Bulletin of TUIT: Management and Communication Technologies

Volume 4 *2021-fourth number* Article 2

April 2021

Analysis of algorithms and implementation of real time speaker identification system

Kamoliddin Shukurov
"Bulletin of TUIT: Management and Communication Technologies", keshukurov@gmail.com

Analysis of algorithms and implementation of real time speaker identification system

Shukurov Kamoliddin Elbobo o'g'li

Senior lecturer of "Computer systems" department, Tashkent University of Information Technology named after Muhammad Al-Khwarezmi

Phone: +99890 940 -04-41

E-mail: keshukurov@gmail.com

Abstract — The article describes an implementing a real time speaker identification system by voice for embedded and general purpose computers. A review and analysis of existing speaker identification algorithms are made. The speaker's input speech is recorded in the system, go through the preprocessing stage, extract features and voice parameters for further identification. To recognize the speaker by voice parameters, the Vector quantization and Hidden Markov model algorithms are used. The VQ and HMM algorithms showed recognition accuracy of 96% and 98%, respectively.

Keywords— *speaker identification, pre-processing, filtering, feature extraction, recognition, MFCC, Vector quantization, Hidden Markov model, confusion matrix.*

I. INTRODUCTION

Analysis of speech signal processing areas today focuses on speech command recognition, synthesis of hardware control commands, transmission of speech signals via IP telephony channels and protection of transmitted speech, concise speech recording and speech biometrics, criminology requires new research into the problems of analysis and synthesis of speech signals in the field of human identification (identification). In solving the problem of spectral transformations, it is difficult to implement without new approaches and approaches to filtering in different environments, parameterization of signals, algorithms for feature extraction and methods of recognition [1].

Biometric technology is based on measuring the unique features of a person. Biometric systems are used where identification of person is required. One of the most common biometric characteristics of a human being is his or her voice. Identifying a person by voice or speech involves a set of technical, algorithmic, and mathematical methods that involve complex steps from sound recording to pre-processing of speech data [2,3].

II. RELATED WORKS

Identification of the speaker by voice has been studied by scientists from different countries for many years and is still ongoing. Because, depending on the application, new approaches are required for speaker identification. In work [4] an updated approach of the SVM algorithm is used, [5] speaker identification is implemented based on VQ, MFCC and Inverted MFCC methods with a recognition

accuracy of 98%, the author of [6] works on speaker identification based on VQ and Kekre's Median Codebook Generation Algorithm with accuracy 84%. In [4], an updated approach of the SVM algorithm was used, in [5], speaker identification was implemented based on VQ, MFCC and Inverted MFCC methods with a recognition accuracy of 98%, the author of [6] works to identify the speaker based on VQ and Kekre's Median Codebook Generation Algorithm with an accuracy of 84%. The author of [7] have design of speaker verification using dynamic time warping (DTW) on graphical programming for authentication process with 84% accuracy. In [8] Speaker Verification using Convolutional Neural Networks (CNN) are implemented. The authors of [9,10] are implemented real time speaker identification on FPGA, but here the emphasis is mainly on the preprocessing stage of signal processing, but the recognition accuracy is not given.

For the implementation of the speaker identification system in biometric systems, control of technical units and voice identification in smart homes. The above algorithms are not sufficient. Because in real conditions our voice is always accompanied by different noises and the implementation of a complex recognition and identification algorithm in Embedded devices requires new approaches and optimal algorithms.

III. MAIN PART

Voice or speech identification is a separate scientific discipline that is part of the processing of speech signals [11]. Human speech identification prohibits access to various information resources and physical objects [12], the management of voice services in mobile communication in systems based on telecommunication channels, protection against fraud through the introduction of voice identification [13,14], can be widely used in investigative processes, in the protection of verbal information [15] and in the identification of offenders through digital speech in digital criminology. In addition, a person has the opportunity to learn about his age, gender and ethnicity or dialect, emotional state through speech [16-18]. Especially nowadays, there is great interest in the application and development of voice identification and speech recognition methods in the areas of voice control and speech recognition in smart home, smart and safe city systems [19, 20].

It should also be noted that the use of a combination of different technologies and methods is required to ensure the safest and most accurate identification and authentication, especially when performing important work. In other words, in these cases, it is advisable to use biometric methods of speech identification with special input / output devices with fast memory and high-performance microprocessors.

Implementing a voice or speech identification system in a real-world setting (other than a recording studio or laboratory) can present serious challenges and barriers:

- in such an identification, various changes and noises occur in the input signal due to the peculiarities of the equipment and devices for recording, processing and storing information.
- external acoustic noises inevitably affect the speech signal, which can

significantly change the individual information properties of the signal.

Therefore, it is difficult to demonstrate a sufficiently high efficiency in the processing and analysis of speech data with external noise or in real conditions. High-performance speech identification systems in the laboratory can show much lower reliability in real conditions.

Identifying a person through speech or voice requires sophisticated hardware and software solutions that incorporate a complex set of technical, algorithmic, and mathematical methods.

Identifying a speaker through speech is divided into 2 parts:

- the method of identification, which depends on the text spoken by the speaker;
- the method of identification, which is not depend on the text spoken by the speaker.

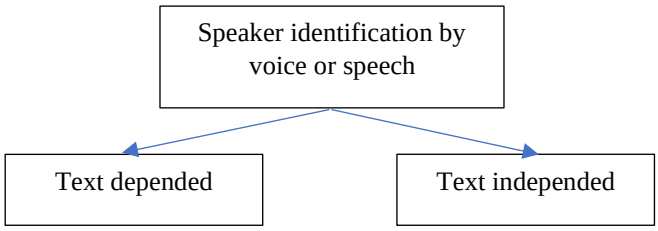

Fig. 1. Methods of speaker identification by voice.

In the non-text-based method of identifying a speaker through speech, the speaker's speech is converted to a digital signal and the formant frequencies are determined from that signal after pre-processing. As mentioned above, the human voice is a biometric property, and each person's voice has its own formant frequency.

In this paper, we will look at text depended identification. This is because most voice control and certain security systems use a specific command word or a specific keyword to identify a person through speech. The following figure illustrates the steps in implementing the structure of a speech identification system.

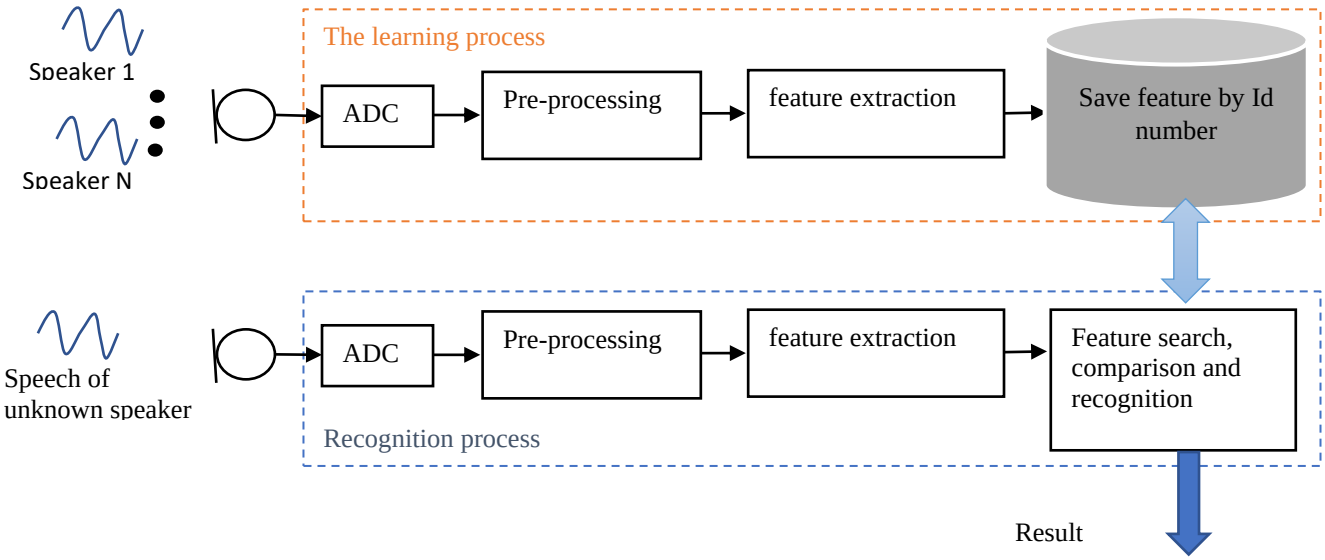

Fig. 2. The structure of the speaker identification system.

There are two steps of identifying a speaker by voice:

- learning;
- recognition.

The following algorithmic steps are performed in the learning process:

1. The analog signal coming from the microphone is converted to digital.
2. Pre-processing operations are performed. The input signal is adaptively filtered, segmented, framed and passed through windows.
3. The properties of MFCC (Mel Frequency Cepstrum Coefficient) are distinguished.
4. An identification number (id) is assigned to the speaker, and the voice features of that speaker are stored in memory.

The following algorithmic steps are performed in the recognition process:

1. The speech of the speaker, who speaks voluntarily from the microphone, is converted from an input analog signal to a digital one.
2. Pre-processing operations are performed. The input signal is adaptively filtered, segmented, framed and passed through windows.
3. The features of MFCC (Mel Frequency Cepstrum Coefficient) are distinguished.
4. The obtained features are compared with the previous features of the speakers stored in the memory of computer(or embedded system) using intelligent algorithms VQ or HMM.
5. If the features of the speaker's voice match with the features stored in the memory, the result is that the speaker has been identified.

Digitization of the input signal from the microphone. The signal read from the microphone is converted from analog to digital according to Kafelnikov's formula. There were no additional requirements for the microphone in this system, and the existing microphone of the personal computer was used. Input speech is defined as a sampling frequency of 16000Hz.

Pre-processing stage. At the pre-processing stage, adaptive filtering is initially performed [21]. Speech sounds are a low frequency signal. Typically, human speech sounds range from 80-180 Hz in men to 165-280 Hz in women. The frequency of common speech signals is in the range of 80 to 3000 Hz. In addition, the addition of a number of external interferences to normal speech sounds can lead to poor signal quality. The frequency of speech signals also varies depending on the person's condition and movement. Filtering variable speech signals is a very complex process, and filtering using simple digital filters is inefficient. Filtering speech signals using adaptive filters serves to improve signal quality.

The output signal y(n) is determined using the following formulas.

$$y(n) = h(n-1)x(n) \quad (1)$$

where *h(n)* is the vector of filter coefficients, *x(n)* is the input signal. Adaptive filter algorithms are used to reduce the error between the *y(n)* filtered signal output and the *d(n)* recording signal during the filtering process.

$$e(n) = y(n) - d(n),$$

$$h(n+1) = h(n) + 2\mu e(n)x(n) \quad (2)$$

Here y(n) - output signal, d(n) - record signal, e(n) - signal error, μ-step size of the adaptive filter. Filter pitch size is an important parameter that has the ability to improve the approximation speed of adaptive filters.

In this system, an adaptive filter of speech signals with LMS (Least Mean Square) algorithm was used (Figure 3). One of the advantages of the LMS adaptive filter is that it allows you to amplify the pure part of speech signals by clearing them of noise. The coefficients are also simple in terms of variability.

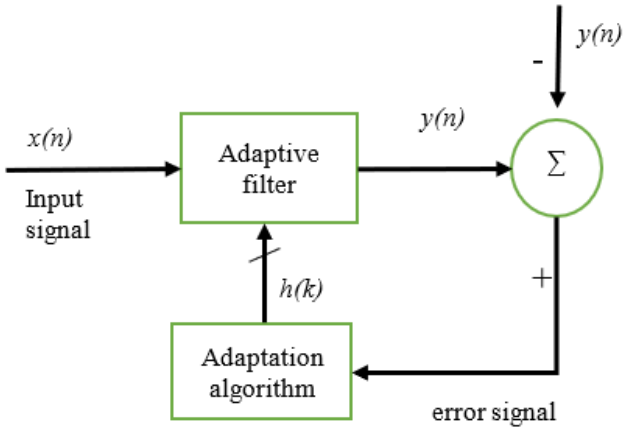

Fig. 3. Adaptive filter structure.

In framing, the input signal is divided into 256 equal parts. Frame values are determined by multiplying the following Hamming windows by the formula:

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \quad (3)$$

where N is the frame length.

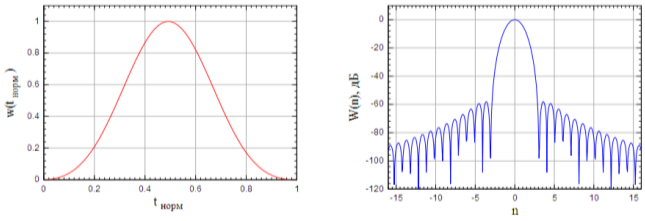

Fig. 4. Hamming window.

Fure transform is applied to the result.

$$X_j(k) = \sum_{n=0}^{N-1} x_j(n)\omega(n)e^{-\frac{2\pi i}{N}kn}, \quad 0 \le k \le N \quad (4)$$

where, j is the number of frames.

Extraction of MFCC features. Mel frequency capstrum coefficients are a common algorithm for extracting the desired features from speech signals, mainly by converting the polished spectral vocal tract data of the speech signal into small number of coefficients.

The time interval for each frame is calculated using the following formula:

$$P_i(k) = \frac{|X_j(k)|^2}{N} \quad (5)$$

We calculate the block of Mel filters.

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1) \le k < f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \quad (6)$$

where, m is the number of filters.

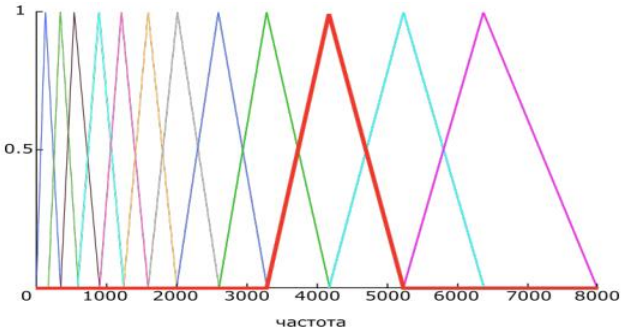

Fig. 5. Mel-filter scale.

The energy values obtained are logarithmized. It's basically brought closer to the human hearing system, where people don't receive loud sounds in a linear fashion, and we use more energy to hear louder sounds.

$$S_j(m) = ln\sum_{k=0}^{N-1} P_j(k)H_m(k), \quad 0 \le m < M \quad (7)$$

We obtain the MFCC features by applying a discrete cosine transform to the values generated above.

$$c_j(n) = \sum_{m=0}^{M-1} S_m(m)\cos(\pi n(m + \frac{1}{2})/M), \quad 0 \le n < M \quad (8)$$

Future comparison and recognition. At this stage, the MFCC features derived from the speaker's speech are compared with the MFCC features stored in the memory of the pre-training computing technique. Algorithms such as the Hidden Markov model of intelligent data processing, artificial neural networks, DTW (dynamic time warping), SVM (support vector machine), kNN (k-nearest neighbors algorithm), VQ (vector quantization) can be used [22-25].

In implementing this system, the VQ algorithm was used to identify the speaker and determine the compatibility of the feature. The VQ algorithm is relatively simple compared to other classification algorithms and is easy to identify in a speaker and implement in speech-dependent speech control hardware and software systems. In addition, the features of this algorithm have the ability to compress and save computer memory.

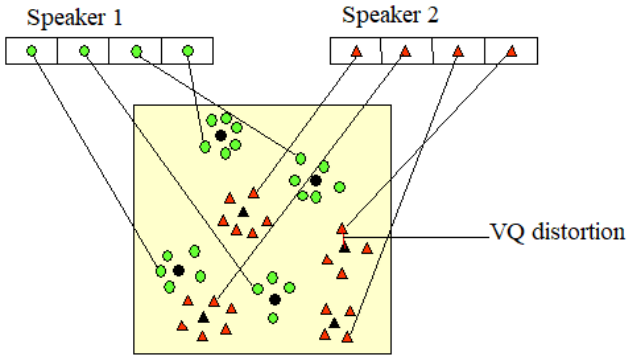

Fig. 6. The structure of the process of processing the VQ algorithm.

The average vector of the VQ value is divided into two parts, and the error value (distortion) of the two mean values is leanrt until the VQ model changes. Each of the two averages is then divided by two again, and the process continues until the required number of centroids is obtained. During testing and recognition, the dynamics of each speaker feature vector are compared with a coding book, and identification is performed by selecting a speaker parameter with minimal error. If the coding book size is C and the coding book vector is ym $1 \le m \le C$, the best matching randomness for each v vector is determined as follows.

$$n *= \arg \min_{1 \le m \le M} d(v, y_m) \quad (9)$$

Currently, Hidden Markov Models (HMM) are the backbone of most successful automatic speech recognition and speaker identification systems. The HMM is based on a finite state machine consisting of N states. Transitions between states at each discrete time moment t are not deterministic, and occur in accordance with a certain probability law and are described by the A_{NN} transition probability matrix.

Thus, the main parameters of the HMM are:

- N is the number of states;
- matrix of probabilities of A_{NN} transitions between states;
- N probability density functions $f_i(x)$.

The probability density function $f_i(x)$ is described, as a rule, by a weighted Gaussian mixture:

$$f(x) = \sum_{i=1}^{M} w_i p_i(x), \quad (10)$$

where M - is the number of components of the mixture; w_i - is the weight of the component of the mixture, and the value $p_i(x)$ is the normal probability distribution for the D-dimensional case:

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\sigma_i|^{1/2}} exp\left(-\frac{1}{2(x-\mu_i)^t \sigma_i^{-1}(x-\mu_i)}\right) 1)$$

where μ_i is the vector of mathematical expectation; σ_i is the covariance matrix.

Working with Hidden Markov Models, as with any other adaptive system, is carried out in 2 stages: training - using the Baum – Welch re-estemation algorithm, decoding - using the maximum likelihood algorithm (Viterbi).

IV. EXPERIMENT RESULTS

Based on the above algorithms, the speech identification algorithm was implemented and tested an application running on the Windows operating system. The tests were carried out for two algorithms (VQ, HMM). The results of speaker identification were obtained according to the following conditions: text-dependent speaker identification, text-independent speaker identification and speed of algorithm execution speech on embedded devices.

Table 1. Identification accuracy of VQ and HMM.

Algorithm	Text depended speaker identification accuracy (%)	Text independed speaker identification accuracy (%)	Algorithm execution speed (second)
VQ	96	89	3
HMM	98	93	8

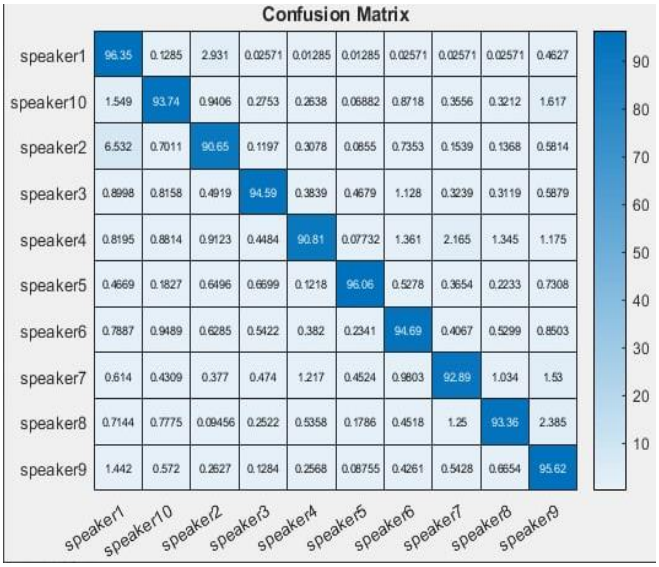

Fig. 7. A Confusion matrix of speaker identification algorithm.

In figure below the main window of the speaker identification system is shown.

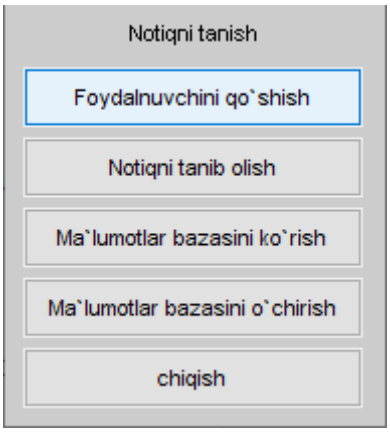

Fig. 8. The main window of the program.

The application has the following sections:

- add a user, in which the user joins the system by pronouncing special speech keywords;
- speaker recognition, in which any speaker may pronounce the keyword into the microphone if it is among the pre-trained and stored into the system memory and system recognizes it;
- in the Database View and Delete sections, you can see pre-trained and

logged-in users and delete them from the system as it needed.

III. CONCLUSION

Biometric identification is evolving in several ways. Methods of identification of speakers through biometric parameters, such as hand writing or by voice, becomes a popular. Hence, the identification of a speaker by voice or speech is widely used in smart home, safe city, smart car and remote voice control systems.

These systems require the development of complex mathematical hardware and software solutions. Intellectual algorithms with varying degrees of complexity are used to identify the speaker by speech.

The results showed that the identification of the speaker can be easily performed in hardware and software by taking the speech signal MFCC information features and classifying it using machine learning algorithms. Moreover, the accuracy level of this algorithm is not less than the accuracy level of other complex algorithms. For implementing a real-time speaker identification system on a hardware-software platform(embedded system), with a less computing resources is more suitable VQ algorithm. If the task of identifying the speaker requires a relatively easy algorithm, but with a high recognition accuracy, then it would be more correct to use the HMM.

REFERENCES

[1] Musayev M.M. Sovremennyye metody tsifrovoy obrabotki rechevykh signalov.// Vestnik TUIT 2(42)/2017. s. 2-13.

[2] Shukurov K.E. Raspberry pi qurilmasida o‘zbek tili nutq buyruqlarini tanib olish tizimini amalga oshirish.// TATU xabarlari 2(54)/2020. 45-61 b.

[3] Shukurov K.E., Ergashev S.B. Biometrik boshqaruv tizimida suxandonni aniqlash masalalariga bo‘lgan yondashuv.// Iqtisodiyotning tarmoqlarini innovasion rivojlanishida axborot-kommunikasiya texnologiyalarining ahamiyati Respublika ilmiy-texnik anjumanining. Ma’ruzalar to‘plami 1-qism. 14-15 mart Toshkent 2019 yil. 458-460 b.

[4] Sahoo, J. K. Deepak R. “Speaker recognition using support vector machines.” *International Journal of Electrical, Electronics and Data Communication*, ISSN: 2320-2084 Volume-2, Issue-2, Feb.-2014.

[5] Singh, S. and E. Rajan. “Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC.” *International Journal of Computer Applications* 17 (2011): 1-7.

[6] H. B. Kekre, V. A. Bharadi, A. R. Sawant, O. Kadam, P. Lanke and R. Lodhiya. “Speaker recognition using Vector Quantization by MFCC and KMCG clustering algorithm,” *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*, 2012, pp. 1-5, doi: 10.1109/ICCICT.2012.6398146.

[7] Barlian H., Dahnial S. “Design of Speaker Verification using Dynamic Time Warping (DTW) on Graphical Programming for Authentication Process.” *JITeCS Volume 2*, Number 1, 2017, pp 11-18

[8] Hossein, S. “Speaker Verification using Convolutional Neural Networks.” https://doi.org/arXiv:1803.05427v2

[9] S.Gourav, S.Goutam, “Real Time Implementation of Speaker Identification System with Frame Picking Algorithm.” *Procedia*

Computer Science Volume 2, 2010, Pages 173-180. doi:10.1016/j.procs.2010.11.022

[10] Ramos-Lara, R., López-García, M., Cantó-Navarro, E. *et al.* Real-Time Speaker Verification System Implemented on Reconfigurable Hardware. *J Sign Process Syst* **71,** 89–103 (2013). https://doi.org/10.1007

[11] Flanagan, Dzh.L. Analiz, sintez i vospriyatiye rechi / Dzh.L. Flanagan; per. s angl. A.A. Pirogova. – M. : Svyaz', 1968. – 396 s.

[12] Mariethoz, J. "Speaker Verification Based on User-Customized Password." / J.Mariethoz, B. Herve, M.F. BenZeghiba // IDIAP Research Report 01-13. – Martigny, 2001. – 22 p.

[13] Pellandini, F. "GSM Speech Coding And Speaker Recognition" / F. Pellandini, M. Ansorge, A. Dufaux [at al.] // International Conference on Acoustics, Speech, and Signal Processing (ICASSP): Book of abstracts. – Istanbul, 2000. – vol. 2. – pp.1085–1088.

[14] Amrouche, A. "Effect of GSM speech coding on the performance of Speaker Recognition System." / A. Amrouche, A. Krobba, M. Debyeche // 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA): Book of abstracts. – Kuala Lumpur, 2010. – pp. 137–140.

[15] Dvoryankin, S.V. O neobkhodimosti novykh podkhodov k otsenke effektivnosti tekhnicheskikh sredstv akustozashchity / S.V. Dvoryankin // Informatsiya i bezopasnost'. – 2002. – №2. – s. 244–245.

[16] Sorokin, V.N. Opredeleniye pola diktora po golosu / V.N. Sorokin, I.S. Makarov // Akusticheskiy zhurnal. – 2008. – T. 54. – № 4. – S. 659–668.

[17] Galunov, V.I. O vozmozhnosti opredeleniya emotsional'nogo sostoyaniya govoryashchego po rechi / V.I. Galunov // Rechevyye tekhnologii. – 2008. – № 1. – s.60–66.

[18] Romashkin, Yu.N. Raspoznavaniye pola diktora na osnove gmm-modeli golosa / Yu.N. Romashkin, Yu.O. Petrov // Rechevyye tekhnologii. – 2009. – № 2. – s. 31–38.

[19] Sorokin, V.N. Fundamental'nyye issledovaniya rechi i prikladnyye zadachi rechevykh tekhnologiy / V.N. Sorokin // Rechevyye tekhnologii. – 2008. – № 1. – s.18–48.

[20] Grebnov, S.V. Razrabotka i realizatsiya dvukhurovnevogo metoda golosovogo upravleniya na osnove skrytykh markovskikh modeley / S.V. Grebnov // Informatsionnyye tekhnologii. – 2009. – № 9. – S. 40–46.

[21] Douglas, S.C. "Introduction to Adaptive Filters" *Digital Signal Processing Handbook Ed*. Vijay K. Madisetti and Douglas B. Williams Boca Raton: CRC Press LLC, 1999

[22] Romanyuk A.G, Smirnov A.N., Antonova V.M. Ispol'zovaniye glubokogo obucheniya neyroseti dlya raspoznavaniya golosovykh komand pol'zovatelya. Zhurnal radioelektroniki [electronic journal]. 2019. № 11. Access: http://jre.cplire.ru/jre/nov19/18/text.pdf. DOI 10.30898/1684-1719.2019.11.18

[23] Rabiner, L., Juang, N-H, "Fundamental of speech recognition." *Pearson Education*, 2007.

[24] Reynolds D.A. “An Overview of Automatic Speaker Recognition Technology.” *The International Conference on Acoustics, Speech, and Signal Processing ICASSP 02. 2002*. P. 4072–4075.

[25] Pervushin YE.A. Obzor osnovnykh metodov raspoznavaniya diktorov / YE.A. Pervushin // Matematicheskiye struktury i modelirovaniye 2011, vyp. 24, s. 41–54.