Voice User Identification in Access Control Systems

Menshakou P.A,
Murashko I.A.
Gomel State Technical University
named after P.O. Suhoy
Gomel, Belarus
Email: pmenshakov@gmail.com,
Email: iamurashko@tut.by

***Abstract*—In view of the constant development of any business, we have a lot of data that must be protected. At the moment, the access control checkpoints are equipped with various types of access control. But most of the access control devices have a high price. Moreover, a large part of the costs fall on the allocation of personal identification for each user agent. The solution to this problem is voice recognition. The use of biometrics eliminates the chips and access cards. Avoid loss of means of identification and theft. And the use of the voice will eliminate the expensive equipment to read the data.**

Keywords*—*Voice, identification, neural networks.

I. Introduction

Currently, voice recognition and biometrics as a whole, already widespread. The simplest example - fingerprint scanners installed on almost every laptop.

"Biometrics" assumes people recognition system on one or more physical or behavioral traits. In the field of information technology, biometric data is used as a form of identifiers of access control. Also, biometric analysis is used to identify people who are under the supervision[1]. Just biometrics provides the behavioral analysis of the object. These include walking, gestures, etc.

Authorization process using biometrics is quite simple. Using an apparatus for varying the characteristics identified by the current data are scanned and compared with previous data.

Biometric systems have a number of important advantages: biometrics uses characteristics of the human body and its behavior, which makes the data unique (give someone else a fingerprint or make the iris of your eye like a someone else's It requires quite rare and sophisticated technologies); Unlike paper-based IDs (passport, driving license, identity card), by a password or personal identification number (PIN), biometric characteristics can not be subjected to theft, can not be lost or forgotten. Quite a long time, fingerprints are used to identify criminals and prevent theft or fraud. Some people are able to imitate the voice, but it requires special skills that are not often met in everyday life.

The latter include voice, gait, gestures, handwriting, etc. Until recently, little behavioral characteristics in identification systems used in connection with the obvious drawbacks. Over time, the human gait can vary. The voice may change as a result of the disease, age-related changes or the environment (for example, high noise level). Currently, however, with the advent of effective methods of digital signal processing interest in this subject has increased significantly in the world.

We propose to use voice recognition to control access to devices on the premises. The advantage of this solution is the simplicity of the hardware and software implementation. For voice print only requires a microphone and an analog-digital converter. These devices are equipped with almost all modern desktop and mobile computers.

Task voice recognition or speaker recognition voice boils down to, to identify, classify and appropriately respond to human speech from the input audio stream. It is usually divided into three sub-tasks: receiving a voice print, identification and verification.

Receiving voice imprint - the process of obtaining the sample, representing the vector characteristics of the speaker's voice. Identification - the process of determining the identity modeled voices by comparing the sample template files saved in the database.

Verification - a process in which by comparing the submitted sample template stored in the database the requested identity is verified. The result is a proof of identity or a negative response system. Implementation of these procedures takes quite a long time, so difficult simultaneous identification of several persons.

The aim is to reduce the time of receipt of voice prints, as well as a decrease in the time of identification and verification of the voice print. To achieve this goal is proposed to use the Kohonen self-organizing map (SOM - Self-Organized Map), where the processing speed has been increased through the provision of neurons with maximum activity, while receiving minimal loss of accuracy.

II. Voice recognition principle

To implement voice recognition is necessary to make a specific course of action.

With microphone voice recording turns out to be identified and sent to the computer. It is to obtain the optimal WAV file, since ease of handling.

The resulting voice recording must be divided into frames. Separation of the footage shown in Figure 1.

This action is necessary for easier work with the recorded soundtrack. Further, all calculations will be made with each frame individually.

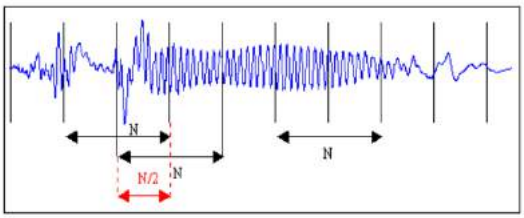

Figure 1. Audio wave

The next step is to remove noise and unwanted effects. This is necessary in order to record obtained at various times correspond to each other, regardless of external factors. There are many ways by which you can reduce the noise effects. I used the multiplication of each frame on the weighty special feature – "Hamming window":

$$\omega(n) = 0.53836 - 0.46164 * cos(\frac{2\pi n}{N-1}) \quad (1)$$

where n – ordinal number of the element in the frame for which the new value of the amplitude is calculated,

N – frame length (number of signal values measured for the period).

The resulting footage is converted to its frequency response using a sweep through the "Fast Fourier Transformation":

$$X_k = \sum_{i=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad (2)$$

where N – frame length (number of signal values measured for the period)

X_n – the amplitude of the n-th signal

$X_k - N$ – complex amplitudes of the sinusoidal signals composing the original signal.

Today the most successful voice recognition system using knowledge about the hearing aid. They are based on the fact that the ear interprets sounds not linearly but in a logarithmic scale. In view of these features is necessary to bring the frequency response for each frame "mels". The dependence is shown in Figure 2.

To go to the "mel" characteristics following relationship is used:

$$m = 2595 log_{10}(1 + \frac{f}{700}) = 1127 log_e(1 + \frac{f}{700}) \quad (3)$$

where m – frequency Melachim, F – frequency in hertz.

This last action is required for the subsequent conversion to vector characteristics which, subsequently, is compared with the database of voice records. The vector will comprise mel-cepstral coefficients, which can be obtained by the following formula:

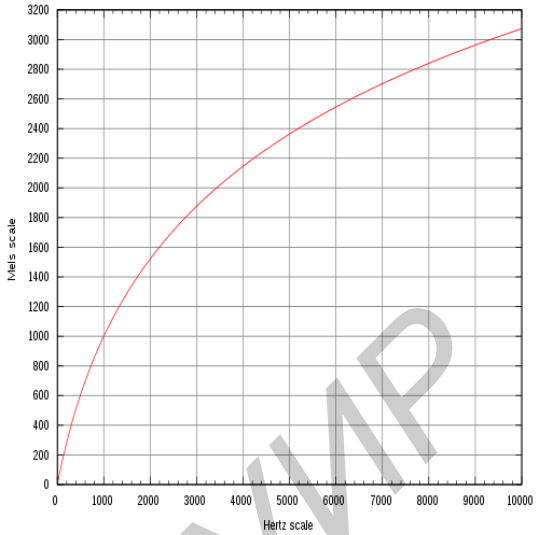

Figure 2. Audio wave

$$c_n = \sum_{k=1}^{K} (logS_k)[n(k - \frac{1}{2})\frac{\pi}{K}] \quad (4)$$

where c_n – mel-cepstrum coefficient at number n,

S_k — amplitude of k item in mel frame,

K – prescribed number of mel-cepstral coefficients $n\varepsilon[1, K]$.

The resulting feature vector is added to the database for comparison with it.

It is recommended to use multiple entries of the same voice. Then, the resulting vector is presented as the arithmetic mean vectors characterizing individual frames of speech. To improve recognition accuracy just need to average the results not only between shots, but also take into account the performance of several speech samples. With a few voice recordings, it is reasonable not to average performance to one vector, and spend clustering, for example using the method of k-means.

III. Neural network comparison

The self-organizing map (SOM) (Kohonen, 1982) is one of the most important neural network architecture. Since its invention it has been applied to so many areas of Science and Engineering that it is virtually impossible to list all the applications available to date (van Hulle, 2010; Yin, 2008). In most of these applications, such as image compression (Amerijckx et al., 1998), time series prediction (Guillen et al., 2010; Lendasse et al., 2002), control systems (Cho et al., 2006; Barreto and Araújo, 2004), novelty detection (Frota et al., 2007), speech recognition and modeling (Gas et al., 2005), robotics (Barreto et al., 2003) and bioinformatics (Martin et al., 2008), the SOM is designed to be used by systems whose computational resources (e.g. memory space and CPU speed) are fully available. However, in applications where such resources are limited (e.g. embedded software systems, such as mobile phones), the SOM is rarely used, especially due

to the cost of the best-matching unit (BMU) search (Sagheer et al., 2006). Essentially, the process of developing automatic speech recognition (ASR) systems is a challenging tasks due to many factors, such as variability of speaker accents, level of background noise, and large quantity of phonemes or words to deal with, voice coding and parameterization, among others. Concerning the development of ASR applications to mobile phones, to all the aforementioned problems, others are added, such as battery consumption requirements and low microphone quality. Despite those difficulties, with the significant growth of the information processing capacity of mobile phones, they are being used to perform tasks previously carried out only on personal computers. However, the standard user interface still limits their usability, since conventional keyboards are becoming smaller and smaller. A natural way to handle this new demand of embedded applications is through speech/voice commands. Since the neural phonetic typewriter (Kohonen, 1988), the SOM has been used in a standalone fashion for speech coding and recognition (see Kohonen, 2001, pp. 360-362). Hybrid architectures, such as SOM with MultiLayer Perceptrons (SOM-MLP) and SOM with Hidden Markov Models (SOM-HMM), have also been proposed (Gas et al., 2005; Somervuo, 2000). More specifically, studies involving speech recognition in mobile devices systems include those by Olsen et al. (2008); Alhonen et al. (2007) and Varga and Kiss (2008). It is worth noticing that Portuguese is the eighth, perhaps, the seventh most spoken language worldwide and the third among the Western countries, after English and Spanish. Despite that, few automatic speech recognition (ASR) systems, specially commercially available ones, have beendevelopedand it isavailable worldwidefor the Portugueselanguage. This scenario is particularly true for the Brazilian variant of the Portuguese language, due its large amount of accent variation within the country. Scanzio et al. (2010), for example, report experiments with a neural network based speech recognition system and include tests with the Brazilian Portuguese language. Their work is focused on a hardware-oriented implementation of the MLP network. In this context, the current paper addresses the application of self-organizing maps to the Brazilian Portuguese isolated spoken word recognition in embedded systems. For this purpose, we are particularly interested in evaluating several software strategies to speedup SOM computations in order to foster its use in real-time applications. The end-user application is a speaker-independent voice-driven software calculator which is embedded in smartphones.

We used learning without a teacher, because it is much more plausible model of learning in the biological system. Kohonen developed and many others, it does not need to output the target vector and therefore, does not require comparison with predetermined ideal responses, and learning set consists only of the input vectors. The training algorithm adjusts network weights so as to produce consistent output vectors, ie, to sufficiently close the presentation of input vectors produce the same outputs. The learning process, therefore, highlights the statistical properties of the training set and groups similar vectors in the classes. Presentation of the input vector of this class will give a certain output vector. The spread signal in such a network is as follows: input vector is normalized to 1.0 and applied to the input, which distributes it on through the matrix of weights W. Each neuron in layer Kohonen calculates the sum at its input and depending on the condition of the surrounding neurons becoming active layer or inactive (1.0 and 0.0). Neurons in this layer operate on the principle of competition, ie. E. As a result of a certain number of iterations is still an active one neuron or a small group. This mechanism is called lateral. Since testing of this mechanism requires significant computing resources, in my model it replaced by finding the maximum neuron activity and awarding him the activity 1.0, and 0.0 all other neurons. Thus, the neuron is activated for which the input vector closest to the vector of the weights. As a sigmoid activation function is used, which is as follows:

$$f(x) = 1/(1 + e^{-a*x}) \tag{5}$$

where a – slope parameter.

Geometrically, this rule shows next picture:

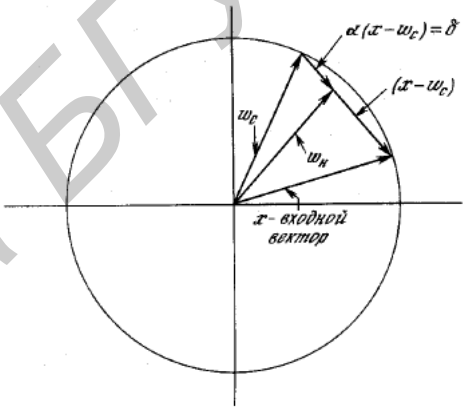

Figure 3. Correction weights of Kohonen neuron

Since the input vector x is normalized, ie. E. Is on a hypersphere of unit radius in the space of weights, then the correction weights on this rule is rotated vector weights toward the input that allows to produce statistical averaging of input vectors, which reacts active neuron. Thus, the study was replaced lateral approach leading to the activation of neurons.

IV. RESULTS

The effectiveness of the proposed approach to the identification of the employee's voice, implemented in the access control system in the room, was estimated on the basis of techniques proposed in [20]. word composed of figures were recorded for the experiment. The data set includes the voice data of 14 speakers. 30 words (10 different words, on 3 samples each) were recorded for each speaker. Table 1 shows the selected words.

Table I. THE WORDS, TAKEN FOR THE EXPERIMENT

One	Two	Three	Four	Five
Six	Seven	Eight	Nine	Zero

Just as in the study described in the article, all the words have been written to the indoors, and used as a source of noise conditioner. Attracted by the speakers (11 men and 3 women) spoke freely, while maintaining their respective accents and

pronunciation defects. It was necessary for the complexity of classification problems, since even the same statements have different durations after detecting the endpoint. Were used for simulation and SOM TS-SOM, sluduet configuration having 10 inputs and 256 neurons located in the 16 * 16 array.

Analysis of results shows that the resulting modification has improved the effectiveness of the implementation is two times as compared to SOM: PDS, a fall of 3 per cent accuracy.

V. Conclusion

The result of this study was to modular application performing voice user authentication. The program consists of three main parts. The first carries the addition of users, the second and the third carries the identification sends information to identify the user.

References

[1] ГОСТ Р 51241-2008 Средства и системы контроля и управления доступом. Классификация. Общие технические требования. Методы испытаний. – М.: Стандартинформ, 2009.

[2] Universal Developing of Persons Identification Based on RFID / G.I. Raho [et al.] // Journal of Emerging Trends in Computing and Information Sciences. – 2015. – Vol. 6, №. 10. – P. 592-597.

[3] Adeyemo, Z.K. Development of Hybrid Radio Frequency Identification and Biometric Security Attendance System / Z.K Adeyemo, O. J. Oyeyemi, I.A Akanbi // International Journal of Applied Science and Technology. – 2014. – Vol. 4, No. 5. – P. 190-197.

[4] Гарафутдинова, Ф.М. Истоки дактилоскопии / Ф.М. Гарафутдинова // Публичное и частное право. –2014. – №II (XXII). – С.173.

[5] Biometrics, Computer Security Systems and Artificial Intelligence Applications / Ed. K. Saeed, J. Pejas, R. Mosdorf. – Springer, 2006. – 345p.

[6] You, Y. Audio Coding: Theory and Applications / Y. You. – NY: Springer, 2010. – 349 p.

[7] Herbig, T. Self-Learning Speaker Identification: A System for Enhanced Speech Recognition / T. Herbig, F. Gerl, W. Minker. – Berlin:Springer Verlag Gmbh, 2011.

[8] Al-Shayea Q.K., Speaker Identification: A Novel Fusion Samples Approach / Qeethara Kadhim Al-Shayea, Muzhir Shaban Al-Ani // International Journal of Computer Science and Information Security (IJCSIS). – 2016. – Vol. 14, № 7. – P.423-427.

[9] Ding Ing-Jr., Chih-Ta Yen, Yen-Ming Hsu. Developments of Machine Learning Schemes for Dynamic Time-Wrapping- Based Speech Recognition/ Ing-Jr. Ding, Chih-Ta Yen, Yen-Ming Hsu // Mathematical Problems in Engineering. – 2013. – P. 5.

[10] Kohonen, T. Self-Organizing Maps / T. Kohonen. – Berlin: Springer, 2001.

[11] Bosi, M. Introduction to digital audio coding and standards / M. Bosi, R.E. Goldberg // Springer Science+Business, Media USA. – 2010. – 434 p.

[12] Меньшаков, П.А. Методика голосовой идентификации на основе нейронных сетей / П.А. Меньшаков, И.А. Мурашко // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2016) : материалы VI междунар. науч.-техн. конф. (Минск, 18-20 февраля 2016 года) / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2016. – С. 411-414.

[13] Сергиенко, А. Б. Цифровая обработка сигналов. – 2-е издание. – СПб: Питер, 2006. –751 с.

[14] Harris, F.J. On the use of windows for harmonic analysis with the discrete Fourier transforms / F.J. Harris// Proceedings of the IEEE. – Vol. 66, Jan 1978. – P. 51-83.

[15] Азаров, И.С. Применение мгновенного гармонического анализа для антропоморфической обработки речевых сигналов/ И.С. Азаров, А.А. Петровский // Вестник МГТУ им. Н.Э. Баумана. Сер. “Приборостроение”. – 2012. – С. 51.

[16] Using self-organizing maps to cluster music files based on lyrics and audio features / Research Congress 2013 De La Salle University, Manila, March 7-9, 2013.

[17] Ghitza, O. Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition / O. Ghitza // IEEE Transactions on Speech and Audio Processing. – 1994. – Vol. 2, № 1. – P. 115–132.

[18] Азаров, И.С. Применение мгновенного гармонического анализа для антропоморфической обработки речевых сигналов/ И.С. Азаров, А.А. Петровский // Доклады БГУИР. – 2011. – № 4. – С. 59–70.

[19] Mwasiagi, I. Self Organizing Maps – Applications and Novel Algorithm Design / Josphat Igadwa Mwasiagi// InTech. – January 21, 2011. – P. 91.

[20] Alejandro, C. Analysis of Kohonen’s Neural Network with application to speech recognition / C. Alejandro // Mexican International Conference on Artificial Intelligence. – Mexico: Guanajuato, 2009. – P. 8.

[21] Koikkalainen, P. Self-Organizing hierarchical feature maps/ P. Koikkalainen, E. Oja// Proceedings of 1990 International Joint Conference on Neural Networks vol. II. – 1990. – San Diego, pp. 279–284.

[22] Загуменнов, А. П. Компьютерная обработка звука./ А. П. Загуменнов - М.: ДМК, 1999. - 384 с

ГОЛОСОВАЯ ИДЕНТИФИКАЦИЯ ПОЛЬЗОВАТЕЛЯ В СИСТЕМАХ КОНТРОЛЯ ДОСТУПА

Меньшаков П.А., Мурашко И.А.

Задача голосовой идентификации или распознавания диктора по голосу сводится к тому, чтобы выделить, классифицировать и соответствующим образом отреагировать на человеческую речь из входного звукового потока. При этом обычно выделяют три подзадачи: получение голосового отпечатка, идентификация и верификация. Используя самоорганизующуюся карту Кохонена, с небольшими модификациями, вместо простых алгоритмов сравнения, можно значительно ускорить данные операции.