The Method of User Identification by Speech Signal

Vyacheslav Nykytyuk[a], Vasil Dozorskyi[a], Oksana Dozorska[a], Andrii Karnaukhov[a] and Liubomyr Matiichuk[a]

[a] *Ternopil Ivan Puluj National Technical University, Ruska str., 56, Ternopil, 46001, Ukraine*

Abstract

The paper proposes a method of four-parameter identification of users by speech signal, which is based on a mathematical model of speech signals in the form of a piecewise stationary random process. Based on the application of the proposed method, three individual parameters of the speech signal are evaluated, such as the value of the main tone frequency, the time durations of the segments of the speech signal that correspond to the vowel and consonant sounds, their alternation in the speech signal - password. The value of the threshold function is used as the fourth parameter. Such time durations are actually determined on the basis of this function, and its value can be tied by a certain dependence to the value of the main tone frequency and the duration of the recording of word - password. This will practically make false identification of a third party impossible and increase the reliability of the identification itself.

Keywords 1

Identification, speech signal, processing method, sliding window, piecewise stationary random process

1. Introduction

The paper considers the method of biometric identification of the user by speech signal. The task of such identification is particularly relevant today in the field of IT technologies [1], in particular for controlling the provision of access to information resources, databases or individual services only to certain users.

Traditional identification methods such as username and password, knowledge-based identification, and SMS-based two-factor identification have disadvantages due to security vulnerabilities ranging from account hijacking and phishing to social engineering. Accordingly, IT departments today are researching and developing more reliable identification systems that reduce the likelihood of theft and fraud [1, 2].

Among the analyzed methods, biometric identification methods are considered the most reliable, since the identifier itself is often an individual part of the user, which practically cannot be used without his knowledge by third parties or forged. Biometric identification refers to security processes that confirm a user's identity using unique biological signs such as retina, iris, voice, facial characteristics, and fingerprints, etc. [3-7]. Biometric identification systems store this biometric data to verify a user's identity when that user accesses their account. Because these data are unique to individual users, biometric identification is generally more secure than traditional forms of multifactor identification.

A comparison of common biometric identification methods was also carried out, in particular on the errors of false identification and the cost of the technical implementation of the method. Summary data are shown in Table 1.

ITTAP'2022: 2nd International Workshop on Information Technologies: Theoretical and Applied Problems, November 22–24, 2022, Ternopil, Ukraine

EMAIL: slavikvv89@gmail.com (A. 1); vasildozorskij1985@gmail.com (A. 2); oksana4elka@gmail.com (A. 3); angryfallenangel@gmail.com (A. 4); mlpstat@gmail.com (A. 5)

ORCID: 0000-0003-1547-8042 (A. 1); 0000-0001-6744-3015 (A. 2); 0000-0001-7053-863X (A. 3); 0000-0003-3042-3066 (A. 4); 0000-0001-6701-4683 (A. 5)

Analyzing the data in Table 1, we can come to the conclusion that the method of identification based on the characteristics and features of speech is optimal, as it does not require expensive equipment and has a low value of false identification of a person [8, 9].

Table 1

Basic characteristics of common biometric identification methods

The method of obtaining biometric parameters	Probability of false identification, %	The cost of technical implementation, USD
Geometric structure of the hand	0,2…1	600-3000
Fingerprints	0,0001	60-600
Peculiarities of the retina pattern	6…10	4000
The iris of the eye	0,0001	500-6000
Face portrait	-	55000
Handwriting	0,5…5	-
Keyboard and computer handwriting	3…9	-
Speech characteristics and features	0,5…5	1-60

However, with the development of computer programs of speech signals generation, common speech identification algorithms become vulnerable, which requires the development of better, more efficient and reliable speech identification algorithms based on the analysis of several different parameters of speech signals. For this, it is necessary to carry out mathematical modeling of speech signals, which consists in choosing a mathematical model of such signals that is adequate for the task of biometric identification of the user and developing methods of their processing.

2. Choice of speech signals mathematical model

The mathematical model of speech signals should take into account the nature of their formation and open up new opportunities in the field of user identification by applying new methods of speech signals processing and obtaining new informative features from them.

It is known [10-13] that the speech signal is a complex non-stationary process, but it can be taken as a stationary random process on short intervals equal from units to several tens of main tone periods. Thus, in the research the class of piecewise stationary random processes [14-16] was used as the speech signals mathematical model.

Let the stationary random process $\xi_1(t)$ be realized on the interval $[0,t_1)$, the stationary random process $\xi_2(t)$ on the next interval $[t_1,t_2)$ and so on, on the interval $[t_{n-1},t_n)$ – the stationary random process $\xi_n(t)$. In general, such a process can be presented in the form $\Xi_n(t)=(\xi_1(t),\xi_2(t),...,\xi_n(t))$, where: $\Xi_n(t)$ – random process – speech signal.

For such a process, the concept of disorder is introduced, which characterizes the transitions between areas where stationary random processes $\xi_1(t),\xi_2(t),...,\xi_n(t)$ are present [14-16]. At the same time, it is possible to define such transitions [16], which will make it possible to segment the password into sections corresponding to individual sounds.

At the same time, taking into account the fact that speech sounds can be divided into vowels, vocalized consonants and noisy consonants, it becomes possible to identify areas in the structure of the speech signal that correspond to such sounds. The duration of these areas and their ratio in the words used for identification can be used as additional informative signs to increase the reliability and accuracy of user identification itself.

On the basis of the selected mathematical model of the speech signal, it becomes possible to apply the processing method, the essence of which is as follows. Actually, the speech signal processing should be carried out on the short time intervals, which are equal to several units or tens of main tone periods - within the limits of the sliding window [16]. By shifting the window in time according to the

jump-like change of the probability characteristics calculated within each window, it becomes possible to detect transitions between individual sounds and select areas corresponding to individual sounds. In this way, it is possible to identify users based on several individual parameters of the speech signal.

3. The method of speech signals processing for the task of user identification

A four-parameter identification method is proposed, which consists in carrying out such identification based on four parameters of the speech signal, three of which are individual biometric characteristics and, in combination, are not suitable for forgery. The actual identification process is carried out in two stages. At the initial stage, in the process of registering a new user, his speech signal is registered, which is a password spoken by a person, that will be used to identify the person at each subsequent request. Both the word - password itself and individual biometric parameters of a person's speech will be unique for identification. In particular, the areas of the speech signal that correspond to vowels and vocalized consonant sounds and the evaluation of the value of main tone frequency are selecting. For identification at the next stages, the values of this frequency and the duration of areas of vowels and vocalized consonant sounds will be used, which are the biometric parameters of the user.

When registering a new user with a certain service or database, he/she (the user) will need to pronounce a sequence of vowel sounds into the microphone. This sounds will be used by the system to determine the approximate value of the main tone frequency, which will be the first individual biometric parameter of the user. Next, the user will need to come up with and say into the microphone a certain test word - a password that will be known only to the user. Based on the pre-received value of the main tone frequency, the system will process the record of this word and select the areas in it where this frequency will be present. These areas will correspond to vowels and vocalized consonant sounds. The value of the main tone frequency as well as the duration of the calculated time intervals of such areas in the test word and their sequence for a separate word will be used in the next identification of the user.

To determine the main tone frequency, the method of formant analysis has proven itself well, according to which the frequency of placement of the first maximum in the spectrum of the vowel sound will correspond to the main tone frequency [13]. At the first stage, after recording the vowel sound or sounds, the system will calculate the signal power spectral density distribution and determine the main tone frequency. But, since the speech signal itself is a random process, the value of this frequency will also change. However, as stated in [13], at short time intervals (less than 0.1 s) the speech signal can be taken as a stationary random process. Thus, it is possible to select sections of the speech signal with a duration of about 0.1 s, calculate the power spectral density distribution for each section, determine the main tone frequency for each such spectrum, and calculate the average value of this frequency. It is in this way that it is planned to determine the main tone frequency in the proven method of speech identification.

At the same time, such an analysis of the speech signal corresponds to the presentation of the latter as a piecewise stationary random process, in particular, when analyzing not only individual vowel sounds, but also sequences of different sounds in the spoken password, for which individual stationary areas will correspond to individual sounds. The processing method itself is based on the application of a sliding window, which is broadcast on the recorded speech signal. Within each window, the presence of the main tone signs is assessed using formant analysis methods. In particular, within each window, estimates of the power spectral density distribution are calculated and the presence of maxima in the frequency range, which can take the value of the main tone frequency, is detected. This range can roughly be from 85 to 450 Hz [12, 13]. The frequency of placement of the first maximum will correspond to the main tone frequency.

Now, knowing the approximate value of the main tone frequency, we can determine the range of its change according to the rule of three sigma [14], according to which the main tone frequency can vary from -3σ to $+3\sigma$, where σ is the root mean square deviation, which can be found from the set of calculated values of the main tone frequency for samples from the speech signal.

Further, the system analyzes the record of the password. Within each sliding window, the power spectral density distribution is calculated and the presence of a maximum within the previously

calculated range of main tone frequency change is determined. By plotting the values of these maxima on the time axis according to the position of each window, a curve can be obtained that will display, for a given audio signal, the areas corresponding to the vowels and vocalized consonant sounds. At the same time, it becomes possible to determine the durations of such areas.

In Figure 1 shows the recording of the speech signal, which is a sequence of vowel sounds [*a*] (the first 7 sounds) and vocalized consonant sounds [*l*] (the last 3 sounds).

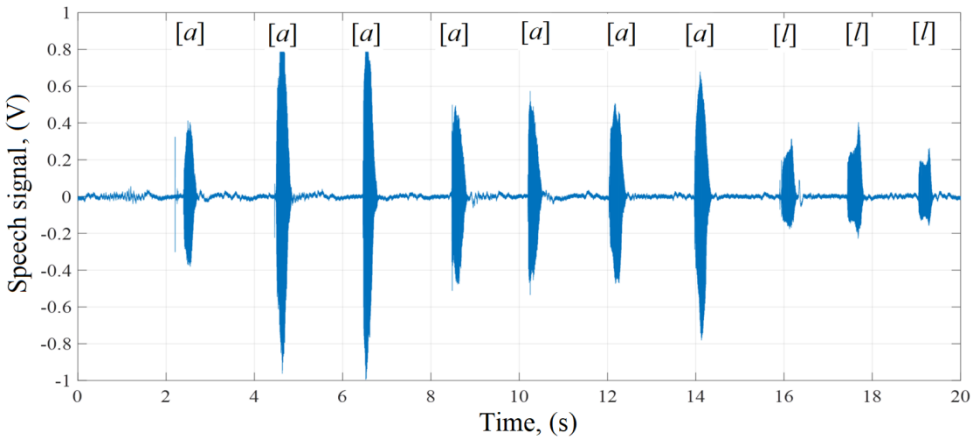

Figure 1: Recording of the speech signal, which is a sequence of vowel sounds [*a*] (first 7 sounds) and vocalized consonant sounds [*l*] (last 3 sounds)

For each individual sound, power spectral density distributions were calculated and individual values of the main tone frequency were determined. As an example, some of them are shown in Figure 2.

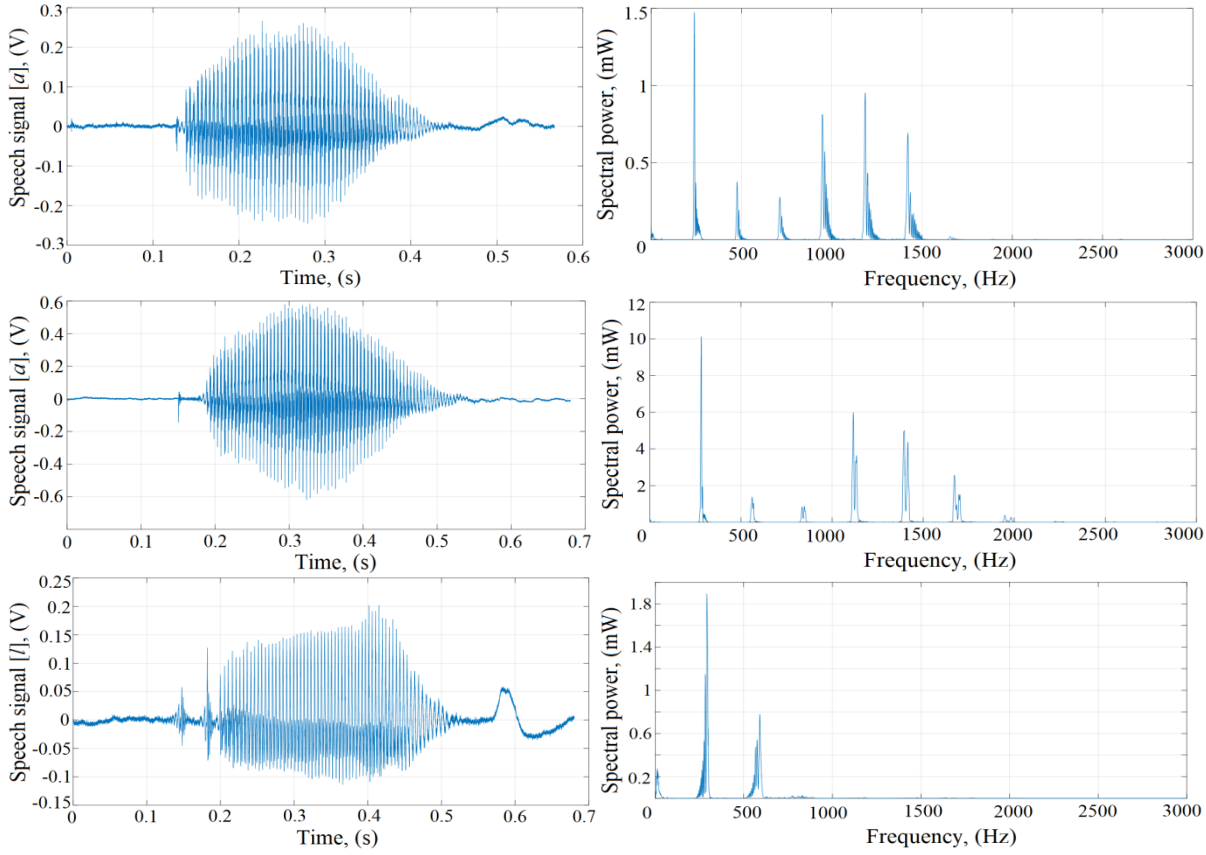

Figure 2: Individual sounds and their power spectral density distributions

Further, the main tone frequency was calculated for each sound within a sliding window with a width of 0.1 s. Based on these values, the average value of the main tone frequency and the range of change of this value were calculated. The average value of the main tone frequency was 206 Hz, and the range of change according to the three-sigma rule was from 166 to 246 Hz.

At the next stage, the analysis of time durations of the areas corresponding to vowel sounds was carried out. For this, the recording of the sequence of vowel sounds [*a*] was used, shown in Figure 3.

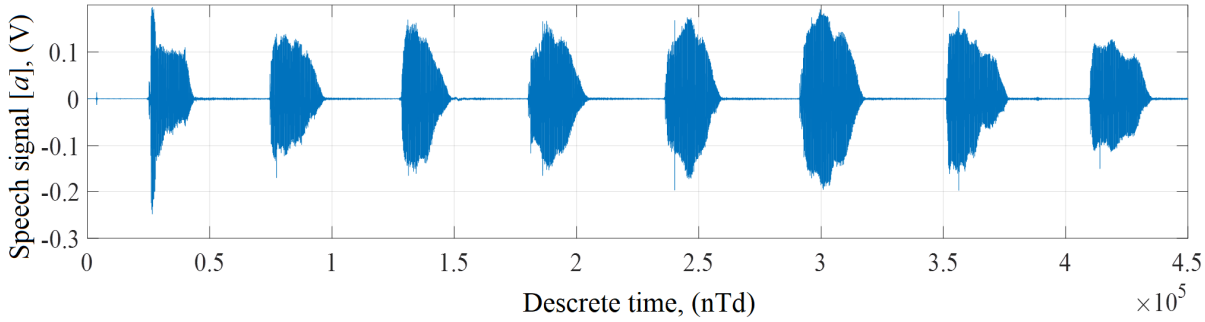

Figure 3: Recording of the sequence of vowel sounds [*a*]

Knowing the value of the main tone frequency and the range of its change, this signal was processed by the proposed method in the following way. A sliding window with a width of 0.1 s was formed. It was broadcast by signal. Within each sliding window, a power spectral density distribution was calculated and the maximum in this spectrum was found in the range of changes of main tone frequency. The value of this maximum was delayed on one time axis according to the placement of the sliding window on the speech signal recording. The view of the resulting graph is shown in Figure 4.

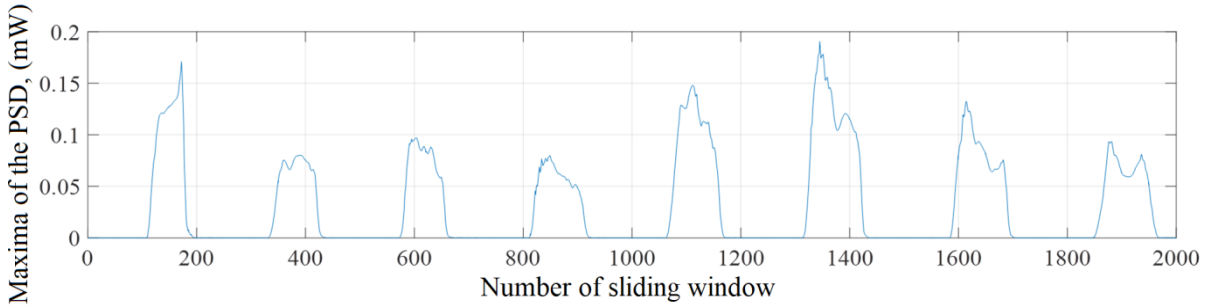

Figure 4: Graph of the presence of maxima in the power spectral density (PSD) distribution within the corresponding sliding window

Comparing the graphs in Figure 3 and Figure 4 we can come to the conclusion that with the help of this processing method it is possible to determine the areas that correspond to vowel sounds by the value of the main tone frequency . However, the fronts of such areas are tilted, so it is necessary to develop a method for establishing the duration of such areas. For this, a threshold function was used, which takes two values, namely: 0 – if the value of the signal in Figure 4 is less than or equal to a certain threshold, and 1 – if the signal value is greater than the threshold. In this way, the graph in Figure 4 will turn into a sequence of rectangular pulses, the width of which will correspond to the duration of the corresponding speech sound.

The threshold value can be selected from the relationship between the main tone frequency and, for example, the duration of the password. Thus, this value will also be individual for each individual user.

For analysis was used the speech signal shown in Figure 1. A graph of the presence of maxima in the power spectral density distribution within the corresponding sliding window and a graph of the threshold function was constructed for it. They are shown in Figure 5. The threshold value was chosen a priori at the level of 0.1.

From Figure 5, it is already possible to determine with sufficient accuracy the durations of areas that correspond to vowels and vocalized consonants, as well as the sequence of these durations in the password.

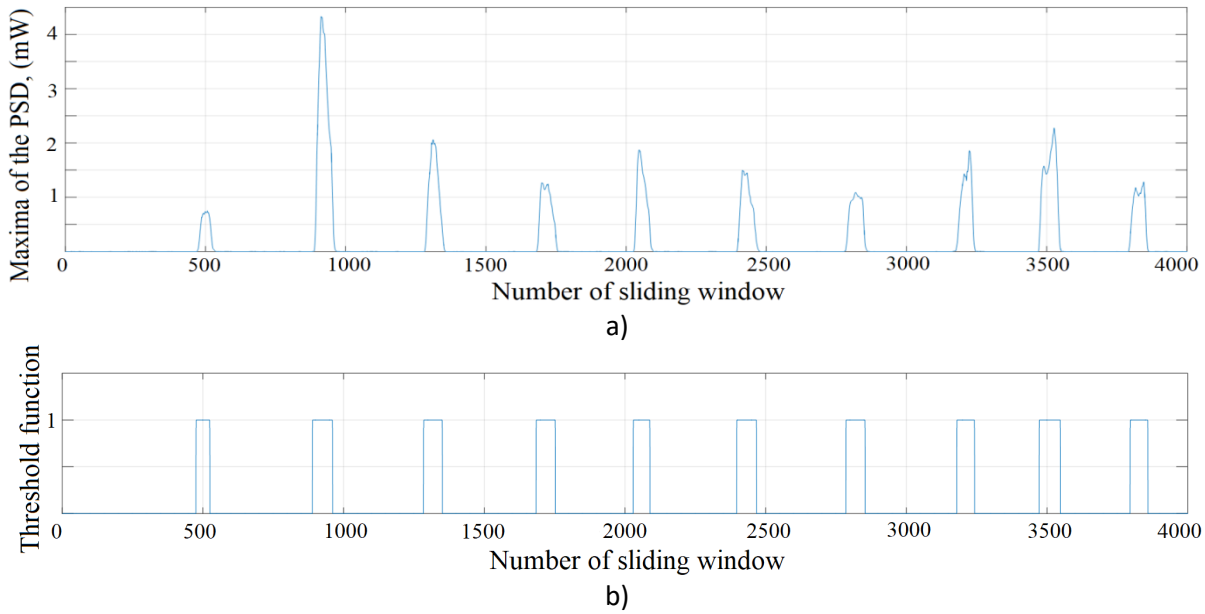

Figure 5: Graph of the presence of maxima in the power spectral density distribution within the corresponding sliding window (a) and graph of the threshold function (b)

However, an important question is how difficult it will be to deceive this method when, for example, the test word is spoken by another person. For this, the range of values of the main tone frequency was changed. The corresponding graphs of the presence of maxima in the power spectral density distribution within the corresponding sliding window and the threshold function in the case of reducing this range to (120-194) Hz are shown in Figure 6.

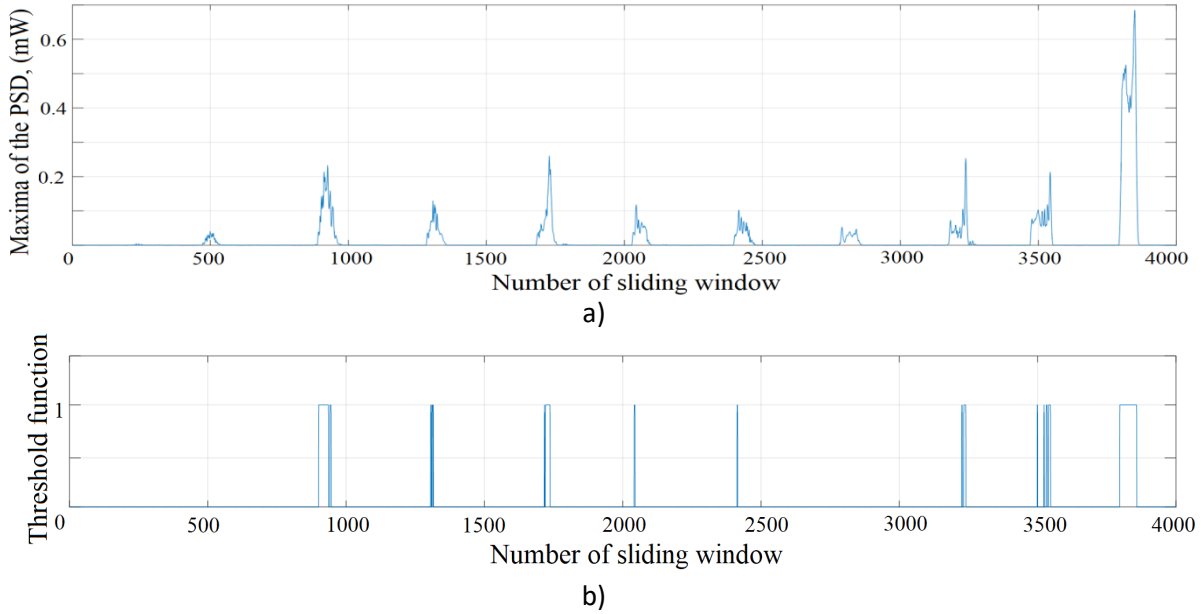

Figure 6: Graphs of the presence of maxima in the power spectral density distribution within the corresponding sliding window (a) and threshold function for the main tone frequency range (120-194) Hz (b)

From Figure 6 it can be seen that when the range of the main tone frequency does not correspond to the actual range, the graphs of the presence of maxima in the power spectral density distribution

within the corresponding sliding window and the threshold function are significantly distorted, there are gaps corresponding to the non-detection of sounds, a significant change in the duration of the corresponding intervals of the threshold function or their loss .

In Figure 7 shows the corresponding graphs of the presence of maxima in the power spectral density distribution within the corresponding sliding window and the threshold function in the case of an increase in the main tone frequency range to (220-294) Hz.

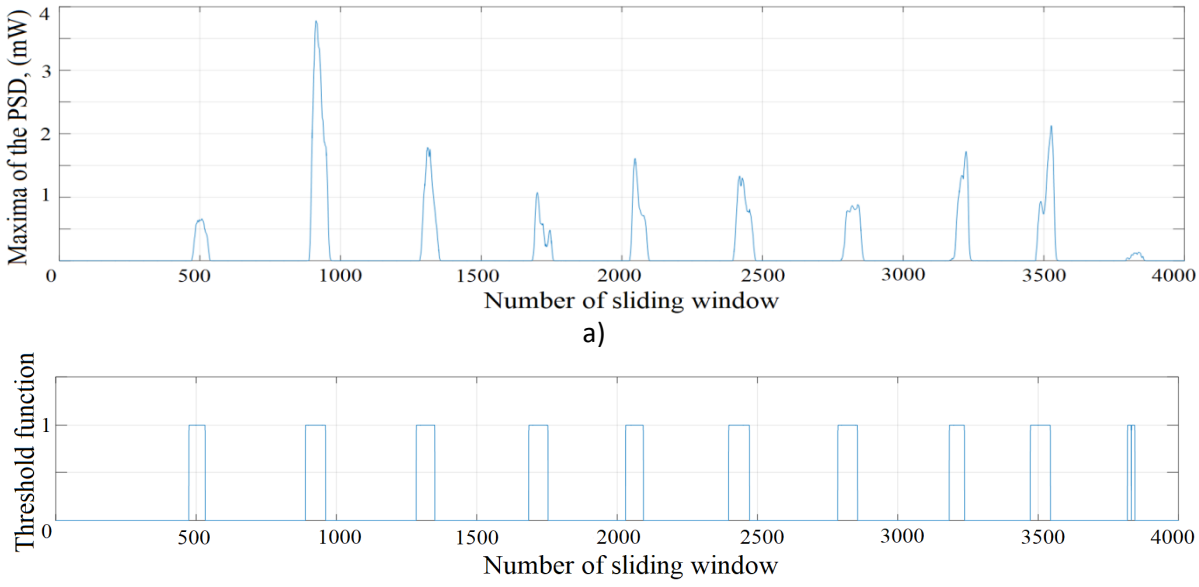

Figure 7: Graphs of the presence of maxima in the power spectral density distribution within the corresponding sliding window (a) and the threshold function for the main tone frequency range (220-294) Hz (b)

Similarly, as in the previous case, the graphs are distorted, and the intervals of the threshold function are significantly shortened and bifurcated. From the analysis of the graphs in Figure 5, Figure 6 and Figure 7, it can be stated that the developed method is efficient, sensitive and makes it possible to identify a person based on such individual biometric parameters of the speech as the main tone frequency, the duration of vowels and consonants located in the password, as well as the sequence of these durations in the password.

Additionally, Figure 8 shows the recording of the word "*s o f a*", which includes two vowel sounds [*o*] and [*a*], and two noised consonants [*s*] and [*f*]. At the top of this figure is shown the threshold function.

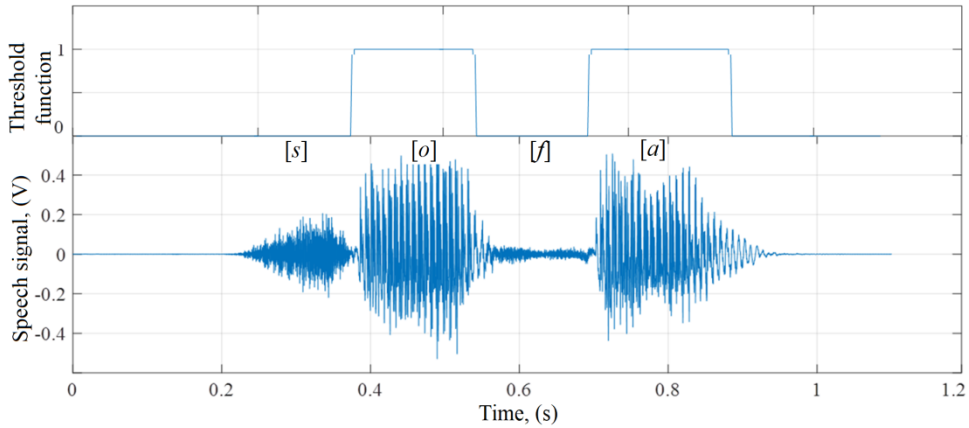

Figure 8: View of the threshold function (top) for the speech signal - the word "*s o f a*"

As expected, the method is sensitive and makes it possible to distinguish areas of vowels and vocalized consonant sounds and is not sensitive to noised sounds.

4. Conclusion

The described method of user identification includes the evaluation of three parameters of the speech signal, which are individual informative signs of the user, namely: the value of the main tone frequency, the duration of the areas of the speech signal that correspond to vowels and vocalized consonant sounds and their ratio in the password (individual for each person). The fourth parameter is the value of the threshold function, which is set in a certain way during the registration of a new user, and the calculated values of the durations of vowels and vocalized consonant sounds will depend on its value. Algorithms for implementing the method at various stages of its implementation are simple and can be integrated into various services and implemented in various software environments with the ability to work in real time. The actual calculations used in the developed method are not complicated.

However, in the future, the method can be improved by expanding the number of informative signs, which are used to identify the user, in particular by using the parameters of the speech signal in the time and spectral domain. However, this would require an additional signal centering and normalization procedure and would be extra sensitive to external noises and interferences.

5. References

[1] "What is Authentication? Definition of Authentication, Authentication Meaning". The Economic Times. Retrieved, 2020-11-15.

[2] Tardo J. and K. Alagappan. SPX: Global Authentication Using Public Key Certificates. M.California, 1991, pp.232-244.

[3] R. Manjula Devi, P. Keerthika, P. Suresh, Partha Pratim Sarangi, M. Sangeetha, C. Sagana, K. Devendran, “Retina biometrics for personal authentication”, Machine Learning for Biometrics. Concepts, Algorithms and Applications. Cognitive Data Science in Sustainable Computing. 2022, pp. 87-104

[4] Biometrics: definition, use cases, latest news, 2022, URL: https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/inspired/biometrics

[5] What is Biometric Authentication? By Dean Nicolls, 2019. URL: https://www.jumio.com/what-is-biometric-authentication/

[6] 5 common biometric techniques compared, 2022. URL: https://www.recogtech.com/en/knowledge-base/5-common-biometric-techniques-compared

[7] Paul Benjamin Lowry, Jackson Stephens, Aaron Moyes, Sean Wilson, and Mark Mitchell. Biometrics, a critical consideration in information security management. Margherita Pagani, ed. Encyclopedia of Multimedia Technology and Networks, Idea Group Inc., 2005 pp. 69–75.

[8] Poddar, Arnab; Sahidullah, Md; Saha, Goutam. Speaker Verification with Short Utterances: A Review of Challenges, Trends and Opportunities. IET Biometrics. 7 (2), March 2018, pp. 91–101.

[9] Pollack, Pickett, Sumby. Experimental phonetics. MSS Information Corporation, 1974, pp. 251–258.

[10] Fant G. Acoustic theory of speech production. The Hague: Mouton, 1970.

[11] Jafek B, Stark A. ENT secrets. Philadelphia, PA: Hanley & Belfus, 1995.

[12] Sadaoki F. Digital speech. Processing, synthesis and recognition. Tokyo: Tokyo institute of technology, 2000.

[13] Rabiner LR, Shafer RW. Digital processing of speech signal. New Jersey: Prentice-Hall, 1978.

[14] J. Bendat, A. Pirsol. Applied analysis of random data. 1989, 540 p.

[15] N.B. Marchenko, V.V. Nechiporuk, O.P. Nechiporuk, Yu.V. Pepa. Methodology of accuracy of information and information-viral systems of diagnostics. NAU, 2014, 377 p.

[16] Vyacheslav Nykytyuk, Vasyl Dozorskyi, Oksana Dozorska. Detection of biomedical signals disruption using a sliding window. Scientific jornal of the Ternopil National Technical University, 2018, Vol. 91, № 3, pp. 125–133.