BULLETIN OF NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
ISSN 1991-3494
Volume 5, Number 387 (2020), 6 – 15 **https://doi.org/10.32014/2020.2518-1467.137**

UDC 004.932.75

**O. Mamyrbayev[1,2], A. Akhmediyarova[1,2], A. Kydyrbekova[1,2],
N. O. Mekebayev[1,2], B. Zhumazhanov[1]**

[1]Institute of Information and Computing Technologies, Almaty, Kazakhstan;
[2]Kazakh National University named after Al-Farabi, Almaty, Kazakhstan.
E-mail: morkenj@mail.ru, aat.78@mail.ru, kas.aizat@mail.ru, nurbapa@mail.ru, bagasharj@mail.ru

BIOMETRIC HUMAN AUTHENTICATION SYSTEM THROUGH SPEECH USING DEEP NEURAL NETWORKS (DNN)

Abstract. Biometrics offers more security and convenience than traditional methods of identification. Recently, DNN has become a means of a more reliable and efficient authentication scheme. In this work, we compare two modern teaching methods: these two methods are methods based on the Gaussian mixture model (GMM) (denoted by the GMM i-vector) and methods based on deep neural networks (DNN) (denoted as the i-vector DNN). The results show that the DNN system with an i-vector is superior to the GMM system with an i-vector for various durations (from full length to 5s). DNNs have proven to be the most effective features for text-independent speaker verification in recent studies. In this paper, a new scheme is proposed that allows using DNN when checking text using hints in a simple and effective way. Experiments show that the proposed scheme reduces EER by 24.32% compared with the modern method and is evaluated for its reliability using noisy data, as well as data collected in real conditions. In addition, it is shown that the use of DNN instead of GMM for universal background modeling leads to a decrease in EER by 15.7%.

Key words: biometrics, speaker verification, short sentences, i-vector, DNN.

1. Introduction. Biometry is the development of statistical and mathematical methods applicable to the problems of data analysis in the biological sciences. The introduction of this technology brings new approaches to the security of computer systems. Identification and verification are two ways to use biometrics to authenticate a person. Biometrics refers to the use of physical or physiological, biological or behavioral characteristics to establish a person's identity. These characteristics are unique to each person and remain partially unchanged during the course of a person's life [1]. A biometric security system is becoming a powerful tool compared to electronic security [2]. Any physiological or behavioral characteristic of a person can be used as a biometric characteristic, provided that it has the following properties: universality, distinctiveness, constancy, collectability, bypass, acceptability and performance [3]. Physiological biometry is associated with body shape. Behavioral biometrics related to human behavior. Speech is a unique biometric feature that falls into both categories [4]. Based on the application, choosing the right biometrics is a crucial part. For example, speech is a biometric feature whose characteristics will differ significantly if a person is affected by a cold or other emotional status. Some of these problems can be solved using a biometric system. Although some of the features offer good performance in terms of reliability and accuracy, none of the biometric features are 100% accurate. With the growing global need for security, the need for reliable face recognition systems is becoming apparent.

For applications related to the flow of confidential information, the accuracy of system authentication is always a priority. Examples of such applications include secure access to buildings, computer systems, laptops, cell phones and ATMs. In the absence of reliable personality recognition schemes, these systems are vulnerable to the cunning of an impostor. The aim of this work is to develop a biometric system using speech technologies for personal identification.

The present work is mainly devoted to the implementation of a biometric system using speech and we use i-vector DNNs, which are usually used to recognize a speaker in a VP.

In this work, various methods are proposed for improving the speaker's gender classification based on the i-vector and deep neural networks (DNN) as an attribute extractor and classifier. First, a model is proposed for generating new functions from DNN. DNN with a bottleneck layer is trained in an uncontrolled way to calculate the initial weights between the layers, then it is trained and not controlled in a controlled way to generate converted low frequency cepstral coefficients (T-MFCC). Secondly, the label method of general classes is introduced among incorrectly classified classes to regularize weights in DNN. Thirdly, DNN-based speaker models using the SDC feature set are offered. A speaker-supported model can more effectively capture the speaker's age and gender characteristics than a model representing a group of speakers. Moreover, the new T-MFCC feature set is used as input to a two-system merger model. The first system is a class model based on the GNN vector, and the second system is a speaker model based on the DNN i-vector [5]. Using T-MFCC as input and combining the final grade with the grade model based on the DNN vector improved the classification accuracy.

2. Human authentication system through speech. Like any other pattern recognition system, a speech-based personality authentication system also consists of three components: (1) feature extraction, which converts the speech waveform into a set of parameters that carry essential information about the speaker; (2) Generating a template that generates a template representing an individual speaker from the object's parameters: and (3) Matching and classifying a template that compares the similarity between the extracted objects and the previously saved template or the number of previously saved templates, respectively providing the speaker's identity. The speaker recognition system consists of two stages: training and testing. At the training stage, speaker models (or patterns) are generated from speech patterns using some selection and modeling methods. At the testing stage, feature vectors are generated from the speech signal using the same extraction procedure as during training. Then a classification decision is made using some matching method. Identity authentication is a binary classification task [6]. The characteristics of the test signal are compared with the claimed speaker circuit, and a decision is made to accept or reject the claim [7]. Depending on the operating mode, speaker recognition can be classified as text-dependent recognition and independent text recognition. Text-dependent recognition requires that the speaker make a speech over the same text both during training and testing, while independent text recognition does not affect the specific text spoken. In this paper, we use the method of recognition of text-dependent texts. In this paper, we use feature extraction methods based on (1) Mel frequency coefficient coefficients (MFCC) obtained from Cepstral speech analysis, and (2) wavelet-octave remainder coefficients (WOCOR) obtained from Linear Prediction (LP) residual. A time-frequency analysis of the LP residual signal is performed to obtain WOCOR [8]. WOCORs are generated by applying a pitch-synchronous wavelet transform to the residual signal. Experimental results show that WOCOR parameters provide additional information to the usual MFCC functions for speaker recognition [9]. Vector quantization (VQ) and Gaussian mixture modeling (GMM) are used to model information about a person from these MFCC and WOCOR functions [10,11]. The modern system uses MFCC derived from speech as feature vectors, and GMM as a modeling method.

2.1 Pretreatment. Voice activity detection (VAD) is an important step in most speech processing applications, especially if there is background noise. The importance of VAD is that it improves intelligibility and speech recognition. Since the speech utterances used in this work were recorded in a public call center, the recording utterances were exposed to noise and other interference. As a result, the VAD algorithm is needed to reduce background noises and quiet eras in utterances in order to prepare them for identifying features. In addition, normalization of cepstral mean dispersion (CMVN) is used to eliminate convolutional distortions and linear channel effects. CMVN can be applied globally or locally. In this work, it is applied globally to obtain a normal distribution with zero mean and unit dispersion. The

MFCC set is one of the most famous sets of spectral characteristics and is widely used in many speech applications [12].

2.2 Extracting functions from speech information. Information about the dynamics is present both in the vocal tract and in the excitation parameters. The voice path system can correspond to speech processing in short (10-30ms) overlapping (5-15ms) windows. It is assumed that the vocal tract system is stationary within the window and can be modeled as an all-pole filter using analysis. The most used form of speech for feature extraction is Cepstrum. Various forms of presentation of Cepstral include complex Cepstral coefficients (CCC), real Cepstral coefficients (RCC), Mel frequency Cepstral coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). Among them, the most commonly used MFCC [13].

2.3 Extracting MFCC Feature Vectors. The state of the system builds a unimodal system by analyzing speech in blocks of 10-30ms with a shift of half the block size. MFCCs are used as feature vectors extracted from each of the blocks. MFCCs are mainly the voice path of the speaker's information and therefore only care about the physiological aspect of the biometric characteristics of speech. The speech signal is obtained by convolution of the voice parameters $v(n)$ and excitation parameters $x(n)$ given by equation (1). We cannot separate these parameters in the time domain. Hence we go for the cepstral domain. Cepstral analysis is used to separate the speech path parameter v (n) and the excitation parameters $x(n)$ from the speech signal $s(n)$.

$$s(n) \ = \ v(n) * \ x(n) \tag{1}$$

Cepstral analysis provides a fundamental convolution property used to separate the parameters of the vocal tract and the excitation parameters. Cepstral coefficients C of length M can be obtained using equation (3.2).

$$C = \ real(IFFT(log|FFT(s(n))|)) \tag{2}$$

A non-linear scale, that is, the relationship between the Mel frequency (f_{Mel}) and the physical frequency (fH_z), is used to extract spectral information from the speech signal using Cepstral analysis.

$$f_{Mel} = 2595 \ log_{10}\left(1 + \frac{fH_z}{700}\right) \tag{3}$$

Using equation (3), we construct a spectrum with critical bands that are overlapping triangular banks, i.e. we map the linearly spaced frequency spectrum (fH_z) to the nonlinearly spaced frequency spectrum (f_{Mel}). In this way, we can imitate the human auditory system and, based on this concept, MFCC feature vectors are obtained. Window control eliminates the Gibbs vibrations that occur by cutting the speech signal. Using equation (4), Hamming window coefficients are generated with which the corresponding frame speech is scaled.

$$w(n) = 5.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \tag{4}$$

But due to working with the Hamming window, samples that are on the verge of the window are weighed with lower values. To compensate for this, we will try to block the frame by 50%. After the window, we calculate the logarithmic spectrum of each frame to find the energy coefficients using equation (5).

$$Y(i) = \sum_{k=0}^{\frac{N}{2}} \log \ |S(k,m)|Hi\left(k\frac{2\pi}{N-1}\right) \tag{5}$$

where $Hi\left(k\frac{2\pi}{N}\right)$ is the ith spectrum of the critical Mel bank, and N is the number of points used to calculate the discrete Fourier transform (DFT). The number M of frequency coefficients Mel calculated using discrete cosine transforms (DCT) using equation (6), which is nothing more than a real IDFT critical-band filter that records the output energy.

$$C(n,m) = \left(\frac{2}{N}\right)\sum_{k=1}^{\frac{N}{2}-1} Y(k)cos\left(k\frac{2\pi}{N}n\right) \tag{6}$$

where n = 1, 2, 3 ……… ..M.

2.4 Fully connected neural network modeling. The neural network used in this work is a fully connected neural network with a straightened linear unit [14] as an activation function. A straightened linear unit is essentially a piecewise function that turns all negative values into 0, while positive values remain unchanged. This practice is called one-way inhibition and makes neurons rarely activated. Its advantage is that the preliminary preparation process can be omitted without the problem of the disappearance of the gradient. In addition, a straightened linear unit only requires addition and multiplication, so it is faster and more efficient than other functions such as sigmoid and tang. The basic unit of a neural network is a neuron, and the direction of data flow in a neuron using a straightened linear unit is shown in figure 1.

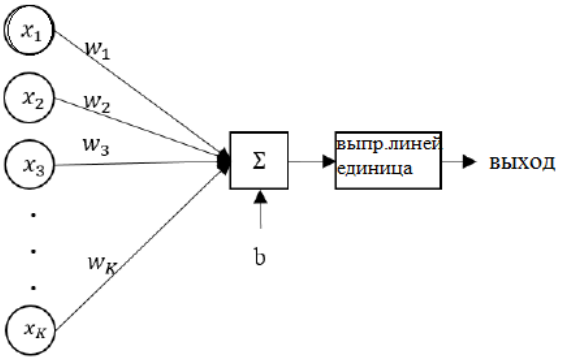

Figure 1 – Schematic diagram of the data flow in a neuron

The calculation formula corresponding to figure 2:

$$y = ВЛЕ(x) = \max(0, x) \quad (7)$$

$$x = b + \sum_K w_k x_k \quad (8)$$

A large number of neurons form a neural network through an extensive relationship, and its structure is shown in figure 2.

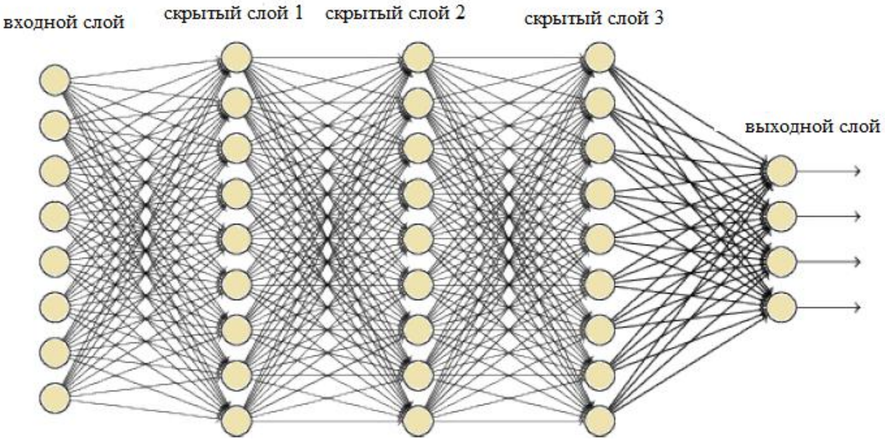

Figure 2 – Fully connected neural network structure

As input to a neural network, the i-vector function is best suited for a linear classifier. However, for neural networks, more hidden layers mean a weaker degree of linearity. Therefore, we use fewer hidden layers in this work. At the same time, when there are too many network parameters and too few training samples, it is very likely that re-equipment will occur [15,16]. Retooling means that the model has small losses and high accuracy of forecasting training data, but large losses and low accuracy of forecasting test data. The dropout strategy is currently an effective method of suppressing retooling. Dropout means that when training a neural network, the values of activation of neurons change by 0 at a certain ratio v, that is, they cancel part of the nodes of the hidden layer in accordance with the ratio v. During testing, the output values of the hidden nodes of the layer should be reduced to $(1-v)$ times, for example, if the normal output signal is a, it should be reduced to$(1-v)$. Screening can be considered as a method of averaging a model that averages estimates or forecasts from different models by a certain weight, which is also referred to as a combination of models in some literature. In each learning process, since the nodes are accidentally ignored, the resulting network is different and can be considered as a "new" model. In addition, the nodes are activated randomly with a certain probability, therefore there is no guarantee that every 2 nodes will be valid together every time, then the update of the weights no longer depends on the joint action of the nodes with fixed relations, which prevents a situation where some functions are effective only if there are other functions [17].

3. Experimental Results

3.1 Comparison of recognition results for GMM vector systems and DNN i-vector. In this section, we show the mapping results for the GMM i-vector system. From table 1 we see that the proposed mapping methods provide a significant improvement for both systems. After mapping, DNN i-vector systems still outperform GMM i-vector systems, and the superiority of DNN i-vector systems becomes even more significant. We also compare mapping results when adding phoneme vectors. The table shows that the effect of adding phoneme information is more important for GMM-i-vectors, and it can achieve a relative improvement of 10% compared to the best basic level of DNN mapping. The reason is that DNN-i-vectors already contain some information about phonemes, while GMM-i-vectors can greatly benefit from adding phoneme vectors. As a result, we summarize the baseline and the best mapping results for both systems in figure 3. Curves DET (Trade Error Error Tradeo) are presented for both women and men. The numbers show that the proposed mapping algorithms provide a significant improvement over the baseline at all operating points.

Table 1 – Results for GMM i-vector systems and i-vector DNNs

	female		male	
	EER (Rel Imp)	DCF08/DCF10	EER (Rel Imp)	DCF08/DCF10
GMM i-vector				
A basic level of	13,6	0,063/0,097	14,3	0,057/0,099
DNN mapping	20,27%	0,054/0,095	22,37%	0,051/0,096
DNN mapping + phoneme information	25,54%	0,053/0,094	27,90%	0,048/0,096
i-vector DNN				
A basic level of	13,4	0,054/0,093	11,5	0,048/0,095
DNN mapping	26,53%	0,046/0,091	27,57%	0,038/0,089
DNN mapping + phoneme information	28,75%	0,046/0,090	29,47%	0,037/0,090

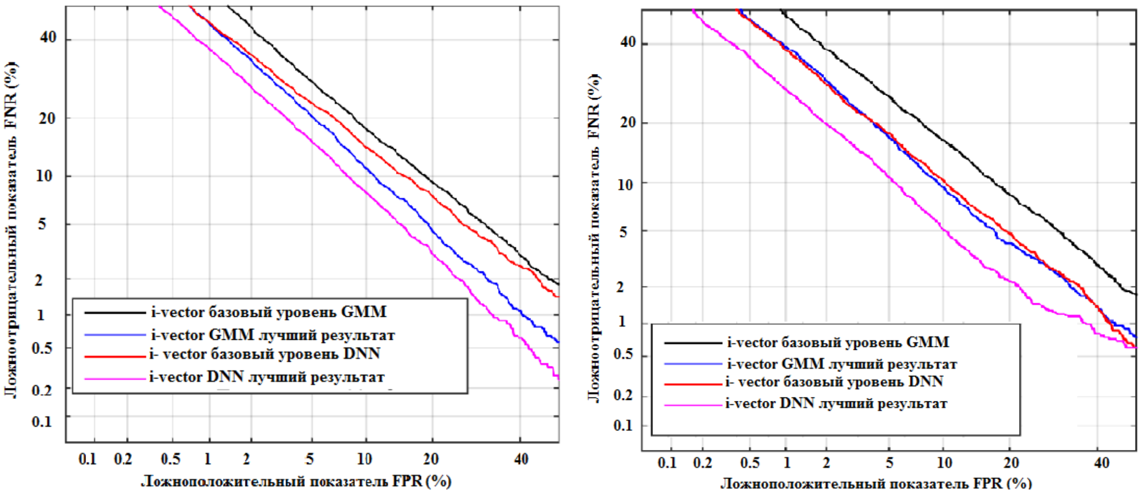

Figure 3 – DET curves for the mapping results of GMM i-vector systems and i-vector DNNs.
The left digit corresponds to the female language, and the right one corresponds to the male language

3.2 Database performance. In this subsection, we apply our technique to a database that contains real audio files collected from open media channels with significant discrepancies. To generate a large number of short statements of random duration, we first combined the dev and eval datasets, and then selected 250 statements from a relatively clean state. We truncated each of 250 statements into several non-overlapping short statements lasting 5 seconds, 3.5 seconds, 2.5 seconds (including both speech and non-speech parts). As a result, 1956 statements were received. In total, we developed 664,746 tests for our task of verifying a speaker with an arbitrary word length. For each short statement, we first lower the sampling frequency of the audio files to a sampling frequency of 8 kHz, and then extract the i-vectors using the previously trained DNN i-vector system. To display the i-vector, we use the most tested models in the SRE10 data set (condition 5s) to apply to the SITW data set. The results of the EER and minDCF estimates are shown in table 2. The table shows that the best models tested on the SRE10 dataset generalize the SITW dataset well, which gives a relative EER improvement of 24.32% for women speaking and 22.87% relative improvement for talking men. The results also show that the proposed methods can be used in real conditions, such as smart home and forensic applications.

Table 2 – DNN-based mapping results for SITW using arbitrary durations of short sentences

	female		male	
	EER (Rel Imp)	DCF10	EER (Rel Imp)	DCF10
Custom Duration				
basic level	16,35	0,079	11,9	0,084
DNN mapping (best models from SRE10)	14,2	0,076	10,8	0,081

3.3 Cartographic effects. To investigate the effect of the proposed i-vector mapping algorithms, we first calculate the mean square Euclidean distance between the short and long pairs of the sentence i-vector in the assessment data set before and after the mapping. The RMS Euclidean distance D_{sl} between the short and long pronunciation i −vector is determined as follows:

$$D_{sl} = \frac{1}{N}\sum_{s=1}^{N}\left(\sum_{i=1}^{L}\left(w_s(i) - w_l(i)\right)^2\right) \tag{9}$$

where w_s and w_l represent the i-vector of the short statement and the long statement, respectively, L is the length of the i-vector, and N is the number of short and long pairs of the i-vector. We compare the D_{sl}values for 10-second and 5-second i-vectors with short sentences, as well as the associated 10-second and 5-second vector expressions for women and men in table 3. From the table we see that after displaying the displayed i- short pronunciation vectors have significantly lower D_{sl}values compared to what they were before the display. After displaying D_{sl} in state 10s less than in state 5s. In addition, we calculate and compare the J-factor of short-pronounced i-vectors before and after the comparison in table 4, which measures the ability to split votes. Given the i-vectors for M votes, the J-coefficient can be calculated using equations 10-12:

$$S_w = \frac{1}{M}\sum_{s=1}^{M} R_i \tag{10}$$

$$S_w = \frac{1}{M}\sum_{s=1}^{M}(w_i - w_0)(\,(w_i - w_0)^T \tag{11}$$

$$J = Tr((S_b + S_w)^{-1}S_b) \tag{12}$$

where S_w is the scattering matrix inside the class, S_b is the scattering matrix between classes, w_i is the average i-vector for the i-th speaker, w_o the average value of all w_i, and R_i is the covariance matrix for the i-th speaker (note that more high J-ratio means better separation). From table 4 we can observe that the mapped i-vectors have significantly higher J-ratios compared to the original i-vectors with short intervals for conditions of 5s and 10s. These results indicate that the proposed DNN-based mapping methods can generalize well to invisible speakers and utterances and improve the ability of i-vectors to split voices.

Table 3 – Square Euclidean distance (D_{sl}) between short and long pairs of sentence i-vectors from SRE10 before and after display

	D_{sl}			
	10 s		5 s	
	original	mapped	original	mapped
female	549,7	304,9	618,8	357,3
male	492,4	305,8	557,4	339,9

Table 4 – J-factor for short-term i-vectors before and after mapping

	J-ratio			
	10 s		5 s	
	original	mapped	original	mapped
female	86,39	94,81	81,85	86,91
male	88,25	91,37	81,15	86,51

4. Conclusion. In this paper, we show how the performance of GMM and DNN-based i-vector speaker verification systems deteriorates rapidly with a decrease in the duration of evaluative statements. This work explains and analyzes the causes of deterioration and offers several DNN-based methods for teaching non-linear mapping of short-sentence i-vectors to their long version in order to improve the performance of short sentence estimation. The proposed mapping method based on DNN is used to model joint representations of i-vectors with short and long statements.

When evaluated using a dataset, mapped short i-vectors can provide a relative improvement of about 24.32% to test short sentences in inappropriate learning conditions, and also exceed the learning conditions of an agreed PLDA using short sentences. We studied several key factors of DNN models and conclude the following: 1) for a trained DNN model with semi-control, uncontrolled learning plays a more important role than controlled learning in the speaker verification task; 2) by increasing the depth of neural networks using residual blocks, we can alleviate the problem of tight optimization of deep neural networks and get an improvement over a shallow network, especially for DNN; 3) the addition of phoneme information can help in the study of nonlinear mapping and provide further improvement in performance, and this effect is more significant for GMM i-vectors; 4) the proposed DNN-based mapping methods work well for short sentences with different and mixed durations; 5) the proposed models can also improve the GMM i-vector system and DNN i-vector, and after matching, the DNN i-vector system still works better than the GMM-i-vector system; and 6) the best-tested models are well generalized to a dataset and provide significant improvement for short sentences of arbitrary length.

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.

О. Ж. Мамырбаев[1,2], А. Т. Ахмедиярова[1,2], А. С. Кыдырбекова[1,2],
Н. О. Мекебаев[1,2], Б. Ж. Жумажанов[1]

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

ТЕРЕҢ НЕЙРОНДЫҚ ЖЕЛІЛЕРДІ (DNN) ҚОЛДАНУ АРҚЫЛЫ АДАМНЫҢ СӨЙЛЕУ АУТЕНТИФИКАЦИЯСЫНЫҢ БИОМЕТРИЯЛЫҚ ЖҮЙЕСІ

Аннотация. Биометрика дәстүрлі сәйкестендіру әдістеріне қарағанда қауіпсіз әрі қолайлы болып келеді. Биометрия – биологиялық ғылымдардағы деректерді талдау мәселелеріне қолданылатын статистикалық және математикалық әдістердің дамуы. Бұл технологияны енгізу компьютерлік жүйе қауіпсіздігіне жаңа көзқарастар әкеледі. Идентификация және растау – жеке куәлікті растау үшін биометриканы қолданудың екі әдісі болып саналады. Биометрика дегеніміз адамның жеке басын анықтау үшін физикалық немесе физиологиялық, биологиялық немесе мінез-құлық сипаттамаларын пайдалануды білдіреді. Соңғы уақытта DNN сенімді және тиімді аутентификация схемасының құралына айналды. Бұл жұмыста келесідей оқытудың екі заманауи әдісін салыстырамыз: Гаусс араласу моделіне негізделген әдіс (GMM) (GMM i-векторымен белгіленеді) және терең нейрондық желілерге негізделген әдістер (DNN) (i-векторлық DNN депбелгіленген). Нәтижелер i-векторы бар DNN жүйесінің әртүрлі ұзындықтарға (толық ұзындығынан 5 сек дейін) арналған *i*-векторы бар GMM жүйесінен жоғары екенін көрсетеді. Соңғы зерттеулерде DNN мәтіндік тәуелсіз дауыстарды тексерудің тиімді функциялары екендігі дәлелденді. Бұл жұмыста мәтінді қарапайым және тиімді әдіспен тексеру кезінде DNN қолдануға мүмкіндік беретін жаңа схема ұсынылған. Тәжірибе көрсеткендей, ұсынылған схема EER технологиясын заманауи әдіспен салыстырғанда 24,32%-ға төмендетеді және оның сенімділігі даулы мәліметтерді, сондай-ақ нақты жағдайда жиналған деректерді пайдалану арқылы бағаланады. Сонымен қатар, әмбебап фондық модельдеу үшін GMM орнына DNN қолдану EER деңгейінің 15,7% төмендейтіні көрсетілген.

Биометриялық қауіпсіздік жүйесі электрондық қауіпсіздікпен салыстырғанда мықты құралға айналуда [2]. Адамның кез-келген физиологиялық немесе мінез-құлық сипаттамасын биометриялық сипаттама ретінде пайдалануға болады, ол үшін келесідей сипаттасы болуы қажет: әмбебаптылық, ерекшелілік, тұрақтылық, жинақталу, айналып өту, қабылдану және өнімділік [3]. Физиологиялық биометрия дене пішінімен байланысты. Адамның мінез-құлқына байланысты мінез-құлық биометрикасы. Сөйлеу – екі санатқа жататын ерекше биометриялық ерекшелік [4]. Қосымша негізінде дұрыс биометрияны таңдау маңызды бөлік болып саналады. Мәселен, сөйлеу дегеніміз – биометриялық сипат, егер адамға суық немесе басқа эмоционалды жағдай әсер етсе, оның сипаттамалары айтарлықтай ерекшеленеді. Осы мәселелердің кейбірін биометриялық жүйенің көмегімен шешуге болады.

Бұл жұмыста атрибуттың экстракторы және жіктеуіші ретінде i-векторлы және терең нейрондық желілерге (DNN) негізделген спикердің гендерлік классификациясын жақсартудың түрлі әдістері ұсынылған. Біріншіден, DNN-ден жаңа функцияларды құру үшін модель ұсынылады. Тығыз қабаты бар DNN қабаттар арасындағы бастапқы салмақты есептеу үшін бақыланбайтын әдіспен оқытылады, содан кейін ол төмен жиілікті цептралды коэффициенттерді (T-MFCC) құру үшін бақыланады және бақыланбайды. Екіншіден, жалпы сыныптардың этикеткалық әдісі DNN-де салмақтарды реттеу үшін қате жіктелген сыныптар арасында енгізіледі. Үшіншіден, SDC мүмкіндіктер жинағын пайдаланатын DNN-негізделген динамик модельдері ұсынылады. Динамик қолдайтын модель спикердің тобын білдіретін модельге қарағанда динамиктің жас және жыныстық сипаттамаларын анағұрлым тиімді көрсете алады. Сонымен қатар, T-MFCC жаңа жиынтығы екі жүйелік біріктіру моделіне кіріс ретінде қолданылады. Бірінші жүйе – GNN векторына негізделген сыныптық модель, ал екінші жүйе – DNN i-векторына негізделген динамикалық модель [5]. T-MFCC енгізу және қорытынды бағаны DNN векторына негізделген градация моделімен біріктіру жіктеудің дәлдігін жақсартты.

Түйін сөздер: биометрика, дауысты тану, қысқа сөйлемдер, i-вектор, DNN.

О. Ж. Мамырбаев^{1,2}, А. Т. Ахмедиярова^{1,2}, А. С. Кыдырбекова^{1,2},
Н. О. Мекебаев^{1,2}, Б. Ж. Жумажанов^{1}

^{1}Института информационных и вычислительных технологий, Алматы, Казахстан;
^{2}Казахский национальный университет им. аль-Фараби, Алматы, Казахстан

БИОМЕТРИЧЕСКАЯ СИСТЕМА АУТЕНТИФИКАЦИИ ЧЕЛОВЕКА ЧЕРЕЗ РЕЧИ С ИСПОЛЬЗОВАНИЕМ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ (DNN)

Аннотация. Биометрия предлагает большую безопасность и удобство, чем традиционные методы идентификации личности. Биометрия – это развитие статистических и математических методов, применимых к задачам анализа данных в биологических науках. Внедрение этой технологии приносит новые подходы к безопасности компьютерных систем. Идентификация и проверка – это два способа использования биометрии для аутентификации человека. Биометрия относится к использованию физических или физиологических, биологических или поведенческих характеристик для установления личности человека. В последнее время DNN стала средством более надежной и эффективной схемы аутентификации личности. В этой работе мы сравниваем два современных метода обучения: этими двумя методами являются методы, основанные на модели гауссовой смеси (GMM) (обозначаемые -вектором GMM) и методы, основанные на глубоких нейронных сетях (DNN) (обозначаемые как *i*-вектор DNN). Результаты показывают, что система DNN с

-вектором превосходит систему GMM с i-вектором при различной длительности (от полной длины до 5с). DNN оказались наиболее эффективными функциями для независимой от текста проверки говорящего в последних исследованиях. В этой работе предлагается новая схема, позволяющая использовать DNN при проверке текста с помощью подсказок простым и эффективным способом. Эксперименты показывают, что предлагаемая схема снижает EER на 24,32% по сравнению с современным методом и оценивается на предмет ее надежности с использованием зашумленных данных, а также данных, собранных в реальных условиях. Кроме того, показано, что использование DNN вместо GMM для универсального фонового моделирования приводит к снижению EER на 15,7%.

Биометрическая система безопасности становится мощным инструментом по сравнению с электронной безопасностью [2]. Любая физиологическая или поведенческая характеристика человека может быть использована в качестве биометрической характеристики при условии, что она обладает следующими свойствами: универсальность, отличительность, постоянство, собираемость, обход, приемлемость и производительность [3]. Физиологическая биометрия связана с формой тела. Поведенческая биометрия связана с поведением человека. Речь является уникальной биометрической характеристикой, которая подпадает под обе категории [4]. Основываясь на приложении, выбор правильной биометрии является важной частью. Например, речь - это биометрический признак, характеристики которого будут значительно отличаться, если на человека влияет холод или другой эмоциональный статус. Некоторые из этих проблем можно решить с помощью биометрической системы.

В этой работе предлагаются различные методы улучшения гендерной классификации говорящего на основе i-вектора и глубоких нейронных сетей (DNN) в качестве экстрактора атрибутов и классификатора. Сначала предлагается модель для генерации новых функций из DNN. DNN со слоем узкого места обучается неконтролируемым образом для вычисления начальных весов между слоями, затем оно обучается и не контролируется контролируемым образом для генерации преобразованных низкочастотных кепстральных коэффициентов (T-MFCC). Во-вторых, метод меток общих классов вводится среди неправильно классифицированных классов для регуляризации весов в DNN. В-третьих, предлагаются модели динамиков на базе DNN, использующие набор функций SDC. Модель с поддержкой динамиков может более эффективно отражать возрастные и гендерные характеристики говорящего, чем модель, представляющая группу говорящих. Более того, новый набор функций T-MFCC используется в качестве входных данных для двухсистемной модели слияния. Первая система представляет собой модель класса, основанную на векторе GNN, а вторая система представляет собой модель динамика, основанную на i-векторе DNN [5]. Использование T-MFCC в качестве входных данных и объединение итоговой оценки с моделью оценки, основанной на векторе DNN, повысило точность классификации.

Ключевые слова: биометрия, верификация диктора, короткие высказывания, вектор, DNN.

Information about authors:

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

Akhmediyarova A.T., PhD, Institute of Information and Computational Technology, Almaty, Kazakhstan; aat.78@mail.ru; https://orcid.org/0000-0003-4439-7313

Kydyrbekova A.S., PhD doctoral student, al-Farabi Kazakh National University, Almaty, Kazakhstan; kas.aizat@mail.ru;

Mekebayev N.O., PhD doctoral student al-Farabi Kazakh National University, Almaty, Kazakhstan; nyrbapakas.aizat@mail.ru; https://orcid.org/0000-0002-9117-4369

Zhumazhanov B., PhD, institute of Information and Computational Technology, Almaty, Kazakhstan; bagasharj@mail.ru; https://orcid.org/0000-0002-5035-9076

REFERENCES

[1] Jain A., Hang L. and Pankanti S. Can Multibiometric Metrics Improve Performance // Hearings Auto ID, 59-64, 1999.

[2] Gorman L. Comparing Passwords, Tokens, and Biometrics for User Authentication // IEEE Proceedings. Vol. 91, N 12, Dec 2003.

[3] Jane A.K., Ross A. and Prabhaker, Introduction to Biometric Recognition // IEEE Transaction in Video Circuits and Systems, Vol. 14, N 1, 4-20, January 2004.

[4] Atal B.S. Automatically Recognizing Speakers by Their Votes // IEEE Proceedings. Vol. 64, N 4, 460-75, April 1976.

[5] Kalimoldayev M.N., Mamyrbayev O.Zh., Kydyrbekova A.S., Mekebayev N.O. Voice verification and identification using i-vector representation // International Journal of Mathematics and Physics 10. N 1, 66 (2019).

[6] Rabiner L. and Jung B.Kh. Fundamentals of Speech Recognition // Pearson Education. 326-396 (1993).

[7] Rosenberg A.E. Automatic Speaker Verification: A Review // IEEE Proceedings. Vol. 64, N 4. P. 475-487. April 1976.

[8] Prasanna S.R.M., Gupta S.S., Yegnanarayan B. Extraction of information about the excitement characteristic of the speaker from the remainder of speech with linear prediction // Speech Communication. Vol. 48, 1243-1261, October 2006.

[9] Mamyrbayev O.Zh., Kydyrbekova AS, Turdalyuly M., Mekebaev NO, Review of User Identification and Authentication Methods by Voice, materials of the scientific conference "Innovative IT and Smart Technologies", 2019. P. 315- 321.

[10] Reynolds D.A. Identification and Verification of Loudspeakers Using Gaussian Mixture Models // Speech Communication. Vol. 17, N 1-2, 91-108, 1995.

[11] Prasanna S.R.M., Gupta S.S., Yegnanarayan B. Extraction of information about the excitement characteristic of the speaker from the remainder of speech with linear prediction // Speech Communication. Vol. 48, 1243-1261, October 2006.

[12] Mamyrbayev O.Zh., Turdalyuly M., Mekebaev N.O., Kydyrbekova A.S. Automatic Recognition of the Speech Using Digital Neural Networks // ACIIDS, Indonesia, Proceedings. Part II, 2019.

[13] Mamyrbayev O., Shayakhmetova A., Kydyrbekova A., Turdalyuly M. Integral speech recognition approach for agglutinative languages // Bulletin AUES. N 1 (48), 2020.

[14] Mamyrbayev O., Toleu A., Tolegen G., Mekebayev N. Neural architectures for gender detection and speaker identification, Cogent Engineering 7 (1), 1727168, 2020.

[15] Mamyrbayev O., Mekebayev N., Turdalyuly M., Oshanova N., Medeni T.I. Voice Identification Using Classification Algorithms // Intelligent System and Computing, 2019.

[16] Kalimoldayev M.N., Mamyrbayev O.Zh., Kydyrbekova A.S., Mekebayev N.O. *Algorithms for Detection Gender Using Neural Networks. ISSN: 1998-4464. 2020. Vol. 14.* P. *154-159.*

[17] Mamyrbayev O.Zh., Othman M., Akhmediyarova A., Kydyrbekova A.S., Zhumazhanov B. Identification and authentication of user voice using DNN features and *i*-vector // Cogent Engineering. Vol. 7, Issue 1, 2020.