On Voice Autentication Algorithm Development

Patryk Bąkowski[0000−0001−8325−1873]
and Dmitry Muromtsev[0000−0002−0644−9242]

ITMO University, Saint Petersburg, 197101, Russian Federation
{baski, mouromtsev}@ifmo.ru
https://en.itmo.ru/

Abstract. This article is devoted to the development of voice authentication algorithm for access control in automated control systems. The existing methods of allocation of individual voice characteristics and construction of voice models are considered. The algorithm of voice authentication for voice control of the Internet of Things system in Russian on the basis of a neural network is offered. The peculiarity of the algorithm is the use of mel-frequency cepstral coefficients and the text independence of the voice message. Experiments aimed at identifying the optimal set of analysed parameters and evaluating the efficiency of the classifier and the authentication system as a whole are described.

1 Problem statement

In today's world, it is necessary to protect multiple sources of sensitive data, both in the industrial environment and in everyday life. Among the many means of ensuring the security of such data, biometric voice authentication systems have a number of advantages. In many situations where it is impossible to get a high-quality image of the user's face or get fingerprints, voice authentication will successfully cope with its task. As a result, such systems are implemented and used in many areas, such as forensics, finance, telecommunications.

The use of voice for speaker recognition tasks has a great potential, in particular due to the fact that to solve such problems there is no need to purchase complex and expensive equipment, it is enough to have a microphone. Such identification and authentication systems can be easily implemented and used both in ACMS (access control and management systems) and on telephone lines and mobile devices. Voice interfaces are the most promising means of interaction with "smart things" due to the naturalness and intuitiveness of this approach for a person: often voice assistants become control centers of smart homes.

At the moment two of the most relevant problems in the field of the Internet of Things with voice interfaces can be formulated:

1. Management of personal data of users.
2. Management of complex systems, such as smart home, smart factory, etc.

In accordance with this, it is necessary to develop an access control system that includes authentication by voice characteristics without the use of code words or phrases.

2 Development of voice authentication algorithm

The system of automatic user authentication by voice is being developed within the voice control system of Internet of things by PICT faculty of the ITMO University. The purpose of this work is to create the voice control system that can, on the basis of voice commands and data obtained from ontological descriptions of devices and indications of smart sensors, logically derive and generate scenarios for the smart (automated) system, for example, the smart home or the smart classroom. Voice control is based on local speech recognition.

This approach unlike cloud-based speech recognition has a number of advantages:

1. No issues related to availability, bandwidth and other factors that affect the speed of recognition inherent in cloud solutions.
2. Unlike cloud systems, there is an opportunity to configure the speech recognition system to solve a specific task. The quality of recognition depends on the language model used. In different application areas, different words have different probabilities. Cloud solutions use standard systems that use an average model of the language, or a model designed to solve the problems posed to the creators of the platform, which do not always coincide with the user tasks of the system.
3. Resource-efficient implementation of voice activation. To implement the activation function using a cloud system it is necessary to broadcast everything that the microphone records to the cloud in order to detect the passphrase. This leads to additional loading of the transmission channel and the cost of Internet traffic.
4. No additional financial costs. There are many open source (free) libraries and tools for local speech recognition, while cloud solutions are commercial and provide paid access to their services.

The following steps are implemented to solve the research task:

1. Data collection.
2. Preprocessing of the voice recordings.
3. Extraction of vectors of individual vocal signs.
4. Building the voice model based on voice characteristics.
5. Decision-making and verification.

The data is collected by the software for working with devices that record audio signals (microphones).

The VAD (Voice Activity Detection) algorithm based on the energy [8] is used to pre-process recorded audio data, namely to remove pauses and non-vocalized fragments. This algorithm splits the speech signal into frames of 40 MS, then

removes those frames which average energy is less than the set threshold: the average energy of the entire record, multiplied by a factor k, that is selected empirically.

IF $E_i < k * E$, where $k < 1$,	Silence
ELSE	Voice activity

The k coefficient in this work was 0.25. Figure 1 shows the signal before and after removing noise and pauses.

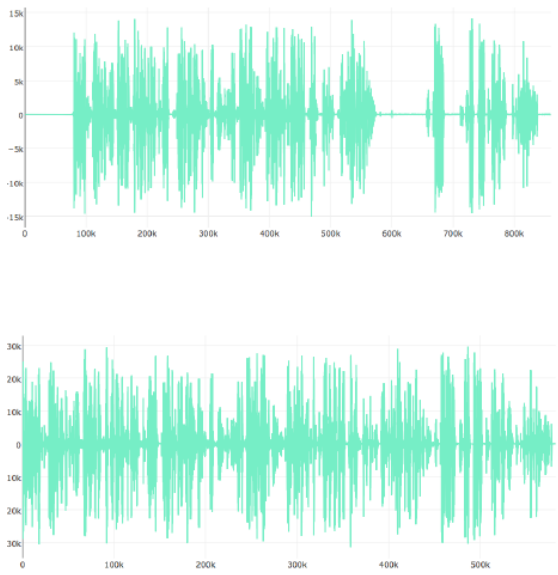

Fig. 1. The signal before and after removing noise and pauses

Features are extracted from the pre-processed voice records. Feature extraction is implemented using the bob.ap library.

As a result 60-element vectors of mel-frequency kepstral coefficients (MFCC) [7] are formed. The process of obtaining vectors is as follows:

1. Splitting the signal into overlapping frames of length 20 ms with an intersection of 10 ms.
2. Obtaining the signal spectrum for each frame by applying Fourier transformation.
3. Decomposition of the spectrum on the mel-scale using triangular filters.
4. Squaring of the obtained values and taking the logarithm.
5. Application of the discrete cosine transformation.

In feature vector-based recognition, the Gaussian mixture model (GMM) [7] or machine learning, such as the SVM support vector method, are most commonly used. In this work, the multilayer neural network [4][5][6] with two hidden layers was used to recognize the speaker by voice. The number of neurons in the input layer $i_1, i_2, \ldots, i_n$, is defined by the dimension of the feature vectors on which learning occurs. In this paper, vectors of dimension n=*60* are used. The number of neurons in the output layer of the network $o_1, o_2, \ldots, o_k$ corresponds to the dimension k of the set of speakers G registered in the system. The architecture of the used neural network is shown in figure 2. Tensorflow [2] and Keras [1] libraries were used to work with neural networks.

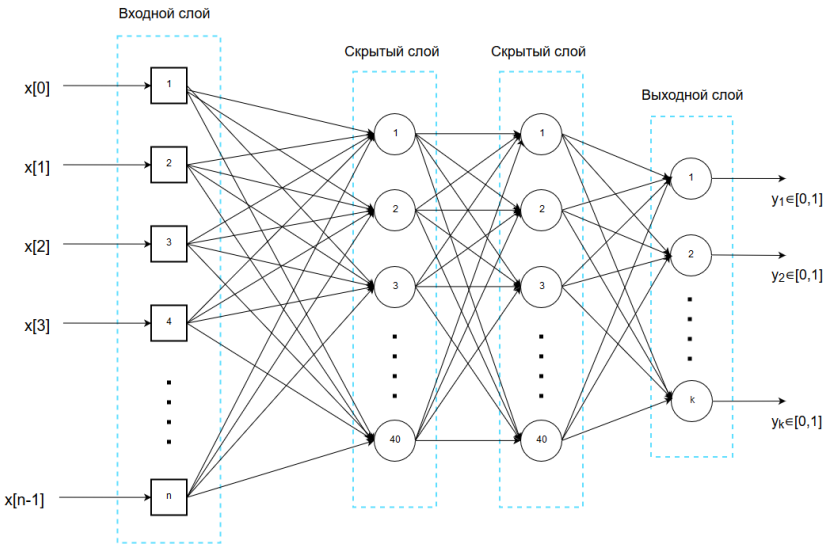

Fig. 2. Architecture of the neural network

Voxforge dataset [3] was used to train the neural network. It contains recordings of various lengths from 500 speakers. During the training of the network and the analysis of its accuracy in solving the problem of speaker classification, the outputs of the last layer of the network $h_\theta(x(m))$ is a K-dimensional vector, where K is the number of speakers, each element of which takes values in the range from 0 to 1. The vector shows with what probability the speaker can be attributed to each of the K classes. The prediction of the speaker class can be carried out using the sum of the logarithms of the probability of M frames. In this case the ID of the predicted speaker $k*$ is the index of the maximum probability value:

$$k^* = \underset{k \in [1,K]}{\arg\max} \left(\sum_{m=1}^{M} log(h_\theta(x^m)k) \right)$$

Further, in the verification and decision-making step, the user is authenticated by comparing the received probability with the threshold value. The threshold value was determined by conducting experiments with speakers who did not take part of the speaker set G formation, known to the classifier (negative experiments), and speaker form the set G (positive experiments). From the obtained values of identification probability for negative and positive experiments, two Gaussian distributions - correct and erroneous identification - were constructed for each user. The intersection point of these graphs of these distributions is the threshold value for a particular user of the system (Fig.3).

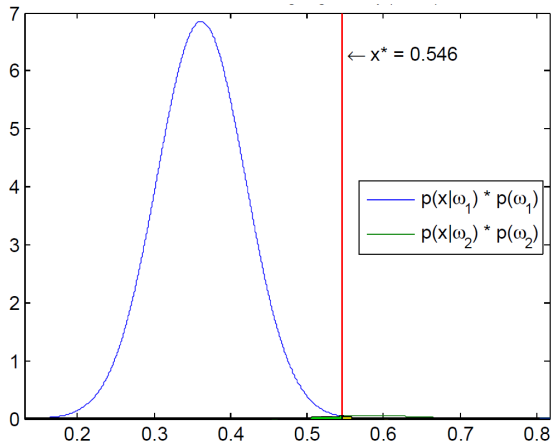

Fig. 3. Determination of the threshold value by the intersection point of Gauss distributions

During the development of the voice authentication system, a number of experiments were conducted to identify the value of the VAD threshold coefficient. The choice of threshold factor is very important when working with energy-based VAD: too high value can cause cutting off frames that contain the speaker's voice and, in turn, too low value can cause many non-vocalized or noisy fragments would not be excluded from the set of frames. Figure 4 shows the dependence of the classifier accuracy on the VAD threshold coefficient. Using VAD with the right threshold value can improve system performance by about 10% compared to raw data.

When testing the developed algorithm of voice authentication results with the following values (shown in figure 5): equal error rate EER = 7%, the coefficient of accurate verification 87.1%.

3 Conclusion

This paper describes the analysis of methods of biometric authentication of the speaker by voice. The existing algorithms and methods of biometric authentica-

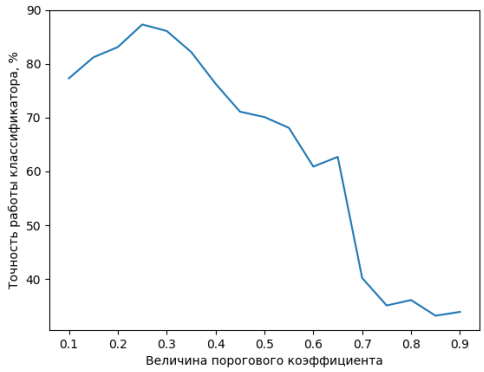

Fig. 4. The dependence of classification accuracy on VAD threshold value

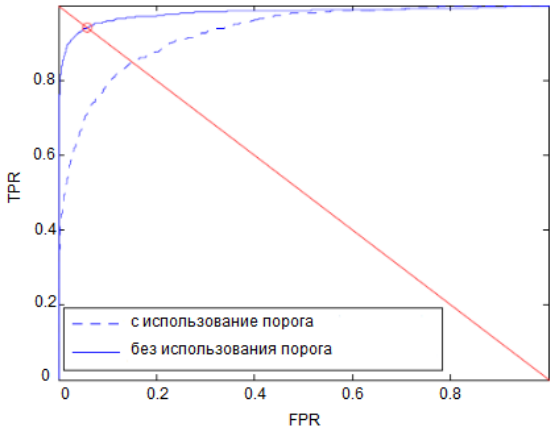

Fig. 5. The dependence of classification accuracy on VAD threshold value

tion by voice, including text-dependent and text-independent algorithms are investigated. The analysis of the tools used in this field is also carried out, and the neural network with two hidden layers and the distribution of neurons 60:40:40:20 is selected. The result is a voice authentication algorithm for access control in automated systems. In addition, specialized software for voice biometric authentication system based on neural networks has been developed. The number of experiments on the selection of the VAD threshold coefficient was carried out. Such metrics as accuracy (87%) and EER (7.1%) were used to evaluate the system. Possible directions of development of this research are formulated.

References

1. Keras. the python deep learning library, https://keras.io/
2. Tensorflow. an end-to-end open source machine learning platform, https://www.tensorflow.org/
3. Voxforge - free speech corpus and acoustic model repository, http://www.voxforge.org/ru/Downloads
4. Buchneva, T., Kudryashov, M.Y.: Neural network in the task of speaker identification by voice. herald of tver state university. series. Applied Mathematics (2), 119–126 (2015)
5. Ge, Z., Iyer, A.N., Cheluvaraja, S., Sundaram, R., Ganapathiraju, A.: Neural network based speaker classification and verification systems with enhanced features. In: 2017 Intelligent Systems Conference (IntelliSys). pp. 1089–1094. IEEE (2017)
6. McLaren, M., Lei, Y., Ferrer, L.: Advances in deep neural network approaches to speaker recognition. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 4814–4818. IEEE (2015)
7. Rakhmanenko, I.A., Mescheriakov, R.V.: Identification features analysis in speech data using gmm-ubm speaker verification system. Trudy SPIIRAN **52**, 32–50 (2017)
8. Verteletskaya, E., Sakhnov, K.: Voice activity detection for speech enhancement applications. Acta Polytechnica **50**(4) (2010)