Proceedings of the Federated Conference on Computer Science and Information Systems pp. 537–540
DOI: 10.15439/2020F43
ISSN 2300-5963 ACSIS, Vol. 21

Voice authentication based on the Russian-language dataset, MFCC method and the anomaly detection algorithm

Anna Sidorova
MEPHI Cryptology and cybersecurity department
Kashirskoe Sh. 31 Moscow, Russia
Email: saa075@campus.mephi.ru

Konstantin Kogos
MEPHI Cryptology and cybersecurity department
Kashirskoe Sh. 31 Moscow, Russia
Email: kgkogos@mephi.ru

***Abstract*—Almost all people's data is stored on their personal devices. There is a need to protect information from unauthorized access. PIN codes, passwords, tokens can be forgotten, lost, transferred, brute-force attacked. For this reason, biometric authentication is gaining in popularity. Biometric data are unchanged for a long time, different for users, and can be measured. This paper explores voice authentication due to the ease of use of this technology, since obtaining voice characteristics of users doesn't require an equipment in addition to the microphone. The method of voice authentication based on an anomaly detection algorithm has been proposed. The software module for text-independent authentication has been implemented on the Python language. It's based on a new Mozilla's open source voice dataset "Common voice". Experimental results confirmed the high accuracy of authentication by the proposed method.**

I. Introduction

VOICE authentication is a biometric authentication method that uses the user's voice as an identifier. It is based on determining the belonging of a given speech signal to some speaker. The voice of the speaker and then the speech signal entering the authentication system are unique. This causes to an interest in him as a biometric object [1].

Most studies on this topic identifies two main types of voice biometric systems: text-dependent and text-independent.

Text-dependent ones are more often used to control access to the system: during authentication, a certain phrase is pronounced, which is compared with the model of user registered in the system. The vulnerability of such systems is obtaining unauthorized access by recording a passphrase using modern tools of acoustic eavesdropping and providing it to the system.

In text-independent systems, almost any fragment of sounding speech can be used. Their accuracy is less, and the complexity of the implementation is greater. When using this mode, along with the mechanism for verifying statements as two-factor authentication, repetition attacks are almost impossible. For this reason, text-independent voice authentication is of special interest.

II. Related works

During the work, the analysis of existing research in the field of voice authentication has performed.

The authors of the work [3] propose a method of text-dependent user authentication according to the phrase, pronounced by him. To extract features, the MFCC (mel-frequency cepstral coefficients) method is used. To compare records, the distance between two sets of coefficients is calculated. If the distance between the two MFCC sets is less than the specified threshold, the entries are considered the same. The best result in experiments based on dataset, collected by the authors: FAR=16,66%, FRR=0%.

The author of the work [4] has developed a text-independent speaker identification system based on GMM (Gaussian mixture models). To calculate features, the author uses MFCC and inverse MFCC. During the recognition, for each segment of the record, the degree of its similarity with the phonemes of each speaker is calculated. A value of the proximity of the speech segment's feature vector (MFCC) and the phoneme is the average logarithmic probability GMM of the phoneme, calculated for this MFCC-vector. When the decision is made on a closed set of speakers, the nearest neighbors method is used. During experiments on the “TIMIT” and “LibriSpeech" speech corpora, the accuracy of 98% was achieved.

The research [5] also considers MFCC and DMFCC to calculate features, and in addition to this method, vector quantization is used. The study uses the "AN4" dataset, which is publicly available on the Carnegie Mallon University website. The highest accuracy obtained by the authors of the work is 89.20%.

The authors of the work [6] have organized a text-independent authentication system "VoizLock" based on their own voice database. The system can check not only the coincidence of votes, but also that the user says what he should. The authors use LPCC — linear predictive cepstral coefficient — to extract features and HMM — the hidden Markov model — method to build models. The highest accuracy in experiments is 86.25%.

In [7], a text-independent authentication system has implemented. The system compares the result of a user's pronunciation with a previously saved voice profile. If the deviation is below the threshold, authentication passes. For feature extraction, instead of LPCC, the authors preferred MFCC as a feature extraction method. It is justified by the fact that it is more reliable and works better. Also integrated in the system is CMS — the cepstral average subtraction — to subtract the estimated channel noise in the spectrum, and DDMFCC to improve voice recognition accuracy. At the last stage, the Euclidean distance between the MFCC is calculated with the help of the DTW— dynamic time warping algorithm. The best accuracy that was obtained on a self-assembled base is 90%.

The peculiarity of work [8] is that authors propose an approach using interactive voice response (IVR) with speaker recognition based on neural networks. After entering the correct password, the user is prompted to enter his voice instance. Since both factors are applied simultaneously, the probability of authentication errors is reduced. In the work characteristics of speakers are extracted using MFCC, and MLP is used to compare the characteristics. Algorithm — Gradient boosting method. The best result in experiments based on dataset, collected by the authors: FAR=14%, FRR=18%.

The authors of the research [9] identify the problem of reducing recognition performance due to strong background noise. It explores the i-vector (the improvement of GMM) system in noisy conditions. Best accuracy in experiments on voice datasets "Switchboard-1" and "NIST SRE" is 80,54%.

The Russian group of companies "Center for Speech Technologies" and its assistant, author of [10], [11] and other works about voice identification have implemented hybrid GMM–JFA–SVM and GMM–TV– SVM systems using MFCC, DMFCC and DDMFCC for calculating features. It based on the «CST-Microphone». The best result in experiments: FRR = 0,3%, FAR = 10%. In [10], the informativeness of speech features for automatic speaker identification systems was studied. The methods MFCC, LPCC, PLP are considered in details, and it is concluded that combining features always provides the lowest EER.

The paper [12] also describes the GMM, proposes an algorithm based on the k-means algorithm, for estimation GMM parameters (Gauss component numbers, etc.) and an algorithm based on vector quantation for initiating GMM parameters to maximize the likelihood function. Experiments based on the statement proposed by the "Oregon Institute of Science and Technology, Centre of Spoken Language Understanding" show that the system has a high recognition rate: about 94%.

There are few works on voice authentication among the publicly available works. Most of the work relates to the classification of many speakers by voice, which is a different task and usually is solved on the basis of algorithms of classification. These systems determine whose voice from the many speakers the new added record is similar most of all. Thus, for the high accuracy of most such systems you need to have a large number of records of people. In practice, usually one person owns some device, one person is registered in the authentication system, and only the data of the genuine user is available to the system. It is not practical to store large voice bases on a device, even if they are collected and stored safely. The most appropriate solution to the authentication problem is to train on the data of a one user and use the resulting model to identify attackers. The above is known as anomaly detection. This paper decided to first investigate the difference between people by voice characteristics, implement a classification method based on proven methods in most works, achieve high accuracy and then implement a voice authentication based on the method of detecting anomalies.

III. DATASET AND FEATURE EXTRACTION

First of all, the implementation of the voice authentication software module requires a dataset to form training and test samples.

Most of works use English-language datasets: "LibriSpeech", the "TED-LIUM" collection, "AN4"; the "TIMIT" dataset.

The multilingual open base "Common Voice", created by Mozilla Corporation, is of greatest interest. It has been decided to use it. Its advantage is that it has a large volume. It is also suitable for authentication because it contains user IDs and it is as close to reality as possible, because Internet users participate in its compilation, and the recording conditions and sound quality are far from ideal. During the work, the Russian-language part of the base was taken.

Data to implement the classification module needs to be pre-processed. Files have been converted from mp3 format to a more universal wav format. The next stage of the work was the detection of speech activity of records, which is necessary to reduce the volume of uninformative data. Speech activity detectors are used for this purpose. The filter filter_silence from the audiosegment software module of the Python language for this purpose is applied, excluding areas of silence. This filter works on the basis of the publicly available WebRTC VAD project.

Similar pre-processing is done on data to implement the authentication module, except that the training sample is a set of legitimate user records, and the test is a set of records of legitimate user and users, who are considered to be attackers.

As a result of the analysis of existing works it is noted that the most frequently used, promising features are MFCC. The reason for this is the simplicity of their calculation and good approximating ability by taking into account the

different human perception of the tone of sound depending on its frequency. In addition, improvements are implemented for them, such as: using their time derivatives DMFCC and DDMCC, reverse MFCC, normalization.

The main stages of feature extraction are as follows [13]:

Fig. 1 Stages of feature extraction from input speech signal — MFCC

At the stage of preprocessing the input speech signal, high frequencies are amplified, noise is attenuated. Filters are applied: triangular, Gaussian, etc. The speech signal is non-stationary, so framing is done to get stationary frames of the signal. Sometimes the speech phrase signal is divided into frames of length N with an overlap, for example, half. When you overlay windows, a window weight function is applied to each individual frame. The Hamming window is used for this purpose. In the next step, the fast Fourier transform is applied. Each frame of N samples is converted from the time domain, in which the signal representation shows the dependence of the amplitude on time, to the frequency domain. Then the frequency is translated into the mel scale. The last step of the method is a discrete cosine transform. It translates the signal back to the time domain, converting it into kepstral coefficients. Cepstral mel coefficients are calculated in each frame.

The MFCC vector of a single speaker audio record is a 13-dimensional vector whose length depends on the length of the record, since the coefficients are calculated from the frames into which the record is split along its length.

IV. Classification

To solve the classification task it was decided to use the GMM method, it has proven itself in many works as a promising method, for which authors achieve improvements. To make a decision about the authorship the maximum likelihood estimation method is used. A detailed mathematical description is given in the work [11]. Thus, the module for classifying speakers in Python is implemented.

V. Authentication

As mentioned earlier, this paper solves the problem of authentication, it is proposed to focus on a legitimate user and when adding new voice records to the system, that is, when the authentication process begins, to determine whether the voice record is abnormal, that is, does not belong to a legitimate user.

When investigating the applicability of anomaly detection algorithms to the problem of voice authentication, various methods were considered. The best results were obtained by algorithms for detecting anomalies when working directly with MFCC. It is important to note that to do this, it was necessary to average the coefficients over time, since each set of coefficients has a different dimension due to different lengths of audio recordings. Among well-known algorithms, the Local Outlier Factor showed the best accuracy on the same data. A software authentication module has implemented using it. It is shown in figure 3.

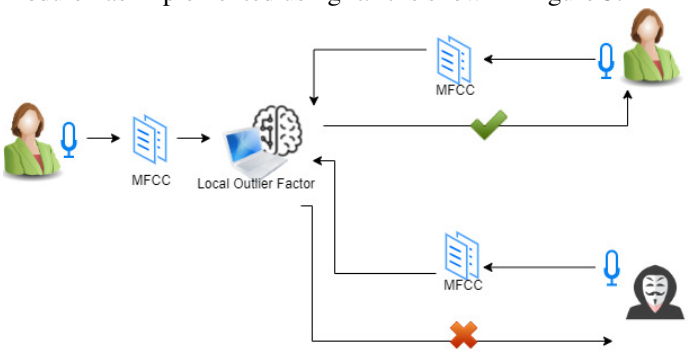

Fig. 3 The proposed scheme of the voice authentication process

To organize the text-independent authentication, it is must to add a speech recognition. This is necessary to check the uttered phrase. The paper considers the existing well-known open mechanisms of speech recognition: DeepSpeech from Mozilla, Kaldi, Vosk, Live Transcribe Speech Engine from Google, etc.

This work the Vosk engine for speech recognition are used due to its ease of use and embedding in the project, and the higher performance of the Kaldi engine that it based on. The re-trained model in Russian is publicly available. Thus, the speech recognition software module is implemented using the above technologies.

VI. Experiment Results

To investigate the possibility of classifying the speakers of the selected dataset by voice characteristics, the classification accuracy was evaluated using the selected methods for features extraction and speaker model construction. The accuracy score was calculated.

The first result obtained was an accuracy of 86%. During experiments the accuracy increases. The best accuracy is equal to 0.95. Table 1 shows the accuracy values when changing the parameters of the algorithms used.

TABLE I.
DEPENDENCE OF ACCURACY ON THE VALUE OF THE PARAMETERS

	Number of gaussians in GMM	The width of the frame in MFCC (FFT)	The value of the threshold, the sound is quieter than which is taken for silence in VAD, %	Accuracy
Set 1	32	1024		0,86
Set 2	16	1024		0,87
Set 3	16	1600		0,88
Set 4	16	1600	6	0,90
Set 5	16	1600	10	0,91
Set 6	16	1600	4; 5	0,95
Set 7	16	1600	1	0,94

After confirming that users are distinguishable by voice characteristics, the software module of voice authentication has implemented. An experimental assessment of authentication accuracy based on another part of the "Common voice" dataset has performed.

As training data and part of the test sample, 66 audio records of a female user who was accepted as legitimate were taken. The test sample also contains 41 records of both male and female attackers for the most reliable experiment.

It should be noted that the number of training sample records was being decreased during the experiment, because it is difficult for the user to record decades of phrases at the registration stage. However, it turned out that high authentication accuracy can be achieved if user records multiple audio records when registering a user, but the minimum recording size is limited. Table 2 shows the results of the experiment.

TABLE II.
RESULTS OF THE EXPERIMENTS, THE ACCURACY

	The test sample includes there are malefact rs of the same gender (the task more difficult).	n_neighbors (Local Outlier Factor)	Training sample size	Limit: audio records over 600kB	Accuracy	FAR	FRR
Set 1	No		6	No	1,00	0,00	0,00
Set 2	Yes	7	66	No	0,98	0,00	0,08
Set 3	Yes	7	50	No	0,95	0,00	0,23
Set 4	Yes	5	7	Yes	0,98	0,00	0,08
Set 5	Yes	5	5	Yes	0,95	0,00	0,23
Set 6	Yes	7	7	Yes	0,78	0,00	0,90

As a result, a software authentication module has obtained. It is clear from the experiment that the value of a type II error is usually higher and in some cases especially large: when the user speaks with a very different intonation, volume, than when registering. For example, the system may not recognize it if the training sample is very small and monotonous. This is not dangerous for the data stored on the device, but it may cause inconvenience for a legitimate user. This problem can be solved by setting the special architecture of authentication service: for example, after a certain time, you can ask the device user to supplement the training sample.

III. CONCLUSION

This paper explores existing approaches to voice authentication of users. The methods of feature extraction from the input signal and training of speaker models, proposed in open works on voice biometrics, are analyzed for the implementation of own system. Open datasets were analyzed and the "Common Voice" base has been selected for the implementation of its own system. It has been decided to first implement the speaker classification module based on the selected Russian-language dataset to make sure that users are distinguishable by voice characteristics, and then to propose and implement an approach based on detecting anomalies. A software module for text-independent identification or classification of speakers has been implemented, and its accuracy is 95%. An approach to voice authentication based on the Local Outlier Factor anomaly detection algorithm, which was not previously used in existing open research, is proposed. During the experiments, the accuracy of 98% was obtained. In the future, it would be interesting to try to implement the developed system in some real application and test it in real conditions.

REFERENCES

[1] Bernstein S.I., Kolokoltsev N.K., Ermolaeva V.V. Voice Authantication. Molodoy uchenyy[Young scientist], 2018, no.25. pp. 93-94. Available at: https://moluch.ru/archive/211/51686/ (accessed 24 April 2019)

[2] Ermilov A.V. Metody, algoritmy i programmy resheniya zadach identifikatsii yazyka i diktora[Methods, algorithms and programs for solving problems of language and speaker identification]: Extended abstract of PhD dissertation (physics and mathematics), 2014, 22 p. (in Russian)

[3] Ivanov D.A., Nikitin A.P. Text-dependent voice authentication method. Istoriya I arkhivy [History and archives], 2016, no. 3 (5). (in Russian)

[4] Zakharova V.V. Razrabotka tekstonezavisimoy sistemy identifikatsii diktora na osnove fonemnogo razbieniya i GMM [Development of a text-independent speaker identification system based on phonemic partitioning and GMM]. Proceedings of the Science Conference of undergraduate and graduate students, Belorusskiy gosudarstvennyy universitet, Minsk, 2017, pp. 28-32. (in Russian)

[5] Nogikh A.A., Solomatin D.I. Text-independent authentication. Sbornik studencheskikh nauchnykh rabot fakul'teta komp'yuternykh nauk VGU [Collection of student research papers of the faculty of computer science of VSU], 2016. pp. 117-123. (in Russian)

[6] Jayamaha R. G. Voizlock-human voice authentication system using hidden markov model. Proceedings of the 4th International Conference on Information and Automation for Sustainability. IEEE, 2008. pp. 330-335.

[7] Yan Z., Zhao S. A Usable Authentication System based on Personal Voice Challenge. Proceedings of the International Conference on Advanced Cloud and Big Data, 2016. pp. 194-199. DOI:10.1109

[8] Shah, S. A. A., Shah S.W., A. ul Asar. Interactive Voice Response with Pattern Recognition Based on Artificial Neural Network Approach. NWFP University of Engineering and Technology, Peshawar, Pakistan, 2007. pp. 249-252.

[9] J. Chang, D. Wang. Robust speaker recognition based on DNN/i-vectors and speech separation. Proceedings of the Acoustics, Speech and Signal Processing (ICASSP) IEEE International Conference, 2017. pp. 5415-5419.

[10] Matveev Yu.N. Research of information content of speech signs for automatic speaker identification systems. Vestn. MGTU im. N. E. Baumana. Ser. Priborostroenie. Spetsial'nyy vypusk. Biometricheskie tekhnologii [Bulletin of the Bauman Moscow state technical University. Series "Instrument making"], 2013. no. 2. pp. 47—51. (in Russian)

[11] Matveev Yu.N. Technologies for biometric identification of an individual by voice and other modalities. Vestn. MGTU im. N. E. Baumana. Ser. Priborostroenie. Spetsial'nyy vypusk. Biometricheskie tekhnologii [Bulletin of the Bauman Moscow state technical University. Series "Instrument making". Special issue. "Biometric technology".], 2012. № 3(3). pp. 46—61. (in Russian)

[12] Sadykhov R.Kh., Rakush V.V. Models of Gaussian mixtures for speaker verification based on arbitrary speech. Doklady BGUIR [BSUIR reports], Minsk, 2003. no. 4 pp. 95-103. (in Russian)

[13] Hundal J.K, Dr. Hamde S. T. Some Feature Extraction Techniques for Voice based Authentication System. Proceedings of the Power, Control, Signals and Instrumentation Engineering (ICPCSI) IEEE International Conference, 2017. pp. 419-421.

[14] Documentation of python-speech-features. Available at: https://python-speech-features.readthedocs.io/en/latest/#welcome-to-python-speech-features-s-documentation (accessed: 20 March 2019).