Cogent Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/oaen20

Identification and authentication of user voice using DNN features and *i*-vector

Kydyrbekova Aizat, Othman Mohamed, Mamyrbayev Orken, Akhmediyarova Ainur & Bagashar Zhumazhanov |

To cite this article: Kydyrbekova Aizat, Othman Mohamed, Mamyrbayev Orken, Akhmediyarova Ainur & Bagashar Zhumazhanov | (2020) Identification and authentication of user voice using DNN features and *i*-vector, Cogent Engineering, 7:1, 1751557, DOI: 10.1080/23311916.2020.1751557

To link to this article: https://doi.org/10.1080/23311916.2020.1751557

Published online: 21 Apr 2020.

COMPUTER SCIENCE | RESEARCH ARTICLE

Identification and authentication of user voice using DNN features and *i*-vector

Kydyrbekova Aizat[1,3]*, Othman Mohamed[2], Mamyrbayev Orken[1,3], Akhmediyarova Ainur[1,3] and Bagashar Zhumazhanov[3]

Received: 09 September 2019
Accepted: 12 March 2020

*Corresponding author: Kydyrbekova Aizat, IICT, Institute of Information and Computational Technology, Almaty 050010, Kazakhstan
E-mail: kas.aizat@mail.ru

Reviewing editor:
Duc Pham, University of Birmingham, UK

Additional information is available at the end of the article

Abstract: Currently, computerized systems, such as language learning, telephone advertising, criminal cases, computerized health care and education systems are rapidly spreading and creating an urgent need for improved productivity. Speech recordings are a rich source of personal, confidential data that can be used to support a wide variety of applications, from health profiling to biometric recognition. Therefore, it is important that the speech recordings are properly protected, so that they cannot be misused. The leakage of encrypted biometric information is irreversible and biometric links are renewable. The article proposes a block diagram of the identification of the users of the systems by individual voice characteristics, based on the joint use of the Deep Neural Network (DNN) method and *i* -vector in the model of the elementary speech units, distinguished by increased security from various types of attacks on the biometric identification system, which allowed identifying the users with probability of first and second errors genus 0.025 and 0.005. The analysis of the vulnerability of the modules of the biometric voice identification system was performed and a structural scheme of the voice identification system of the user identification by voice with enhanced the protection

Kydyrbekova Aizat

ABOUT THE AUTHOR

Orken Mamyrbayev, PhD, Associate Professor, head of the Laboratory of computer engineering of intelligent systems at the Institute of Information and Computational Technologies. In 2002 he received Master's degree at the Kazakh National Pedagogical University named after Abay in specialty Computer Science. In November 2014, he received the PhD degree in Information systems at the Kazakh National Technical University named after K. I. Satpayev. He is a member of the dissertation council "Information Systems" at L.N. Gumilyov Eurasian National University in the specialties Computer Sciences and Information systems. His research team received 1 patent for an invention and 5 copyright certificates for an intellectual property object in software. Currently, he manages 2 scientific projects, such as the Development of multilingual automatic speech recognition technology using deep neural networks and Methods and models for searching and analyzing criminally significant information in unstructured and poorly structured text arrays.

PUBLIC INTEREST STATEMENT

Automatic speech recognition is a dynamically developing area in the field of artificial intelligence. In recent years, the use of biometric technologies is considered to be the most promising for identifying a person's personality, especially in access control systems, when conducting financial transactions, when requesting limited access information by phone, when managing various devices, in forensics, etc. The article considered a common method for solving problems of analysis and identification of voice deep neural networks. A similar method of biometric identification of users of information systems includes voice identification, which allows you to receive and transmit biometric data to the certification center without the use of specialized and expensive biometric information readers: it is enough to have a phone or microphone connected to a computer. The aim of the article is to increase the efficiency of voice identification of users of information systems by developing methods and algorithms for solving this problem based on DNN and i-vector functions.

against attacks was proposed. The use of elementary speech units in the developed identification systems makes it possible to improve computational indicators, reduce subjective decisions in biometric systems, and increase the security against attacks on the voice biometric identification systems.

Subjects: Computing & IT Security; Artificial Intelligence; Information & Communication Technology; ICT

Keywords: voice identification; voice authentication; deep neural network; speech recognition; elementary speech unit (ESU)

1. Introduction

Automatic speech recognition is one of the active research topics that is trying to teach an independent machine ability to recognize and process the human speech. By identifying the speech, the unit can use decoded speech as input for a wide range of the real-world applications. For example, call management, security identification, client request processing and computer dictation. The speech signal carries linguistic information and the information depend on the speaker, such as age, gender, emotional state, and some ethnic characteristics. There are many factors that affect the reliability of any speech identification and authentication system, for example, speech spectral density, speech segments, context-sensitive, stress and pronunciation. Developing a reliable speech identification and authentication system requires a set of reliable methods that play a pivotal role for the successful speech recognition, for example, effective feature extraction methods for capturing the speech variability and the speaker, acoustic modeling techniques, pronunciation modeling methods and various benchmark tests. Speech recognition was previously studied in the literature, as in (Juang & Rabiner, 2005; Mangu et al., 2000; Varga & Steeneken, 1993; Wu et al., 1998), and recently the main research efforts have been focused on improving the speech recognition systems using new methods and ideas, as in (Chan et al., 2016; Gahremani et al., 2014; Gemmeke et al., 2011; Kundu et al., 2016; Yao et al., 2012).

Recently, with the development of technology, the speaker's voice has become a necessity for the speaker verification and identification systems, such as identifying criminal suspects, improving human-machine interaction, and adapting music to wait in line. Although there have been many studies on the extraction of traits and the development of a classifier for improvement, the classification accuracy is still not satisfactory.

This article proposes various methods for improving the voice classification based on the deep neural networks as a feature extractor and a classifier. First, a model is proposed for generating new functions from DNN. The proposed method uses the HMM (Hidden Markov Model) tool to search for triphons in a bound state for all statements that are used as labels for the output layer in DNN.

Over the past decades, voice recognition is one of the most popular areas in the speech research. This field consists of two main subfields: voice identification and voice authentication as described in the following items.

- Voice Identification: This is the process of determining whose speaker provides the speech. In the process of identifying the speakers, the number of the decisions depends on the number of the people in the databases used, so the performance of the voice identification system will decrease, if the size of the votes used to build the system increases. In general, in any voice identification system, a given speech statement (speech signal) is processed and analyzed for the comparison with various models for the well-known speakers. Then the given speech (unknown speaker) is identified as speaking, which best corresponds to the known identified models.

- Voice authentication: this is the process of verifying the identity of the speaker based on his/her speech. With simple words in this area, a given speech statement for an unknown speaker is compared with the speaker's model, whose personality is asserted. If he overcomes the threshold, the claimed identity is verified and accepted, otherwise the identity is rejected. Choosing the optimal threshold for accepting and rejecting a declared personality is one of the most important questions for a speaker to check. Selecting a high threshold results in most unauthenticated users (imposters) accessing the system, but it also increases the risk of rejecting authenticated users to gain access to the system. On the other hand, choosing a low threshold will increase the risk of accepting unauthenticated users, even if in most cases it gives access to authenticated users. Therefore, the choice and optimal threshold should be taken into account based on the distribution of unauthenticated users and authenticated users in the new system.

In recent decades, several voice recognition studies have been carried out using various methods. These methods can be generalized and divided into four categories: systems based on vector quantization (Du et al., 2006; Jitendra & McCowan, 2003; Makovkin, 2006; Valsan et al., 2002), systems based on GMM (Fine et al., 2001; Meuwly & Drygajlo, 2001; Yuan & Lieberman, 2008), systems based on factor analysis (Senor & Lopez-Moreno, 2014; Yu et al., 2014), and more recently systems based on deep neural networks. (Ley et al., 2014; Richardson et al., 2015).

This article proposes three different methods for improving speech classification. Each of the proposed methods focuses on one specific area for problem improvement. The first area is a set of functions, the second area is the classification method, and the third area is the classifier architecture. Each method will be explained in details and will be accompanied by any restrictions or limitations for the application.

2. Converted MFCC feature set for speech classification

For the beginning, the creation of transformed functions and the proposed regularized DNN weights using general class labels are explained. An approach to transforming existing functions into more efficient ones is proposed. Mel-Frequency Cepstral (MFCC), their first and second derivatives are used as input for comparison, as most of the previous studies used MFCCs for voice recognition.

2.1. Generating converted functions

New converted items are generated from input features using DNN. For example, voice and spectral elements can be used to create a new form of elements in the speech field. The main steps in extracting BNF functions from input objects are:

- Input object;
- DBN (Unsupervized Phase Weight Initialization);
- DBF;
- Output (BNF Features).

The DNN, which is used to create these objects, consists of several hidden layers, in which one of them has a very small number of units compared to other layers. The resulting elements can be considered as a low-dimensional representation, since the layer of bottlenecks compresses the input objects and output labels to form new objects. This is like a method of non-linear reduction of dimensionality, since it creates a low-dimensional set of functions from input functions based on nonlinear activation functions used to create outputs of modules in a neural network. Recently, the use of DNN with a bottleneck has shown improved results in the auto-encoder for the reconstruction of input functions. In this study, transformed features are examined further and used to classify speech.

This article deals with the extraction of the phoneme label and the elementary speech unit extractor. First, labels are extracted for each frame for all the statements. Then, based on the extracted labels, the ESU extractor generates Transformed Mel-Frequency Cepstral (T-MFCC) using a bottleneck layer in the trained DNN.

A system combining several methods for voice identification was proposed: the systems were a GMM system based on the functions of the MFCC, SVM based on the average *i*-vector. In addition, using the proposed method improves the classification accuracy.

2.2. Preliminary processing

Voice activity detection (VAD) is an important step in most speech processing applications, especially if background noise is present. The importance of VAD is related to the fact that it improves intelligibility and speech recognition. Since the speech utterances used in this work were recorded in a public place, the recorded utterances were subject to noise and other interference. As a result, the VAD algorithm is necessary to reduce the background noise and quiet epochs in statements, in order to prepare them for feature extraction.

The use of neural network methods in the task of voice verification is shown in (Mamyrbayev et al., 2019). For voice verification systems, DNNs can be binary classifiers that distinguish between “their own” and “another’s”. It is shown that the use of the cluster model of the ESU can significantly reduce computational costs and increase reliability in solving the problem of voice identification compared to existing methods. In real conditions, user speech is often recorded in a whole set of sound files for the convenience of further processing. Therefore, a method was proposed for the formation of ESU clusters from continuous speech recorded in several files. The automatic cluster formation procedure allows you to: highlight the boundaries of the ESU, adjust the initial calculation parameters, calculate the stable features of the ESU, and classify them into clusters. Assume that X is a multidimensional re-sampling from x of a speech signal from some unknown user exactly, and $P_0 = N(K_0)$ is the normal distribution law defined by its $(n \times n)$ matrix of autocovariance (K_0) in the role of a statistical image of the known user. It is required to solve the problem: does the sample (voice) X belong to user P_0. This is one of the pattern recognition tasks for the case of a dichotomy, i.e. binary (“yes”—“no”) set of decisions. The problem in this formulation was studied in detail in (Savchenko, 2009) and a pair of Gaussian distributions according to Kullback—Leibler was also shown.To solve the problems of classification (clustering) a set of unclassified objects is required. Clustering is the organization of objects in classes that satisfy certain quality standards. Ribbon structure expression for optimal crucial statistics is reduced to the form:

$$P_{x,r} = \frac{\Delta 1}{F} \sum_{f-1}^{F} \begin{pmatrix} G_x(f) & & G_1(f) \\ G_r(f) & +m & G_x(f) \end{pmatrix} - 1 \rightarrow min_{r=1,R} \tag{1}$$

Here $G_x(f)$ is the sample estimate of the power spectral density of signal X as a function of the discrete frequency $f; G_x(f)$—power spectral density r-th signal from the dictionary of standards; F—the upper limit of the frequency range of the signal or used communication channel. This is the well-known formulation of the information discrepancy minimum criterion (DMC) based on the speech signal AR-model:

$$X(n) = \sum_{i=1}^{P} a(i)x(n-1) + \varepsilon(n) \tag{2}$$

Here $X(n)$ is the value of the nth reference of the speech signal, $a = \{a(i)\}$is the *i*-vector of its AR—coefficients, P is the order of the AR model, and$\varepsilon(n)$is the generating process of the white Gaussian noise type) with a zero value of the expectation and a fixed variance σ_2.

Numerical taxonomy is one of the first approaches to solving clustering problems. Numerical methods are based on the representation of objects using a set of properties. If there are correct labels for each object (a vector of *n* values of attributes) it can be considered as a point in n—dimensional space. A measure of the similarity of two objects can be considered the distance between them in this space.

Denote the speech signal being analyzed by the vector X of its samples $x = \{x_1, \ldots, x_N\}$, where N is the sample size. We select the first m-samples in it. With a standard sampling rate of 8 kHz, the homogeneity segment we choose will correspond to m = 80–200 discrete samples. We use the selected data segment $x_1 = \{x_1, \ldots, x_N\}$ as a training sample x_1 for evaluating the AKM of the first elementary speech units from our signal:

$$\widehat{K_1} = M_{-1} \sum_{m=1}^{M} X_1 X_1^1 \tag{3}$$

$\widehat{K_1}$ – The corresponding distribution law.$P_1 = N\left(\widehat{K_1}\right)$- This is the first of the peaks of our future "tree". Following the expression for the decisive statistics, we define for it the specific value of the information discrepancy with respect to the first ESU:

$$p\left(x_{2,}x_1\right) = p(x)|_{x_{2x}} \tag{4}$$

The result obtained is comparable with a certain threshold level p_0 of admissible values of mismatches between different implementations of the same-named oral speech backgrounds:

$$p\left(x_{2,}x_1\right) \geq p \tag{5}$$

Provided that inequality Equation (5) is fulfilled, the second vertex will appear in our tree, and after this we equate the number of its vertices $R = 2$. ESU of the signal $X = x(t), t = 1, 2, 3 \ldots$ As applied to the sequence of homogeneous phonemes of the user, calculated by the sum of the realizations $x(t) = c(t) - c(t-1)$, wherec(t)- is the number of repeated phonemes, $t = 1, 2, 3 \ldots$got c to be a segment of 60 realizations, the resulting data segment $x_1 = \{x_1, \ldots, x_c\}$ acts as a training sample x_1Otherwise, the decision is made to combine the x_1 and x_2 samples into one extended sample of the first ESU, after that, we equate $R = 1$ and calculate the second segment of the sample $x_2 = \{x_c + 1, \ldots, x_{2c}\}$. This is a typical formulation of the information $(R+1)$ element (Makovkin, 2006).Calculations according to the scheme of Equations (1), (4), (5) are repeated cyclically for all subsequent data segments from the initial sample of observations *x*, and they will be repeated with a "cumulative total" for the variable $R = 2, 3$, ... As a result, we will get a tree with some fixed number vertices of R^*. Each vertex is the code of one of the phonemes highlighted in the analysis. The greater the number of vertices in a constructed tree for a particular user, the richer his speech is from a fundamental, phonetic point of view. Obviously, using the described tool can be performed phonetic analysis of speech. However, there is also an obvious problem: an excessively large number of phonemes in the user's speech is a sign of its vagueness, or not informativeness. Therefore, after performing all the above calculations, we sort the resulting vertices by volume $\{V_r\}$ of their classified samples into two sets: a set of distinct ESU for which the condition$V_r \geq V_0$and many fuzzy, dubious ESU otherwise. Here, V_0 is some threshold level for the minimum sample size. From the point of view of the quality of oral speech, the primary interest, of course, is a set of clear ESU. In this case, it should be considered the main result of the phonetic analysis of speech.

In addition, normalization of the cepstral average dispersion is used to eliminate convolutional distortion and linear channel effects. The normalization of cepstral mean dispersion can be applied globally or locally. In this article, it is applied globally to obtain a normal distribution with zero mean and unit variance.

The MFCCs are one of the most well-known sets of spectral characteristics and are widely used in many speech applications. This work uses MFCC. MFCC is a set of coefficients that are used as features; they are built using frequency information from vocal tracks. They represent acoustic signals in the cepstral region, which use *i*-Vector to represent window short signals as a real signal cepstrum. It is based on our natural auditory perception mechanism, so the MFCC bands are evenly spaced on the Mel scale (Shahamiri & Binti Salim, 2014). The first approach, MFCC, uses the Fourier transform for the speech signal to obtain a spectrum. The power of the spectrum, called the melting scale, scales and approaches the response of the human ear. Small frequencies take the algorithm and apply it to the discrete cosine transform to smooth the spectral and decorrelated elements.

The Fourier transform of the signal $x[n]$ is defined by the equation:

$$\mathrm{X}[k] = \sum_{n=0}^{N=1} x[n]w[k-n]\exp\left(\frac{2\pi j}{N}kn\right) \tag{6}$$

with N as frame length and Hamming window:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{7}$$

Automated systems that use voice as input and output when interacting with the user. These systems are based on the speech technology such as automatic speech recognition (ASR). In nature, the properties of speech signals change rapidly over time. The discrete Fourier transform is used to calculate the power spectrum of each frame. Elementary speech units of low-pass filters are used for low frequencies, while a wide range of low-pass filters are used for high frequencies. The main point of using ESU low-pass filters is to determine the energy level of different frequency ranges. The discrete cosine transform of the data outputs of the log filters is calculated. In this article, speech statements were divided into frames with a size of 25 ms. 12 MFCC and normalized energy with their first and second derivatives were calculated for each frame, resulting in 39 coefficients representing each frame.

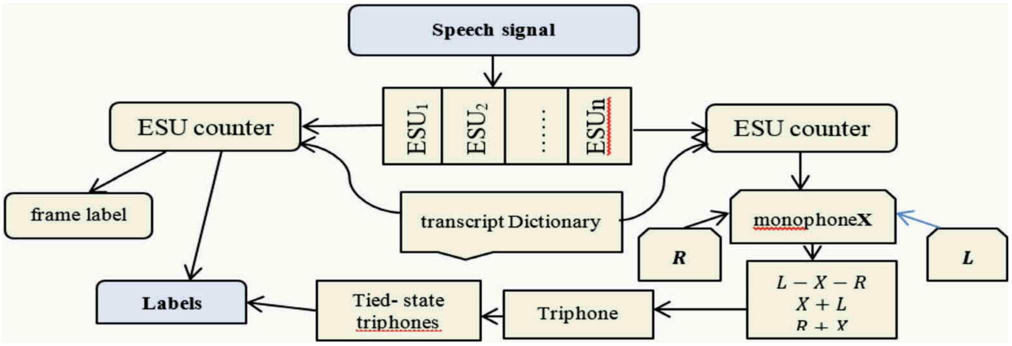
2.3. Extraction of phoneme mark

Typically, each database has a decryption file for each statement that contains the spoken words. Using the transcript together with voice audio files, phonemes are extracted, and this process is called the grapheme to phoneme phase. The main function of the toolkit is the creation of HMM for speech problems, such as discriminators (Young et al., 2006). In the field of speech recognition, speech recognition is performed by matching the sequence of speech vectors with the desired sequence of characters. When performing speech recognition there may be several complications. For example, the mapping between characters and the speech is not one-to-one. In most cases, the speech vector can be displayed on a variety of characters. Another complication is unclear word boundaries in speech. This will result in incorrect display between speech and characters. The principle of accumulation of useful information is based on the results of the comparison of two methods: *i*—vector and DNN on the set of their ESU. In (Hand & Till, 2001; Hinton & Salakhutdinov, 2006), it was concluded that the *i*—vector is an effective method for recognizing speech and was used to recognize the speaker. This method is widely used in many recognition systems when processing information against interference, in particular, in radar systems. Based on this method, a new decision-making algorithm and statistical analysis of user phonemes was developed to solve the voice identification problem. The system is designed to solve such problems with the HMM. HMMs are used to align phonemes with correct labels. It provides word isolation to solve the problem of unclear definition of boundaries. In this article, the created system from (Young et al., 2006) is used to search for triphons in a bound state, which will later be used as labels for the output layer in DNN.

The stages of finding triphons in a bound state are shown in Figure 1:

Step 1: Generate monophones, considering all pronunciations of each statement in the database. The pronunciation that best matches the voice audio will be selected as the output.

Figure 1. Process for extracting phoneme frame labels.



Step 2: Produce the triphons. Monophones are used to make triphons. The current mono *X*, the previous mono *L* and the next mono *R* are processed together.

Step 3: Generate triphons that do not exist in the training data. They are called linked triphons.

Step 4: Find the best match between each frame of a speech and a triphon in a connected state. The best match is the label of the phoneme of the corresponding target frame.

Phoneme labels are used for speech recognition. Here, phoneme labels are used to create converted functions. It stores in the phoneme the specific characteristics of each speaker. Phoneme labels also help DNNs use distinguishing information in transformed functions. In our case, we will use the elementary speech unit as distinctive information in the transformed functions.

3. Extracting transformed features

This section discusses the process of extracting transformed features. First, the DNN learning procedure is performed in two stages: generative (unsupervised) and controlled. Then, the process of extracting the transformed traits based on the trained DNN will be explained in the ESU extractor section.

3.1. DNN training procedure

A neural network is defined as “a system consisting of a series of simple but highly interconnected processing elements that process information through its dynamic state response to external inputs.” As the name implies, the neural network algorithm mimics the biological system, but on a much smaller scale. These highly interconnected processing elements are called neurons and belong to another layer, as shown in Figure 2. The inventor defines a neural network as “a system consisting of a series of simple but highly interrelated processing elements that process information through their dynamic state response to external inputs”. As the name implies, the neural network algorithm mimics the biological system, but on a much smaller scale. These highly interrelated processing elements are called neurons and belong to another layer, as shown in Figure 2.

A neural network can have one or more hidden layers, depending on the complexity of the system. Templates are sent through input data, while targets are sent to output data. Thanks to the exchange of data between hidden layers, the system trains the data, changing the weights of each template in such a way as to ensure maximum performance.

The first stage is generative. DNN is pre-trained using an unsupervised learning methodology that uses RBM. The second stage is discriminatory. DNN is trained using an observable back-propagation algorithm. RBM has an input level *V* (visible level), where$V = \left\{v_{1,}v_{2,}\ldots ,v_V\right\}$ and an

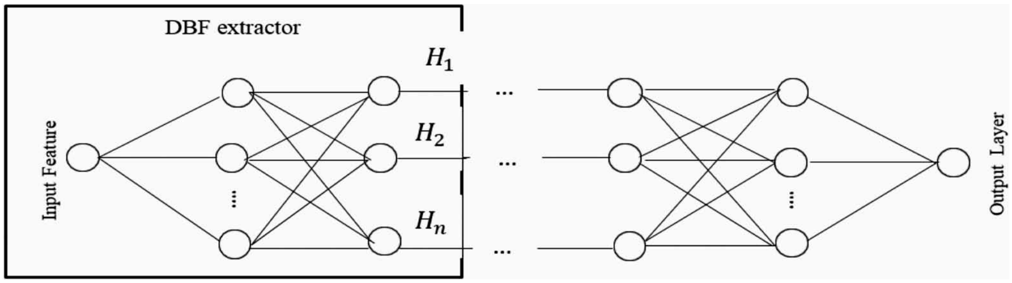


Figure 2. Pre-training phase at DNN.

output level*H* (hidden level), where $h = \{h_1, h_2, \ldots, h_H\}$ (Hand & Till, 2001). Visible and hidden layers are made up of blocks. Each unit in the visible layer is associated with all units in the hidden layer. The limitation of this architecture is that there is no communication between the blocks at the same level. In this work, two types of RBM are used: BB-RBM and GB-RBM (Hinton & Salakhutdinov, 2006). In BB-RBM, the unit values of the visible and hidden layers are binary, $V \in \{0,1\}$and $H \in \{0,1\}$. The BB-RBM energy function is defined in Equation (8).

$$E(v,h) = -\sum_{i=1}^{V}\sum_{j=1}^{H} V_i h_j w_{i,j} - \sum_{i=1}^{V} V_i b_i^v - \sum_{j=1}^{H} h_j b_j^n \tag{8}$$

where V_i is the visible unit in layer *i*, and Hj is the hidden unit in layer *j*. W_{ij} stands for the weight between the visible unit and the hidden unit. b_i^v and b_j^n is the offset of the visible unit *i*,*j* in layer i and the hidden unit in layer*j*, respectively. For GB-RBM, the visible unit values are real, where $V\epsilon R$, and the hidden unit values are binary, where $H\epsilon\{0,1\}$. The energy function of this model is defined as in Equation (9)

$$E(v,h) = -\sum_{i=1}^{V}\sum_{j=1}^{H} \frac{V_i h_j w_{i,j}}{\sigma_i} - \sum_{i=1}^{V} \frac{V_i b_i^v}{\sigma_i^2} - \sum_{j=1}^{H} h_j b_j^n \tag{9}$$

where σ_i is the standard deviation of the Gaussian noise for the visible unit *i*. The joint probability distribution, which is associated with the (v,h) configuration, is defined in Equation (10):

$$\sigma(v,h,\theta) = \frac{\exp(-E(v,h;\theta))}{Z} \tag{10}$$

θ represent weights and displacements, while Z is a function of the separation defined in Equation (11)

$$Z = \sum_{v}\sum_{h} \exp(-E(v,h;\theta)) \tag{11}$$

ESU is the main building block in DNN. It is used as a function detector and is trained without supervision. The output of a trained ESU is used as an input for DNN training. The DNN learning algorithm is layered and uncontrollable. Layered learning helps you find descriptive characteristics that represent the correlation between the input data in each layer. The DNN learning algorithm works to optimize weights between layers. Moreover, it has been proven that initialization of weights between levels in a DNN network improves results to a greater extent than when using random weights. Another advantage of learning DNN is its ability to reduce the impact of fit problems, where both are common problems in models with a large number of parameters and deep architecture (Mamyrbayev, Kydyrbekova et al., 2019). After the DNN learning is completed and the weights between the levels in the DNN stack are optimized, the supervised learning process begins by adding the last label layer on top of the DNN layers. In our work, these labels represent triphons in a bound state for speech data of the utterance.

3.1.1. ESU extractor

The ESU extractor architecture is generated from trained DNN, where each level represents a different internal structure of the input objects. In DNN, the output of each hidden layer creates transformed objects. All layers above the bottleneck layer are removed to obtain an ESU extractor, as shown in Figure 3. Figure 3 shows the proposed DNN voice-base architecture using elementary speech units.

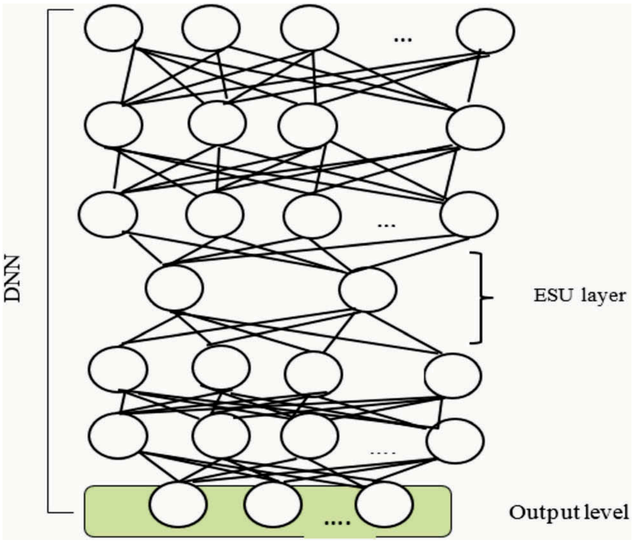
The weights for the DNN are tuned during supervised phase. Introducing elementary speech unit layer has many benefits as reducing the number of units inside the bottleneck layer, getting rid of redundant values from the input feature set, and reflecting the class labels during the classification process. It also helps to capture the descriptive and expressive features of short-time speech utterances (Mamyrbayev, Turdalyuly et al., 2019). Given a ESU extractor with *M* layers, the features at the output layer can be extracted using Equation (12).

$$\begin{cases} l_1(x) = \sigma\left(\sum\limits_{n=1}^{F_1} (w(x_n + b_1))\right) \\ l_2(x) = \sigma\left(\sum\limits_{n=1}^{F_2} (w(x_n + b_2))\right) \\ \vdots \\ l_m(x) = \sigma\left(\sum\limits_{n=1}^{F_m} (w(x_n + b_m))\right) \end{cases} \tag{12}$$

where σ is computed by the logistic function $\sigma(x) = \frac{1}{1+exp(-x)}$.$x = \{x_1, \ldots, x_n\}$ is the feature set vector, and *n* is the number of input features. l_m—is the output of the m^{th} layer. *F*—is a varying number that represents the input for each layer in the ESU extractor. *w*- represents the weights between the input and output nodes in each layer.*b*- represents the bias for each layer.

ESU are used to capture phonetic components of speech. The participation of ESU tags in the generation of transformed MFCC allowed us to understand prosodic features such as intonation, stress, tone and rhythm of the speaker. In addition, the transformed functions are the result of the use of ESU marks in the training data, and this has helped to remove any noise or silent frames so that the transformed functions are calculated without acoustic background noise. To improve the performance of traditional MFCCs, a converted MFCC feature set is generated using an ESU extractor. Work results shows DNN can be designed and trained for seamless adaptation with an ESU extractor, so that new converted properties can be obtained. Using ESU has a number of advantages, since

Figure 3. Voice-base architecture DNN using elementary speech units.



eliminating redundant values from a set of input functions by reducing the number of units in the bottleneck layer and reflecting the label class in the classification process. In addition, the bottleneck layer forces the neural layers to filter input objects in order to preserve descriptive and distinctive features derived from short speech utterances. The input data for the third DNN is the elementwise summation of the output layers of age and gender DNN. The results of the proposed work are compared with two source systems; *i*-Vector and GMM-UBM in a public database.

3.2. The architecture identification system

When forming the process of user presence by voice, the following steps can be distinguished:

- receiving a sample;
- segmentation and feature extraction;
- quality checks (which may reject a sample or signs unsuitable for comparison and requiring the receipt of additional samples);
- comparison with a specific or all database templates that determine the degree of similarity for each comparison;
- making decisions regarding the identity of the patterns to be taken if the degree
- check the result of solving one or several attempts.

The generalized structural scheme identification and authentication systems through speech technologies includes such components as:

- input device;
- subsystem for processing voice data;
- subsystem storage templates;
- subsystem comparison and decision making;
- interface of the application and data transfer subsystem.

The main components of the system are shown in Figure 4.

3.2.1. Possible attacks on the voice identification system

Despite the fact that the recognition accuracy of voice biometric systems has increased significantly in the past few years, in practice only a small number of people trust security systems based on voice biometrics. The most common argument against using such systems is that an attacker can easily bypass the biometric access control system using simple imitation techniques, posing as another user. Consider the main possible intruders aimed at hacking the voice biometric system:

- replacement of a template registered in the system with an unauthorized template;
- deletion of a template registered in the system;
- adding an unauthorized template to the system;
- the impact on the level of the threshold value of decision making;
- the use of modified, unauthorized biometric equipment;

Typical attack on various elements of voice biometric system:

(1) Attack to the input device biometric information.
(2) Attack on the communication channel between the introduction and the data processing component.
(3) Attack on the data processing component.

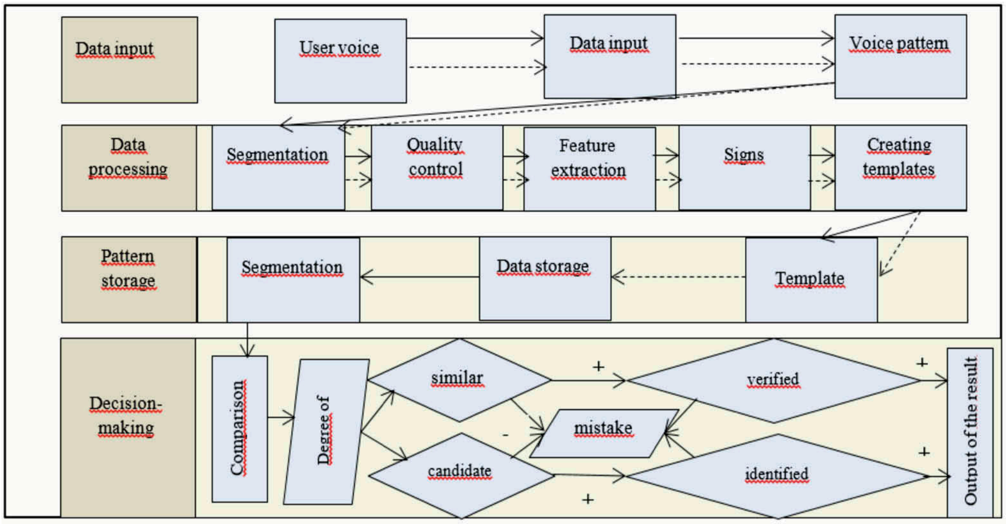


Figure 4. Structural scheme of the main components of voice identification systems.

(4) Attack to the communication channel between the data processing component and the speech pattern database.

(5) Attack to the speech pattern database.

(6) The attack on the communication channel between the data processing component and the decision-making component.

(7) The relationship between the database and the decision component.

(8) Attack on the decision component.

(9) Attack on the communication channel between the decision component and the interface

(10) Attack on the system output interface.

All of the listed attacks, with the exception of the attack on the biometric information input device, are common, regardless of the modality of the biometric system. Effective counteraction to these attacks is achieved by applying digital coding, encrypting an open data transmission channel and using time stamps. Thus, the biometric information input module remains the most vulnerable component of the system. When assessing the resistance of voice biometric systems to attack methods, exclude from consideration the ones based on depersonalization methods, as well as recording and repetition methods, since these types of attacks are independent of the development of technology and have already been studied enough.

The main components of the generalized voice biometric system with their connections. In the scheme shown in Figure 5 any of the elements can be attacked in order to hack the system.

3.2.2. User identification module

The following is an identification module that provides protection against attacks based on the presentation of a randomly generated sequence of keywords from a fixed-size dictionary, see Figure 6.

The advantage of this block diagram Figure 6 is the use in the process of identifying a generator of a random sequence of words recorded by the user in the reference base of the standards and the calculation of the probability of belonging of the spoken voice to the standard, which allows to increase the security from the user's voice recording and playback. In standard identification systems, there is the following drawback: an intruder can record a reference voice message on

Figure 5. Structural scheme voice identification system with possible attacks on system modules.

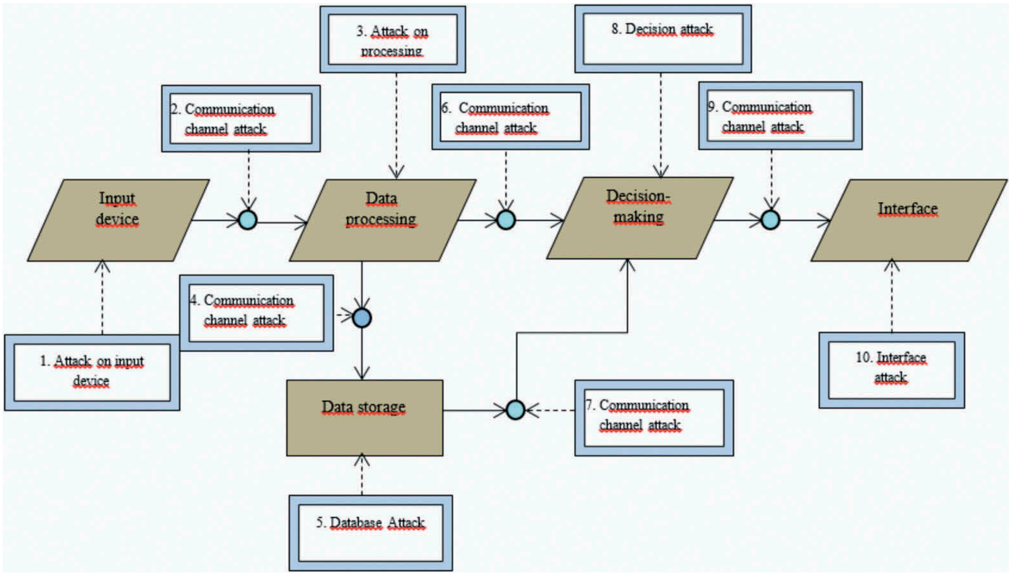
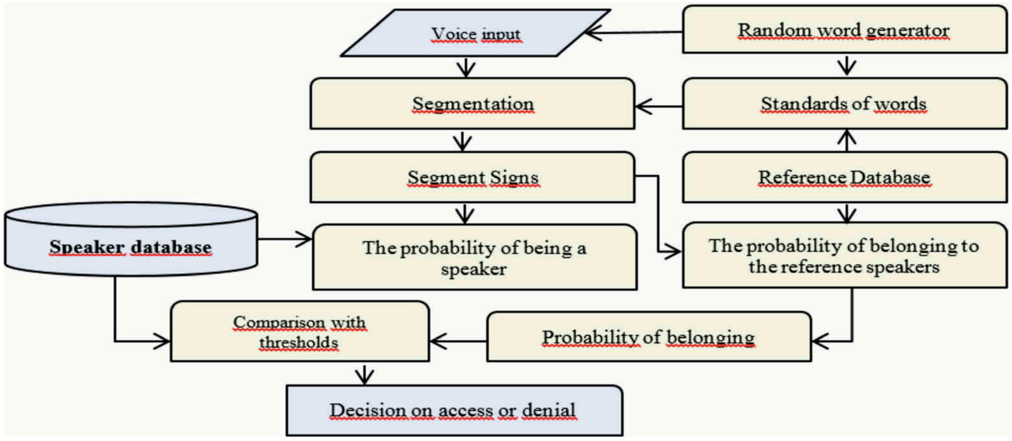

3. Attack on processing
8. Decision attack
2. Communication channel attack
6. Communication channel attack
9. Communication channel attack
Input device
Data processing
Decision-making
Interface
4. Communication channel attack
1. Attack on input device
Data storage
7. Communication channel attack
10. Interface attack
5. Database Attack

Figure 6. Identification module.

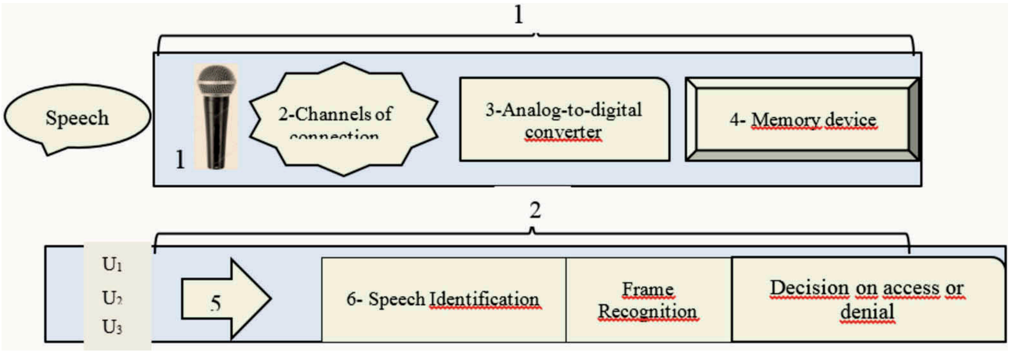

Voice input
Random word generator
Segmentation
Standards of words
Segment Signs
Reference Database
Speaker database
The probability of being a speaker
The probability of belonging to the reference speakers
Comparison with thresholds
Probability of belonging
Decision on access or denial

a Dictaphone, and then get access by reproducing this record. To eliminate this drawback, the following identification scheme is proposed.

The practical implementation of the identification module with protection against attacks provides for the presence of several stages of work. At the 1st stage, the reference voice of the user is recorded, and at the 2nd stage—the comparison of the spoken voice with the standard occurs. In standard identification systems, there is the following drawback: an intruder can record a reference voice message on a Dictaphone, and then get access by reproducing this record. To eliminate this drawback, the following identification scheme is proposed (Figure 7).

At the 1-st stage, unrelated short voice messages are recorded as separate words. Under number 1—the microphone is designated, 2—communication channel, 3—analog-to-digital converter, 4—storage device. At the 2nd stage, the information system selects several audio messages recorded by the user and provides an opportunity to pronounce these words in the chosen order with the text indicated in the system window. The user utters these words for subsequent analysis in the data analysis device—5. Next, the identification device—6, performs the procedure of analyzing and counting the user's repeated frames, checks the recognition of frames in

Figure 7. Block diagram of an identification module with protection against attacks.



accordance with the spoken word. When the frames correspond to a particular user the most (at least 60%) and the recognized words match the words in the system window, the decision is made to grant access to the user.

The developed voice identification module does not depend on the language, nationality, age, gender, emotional state and health of the user. Identification requires a passphrase of only 3–5 seconds, which allows you to significantly save time when passing through this procedure and fully automate it. A dynamically changing passphrase that the user is offered to say (for example, a certain sequence of phrases or numbers) allows the attackers to increase the resistance to hacking the system and eliminate the possibility of an attack.

Next, will evaluate the effectiveness of the proposed identification scheme. When recording phonemes, used *N* phrases. During identification, the user's voice is formed from *n* standards. Then the number of phonemes of user *K* will be equal to:

$$K = N^n \tag{13}$$

If the intruder recorded the user's pronunciation, then when attempting to hack, the probability *P* that the message recorded by him matches the requested identification system will be equal to:

$$P = \frac{1}{N} \tag{14}$$

Below is a Table 1. of 3 probabilities *P*calculated in accordance with Equation (14) for the values of *N*and *n*. The horizontal values of *N* are located, and vertically—*n*, at the intersection of the column and row—the corresponding probability *R*.

4. Experimental results and discussions

Table 2 shows that the overall accuracy of voice recognition using Transformed Mel-Frequency Cepstral Coefficients T-MFCC is 56.08% and 58.92% for the *i*-vector and DNN, respectively. On the other hand, voice recognition accuracy using traditional MFCC is calculated as 43.61% and 46.21%

Table 1. Probabilities *P* calculated in accordance with formula (14) for some values of *N* and *n*

n N	7	14	21	28	35
1	0.142857143	0.071428571	0.047619048	0.035714286	0.028571429
2	0.020408163	0.005102041	0.002267574	0.00127551	0.000816327
3	0.002915452	0.000364431	0.00010798	4.55539E-05	2.33236E-05
4	0.000416493	2.60308E-05	5.14189E-06	1.62693E-06	6.66389E-07
5	5.9499E-05	1.85934E-06	2.44852E-07	5.81045E-08	1.90397E-08

As can be seen from the table, the probability of *P* is small, so the probability of hacking the system, even if the violator was able to overhear and record the pronunciation of the phrase, is also small. Therefore, the scheme of work of this module was introduced into the developed voice identification systems.

Table 2. The overall accuracy of the voice recognition of DNN and I-Vector using traditional and T-MFCC (%)

classifier		U1	U2	U3	U4	U5	U6	U7	Ovll. Acc
i-vector	Traditional MFCC	64.85	57.11	49.02	24.51	27.янв	49.93	32.81	43.61
	T-MFCC	60.35	65.01	48.01	45.47	49.61	57.12	67.02	56.08
DNN	Traditional MFCC	54.43	53.1	44.81	26.1	43.22	46.21	55.61	46.21
	T-MFCC	62.21	61.6	53.48	47.9	51.91	64.25	71.08	58.92

for the same classifiers. Accuracy of voice recognition U1, U2, U3 and U4 has increased dramatically. T-MFFCs, which are generated for the first time in this work, have increased the overall classification accuracy by about 13%. One of the reasons for this improvement is that T-MFCC features are prosodic features in addition to spectral features. The participation of phoneme marks in the generation of T-MFCC allowed us to understand such prosodic features as intonation, stress, tone and rhythm of the speaker. Another reason is that the transformed functions are the result of using phoneme labels in the training data, and this helped remove any noise or silent frames so that the transformed functions are calculated without acoustic background noise.

Figure 8 shows the receiver performance of the converted and traditional MFCC (with random and regularized weights) using the DNN voice and *i*-vector. The receiver curves are calculated using the "one piece" rule. It has been established that the area under the curve for T-MFCC is larger than for traditional MFCC (Table 2 compares the area under the curve for both sets). The values of the area under the curve are calculated as in (Mamyrbayev, Turdalyuly et al., 2019). DNN voice recognition works better than *i*-vector voice recognition in terms of the area under the curve.

Another analysis for comparing T-MFCC and baseline MFCC was done by comparing the variations between the standard deviation of the MFCC and the normalized energy parameter for each class for additional insight, this is shown in Figure 8(a,b). It is observed that the T-MFCC characteristics represent less intra class variation than the original MFCC. It is also observed that there are significant differences between the classes in the characteristics of the T-MFCC. Minimal intra class variation and maximum interclass variation in features are preferred for better classification.

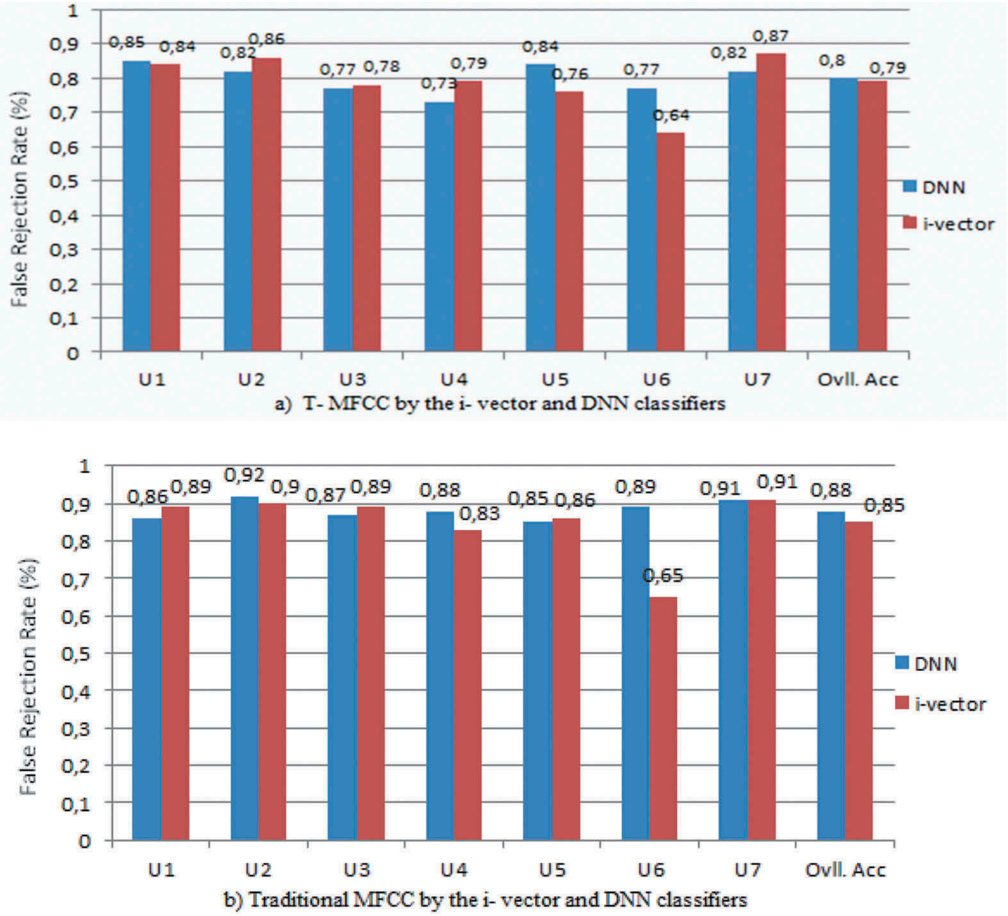
Regarding the comparison of classifiers, the DNN classifier worked slightly better than the *i*-vectorclassifier. The corresponding measurements are the DNN classifier and the i-vector classifier under the voice recognition curve are shown in Table 3. Figure 8 shows the dispersion of weights at each level in the DNN classifier using random weights and regularized weights. A higher difference between the weights in each layer is necessary to distinguish between different classes. As can be seen in Figure 8, the difference between weights using common labels is higher than that of randomly initialized weights, therefore regularized weights converge faster than random weights for most DNN layers.

The purpose of the experiment was to study the probability of correct identification using voice cloning and parodying technologies (voice changing) to modify the "fake" voice of the user. The phonograms of five well-known users (U1, U2, U3, U4, U5) were recorded in monologue mode, the 1st recording was made on a regular microphone without using third-party programs, the 2nd recording was conducted with the involvement of "parodist" voice and using specialized cloning programs Morphvox and Voice changer voices. When creating a phonetic base, all the phonemes were named according to the user being studied (Kalimoldayev et al., 2019).

Table 3. The corresponding measurements of the area under the voice recognition curve

	DNN		***i*-vector**	
Classifier	**Traditional MFCC**	**T-MFCC**	**Traditional MFCC**	**T-MFCC**
U1	0.85	0.86	0.84	0.89
U2	0.82	0.92	0.86	0.9
U3	0.77	0.87	0.78	0.89
U4	0.73	0.88	0.79	0.83
U5	0.84	0.85	0.76	0.86
U6	0.77	0.89	0.64	0.65
U7	0.82	0.91	0.87	0.91
Ovll. Acc	0.80	0.88	0.79	0.85

Figure 8. (a, b) Corresponding measurements for classification of speakers.



a) T- MFCC by the i- vector and DNN classifiers

b) Traditional MFCC by the i- vector and DNN classifiers

For the full experiment, the phonemes of all 5 users (645 phonemes) were recorded into a single common database “IDENTIFICATION known Users … .”.Further, for the text-independent identification of users, continuous text of the studied user was submitted, obviously different from the text used to create the phonetic base.

When conducting experimental tests of the system for identifying users by voice, a personal computer with a processor of at least 2000 MHz and 1 GB of RAM, a Windows operating system, and a sound card with a sampling frequency of 8 kHz and the ability to record sound files are required.

After downloading an unknown audio signal, the phoneme was segmented according to the base used. At the final stage, according to the algorithm, all recognized phonemes were counted and the dominant phonemes were calculated among all the others. If the majority of FBD phonemes are recognized correctly, then the number of units in *N* positions of the analyzed signal is greater than a certain threshold value (60%) defined above, which allows you to decide whether this signal belongs to a specific user or modify it, i.e. to identify the true voice of a known user or his “clone”.

Table 4. shows that in the spoken phrase, 317 phonemes are allocated in total, of which 274 phonemes belong to the user Urgant, and 43 phonemes are recognized as “false” phonemes that resemble phonemes of other users, that at least 60% of the total number of phonemes, which allows us to identify the user “P1”, and in Table 4 313 phonemes are allocated, of which 282 phonemes belong to the user “U1 clone”, and 31 phonemes are recognized as a “false” phoneme, which allows us to identify the user “U1 clone”.

Table 4. User identification “U1”, “U1 clone” the system

Signal segmentation	U1	U1 clone
The number of selected phonemes	317	313
Number of recognized phonemes	317	313
Speaker phoneme number	274	282
Number of phonemes recognized as “false”	43	31

Figure 9.Percentage ratio of user phonemes “U1” and “U1 clone”.

Table 5 summarizes the results of identification by users, which are fully consistent with the above logic.

Thus, according to the larger number of phonemes belonging to a particular user, you can identify the voice of the “original” user and his clone (fake). The results of this experiment show that the developed identification algorithm allows you to uniquely identify the user, even when trying to “fake” his voice, which certainly can be used to improve the reliability of identification in access control systems using speech technologies.

As a result of the analysis of the above voice identification systems, Table 6 was constructed with comparative characteristics of biometric voice identification and identification systems.

After analyzing the results, we can conclude that the developed system within the framework of research has several advantages over similar systems, showing a low probability of errors with equal values of False Acceptance Rate (FAR) and False Rejection Rate (FRR), a small percentage of registration errors, a low probability of false acceptance, but the system is inferior to its counterparts in terms of the probability of a false deviation. It can be concluded about the competitiveness of the developed system, given the high performance and relatively low cost of the system.

Biometric authentication must make a decision from some organic input. Decisions must be made with two main indicators. These indicators used for making decisions are the false acceptance rate and the false failure rate.

Figure 9. Shows a diagram of the percentage ratio of the phonemes of users “U1” and “U1 clone”.

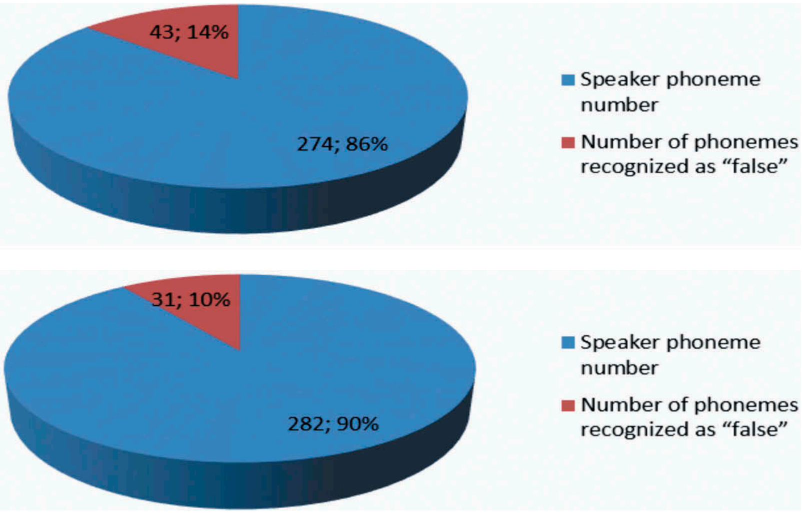


Table 5. Summary table of identification results by the number of phonemes of 10 users in the original and "cloned" sound

User	total number of phonemes	user phonemes	other phonemes
U1	317	274	43
U1 clone	313	282	31
U2	167	164	3
U2 clone	25	25	0
U3	351	278	73
U3 clone	347	247	100
U4	73	73	0
U4 clone	29	27	2
U5	322	265	57
U5 clone	313	215	98

Table 6. Voice Identification Characteristics

System scan options	Identification system
Equal Error Rate (EER)- probability of biometric access system errors, in which FAR and FRR are equal	1%
Registration Denied	2%
FAR—probability of false acceptance. The probability of making a "foreign" user for "his" (the error of the second kind)	0.025
FRR—probability of false rejection. The probability of rejection of "their" user, taking him for "someone else" (the error of the first kind)	0.005
System cost	low

5. Conclusion and future works

In the modern world, any enterprise faces the challenge of protecting against unauthorized access to its material and computer resources. An example of practical application is the introduction of developed voice identification technologies into access of control systems.

The developed identification and authentication system is a unique voice biometric identification system. The voice is the only biometric feature that can be installed and confirmed from a distance: by telephone or via the Internet. This is the reason for the key advantage of the system. The system is an effective text-independent solution that complements the information security system of an enterprise with modern biometric identification tools.

User identification is a biometric proof of a person's identity based on their voice. Each person has his own individual "voice imprint". Identification is carried out automatically on the basis of a person pronouncing a certain phrase and comparison with the standard of his voice that was previously created for this person. The technology can be used in Web-systems and the Internet, in interactive self-service systems that are modern when receiving information by telephone, in access control systems while limiting physical access to the premises.

The proposed algorithms for identifying users of systems based on individual voice characteristics, based on the combined use of the DNN method and the cluster model of elementary speech units in the *i*—vector, differing in increased security from various types of attacks on the biometric identification system, which allowed identifying users with the first and second kind of error probability of 0.025 and 0.005.

The main modules and the voice identification subsystems are described. A structural diagram of the interaction of the main components of the systems is presented. The analysis of the vulnerability of the modules of the system of biometric identification by voice is performed and a block diagram of the identification of users by voice with enhanced protection against attacks is proposed.

The use of elementary speech units in the developed identification algorithms allows to increase computational indicators, reduce subjective decisions in the biometric systems and increase security against attacks on the voice biometric identification systems.

The relevance of the application of the system is associated with the possible implementation of the following security threats:

(1) The problem of “identity theft”: Financial theft of personal data (using other people’s personal data, you can purchase goods and services);
(2) Identity theft with criminal intent, personal cloning, business and commercial identity theft.
(3) Losses of companies as a result of theft of corporate (access to internal databases) data by employees of the companies themselves or by external intruders.

Prospects for **further research** are related to the expansion of the experimental base in the field of voice identification, for example, more in-depth studies on the application of the developed method and algorithm for forensic examination.

Acknowledgements
This work was supported by the Ministry of Education and Science of the Republic of Kazakhstan. IRN AP05131207 Development of technologies for multilingual automatic speech recognition using deep neural networks.

Funding
The authors received no direct funding for this research.

Author details
Kydyrbekova Aizat[1,3]
E-mail: kas.aizat@mail.ru
Othman Mohamed[2]
E-mail: mothmanupm@gmail.com
Mamyrbayev Orken[1,3]
E-mail: morkenj@mail.ru
ORCID ID: http://orcid.org/0000-0001-8318-3794
Akhmediyarova Ainur[1,3]
E-mail: aat.78@mail.ru
Bagashar Zhumazhanov[3]
E-mail: bagashar.zhumazhanov@narxoz.kz
[1] Institute of Information and Computational Technology, Almaty 050010, Kazakhstan.
[2] Department of Communication Tech and Network, Universiti Putra Malaysia, UPM 43400, Serdang, Selangor D.E., Malaysia.
[3] al-Farabi Kazakh National University, Almaty 050040, Kazakhstan.

Citation information
Cite this article as: Identification and authentication of user voice using DNN features and *i*-vector, Kydyrbekova Aizat, Othman Mohamed, Mamyrbayev Orken, Akhmediyarova Ainur & Bagashar Zhumazhanov, *Cogent Engineering* (2020), 7: 1751557.

References

Chan, V., Jaitli, N., Le, K., Vinhals O. 2016. Listen, listen and say: A neural network for recognizing conversational speech with a large vocabulary. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4960–4964).

Jun Du, Peng Liu, Frank Soong, Jian-Lai Zhou, Ren-Hua Wang (2006). Speech recognition performance with noise in HMM recognition. *Processing Chinese Spoken Language (Lecture Notes in the Field of Computer Science)*, *4274*, 358–369.

Fine, S., Navratil J., Gopinath R.A. 2001. GMM/SVM hybrid approach to speaker identification. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’01)* (pp. 417–420).

P. Ghahremani, B. Baba Ali, D. Povey, K. Riedhammer, J. Trmal, S. Khudanpur 2014. Algorithm for extracting pitch, configured for automatic speech recognition. In *The IEEE International Conference on Acoustics, Speech and Signals Processing (ICASSP)* (pp. 2494–-2498), Florence, Italy .

Gemmeke, J. F., Virtanen, T., Hurmalainen, A. (2011). Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Operations on Sound, Speech and Language Processing*, *19*(7), 2067–2080. https://doi.org/10.1109/TASL.2011.2112350

David J. Hand, Robert J. Till. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45* (2), 171–186. https://doi.org/10.1023/A:1010920819831

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimension of data using neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Juang B.H., Rabiner L.R. (2005) “Automatic Speech Recognition—A Brief History of the Technology,” Elsevier Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, *11*, 806–819

Jitendra, A., McCowan, I. (2003). Speech/music segmentation using the functions of entropy and dynamism within the framework of the HMM classification. *Speech Communication*, *40*(3), 351–363. https://doi.org/10.1016/S0167-6393(02)00087-0

Kalimoldayev, M. N., Mamyrbayev, O. Z., Kydyrbekova, A. S., Mekebayev, N. O. (2019). Voice verification and identification using i-vector representation. *International Journal of Mathematics and Physics*, *10*(1), 66. https://ijmph.kaznu.kz/index.php/kaznu/article/view/280 https://doi.org/10.26577/ijmph-2019-i1-9

Kundu, S., Mantena, G., Qian, Y., Tan, T., Delcroix, M., Sim, K. S. 2016. A joint study of the acoustic factor for reliable automatic speech recognition on the basis of deep neural networks. In *The IEEE International Conference on acoustics, Voice and Signal Processing (ICASSP)* (pp. 5025–5029).

Ley, Y., Scheffer, N., Ferrer, L., McLaren, M. 2014. Speaker recognition using phonetically aware deep neural network. In *Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference* (pp.1695–1699).

Makovkin, K. A. (2006). Hybrid models: Hidden Markov models and neural networks, their application in speech recognition. In *Coll.: Modeling, algorithms and architecture of speech recognition systems* (pp. 96–118). EC of the Russian Academy of Sciences Moscow: Computing Center RAS.

Mamyrbayev, O. Z., Kydyrbekova, A. S., Turdalyuly, M., Mekebaev, N. O. 2019. Review of user identification and authentication methods by voice. In *Materials of the scientific conference "Innovative IT and Smart Technologies"* : Mat. scientific conf. - Almaty: IITT MON RK, (pp.315–321).

Mamyrbayev, O. Z., Othman, M., Akhmediyarova, A. T., Kydyrbekova, A. S., Mekebayev, N. O. (2019). Voice verification using -vectors and neural networks with limited training data. *Bulletin of the National Academy of Sciences of the RK Issue*, *3*, 36–43. https://www.researchgate.net/publication/333891112

Mamyrbayev, O. Z., Turdalyuly, M., Mekebaev, N. O., Kydyrbekova, A. S., Turdalykyzy T., Keylan A. 2019. Automatic recognition of the speech using digital neural networks. In 11th Asian Conference *ACIIDS, Indonesia, Proceedings, Part II*

Mangu, L., Brill, E., & Tolke, A. S. (2000). Finding consensus in speech recognition: Minimizing word errors and other applications of confusion networks. *Computer Speech & Language*, *14*(4), 373–400. https://doi.org/10.1006/csla.2000.0152

Meuwly D., Drygajlo A. 2001. Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM), in A Speaker Odyssey-The Speaker Rechoping Workshop, Crete, Greece, 52–55, http://www.isca-speech.org/archive .

Richardson, F., Reynolds, D., & Dehak, N. (2015). The unified deep neural network for speech and language recognition. *Preprint arXiv arXiv: 1504.00923*.

Savchenko, V. V. (2009). The phonetic decoding method of words in the task of automatic speech recognition based on the principle of minimum information mismatch. *Proceedings of Russian Universities. Radio Electronics*,Issue 5. from. 41–49.https://elibrary.ru/item

Senor, A., & Lopez-Moreno, I. 2014. Improving the independence of DNN carriers using i-vector inputs. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* New York: Google Inc. (pp. 225–229).

Shahamiri, S. R., & Binti Salim, S. S. (2014). Artificial neural networks as speech recognizers for dysarthric speech: Identifying the best- performing set of MFCC parameters and studying a speaker-independent approach. *Advanced Engineering Informatics*, *28*(1), 102–110. https://doi.org/10.1016/j.aei.2014.01.001

Valsan, Z., Gavat, I., & Sabach, B. (2002). Statistical and hybrid methods of speech recognition in Romanian. *International Journal of Speech Technologies*, *5*(3), 259–268. https://doi.org/10.1023/A:1020249008539

Varga, A., & Steeneken, H. J. (1993). Evaluation for automatic speech recognition: II. NOISEX-92: Database and experiment on the study of the effect of additive noise on speech recognition systems. *Speech Communication*, *12*(3), 247–251. https://doi.org/10.1016/0167-6393(93)90095-3

Wu, S. L., Kingsbury, E., Morgan, N., & Greenberg, S. 1998. Incorporating information from syllable-length time scales into automatic speech recognition. In *The IEEE International Conference on Acoustics, Speech and Signal Processing*, Seatle, (pp. 721–724).

Yao, K., Yu., D., Seyde, F., Su, H., Deng, L., & Gong, Y. 2012. Adaptation of context-dependent deep neural networks for automatic speech recognition. IIn: Proceeding of the IEEE Spoken Language Technology Workshop (SLT), (pp. 366–369).

Young, S., Evermann, G., Gales, M., Hein, T., Kershaw, D., Liu, H. (2006, December). *The Book of the CTC*. Faculty of Engineering, University of Cambridge, edition of the CTC, version 3.4.

Yu, K., Liu, G., Ham, S., & Hansen, J. 2014. Spreading uncertainty in foreground analysis for noise-resistant speaker recognition. In *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4017–4021). https://doi.org/10.1109/ICASSP.2014.6854356

Yuan, J., & Lieberman, M. (2008). Speaker identification in the SCOTUS building. *Journal of the Acoustic Society of America*, *123*(5), 3878. https://doi.org/10.1121/1.2935783

***Cogent Engineering* (ISSN: 2331-1916) is published by Cogent OA, part of Taylor & Francis Group.**