] infrastructure has been the latest advancement in automatic speaker recognition for several years. Recently, a new approach to speaker recognition based on deep neural networks (DNN) was presented, namely based on the *x*-vector [8]. This new framework has been shown to provide a significant improvement in speaker recognition performance over the *i*-vector approach in various databases. Conceptually, the *i*-vector and *x*-vector approaches are similar – they both transform the speech recording into a fixed-length vector representing the speaker. While the *i*-vector approach achieves this by factorizing the space of total variability based on the Gaussian mixture model [7], the *x*-vector approach uses the straight DNN [9]. Once extracted, *x*-vectors can be compared in the same way as *i*-vectors. In [7, 8], data augmentation is used to improve the performance of deep neural networks (DNN) implementations for speaker recognition. The results show DNN *x*-vectors leverage data augmentation through supervised learning. However, it is difficult to collect a significant amount of labeled data for training. As a result, *x*-vectors provide superior performance on estimated datasets.

In [9], the authors present the VOCALISE *x*-vector system [9] and some experiments on speaker recognition on several complex subsets of the forensic NFI-FRIDA database [9]. The paper discusses a database of critical speech recordings – NFI-FRIDA, which were obtained simultaneously by several recording devices, many forensic databases are collected, including Ramos and others.

Hidden Markov models are statistical models in which the system is modeled as a Markov process with unknown parameters. The goal is to determine the most likely state of the test case sequence relative to the pre-training models. For speaker recognition applications, each state of the hidden Markov model can be represented by different elements of speech. Modern text-independent speaker verification systems using Gaussian mixture models do not take into account the temporal ordering of feature vectors. Hidden Markov models have certain advantages. In the problem of text-dependent speaker recognition, a priori knowledge of the text content is used, while the hidden Markov models are more accurate than the models of Gaussian mixtures. In [10], some advantages of using a combination of speech recognition on the syllabic hidden Markov model and speaker-oriented recognition based on the Gaussian mixture model are shown. The systems under consideration, based on traditional Gaussian mixture models (GMM), have achieved satisfactory results for speaker recognition only when the speech length is large enough. Using an optimal feature set based on the Mel Frequency Cepstral Coefficients (MFCC), we achieved an equal error rate (EER) of 3.21 % compared to the previous reference EER of 4.01 % in the THUYG-20 database.

Multilayer perceptron is a type of trained neural networks [11]. The use of neural network methods in the problem of speaker verification is shown in [12]. For speaker verification systems, multilayer perceptron can be binary classifiers that distinguish between "friend" and "alien" classes.

Support Vector Machine is a binary classifier [13]. The basic principle is the projection of nonlinearly separable

multidimensional data into hyperspace, where it can be linearly separable. A significant problem for voice identification systems based on the above methods is the strong effect of external ambient noise on the original voice recordings, from which informative features are distinguished. Conditional on these signs causes a high level of identification errors. This problem is investigated in the following works.

In recent years, the support vector machine has been considered one of the most effective discrimination methods [14]. In the speaker verification problem, the support vector machine can be used separately or in combination with other classification methods. For example, in [15], a combination of methods based on a Gaussian mixture model and a support vector machine is used. A five-level DNN spoofing detection classifier is trained using dynamic acoustic characteristics and a new simple estimation method is proposed using only log probability (HLL) for spoofing detection. Here it is mathematically proven that the new HLL estimation method is more suitable for the spoofing detection problem than the classical LLR estimation method, especially when spoofing speech is very similar to human speech. DNN-HLL spoofing detection systems with ASV systems can significantly reduce the false acceptance rate of spoofing attacks.

Speaker recognition was investigated in [16]. Gaussian mixtures were used as the main classification method for speaker recognition. The impact of noise is also investigated in [17]. Gaussian mixtures are used as the main classification method for speaker recognition. In [18], the identification of a speaker by voice was also investigated. The basic speaker identification method was based on Gaussian mixtures. The proposed methods train a convolutional neural network (CNN) model to display *i*-vectors. The trained CNN model is then used to generate a matched version of the short utterance *i*-vectors during the evaluation phase. The improved systems also perform better than the condition of training PLDA of consistent length using short sentences. The proposed methodology achieves a relative improvement of up to 30 % under mismatched PLDA learning conditions and outperforms the GMM-based method.

The review shows that all classification methods in the problem of speaker identification have certain shortcomings: some are sensitive to noise, others are not clear, and others are too difficult for practical use. In this regard, practical implementations of speaker identification systems are not widespread enough. They are largely devoid of the noted shortcomings in the voice recognition system using DNN and *i*- and *x*-vector classifiers. The use of these methods in the problem of voice identification is discussed in detail later in this work.

The most common methods used for speech recognition are briefly described. These systems include complex mathematical functions and extract hidden information from the input processed signal.

Hidden Markov Modeling (HMM) is a widely used pattern recognition method in the field of speech recognition. It is the Markov mathematical model defined by a set of output distributions. This method is more general and has a stronger mathematical foundation than the knowledge-based approach and the model-based approach. In this method, speech is divided into small audible objects, and these objects represent the state in the Markov model. There is a transition from one state to another in terms of the transition probability. [19] defines an approach to a lightweight speech recognition system using a hidden Markov model (HMM) and achieved an accuracy of 87.23 % at an average error rate of 12.7 % based on word error rate (WER) calculations.

Dynamic Time Warping (DTW) method compares words to reference words. It is an algorithm for measuring the similarity of two sequences that can vary in time and speed. In this method, the time dimensions of unknown words are changed until they match the reference word, and a dynamic time algorithm is implemented to compare the human voice with the reference voice as a reference authentication process. In [20], the test results show that the system accuracy of speech recognition is on average 86.785 %.

Vector Quantization (VQ) is a method of mapping a vector from a large vector space to a finite number of regions in that space. Vector quantization is based on the principle of block coding. Each region is called a cluster and may be called a central word, which is called a codeword. A codebook is a set of all codewords. With this training, the error rate is about 13 % [21], the average accuracy of speech recognition is 87 %.

Deep neural networks (DNNs) are machine-learning tools that allow learning complex nonlinear functions of a given input to reduce the cost of errors. A graphical example of a standard deep neural network can be seen in Fig. 1.

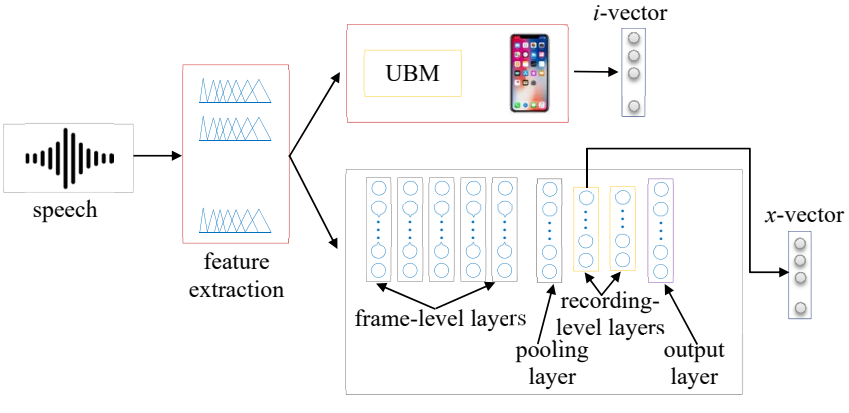

Fig. 1. Deep neural network

Thus, the forwarding DNN used for the classification task may have the following general structure: an input layer provided with some input vectors representing the data; two or more hidden layers (as opposed to a single latent layer shallow architecture) where the transformation is applied to the output signal from the previous layer, taking the appearance of the previous layer as it moves away from the input layer; and an output layer that calculates the output DNN. At this last level, the error criterion is used to calculate the costs and the result is compared to a benchmark (for supervised learning). An input vector x of size D is fed to DNN, which is transformed by hidden layers h_j (consisting of N_j hidden units) in accordance with the function g and DNN parameters (weight matrices W and displacement vectors b). Finally, the output layer O provides DNN output for the

target task (in the case of classification, the probability that the input vector belongs to each class C).

It should be noted that a joint model based on i & x-vectors has not been developed for the identification of users of information systems by voice for the Kazakh language. In this work, we have combined two popular models to improve the performance of the voice verification system.

3. The aim and objectives of the study

The aim of the study is intended to increase the security of the process of identifying users in information systems by voice by developing methods and algorithms for solving this problem using DNN and i- and x-vector classifiers.

To achieve this aim, the following objectives are accomplished:

– research of methods, algorithms, systems of user identification by voice and analysis of the vulnerability of modern voice biometric systems to various methods of falsification of individual voice characteristics;

– development of an algorithm for user identification based on individual voice characteristics in terms of speech variability, taking into account the possibility of protection against various types of attacks on the biometric identification system;

– calculation of errors of the proposed method to assess the reliability of the developed approach.

4. Materials and methods

Development of a software package for biometric identification of users of information systems by voice, experimental studies of the developed identification complex, presentation of recommendations for its practical application in real conditions.

Each person has a unique tone, rhythm, frequency and pitch to express including where they stop in phrases and how quickly they speak depending on where they are in the phrase [22]. Obviously, the average man has a lower voice than the average woman, but the average voice range of each person is unique. People have an interesting characteristic of different accents when they speak. Even through a single specific word, there are several differences in the ways of word and, in turn, sound is produced. The highest frequency that a person can produce is about 10 kHz, while the lowest is about 70 Hz [23, 24]. The speaker recognition system is used for automatic and computational feasibility. The HMM isolates the unwanted noise signal and simulates the spectral representation to form a mixture of the Gaussian function. The spectral envelope corresponds to the Gaussian number defined in the installed system. It will consist of Cepstral coefficients. Another method used in voice recognition is the Fusion classifier system, which uses a minimum amount of input data to get the right solution [1]. Each of the different classifiers shows information about voice patterns and is combined with other classifiers; the system can achieve a higher level of security. Enter the voice of the authenticated user as voice data, x. The Perceptual Linear Prediction (PLP) Coefficients are used as attributes. Model S is set as an authenticated user. The voice is recorded and the function parameter is extracted using three different algorithms, namely GMM, MFN, and SVM. These three different algorithms are used to calculate the match estimate between each authenticated user. Each individual classifier will display different user function parameters and combine all classifier merge weights match estimates to decide whether the user will be accepted or rejected. The system is determined using a false acceptance rate (FAR) and a false rejection rate (FRR).

In this project, MATLAB and ARDUINO will be used. MATLAB software is used for the voice recognition part while the ARDUINO software focuses on the communication system part such as control of the LED indicator switch, LCD screen display and the on/off of the magnet door. During the training phase, the input voice from the microphone will be extracted from the actual uttered speech by the silence detection, then hamming is applied to smooth the voice signal. By using the MFCC, the energy feature of the user is extracted and saved as the reference template. The input voice signal from the testing phase will be checked whether it matches the reference template or not, then the result is calculated. If the result is in the range with the reference template, then the voice is accepted and otherwise. Fig. 2 shows the block diagram of the voice recognition system.

The actual uttered speech is extracted with the silence detection and the others will be ignored by filtering [25]. Hamming window is applied to each window in order to decrease the spectral distortion created by the overlap window. Hamming window can improve the sharpness of harmonics and removes discontinuities on the edges using 1.

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad n \in [0, N-1]. \tag{1}$$

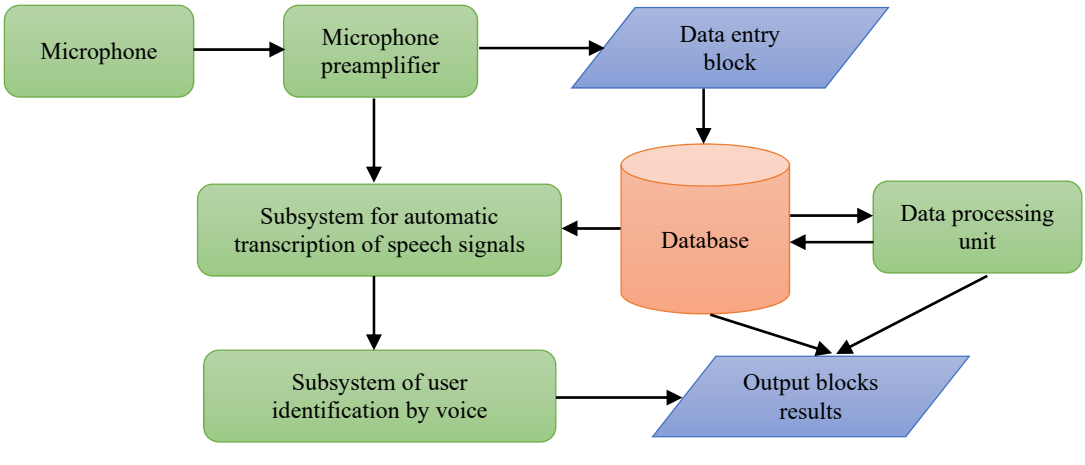

Fig. 2. Block diagram of the voice recognition system

The Fast Fourier Transform (FFT) is a powerful algorithm to calculate the discrete Fourier Transform and convert the signal from the time domain to the frequency domain. The FFT calculation time is 10 times lower than a classic DCT. The horizontal axis represents time while the vertical axis represents frequency [26].

Mel Filter Bank is used to determine the frequency content across each filter. The Mel filter bank is built from triangular filters. The filters are overlapped in such a way that the lower boundary of one filter is situated at the center frequency of the next filter. 1,000 Hz was defined as 1,000 mels. An approximate formula to compute the Mels for a given frequency in Hz is 2.

$$Mel(f)=2,595*\log_{10}(1+f\ /\ 700). \quad (2)$$

The overlapping windows in the frequency domain can be directly used. The energy within each triangular window is obtained and followed by the DCT to achieve better compaction within a small number of coefficients and results known as MFCC. The data will be stored in the database and will be compared with voice input during the testing phase with the same process steps [27].

In this work, the language recognition system consists of two clearly separated parts that need to be trained: DNN, used as a feature extraction tool, and the *x*&*i*-vector pipeline. We use a database to train DNN for later use as a bottleneck debugging tool. This database contains about 225 hours of speech (telephone conversations in Kazakh) of 100 speakers. 10 % of this dataset is reserved for DNN performance testing. This dataset is tagged for word-level speech recognition purposes and will be used to train the ASR system. These alignments are then used to train DNNs with a bottleneck layer.

To develop a bottleneck-based language recognition system, many parameters need to be configured. The configuration parameters of the UBM/*i*-vector system are used as the voice recognition backend captured through experiments in this work. We then investigate various DNN configurations used as a bottleneck extractor. Thus, in this section, we describe these experiments varying the DNN architecture and present the results in terms of DNN performance (phoneme state classification frame accuracy) and ultimate voice recognition performance (average EER). We evaluate the language recognition system on the test design dataset, where we examine the impact of variations in the DNN topology, and finally, we show the results on the evaluation dataset.

5. Results of research of voice recognition methods

5. 1. Methods of modern voice biometric systems for identifying users by voice

Two experiments are carried out to analyze the performance of the voice recognition system. One experiment is to test the accuracy of one's own voice, while another experiment is to test the accuracy of other people's voices when the administrator's voice is set as the reference template. During a speaker uttered speech, his voice will produce a waveform known as voice pattern. Every of the voice patterns is unique and different from other users. Therefore, the first and second experiments are used to analyze the accuracy of the verification process.

We need a hardware setup for the voice recognition security system. With this system, the output data can be read and transferred from the MATLAB by setting the baud rate at 9,600 and all the I/O pins. 5 V are supplied to the Arduino Uno. All pins are set as an output pin to connect with the LED indicator, buzzer, LCD display and magnetic door lock. Arduino Uno is used for automatically reading the output data according to the condition of MATLAB software.

After the voice input is accepted and the systems are identified as a user with user rights, the Arduino Uno will activate the LCD display to display "Qosh keldiniz/Welcome/", green LED indicator will turn on and the magnetic door lock will open.

At the same time, if the voice input is rejected and the system is determined as an impostor user, the Arduino Uno will activate the buzzer to turn on, red LED indicator will turn on, and LCD display will display "Keshiriniz, qaıta qosylynyz/Sorry, please reconnect/", and the magnetic door lock remains locked.

A microphone is used to record the user's voice. The voice sampling frequency is set at 10,000 Hz and the duration period is set as 1 second. Fig. 3 shows the word "Salem/Hello/" is uttered by the user from the microphone.

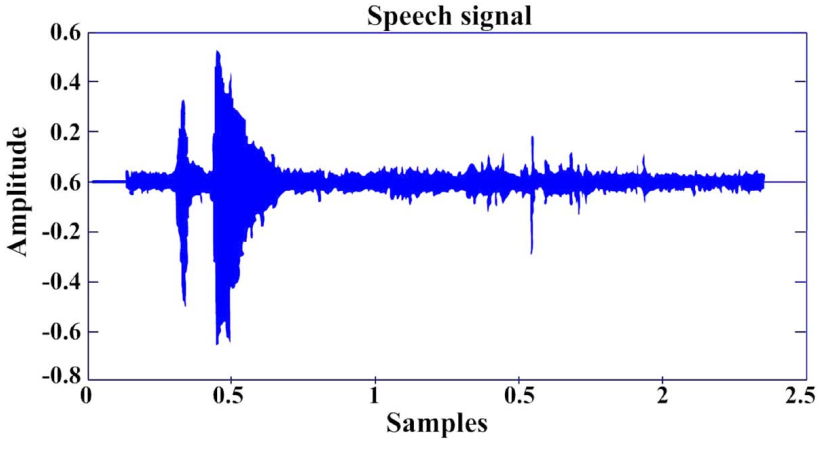

Fig. 3. The word "Salem/Hello/" pronounced by the user from the microphone

The input voice signal is recorded for 1 second then the silence detection will extract only the uttered speech out and ignore the noise signal. After the actual speech signal is extracted by the silence detection, the Hamming window is used to smooth the input voice signal. Fig. 4 shows the word "Salem/Hello/" after smoothing by the Hamming window.

After smoothing the input speech signal, the signal enters the time domain. A fast Fourier transform (FFT) changes the input speech signal from the time domain to the frequency domain. Fig. 5 shows the word "Salem /Hello/" in the frequency domain.

Changing from the time domain to the frequency domain, the input voice signal and the energy are calculated using a formula. The overlapping triangle window and the energy in each window are determined. Fig. 6 shows the word "Salem/Hello/" after Mel-warping.

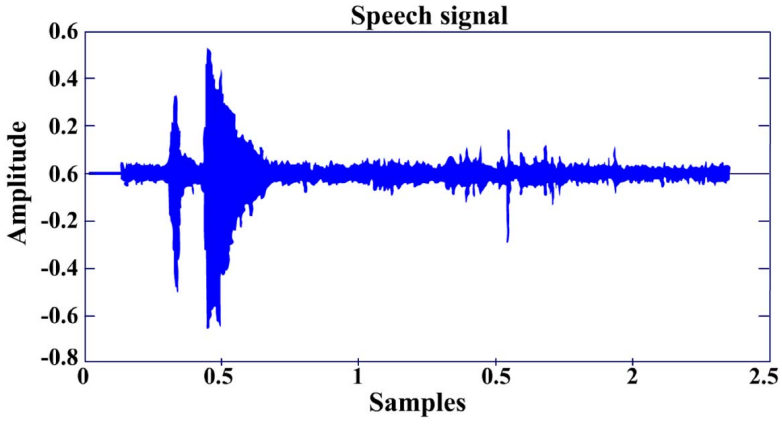

Fig. 4. The word "Salem/Hello/"after using the Hamming window

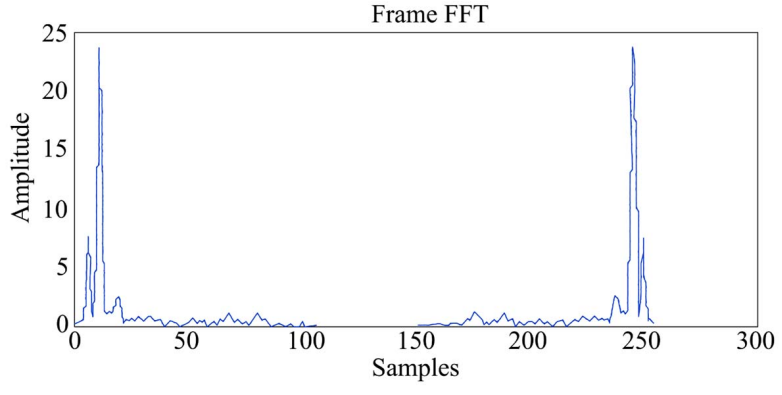

Fig. 5. The word "Salem/Hello/" after the FFT

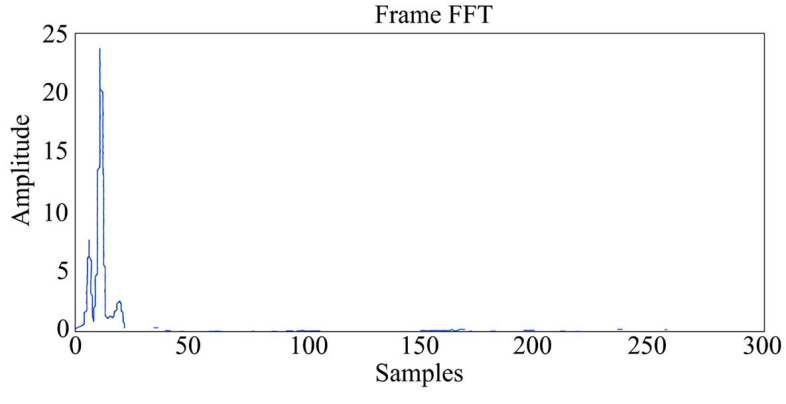

Fig. 6. The word "Salem/Hello/" after Mel-warping

By repeating the experiment to recognize the admin's voice 20 times, 2 times it failed to recognize the administrator's voice. Thus, our research study is able to achieve 90 % accuracy for this voice recognition system. The failure of the system to recognize the authenticated user's voice is due to energy depth variability of the speech uttered by the speaker. In recognition algorithms, it calculates the summation energy in each window and the energy value regardless of whether the spectrum reaches peaks at a specific frequency. The user spoke loudly or soft, will affect the energy of the voice signal. This will affect the output being accepted or rejected too. However, it would be better to improve the accuracy of voice verification.

5. 2. Development of user identification algorithms for a biometric identification system

Hidden Markov Model (HMM) is one of the text-dependent methods. First, the user's voice through the microphone is recorded into a *.wav* file. The speech signal is then converted to a digital signal using an A2D converter [28]. Each statement is converted into a cepstra domain at the training stage. The user's features parameter is then extracted and compared with the reference voice pattern to obtain a probability factor. A probability factor is a comparison of the correspondence of two models to express how many times the probability that data is under one model exceeds the other. After determining the probability, a decision is made: either the user's voice is accepted, or the impostor's voice is rejected. Fig. 7 shows the user verification process.

Preprocessing.

In preprocessing, we remove the maximum part of silence present in the signal. To achieve this, we will use the theory of the probability density function to remove noise and the signal silence part. Typically, the first 200 ms of any recorded speech signal corresponds to silence, since there is always a time interval between the start point of the conversation and the start of voice recording, which is usually at least 200 ms. The normal density function is used to remove silence and find the endpoints of the signal defined as follows (3):

$$u = \frac{x - \mu}{\sigma}, \tag{3}$$

where μ, σ – average value and dispersion of the first 200 ms of the speech signal. Speech preprocessing serves different purposes in any voice processing application. It includes noise removal, endpoint detection, etc.

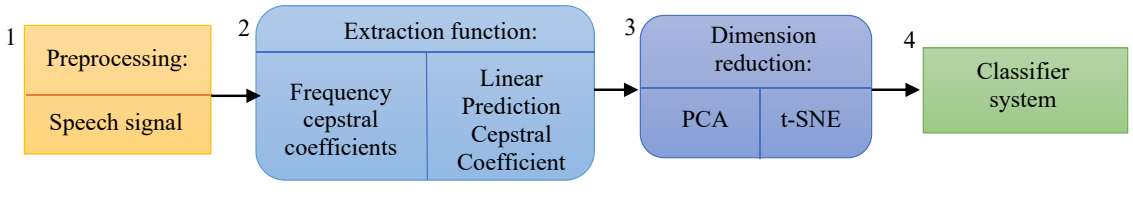

Fig. 7. User verification flow chart

Extraction function.

Voice algorithms consist of two parallel paths. The first is training, in this part we send voice signals along with their identification to the algorithm so that it is possible to classify the extracted functions, the second is testing, where it is the one used to identify a person. In the extraction of voice features, it plays an important role in extracting features from infinite information containing a speech signal that can be used to identify a speaker among a group of *N* speakers. MFCC, LPCC, and PLP methods are used to extract short-term spectral characteristics that will be compared to find the best possible extraction method for different applications. Voice signals are not constant for long duration and are constant when we accept them for a short duration of 20–25 ms.

Mel-frequency cepstral coefficients (MFCC).

MFCC uses an all-zero model to calculate spectra. The output of the preprocessing unit is taken as input in the feature extraction step, where a pre-emphasis is placed on the signal to increase the signal energy at higher frequencies [29], since it also eliminates the DC bias present in the signal. The transfer function of this step is as follows (4):

$$Y[n]=x[n]-a*x[n-1], \tag{4}$$

where is between [0.9, 1], and signal strength at a higher frequency varies with from 0.9 to 1.

Dimensionality reduction.

From the theoretical point of view, the more the features, the better the performance, but as the number of features increases, the performance of the system decreases. So, in order to increase the performance of the algorithm, we use dimensionality reduction techniques. Dimensionality reduction means information loss so our main objective in choosing a dimensionality reduction technique is to preserve as much information as possible while reducing the dimension of the voice signal. In this paper, we are going to use two dimensionality reduction techniques.

T-Distributed Stochastic Neighbor Embedding (t-SNE).

T-SNE is a dimensionality reduction technique that attempts to convert data point nearby into clusters and sends points that are beyond the threshold to a very far distance. Let x correspond to the data point in the high-dimensionality space and y denote the data points corresponding to the low dimensionality space. Then we find the conditional probability between the points denoted by $p_{i/j}$ (5), (6).

$$p_{i/j}=\frac{e^{\frac{-d\,ij^2}{2\sigma i^2}}}{\sum ke^{\frac{-dik^2}{2\sigma i^2}}}, \tag{5}$$

$$p_{ij}=\frac{p_{i/j}+p_{j/i}}{2n}. \tag{6}$$

If d_{ij} represents the distance of the i-th feature from the j-th feature, and σ_i is the Gaussian variance centered at the i-th feature, then p_{ij} is found in a similar way, the conditional probability of low dimensionality features is represented as q_{ij} (7). We choose q_{ij} in such a way that the resulted cost function is minimum (8), (9).

$$q_{j/i}=\frac{e^{-dij^2}}{\sum ke^{-dik^2}}, \tag{7}$$

$$\cos t=\sum_i KL(P_i\|Q_i)=\sum_i\sum_j \log\frac{p_{j/i}}{q_{j/i}}, \tag{8}$$

$$\frac{\partial \mathrm{Cos}t}{\partial y_i}=2\sum_j (y_j-y_i)\left(p_{j/i}-q_{j/i}-p_{i/j}-q_{i/j}\right). \tag{9}$$

Classifier system.

Deep neural networks.

The model is determined by its parameters: weight matrices $W_{j,j-1}$ and displacement vectors b_j, where j is converted from 1 to the number of hidden layers. These parameters are usually adjusted repeatedly to decrease the cost function as the stochastic descent gradient decreases. Thus, for a given set of exercises, where $x^{(i)}$ is a given feature vector and $y^{(i)}$ corresponds to class (true), each hidden layer uses a nonlinear transformation function g to the result of the previous level [10]. This transformation takes into account the parameters W and b, which link one layer to the previous one, and provides the neuron activation values with the following (10), (11):

$$h_j\left(x^{(i)}\right)=g\left(W_{j,j-1}h_{j-1}\left(x^{(i)}\right)+b_j\right),\quad j=2,...,N-1, \tag{10}$$

$$h_1\left(x^{(i)}\right)=g\left(W_{0,1}\left(x^{(i)}\right)+b_1\right). \tag{11}$$

Finally, for the classification task, the output layer computes the softmax function, which outputs the probability P of this input x to belong to a certain class (12).

$$P\left(c\,|\,h(x)\right)=\frac{\exp\left(W_l^c h_l(x)+b_l^c\right)}{\sum_{k=1}^{C}\exp\left(W_l^k h_l(x)+b_l^k\right)}. \tag{12}$$

where $h_l(x)$ refers to the last hidden layer activation for inputs x, W_l^c and b_l^c denote the weight matrix and the bias vector, receptively, that connect the output module for class c to the last hidden layer, and C is the total number of classes. To configure parameters for the task, we consider a cost function that tries to minimize the error between the forecast (network output) and the true class, and the parameters are changed step by step through back propagation [30].

Language recognition:

a) *i*-vector approach.

The *i*-vector system in the classic pipeline can be divided into several stages:

– UBM-GMM modeling. The first step in voice recognition is to model the space using the Gaussian mixing model (GMM) [31]. GMM is based on function vectors using an expectation-maximization (EM) algorithm with data from different operators belonging to different languages (for example, in our case, MFCC functions or bottlenecks) [32]. The GMM results are called the universal background model and are determined by the mean vector (μ-supervector, sequence of mean vectors of each Gaussian component) and the covariance matrix (S, covariance matrices of each Gaussian component).

After *i*-vectors are calculated, classification is performed. The higher the final score, the higher the likelihood of belonging to the same class, since these data points are closer in the *i*-vector space. In this work, we will consider this classifier. UBM training database: speech data used for UBM

and T-matrix training were selected from the corpus of Kazakh speech. A total of 20 audio recordings of speakers are in the training and test set for the gender determination task. The number of audios for men and women is 8 and 12 in the training set. The remaining 50 audio for men and women are used as a test set. For comparison, an *i*-vector system based on GMM was also built. The training was based on a UBM model, such as the GMM-UBM system. For a DNN-based *i*-vector, the DNN model was studied using a system procedure. The dimension of the DNN-based *i*-vector was set to 400.

The results from the EER point of view are presented in Table 1, where "GMM-UBM" is the basic GMM-UBM system. "GMM *i*-vector" denotes a conventional *i*-vector system with a cosine distance metric based on GMM, and "DNN *i*-vector" denotes an *i*-vector system based on DNN with a cosine distance metric. The DNN-based *i*-vector system significantly exceeds its relative baseline. This confirms the effectiveness of recognition methods [33]. In addition, it can be seen that the GMM-UBM baseline is superior to the two *i*-vector systems, but after using probabilistic linear discriminant analysis (PLDA) [34], the *i*-vector system is improved and outperforms the GMM-UBM system.

Table 1

Combined system results

System	EER %
GMM-UBM	30.01
GMM *i*-vector	41.78
DNN *i*-vector	31.12
DNN *i*-vector+PLDA	18.14
Combination system	16.56

We combine the "DNN *i*-vector+PLDA" system and Fig. 8 shows the characteristics at different values. This clearly shows that a combination of systems results in better performance than each individual.

Statistics computation. For a given statement of the trained UBM, defined by its parameters $\lambda=\{\mu, S\}$, the next step is to calculate the Baum-Welch statistics. These statistics represent each frame according to GMM-UBM. Then, for each Gaussian component c and each utterance frame u_t, sufficient zero- and first-order statistics are obtained as follows (13), (14):

$$N_t = \sum_t P(c \mid u_t, \lambda), \tag{13}$$

$$F_t = \sum_t P(c \mid u_t, \lambda)(u_t - \mu), \tag{14}$$

where $P(c|u_t,\lambda)$ is the posterior probability of component c generating the frame u_t.

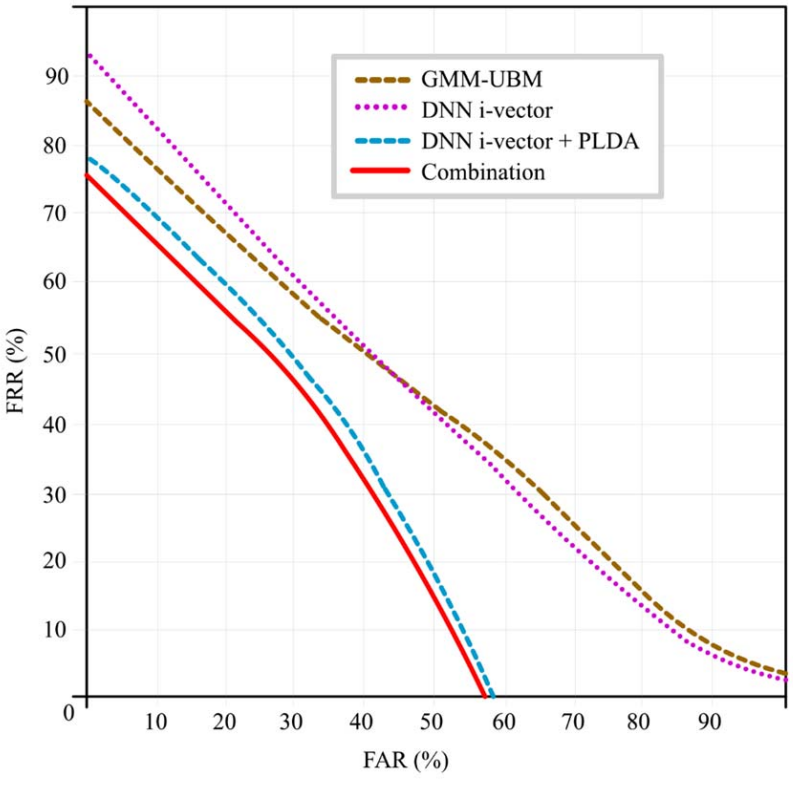

Fig. 8. Characteristics of various values

Total variability subspace training and *i*-vector extraction. In general, the idea of the Total Variable (TV) approach is to project the supervector of means from a given utterance into a subspace *T*. The *T* projection matrix is trained via Expectation-Maximization (EM), with a dataset, which includes variability useful for the target task (language variability, for the language recognition case).

Then, the total variability model can be represented as follows (15):

$$\mu_{UTT} = \mu_{UBM} + Tw, \tag{15}$$

where μ_{UTT} is the utterance-dependent supervector, μ_{UBM} is the UBM supervector of means (language-independent) and is a latent variable, will be an *i*-vector representing each utterance.

To extract the *i*-vector corresponding to a given utterance, the UBM is used to collect the Baum-Welch statistics from the utterance. Once these statistics and the *T* matrix are available, each *i*-vector can be extracted with the following (16):

$$w = \left(I + T^t \sum^{-1} NT\right)^{-1} T^t \sum^{-1} F, \tag{16}$$

where *N* and *F* are the matrices composed of the zero- and first-order statistics, and is the covariance matrix of *F*. These *i*-vectors will have information of the language contained in the utterance they represent, since that is the task for which the *T* matrix has been trained.

Voice recognition bottleneck features [35]. As mentioned earlier, voice recognition systems based on the elementary speech unit (ESU) have become modern in this area. In these systems, a deep neural network (DNN) with an elementary unit (ESU) layer is trained for ASR [36]. This

DNN is provided with input vectors representing frames of audio segments, and the time-dependent output of the bottleneck level is used as a new frame-by-frame representation of the audio signal [32]. With these feature vectors, the classical UBM/*i*-vector scheme is used to perform the voice recognition task. The representation of this structure is shown in Fig. 9.

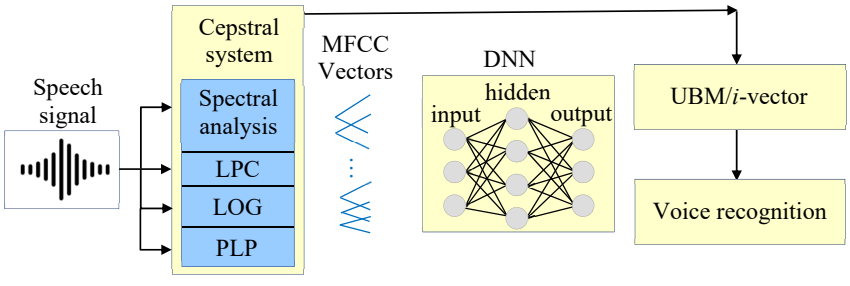

Fig. 9. Representation of the structure of the voice recognition system

Finally, once the *i*-vectors are computed, classification is performed. One of the basic approaches is the cosine distance scoring, in which a score is extracted for each trial or comparison between test and train (language model) *i*-vectors. The higher the resulting score, the higher the probability to belong to the same class, since those data points are closer in the *i*-vector space. This will be the classifier considered in this work.

In the field of speaker recognition, some other scoring approaches in the *i*-vector subspace are usually applied to compensate the variability still existing in the *i*-vector such as Linear Discriminant Analysis (LDA) or Probabilistic LDA (PLDA). However, these approaches were not successfully applied to language recognition, due to the projection into an (N–1) n-dimensional space (where N is the number of languages involved in the task) with the consequent loss of information for LID, where the number of classes is much smaller than in the case of speaker recognition [37].

b) *x*-vector system.

The *x*-vector system [27, 38] is a DNN-based speaker recognition system that is trained by DNN to extract the speaker's representation, and the embedding of the extracted speaker is called an *x*-vector. As shown in Fig. 1, the entire system can be divided into two modules. The first part is a frame-level feature extraction module. Since the speech signal is a time-sequence signal and there is also time-sequence information between frames, the network layer here uses a time-delay neural network [39, 40] to extract the characteristics of the frame layer. The structural information of the temporal sequence of a speech signal can be learned through multilayer DNN, and the end result of this module is the characteristics of the voice at the frame level. The second part is a segment-level feature extraction module. For frame-level voice features extracted by DNN, a statistical layer is used to calculate the mean and standard deviation of these features. This mechanism can convert voice frame-level functions to segment-level functions. At the same time, the statistical layer mechanism can be used to normalize the unequal frame-level feature into a supply level feature of equal length. After the statistics layer, the two full connection layers and the output Softmax layer are connected, and the feature vectors are extracted from the first full connection layer as a voice representation *x*-vector. The *x*-vector voice embedding system uses a time-lagged neural network to compute voice embedding from variable length utterances. After a fixed-length voice embedding (*x*-vector) is obtained from the speech segments, the PLDA is used as a backend for voice classification. It also simplifies a way to increase the amount and variety of training data, called data augmentation. This process adds noise and reverberation to the training samples and includes them in training along with the original samples. The ability to use the same front-end (feature extraction) and back-end (vector comparison) for *i*-vector and *x*-vector systems simplifies system integration and allows for a more direct comparison between the two modeling approaches.

The extracted *x*-vector of speaker characteristics not only contains information about the speaker, but also contains pronunciation information and other noise information such as channel noise and environmental noise. For the speaker recognition task, you only need to pay attention to information that can confirm the identity of the speaker. To reduce the influence of irrelevant information on recognition results, a PLDA model is adopted here to recover the extracted speaker characteristics and separate the speaker information from other irrelevant information.

We define the *j*-th sound of the *i*-th speaker as x_{ij}. Then, according to factor analysis, we define the generation model x as follows (17):

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \varepsilon_{ij}, \tag{17}$$

where w_{ij} denotes the mean of the data, F denotes speaker space, G denotes noise space, μ denotes noise covariance, and h_i denotes an implied variable related to the *i*-th speaker, represents an implied variable associated with the first speaker's speech JTH, i.e. representation of x in the noise space. The vector is a voice-independent vector, F and G represent the voice and channel correlation matrix, respectively, and the diagonal matrix Σ is the residual variable. The variables h and x represent the speaker and channel factor, respectively.

$\varepsilon_{ij} \in (0, \Sigma)$ is the specified noise covariance indicated by h_i associated with the case of hidden voice variables, namely x_{ij} in the speaker space. w_{ij} spoke in the case of the *j*-th speech of the *i*-th speaker about hidden variables, namely x_{ij}, said in the channel space. The vector μ is a speaker-independent vector (global average), F and G are the speaker and channel correlation matrix, respectively. The variables η_l represent the speaker factor.

For speaker recognition tasks, the PLDA model is used for the classifier on the server. To calculate the log likelihood ratio of two sounds for decision making, the formula is as follows (18):

$$\text{score} = \log \frac{p(\eta_1, \eta_2 \mid H_s)}{p(\eta_1 \mid H_d)\, p(\eta_2 \mid H_d)}. \tag{18}$$

In the above, if there are two test sounds, the hypothesis that two sounds come from the same space is H_s, and the hypothesis that two sounds come from different spaces is H. The degree of similarity of two sounds can be measured by calculating the log likelihood ratio. If the score is higher, the probability that two sounds belong to the same speaker is higher.

Short Utterance Variation (SUV) is fixed as follows (19):

$$SUV = \sum_{S=1}^{N} \left(w_{full} - w_{short}\right)\left(w_{full} - w_{short}\right)^{T}, \quad (19)$$

where w_{full} and w_{short} represent full and short i-vectors. N is the total number of x-vectors in the design set. The transformation matrix D is estimated using the Cholesky decomposition $DD^T=SUV$.

The training data and PLDA scores are then transformed as follows (20):

$$w_{SUV} = D^T w, \quad (20)$$

where w is the raw i-vector and w_{SUV} is the x-vector with duration mismatch compensation.

5. 3. Calculation of errors of the first and second kind to assess the reliability of the system

The object of research in the work is the identification of a person by voice data. The subject of research is the characteristics of the voices of the announcers.

To clarify the reliability and validity of the proposed approaches, to determine the boundaries of their applicability, as well as to clarify the factors most significant for identification, quantitative calculations of identification errors were carried out. Errors of the 1st kind meant falsely rejected voice signals, and errors of the 2nd kind – falsely received signals. Note that, for practical purposes, errors of the second kind are more dangerous, since in this case access to confidential information of persons who do not have the right to do so is achieved.

An error of the 1st kind, showing the share of identification refusals, was calculated by (21), (22):

$$\delta^I = \frac{1}{M}\int_{i=1}^{M} \delta_i^I, \quad (21)$$

$$\delta_i^I = \frac{1}{N_i}\sum_{k=1}^{N_k} x_{ik}, \quad x_{ik} = \begin{cases} 1, \text{if } a_{ijk} = 0, i = j \\ 0, \text{if } a_{ijk} = 1, i = j \end{cases}, \quad (22)$$

where i, j is the index of the group of one speaker;

k is the index within the group of one speaker;

a_{ijk} is the result of speaker identification;

x_{ik} is the value characterizing the k-th result of identification of the i-th speaker;

N_k is the number of "samples" in the group of one speaker;

M is the number of speakers;

δ_i^I is the error of the first kind of speaker number i.

Type 2 error showing the share of speakers is wrongly identified as "ours" was calculated by (23)–(25):

$$\delta^{II} = \frac{1}{M}\int_{i=1}^{M} \delta_i^{II}, \quad (23)$$

$$\delta^I = \frac{1}{M}\int_{i=1}^{M} \delta_i^I, \quad (24)$$

$$\delta_i^{II} = \frac{1}{M-1}\sum_{l=1, l\neq i}^{M-1}\left(\frac{1}{N_j}\sum_{j=1}^{N_j} x_{itj}\right), \quad x_{ilj} = \begin{cases} 1, if\ a_{ijk} = 1 \\ 0, if\ a_{ijk} = 0 \end{cases}, \quad (25)$$

where δ_i^{II} is the error of the second kind of speaker number i.

The calculations of errors of the 1st and 2nd kind were carried out depending on the number of speakers included in the database, the duration and repeatability of the word combinations (phrases) that make up this base.

Table 2 shows the results for the accuracy of a voice recognition system. Table 3 shows the results for recognizing the admin user among 10 people, including the admin user and other users. Among the 10 people testing f the voice recognition system, admin user and imposter 5 had been recognized by the voice recognition system, while the others are being rejected.

Table 2

Results for the accuracy of the voice recognition system

Recorded voice	Voice identification testing phase (Hz)	Reference voice template in the learning phase (Hz)	Difference between reference voice and voice ID	Mean Square Error	Middle step	Conclusion
1	0.996	0.0129	0.8801	0.5901	0.3185	accept
2	0.0311	0.0129	0.5801	1.2891	0.4668	accept
3	0.2701	0.0129	0.7511	2.2846	0.5511	reject
4	0.0511	0.0129	0.7481	0.6598	0.1379	accept
5	0.0127	0.0129	−0.0324	0.8451	0.3802	accept
6	0.0746	0.0129	0.8311	1.8261	0.2134	reject
7	0.0954	0.0129	0.8611	1.7491	0.0591	reject
8	0.0301	0.0129	0.5811	0.9685	0.0192	accept
9	0.0311	0.0129	0.5901	0.6545	0.2794	accept
10	0.0314	0.0129	0.2213	0.7541	0.3191	accept
11	0.0171	0.0129	0.8302	0.8401	0.4234	accept
12	0.0743	0.0129	0.2322	0.7754	0.2599	accept
13	0.0167	0.0129	0.2902	0.5701	0.2003	accept
14	0.0179	0.0129	0.2789	0.2911	0.3058	accept
15	0.0297	0.0129	0.5721	0.9901	0.1601	accept
16	−0.1101	0.0129	0.5801	2.1059	−1	reject
17	0.0377	0.0129	1.1161	1.4899	0.2798	accept
18	0.2301	0.0129	0.2811	0.9602	−0.2398	accept
19	0.1102	0.0129	0.9439	3.0259	0.3141	reject
20	0.0781	0.0129	0.8121	1.0745	0.1803	accept

Table 3

Results for recognizing a user-administrator among 10 people

Recorded voice	Second entrance (age, gender)	Voice recognition test phase (Hz)	Reference voice pattern in the learning phase (Hz)	Middle step	Mean Square Error	Conclusion
Admin	32, female	0.0756	0.0127	0.3209	0.9041	accept
Impostor 1	28, man	0.8701	0.0127	0.9801	3.2103	reject
Impostor 2	42, female	0.2109	0.0127	0.2804	2.7312	reject
Impostor 3	25, female	0.6607	0.0127	0.7735	3.4963	reject
Impostor 4	27, man	0.9477	0.0127	0.8731	3.5011	reject
Impostor 5	32, female	0.0311	0.0127	0.1771	0.9018	accept
Impostor 6	31, female	0.2583	0.0127	0.1946	4.0221	reject
Impostor 7	32, female	0.4302	0.0127	0.5432	3.2042	reject
Impostor 8	28, man	0.8372	0.0127	0.8553	6.4532	reject
Impostor 9	28 man	0.3371	0.0127	0.5989	1.9027	reject

In this experiment, testing is performed with 10 different users in Table 3, only one person is an authenticated user, and the rest are other people. Among 10 people of different

genders or ages with an authenticated user, the voice recognition system is able to correctly recognize the administrator's voice. Different gender and different age are a test to see if they can affect the accuracy of voice recognition.

The DNN-based system using *i* & *x* vectors provides a powerful and flexible tool for automatic speaker recognition. It allows the medical examiner to interpret speaker recognition results in terms of a likelihood ratio. Significant performance improvements are seen when using the new *i* & *x*-vector structure with complex data. It has been demonstrated that the performance of *x*-vectors is vastly superior to that of *i*-vectors, especially over short periods of time.

DNN *x*-vectors are trained in a discriminatory manner using speaker labels. DNN *x*-vectors are capable of using large amounts of training data and the *i*-vector structure is saturated after a certain amount of training data. It also simplifies a way to increase the amount and variety of training data, called data augmentation. This process adds noise and reverberation to the training samples and includes them in training along with the original samples. The ability to use the same front-end (feature extraction) and back-end (vector comparison) for *i*-vector and *x*-vector systems simplifies system integration and allows a more direct comparison between the two modeling approaches.

6. Discussion of experimental results of the voice identification system

In this work, for the first time, a comprehensive study of the possibilities of the technology of voice identification of users based on traditional MFCC using DNN and *i* and *x*-vector classifiers was carried out. The solutions for improving the means of voice protection of information are proposed and the architecture of the protection system for voice data using DNN and *i* and *x*-vector classifiers of voice identification is developed.

The results of the work can be used in the development of systems and software and hardware devices for biometric identification of a person, in various access control systems, including those using telecommunication communication channels.

The proposed and tested systems for constructing voice databases can be used to create technical systems for voice identification, to assess the reliability of such systems. The results of the study of the protection system based on voice data using DNN and *i* and *x*-vector classifiers can be used for approbation and verification of methods and technical means for assessing the security of speech information from leakage through various channels.

The system proposed in the work provides the practical implementation of reliable data classification and subsequent identification of the person in the most difficult cases of overlaying voice recordings of a large number of speakers with similar frequency characteristics of the voice. This can be used in the tasks of forensic examination in the investigation of computer crimes, when it becomes necessary to identify unknown voice recordings.

The emergence of new and more advanced technical means intended for illegal access to voice information (for example, to confidential conversations), the improvement of various technical channels of speech information leakage cause a number of other aspects that ensure the importance of research and development of voice identification. Namely, new systems and methods of such identification are in demand in the development and testing of devices for protecting speech information from leakage through acoustic and vibroacoustic channels, as well as in assessing the protection against this leak.

Voice identification of a person is closely intertwined with the forensic tasks associated with the analysis of phonograms that preceded it. These tasks appear during the investigation of crimes when there is a need to identify an unknown voice recording (for example, telephone conversations). Mathematical and technical methods and techniques of identification are quite applicable and useful in carrying out phonoscopic examinations, they contribute to important practical activities in the search for a criminal in the presence of recordings of his voice. It should be especially emphasized that modern computer crime is primarily associated with remote access to information via wireless communication channels. This circumstance determines its latent nature and the very high complexity of the investigation of computer crimes, leaving, as a rule, not material, but so-called virtual traces. In this regard, voice identification systems can provide significant assistance in the investigation and prevention of computer crimes, in particular, they can be involved in the analysis of voice signals in telecommunication channels.

At the same time, further development of personality identification by voice faces significant difficulties. This is due to rather serious shortcomings inherent in this type of biometric identification. Such disadvantages include the low discriminating ability of the method, which can lead to a significant number of errors in identification. These errors can become unacceptable when carrying out identification in difficult conditions.

This paper discussed the requirements of a voice authentication/identification system in modern voice interactive systems. The traditional approach prepares a single speaker model from all registration data for speaker recognition. But we have proposed methods using DNN and *i* and *a* classifiers variable-length *x*-vector system that selects a model depending on the length of test speech samples and performs recognition. The results obtained using our method were better for both verifying the speaker and identifying in an open database. We have also developed a new True Short Speech Database for Kazakh Speaker Recognition. The database consisted of speech commands and instructions that are commonly used in voice interactive systems. The results achieved with the proposed approach to this database were also very encouraging. This suggests that the proposed method can be used to identify the speaker in interactive systems with a telephone voice to make it more convenient. But in critical identification systems, the length of the speech pattern is important. A person will only be recognized if the match result is greater, from the application and can be empirically selected. Experiments using DNN and *i* and *x*-vector classifiers have shown good results compared to classical machine learning algorithms in all cases of cybersecurity experiments. This is because with the use of DNN and *i* and *x*-vector classifiers features are implicitly extracted and generated better, identifying characteristics of the data that lead to greater accuracy. The highest accuracy obtained by DNN when classifying is 0.025 and 0.005, taking into account the possible impact of various types of attacks on the biometric identification system, fraud detection – 0.972 and 0.916, respectively, the proposed model of DNN and

i and *x*-vector classifiers reduces the average equal error rate for all types of attacks to 0.045 %, thus exceeding the performance of previously published approaches. Combining the DNN-HLL spoofing detection system based on the *i* and *x*-vector classifiers with ASV systems can significantly reduce the false acceptance rate of spoofing attacks.

The combination of methods for user identification in Table 1 shows the lowest EER equal to 16.56 %. In addition, during the experiment, it was revealed that the combined use of *i* & *x* vectors in building the model shows a better result than using them separately, and this is displayed in Table 3, where the system accurately recognizes the voices of the Admin and Imposter 5, which has the same rights as admin. Thus, the DNN-based system with *i* & *x* vectors is a powerful and flexible tool for the automatic recognition of Kazakh speakers.

It should be noted that problems were identified in the system when recognizing the voice of an authenticated user when the user spoke quietly, which affected the quality of the system indicators. In this case, it is necessary to conduct additional research and experiments to eliminate this problem.

7. Conclusions

1. The results of experimental tests on user identification obtained in the "information system for user identification by voice" based on the use of the methods and algorithms described above, which differ from existing systems by the possibility of text-based independent identification with security from attacks, are given. This makes it possible to confirm the high reliability of the identification procedure when external factors with speech variability are exposed to the user and to prevent attempts to attack the identification system. These identification methods and algorithms are applicable both in information protection systems against unauthorized access, using voice parameters to identify users, and in systems for differentiating access to premises with voice identification.

2. Algorithms for identifying users of information systems by individual characteristics of the voice based on the proposed methods and an identification algorithm, which makes it possible to increase the security of the identification process from external attacks in access control systems, were developed. The developed identification algorithm can also be used in forensic (phonetic) examination systems that use the voice of a suspect as an evidence base.

3. The calculations of errors of the 1st and 2nd kind were carried out depending on the number of speakers included in the database, the duration and repeatability of the phrases that make up this base. Identification of users of information systems by individual characteristics of the voice in the space of uninformative features is based on the use of DNN and *i* and *x*-vector classifiers, allowing to identify users with the probability of errors of the first and second kind to assess the reliability of the developed approach: 0.025 and 0.005, taking into account the possible impact of various types attacks on the biometric identification system.

Acknowledgments

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09259309).

References

1. Mohamed, S., Martono, W. (2009). Design of fusion classifiers for voice-based access control system of building security. WRI World Congress of Informatics and Information Engineering. Los Angeles, 80–84. doi: http://doi.org/10.1109/csie.2009.983
2. Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., Wang, R. (2017). Speaker identification features extraction methods: A systematic review. Expert Systems with Applications, 90, 250–271. doi: http://doi.org/10.1016/j.eswa.2017.08.015
3. Zeinali, H., BabaAli, B., Hadian, H. (2018). Online signature verification using i-vector representation. IET Biometrics, 7 (5), 405–414. doi: http://doi.org/10.1049/iet-bmt.2017.0059
4. Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S. et. al. (2004). A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Advances in Signal Processing, 2004 (4). doi: http://doi.org/10.1155/s1110865704310024
5. Finnian, K., Anil, A., Forth, O., van der Vloed, D. (2019). From i-vectors to x-vectors – a generational change in speaker recognition illustrated on the NFI-FRIDA database. IAFPA conference. Istanbul.
6. Qi, D., Longmei, N., Jinfu, X. (2018). A Speech Privacy Protection Method Based on Sound Masking and Speech Corpus. Procedia Computer Science, 131, 1269–1274. doi: http://doi.org/10.1016/j.procs.2018.04.342
7. Kelly, F., Forth, O., Kent, S., Gerlach, L., Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. Audio Engineering Society (AES) Forensics Conference 2019. Porto.
8. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi: http://doi.org/10.1109/icassp.2018.8461375
9. Van der Vloed, D., Bouten, J., Kelly, F., and Alexander A. (2018). NFI-FRIDA – Forensically Realistic Inter-Device Audio. IAFPA 2018.
10. Tiwari, V., Hashmi, M. F., Keskar, A., Shivaprakash, N. C. (2019). Speaker identification using multi-modal i-vector approach for varying length speech in voice interactive systems. Cognitive Systems Research, 57, 66–77. doi: http://doi.org/10.1016/j.cogsys.2018.09.028

11. Khaikin, S., Kussul, N. N. (Ed.) (2006). Neural networks: full course. Moscow: Publishing house "Williams", 1104.
12. Eskimez, S. E., Soufleris, P., Duan, Z., Heinzelman, W. (2018). Front-end speech enhancement for commercial speaker verification systems. Speech Communication, 99, 101–113. doi: http://doi.org/10.1016/j.specom.2018.03.008
13. Devan, P., Khare, N. (2020). An efficient XGBoost–DNN-based classification model for network intrusion detection system. Neural Computing and Applications, 32 (16), 12499–12514. doi: http://doi.org/10.1007/s00521-020-04708-x
14. Vapnik, V. N., Chervonenkis, A. Ia. (1974). Teoriia raspoznavaniia obrazov (statisticheskie problemy obucheniia). Moscow: Nauka, 416.
15. Yu, H., Tan, Z.-H., Ma, Z., Martin, R., Guo, J. (2018). Spoofing Detection in Automatic Speaker Verification Systems Using DNN Classifiers and Dynamic Acoustic Features. IEEE Transactions on Neural Networks and Learning Systems, 29 (10), 4633–4644. doi: http://doi.org/10.1109/tnnls.2017.2771947
16. Guo, J., Nookala, U. A., Alwan, A. (2017). CNN-Based Joint Mapping of Short and Long Utterance i-Vectors for Speaker Verification Using Short Utterances. Interspeech 2017. doi: http://doi.org/10.21437/interspeech.2017-430
17. Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., Zue, V. (1993). TIMIT speech data corpus. Philadelphia: Linguistic Data Consortium. doi: https://doi.org/10.35111/17gk-bn40
18. Richardson, F., Reynolds, D., Dehak, N. (2015). Deep Neural Network Approaches to Speaker and Language Recognition. IEEE Signal Processing Letters, 22 (10), 1671–1675. doi: http://doi.org/10.1109/lsp.2015.2420092
19. Tailor, J. H., Shah, D. B. (2017). HMM-Based Lightweight Speech Recognition System for Gujarati Language. Lecture Notes in Networks and Systems, 451–461. doi: http://doi.org/10.1007/978-981-10-3920-1_46
20. Prasetio, B. H., Syauqy, D. (2017). Design of Speaker Verification using Dynamic Time Warping (DTW) on Graphical Programming for Authentication Process. Journal of Information Technology and Computer Science, 2 (1), 11–18. doi: http://doi.org/10.25126/jitecs.20172124
21. Mahalakshmi P., Shayon, Ashok, S. (2015). MFCC and VQ based voice recognition security system. International Journal of Applied Engineering Research, January, 10 (59), 219–233.
22. Krom, G. de. (1994). Consistency and Reliability of Voice Quality Ratings for Different Types of Speech Fragments. Journal of Speech, Language, and Hearing Research, 37 (5), 985–1000. doi: http://doi.org/10.1044/jshr.3705.985
23. Campbell, J. P. (1997). Speaker recognition: a tutorial. Proceedings of the IEEE, 85 (9), 1437–1462. doi: http://doi.org/10.1109/5.628714
24. Yu, Y., He, J., Zhu, N., Cai, F., Pathan, M. S. (2018). A new method for identity authentication using mobile terminals. Procedia Computer Science, 131, 771–778. doi: http://doi.org/10.1016/j.procs.2018.04.323
25. Ranjan, S., Yu, C., Zhang, C., Kelly, F., Hansen, J. H. L. (2016). Language recognition using deep neural networks with very limited training data. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi: http://doi.org/10.1109/icassp.2016.7472795
26. Li, W., Fu, T., You, H., Zhu, J., Chen, N. (2016). Feature sparsity analysis for i-vector based speaker verification. Speech Communication, 80, 60–70. doi: http://doi.org/10.1016/j.specom.2016.02.008
27. Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., Wang, R. (2017). Speaker identification features extraction methods: A systematic review. Expert Systems with Applications, 90, 250–271. doi: http://doi.org/10.1016/j.eswa.2017.08.015
28. Shrawankar, U., Thakare, V. M. (2013). Techniques for feature extraction in speech recognition system: a comparative study. International Journal of Computer Applications in Engineering, Technology and Science, 412–418.
29. Kalimoldayev, M. N., Mamyrbayev, O. Zh., Kydyrbekova, A. S., Mekebayev, N. O. (2020). Algorithms for Detection Gender Using Neural Networks. International journal of circuits, systems and signal processing, 14, 154–159. doi: http://doi.org/10.46300/9106.2020.14.24
30. Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus: Springer-Verlag New York, Inc.
31. Ibrahim, N. S., Ramli, D. A. (2018). I-vector Extraction for Speaker Recognition Based on Dimensionality Reduction. Procedia Computer Science, 126, 1534–1540. doi: http://doi.org/10.1016/j.procs.2018.08.126
32. Lozano-Diez, A., Zazo, R., Toledano, D. T., Gonzalez-Rodriguez, J. (2017). An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition. PLOS ONE, 12(8), e0182580. doi: http://doi.org/10.1371/journal.pone.0182580
33. Li, L., Wang, D., Zhang, X., Zheng, T. F., Jin, P. (2016). System combination for short utterance speaker recognition. 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). doi: http://doi.org/10.1109/apsipa.2016.7820903
34. Prince, S. J. D., Elder, J. H. (2007). Probabilistic Linear Discriminant Analysis for Inferences About Identity. 2007 IEEE 11th International Conference on Computer Vision. doi: http://doi.org/10.1109/iccv.2007.4409052

35. Mamyrbayev, O., Turdalyuly, M., Mekebayev, N., Alimhan, K., Kydyrbekova, A., Turdalykyzy, T. (2019). Automatic Recognition of Kazakh Speech Using Deep Neural Networks. Intelligent Information and Database Systems Proceedings, Part II, 465–474. doi: http://doi.org/10.1007/978-3-030-14802-7_40
36. Kydyrbekova, A., Othman, M., Mamyrbayev, O., Akhmediyarova, A., Zhumazhanov, B. (2020). Identification and authentication of user voice using DNN features and i-vector. Cogent Engineering, 7 (1), 1751557. doi: http://doi.org/10.1080/23311916.2020.1751557
37. Naidu, B. R., Babu, M. S. P. (2018). Biometric authentication data with three traits using compression technique, HOG, GMM and fusion technique. Data in Brief, 18, 1976–1986. doi: http://doi.org/10.1016/j.dib.2018.03.115
38. Richardson, F., Reynolds, D., Dehak, N. (2015). A unified deep neural network for speaker and language recognition. arXiv preprint arXiv:1504.00923.
39. Snyder, D., Garcia-Romero, D., Povey, D. (2015). Time delay deep neural network-based universal background models for speaker recognition. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 92–97. doi: http://doi.org/10.1109/asru.2015.7404779
40. D. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S. (2017). Deep Neural Network Embeddings for Text-Independent Speaker Verification. Interspeech 2017, 999–1003. doi: http://doi.org/10.21437/interspeech.2017-620