Human voice identification based on the detection of fundamental harmonics

Magzhan Aliaskar
Department of Information Systems
Al-Farabi Kazakh National University
Almaty, Kazakhstan
87071160626@mail.ru

Talgat Mazakov
. Department of Artificial Intelligence and Big Data
Al-Farabi Kazakh National University
Almaty, Kazakhstan
tmazakov@mail.ru

Aigerim Mazakova
. Department of Artificial Intelligence and Big Data
Al-Farabi Kazakh National University
Almaty, Kazakhstan
aigerym97@mail.ru

Sholpan Jomartova
Department of Artificial Intelligence and Big Data
Al-Farabi Kazakh National University
Almaty, Kazakhstan
jomartova@mail.ru

Timur Shormanov
Department of Informatics
Al-Farabi Kazakh National University
Almaty, Kazakhstan
tt007@mail.ru

2022 IEEE 7th International Energy Conference (ENERGYCON) | 978-1-6654-7982-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ENERGYCON53164.2022.9830471

Abstract— **The article is devoted to the development of a system for biometric identification of a person by voice. The article discusses algorithms for the analysis of audio recordings for biometric identification of a person by voice. The technique of experimental research is considered, the process of processing the results of identification is described. Used algorithms MFCC and PLP for digital processing and analysis of audio recordings. An algorithm based on multi-criteria optimization has been developed for acoustic speech analysis. Various algorithms are used, such as hidden Markov models (CMM or HMM), as well as a model of a mixture of Gaussian distributions (GMM or GMM); in recent years, Wave Net neural networks have been actively used. The result of determining the tone of speech and the content of speech for the purposes of identification by voice is obtained.**

In the work, to compare the recorded voice with the saved voice for the purpose of personal identification, an unlimited text independent recognition system was applied using the Gaussian mixture model. The recorded voices were processed and stored during the registration phase, and the probing voices were used for comparison during the verification / recognition phase of the system. For biometric identification of a person by voice, the MFCC and PLP algorithms were used for digital processing and analysis of audio recordings. The result obtained makes it possible to determine the fundamental harmonics of speech for the purposes of identification by voice. The "Multiparameter automated system of biometric identification of a person" was developed on the Visual FoxPro DBMS. Today, speech identification and authentication is used in a wide range of applications ranging from smartphone applications to access control systems. Additional confirmation of the relevance of this area is the many research centers.

Keywords— biometric personality identification, information security, WAV file, identification, verification, matching, voice characterization, human speech, acoustic modeling.

I. Introduction

The problem of information protection and information security is one of the most important aspects of the development of modern society. Currently, the solution to this problem in the development and operation of information systems for various purposes is associated with the development of all kinds of requirements for ensuring their security and the creation of software and hardware from unauthorized access [1-3].

Automatic person recognition for identification has a large number of applications in various fields. Public safety problems, the need for remote authentication, the development of human-machine interfaces are causing increased interest in this technology [4-5].

Methods of biometric identification of a person are increasingly used in access control systems to workplaces, mobile devices, local and global information resources. Since the implementation of the systems does not require specialized equipment, and the biometric feature cannot be lost, forgotten or transferred [6].

In the Decree of the President of the Republic of Kazakhstan dated October 10, 2006 N 199 «On the Concept of Information Security of the Republic of Kazakhstan»: «Analysis of the current state of information security in Kazakhstan shows that its level currently does not correspond to the needs of a person, society and the state» and as the main goal of ensuring information security: « Creation and strengthening of a national information protection system, including in state information resources».

On January 31, 2017, the President of the Republic of Kazakhstan Nursultan Abishevich Nazarbayev addressed the people of Kazakhstan with the message «The third modernization of Kazakhstan: global competitiveness». This appeal noted the need for the development and adoption of the program «Digital Kazakhstan». In this regard, on behalf of N.A. Nazarbayev, in order to ensure the information security of society and the state in the field of informatization and communications, as well as to protect the privacy of citizens when they use information and communication technologies, the Concept «Cyber Shield of Kazakhstan» was developed. It notes that special attention is required for the training of personnel in the universities of Kazakhstan on information security and the development of domestic information security tools.

Authentication methods based on the measurement of human biometric parameters provide 100% identification. At the moment, the following biometric characteristics are successfully used in biometric systems for user authentication: iris, fingerprint, palm print, vascular patterns, face geometry, voice print, signature, DNA comparison,

978-1-6654-7982-0/22/$31.00 ©2022 IEEE

which have properties without which their practical application is impossible [7-9].

The interest in identification systems is due to a wide range of practical applications: checking access rights to various systems (databases, communication channels, premises, devices and mechanisms; bank accounts, etc.).

One of the parameters of biometric personality identification is voice, but a person's voice can change depending on age, emotional state, health or other factors, which makes the identification process more difficult to implement. Voice identification technology is used in various areas of information security, access control systems, forensic science and other areas.

Oral speech of a person is an ordered system of acoustic signals that are perceived as a sound image, in the oral speech of a person his individual signs and characteristics are reflected. The individuality of the voice is a consequence of the shape and size of the mouth and nasal cavity of the throat and respiratory organs. Thus, the physical characteristics of sounds – frequency, duration, intensity – are strictly individual for each person. The acoustic characteristic of the voice is relatively stable over time and remains individual even with pathological changes in the organs of speech. The task of identification by voice consists in the selection of human speech from the input audio stream, its classification and recognition.

Since the human voice is the sum of many separate frequencies created by the vocal cords, there are several features that can be observed and analyzed in the speech of each person:

– vocal speech (loudness, tempo, stability – physical components);

– tonality of speech (intonation – psychological components);

– content of speech (vocabulary of a specific personality).

Loudness is a subjective measure of sensation associated with the impact on the hearing organs of sound vibrations and depends on the amplitude and frequency of these vibrations. The rate of speech is a subjective measure related to the speed of pronunciation of certain segments of speech in time. Pace can be related to content, usually the most important words are spoken more slowly. The volume and tempo of speech are individual for each person [10-11].

A person has a certain vocabulary, this stock is determined by his social and mental environment. The peculiarities of speech, voice, intonation, as well as the manner of speaking, formed in adolescence by about twenty years of age, persist throughout life and have a complex of certain, only inherent features. After analyzing the individual elements of speech, you can determine the individual manner of a person's speech.

Also, one of the most important characteristics of a voice identification system is the speed (performance) of personal identification.

There are certain dependencies between the identification signs of the voice of the same person, obtained at different times and in different emotional states, which must be established.

The difference in timbres of different voices is described by different frequency spectra. The mathematical apparatus for analyzing the frequency spectrum is the Fourier transform, as a way to describe a complex sound wave with a spectrogram.

Advantages of voice identification: do not alienate from a person; does not require direct contact; does not require complex technical devices. The speaker's voice, and as a result, the speech signal itself is unique due to the specific physiological structure of his articulatory apparatus and the specificity of his speech [13-14].

The speaker identification problem is divided into three relatively independent parts:

– extraction of informative features (parametrization of the speech signal);

– the procedure for building a standard for a given speaker;

– decision making based on comparison with standards.

The most important element of successful speaker recognition is the selection of informative features that can effectively represent information about the characteristics of a particular speaker. The requirements for them are as follows:

– the efficiency of presenting information about the features of a particular speaker's speech;

– ease of measurement;

– stability over time;

– practical independence from the acoustic environment;

– impervious to imitation.

One of the key features is the pitch frequency – the frequency of the impulses of the vocal source resulting from the vibrations of the vocal cords. In this case, the frequency of oscillations can be violated due to changes in the amplitude, frequency, phase of oscillations, the presence of noise, therefore, the frequency of the fundamental tone is understood as the average estimate over a certain interval. Speech signals are nonlinear and non-stationary signals of complex shape, the amplitude and frequency characteristics of which change rapidly over time. In the field of speech signal processing, the most popular decomposition methods are Fourier transform (FT) and wavelet transform (WP), which have a number of advantages and disadvantages [12].

II. Ease of Use

Improving the reliability of voice authentication systems is an urgent scientific and technical task. The accuracy of identification (establishment) and verification (confirmation) of a person by voice is largely determined by the adequacy of the mathematical model describing the speech signal.

An increase in accuracy within the framework of existing methods for describing speech signals leads, as a rule, to a significant increase in the number of model parameters, which entails an increase in the systematic error and processing time of the received data, as well as a decrease in the significance of such parameters for characterizing the individual characteristics of a human voice. Mathematical models corresponding to the physics of processes have the highest description accuracy, therefore, when developing a mathematical model of a speech signal, it must be adequate to the acoustic theory of speech production.

Let the audio recording be given in WAV format [15]. In order to process it, we consider the data in the form of a digital series. To identify the k fundamental harmonics, we approximate the original series by a polyharmonic process

$$y(t) = A_0 + \sum_{j=1}^{k} \left(A_j cos \tfrac{2\pi}{T_j} t + B_j sin \tfrac{2\pi}{T_j} t \right) \qquad (1)$$

where $t \in [-L, L]$, L – quite a large number.

We transform the known series $x_1(t_1)$ into the series $z(t_1)$:

$$z(t_i) = x_1(t_i - t_n).$$

Then the problem is reduced to determining the coefficients $A_0, A_j, B_j, T_j, j = \overline{1,k}$ from the condition of the minimum of the functional:

$$F(A_0, A_j, B_j, T_j) = \int_{-L}^{L} (z(\tau) - y(\tau))^2 d\tau$$

or

$$F(A_0, A_j, B_j, T_j) = \sum_{i=1}^{n} (z(t_i) - y(t_i))^2 \qquad (2)$$

Here n is the total number of known measurements.

To determine the coefficients $A_0, A_j, B_j, T_j, j = \overline{1,k}$ the following algorithm is proposed:

Step 1. We define A_0 by the formula

$$A_0 = -\frac{1}{n} \sum_{i=1}^{n} z(t_i).$$

Step 2. Center the row $z(t_i)$:

$$\hat{z}\,(t_i) = z(t_i) - A_0.$$

Step 3. Let $j = 1$ be the number of the determined harmonic.

Step 4. We calculate the coefficients of the new series:

$$\tilde{z}\,(t_i) = \hat{z}\,(t_i) - \sum_{m=1}^{j-1} \left(A_m cos \frac{2\pi}{T_m} t_i + B_m sin \frac{2\pi}{T_m} t_i \right)$$

(by this step, the coefficients $A_m, B_m, T_m, m = \overline{1, j-1}$ are assumed to be already calculated).

Step 5. Let us determine the coefficients A_j, B_j, T_j from the condition of the minimum of the functional

$$\tilde{F}\,(A_j, B_j, T_j) = \sum_{i=1}^{m} \left(\tilde{z}\,(t_i) - A_j cos \frac{2\pi}{T_j} t_i - B_j sin \frac{2\pi}{T_j} t_i \right)^2$$

$$\frac{\partial \tilde{F}\,(A_j, B_j, T_j)}{\partial A_j} = 0,$$

$$\frac{\partial \tilde{F}\,(A_j, B_j, T_j)}{\partial B_j} = 0, \qquad (3)$$

$$\frac{\partial \tilde{F}\,(A_j, B_j, T_j)}{\partial T_j} = 0.$$

The system of equations (3) is nonlinear with respect to the unknowns A_j, B_j, T_j. To solve it, we express the coefficients A_j and B_j from the first two equations in terms of T_j:

$$A_j = \frac{(p_1 q_3 - p_2 q_2)}{\Delta B_j} = \frac{(p_2 q_1 - p_1 q_2)}{\Delta} \qquad (4)$$

where $\Delta = q_1 q_3 - q_2^2$,

$$p_1 = \sum_{i=1}^{n} \tilde{z}\,(t_i) cos \frac{2\pi}{T_j} t_i, \quad p_2 = \sum_{i=1}^{n} \tilde{z}\,(t_i) sin \frac{2\pi}{T_j} t_i$$

$$q_1 = \sum_{i=1}^{n} \tilde{z}\,(t_i) \left[cos \frac{2\pi}{T_j} t_i \right]^2, \quad q_3 = \sum_{i=1}^{n} \tilde{z}\,(t_i) \left[sin \frac{2\pi}{T_j} t_i \right]^2$$

$$q_2 = \sum_{i=1}^{n} \tilde{z}\,(t_i) cos \frac{2\pi}{T_j} t_i sin \frac{2\pi}{T_j} t_i.$$

Substituting expressions (4) into the functional $\tilde{F}\,(A_j, B_j, T_j)$, we obtain a new functional

$$\hat{F}\,(T_j) = \sum_{i=1}^{n} \left(\widetilde{z}\,\,(t_i) - f(T_j) \right)^2 \qquad (5)$$

where

$$f(T_j) = A_j(T_j) cos \frac{2\pi}{T_j} t_i + B_j(T_j) sin \frac{2\pi}{T_j} t_i.$$

The minimum of the functional of one variable (5) can be found by various numerical methods [16-18].

Step 6. After finding the minimum of functional (5), we define A_j, B_j, T_j by the formula (4)

Step 7. We calculate the value of the functional

$$F_j = \sum_{i=1}^{n} \left(z(t_i) - A_0 - \sum_{m=1}^{j-1} \left(A_m cos \frac{2\pi}{T_m} t_i + B_m sin \frac{2\pi}{T_m} t_i \right) \right)^2$$

If $j=k$, then go to step 8, otherwise we calculate the next harmonic, that is, we increase j by one and go to step 4.

Step 8. Thus, we determine the number of leading harmonics j and the unknown coefficients $A_0, A_m, B_m, T_m, m = \overline{1,k}$.

In the mode of filling the compared database (DB), the found parameters (together with the identifying data of the person) are entered into the database.

In the search mode, the calculated parameters of the searched person are compared with the database data.

III. Conclusion

An algorithm for calculating harmonics from processed speech information has been developed. The proposed algorithm allows for operational identification of a person with maximum efficiency; the results obtained will provide the ability to reliably recognize an individual by voice.

The developed technology can be used (after appropriate adaptation) for wider application. In particular, they can be used to build a psychological portrait of the speaker, determine his gender and age.

A feature of the proposed algorithm is a possible increase in the number of harmonics to improve the quality of recognition. However, it increases the number of mathematical operations, and this problem will be taken into account in what follows.

In the future, it is also expected to use the proposed algorithm to control the "smart home".

Since our research group is engaged in the design and automation of various technical (including their controllability and optimal control), sound control becomes very relevant for us.

The positive economic effect is that no expensive equipment is required to receive voice information. The social effect is expressed in the breadth of applicability of algorithms for processing voice information.

ACKNOWLEDGMENT

The work was carried out at the expense of program-targeted funding of scientific research for 2021-2022 under the IRN project OR11465437 "Development of a national electronic data bank on the scientific zoological collection of the Republic of Kazakhstan, ensuring their effective use in science and education."

REFERENCES

[1] G.A. Buzov “A practical guide to identifying special technical means of unauthorized obtaining of information”. M.: Hot line - Telecom, 2010. – 240 p.

[2] V.G. Gribunin “Comprehensive information security system at the enterprise”. M.: Publishing house «Academy», 2009. – 416 p.

[3] Yu.F. Katorin, A.V. Razumovsky, A.I. Spivak “Information protection by technical means’. SPb.: NIU ITMO, 2012. – 416 p.

[4] R.M. Ball, J.H. Connel, S. Pancanti, N.K. Ratha, E.W. Senor “Biometrics Guide’. M.: Technosphere, 2007. – 368 p.

[5] V.M. Koleshko, E.A. Sparrow, P.M. Azizov, A.A. Khudnitsky, S.A. Snigerev “Traditional methods of biometric authentication and identification”. Minsk: BNTU, 2009. – 107 p.

[6] A.A. Afanasyev, L.T. Vedenyev, A.A. Vorontsov, E.R. Gazizova “Authentication. Theory and practice of providing secure access to information resources”. M.: Hot line - Telecom, 2012. – 550 p.

[7] A.P.Zaitseva “Technical means and methods of information protection“. M.: Publishing house «Mechanical engineering», 2009. – 508 p.

[8] [8] Shangin V.F. Comprehensive information protection in corporate systems. – M.: Publishing House «Forum», 2012. – 592 p.

[9] G.A. Kukharev, E.I. Kamenskaya, Yu.N. Matveev, N.L. Shchegoleva “Methods for processing and recognizing facial images in biometric problems”. M.: Politekhnika, 2013. – 416 p.

[10] R.A. Vasiliev “Biometric identification of users of information systems based on a cluster model of elementary speech units” Dis. Candidate of Technical Sciences on special: 05.13.19 – Ufa State aviation tech. un-y, 2017. – 153 p.

[11] T.Zh. Mazakov, Sh.A. Jomartova, T.S. Shormanov, G.Z. Ziyatbekova, B.S. Amirkhanov, P. Kisala. “The image processing algorithms for biometric identification by fingerprints” News of the national academy of sciences of the Republic of Kazakhstan. Series of Geology and Technical Sciences, 2020. – Vol. 1, – No 439. – P. 14-22. https://doi.org/10.32014/2020.2518-170X.2

[12] A.V. Agranovskiy, D.A. Lednov “Theoretical aspects of algorithms for processing and classification of speech signals”. M.: Radio and communication, 2004. – 164 p.

[13] T.Kintzel “Programmer's guide to working with sound”. M.: DMK Press, 2000. – 432 p.

[14] T. Mazakov, Sh. Jomartova, G. Ziyatbekova, M. Aliaskar. “Automated system for monitoring the threat of waterworks breakout” Journal of Theoretical and Applied Information Technology. – 2020. – Vol. 98, No. 15. – P. 3176-3189.

[15] V.P. Leontiev “Multimedia: photo, video and sound on the computer”. M.: OLMA Media-Group, 2009. – 379 p.

[16] S.A. Aivazyan, V.M. Bukhshberger, I.S. Enyukov, L.D. Meshalkin “Applied statistics. Classification and dimensionality reduction”. M.: Finance and statistics, 1989. – 607 p.

[17] B.I. Sokil, A.P. Senyk, M.B. Sokil, A.I. Andrukhiv, M.M. Kovtonyuk, K.Gromaszek, G.Ziyatbekova, Y. Turgynbekov “Mathematical models of dynamics of friable media and analytical methods of their research” Przeglad Elektrotechniczny, 2019 doi: 10.15199/48.2019.04.13

[18] T. Mazakov, W. Wójcik, Sh. Jomartova, N. Karymsakova, G. Ziyatbekova, A. Tursynbai “The Stability Interval of the Set of Linear System” Journal of electronics and telecommunications. – 2021. – Vol. 67. – No. 2. – P. 155-161. doi: 10.24425/ijet.2021.135958