INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2020, VOL. 66, NO. 3, PP. 583-588
Manuscript received February 25, 2020; revised July, 2020. DOI: 10.24425/ijet.2020.134015

Development of Speaker Voice Identification Using Main Tone Boundary Statistics for Applying To Robot-Verbal Systems

Yedilkhan Amirgaliyev, Timur Musabayev, Didar Yedilkhan, Waldemar Wojcik, and Zhazira Amirgaliyeva

***Abstract*—Hereby there is given the speaker identification basic system. There is discussed application and usage of the voice interfaces, in particular, speaker voice identification upon robot and human being communication. There is given description of the information system for speaker automatic identification according to the voice to apply to robotic-verbal systems. There is carried out review of algorithms and computer-aided learning libraries and selected the most appropriate, according to the necessary criteria, ALGLIB. There is conducted the research of identification model operation performance assessment at different set of the fundamental voice tone. As the criterion of accuracy there has been used the percentage of improperly classified cases of a speaker identification.**

***Keywords*—speaker voice identification, voice interface (FXO), human being and robot interrelation (HRI), speech recognition, statistics of voice fundamental tone, computer-aided learning, neural network**

I. Introduction

THE article is devoted to the development, research and software implementation of methods for forming a digital voice image of a certain person in order to identify the speaker by voice. Through the use of a digital voice portrait of a person, which allows you to more accurately take into account the individual voice characteristics of individuals, an improvement in quality characteristics is achieved in solving problems in the field of computer speech recognition and identification of a person's personality by his voice. Also, with the use of digital voice portrait of a person, the data on the individual speech and voice characteristics of a certain personality will be unified, which allows the use of a single and unified data set in solving various problems in the field of speech technologies.

A distinctive feature of the proposed technology is the assessment of available options for the algorithmic parameters of the system and the selection of their suitable values by means of developed software tools and based on the analysis of a digital recording of a natural speech signal. In this case, as the selection criteria for the most suitable parameters of the set of algorithms used, the results are the accuracy of the estimates of the results of speaker identification and speech recognition. One of the main principles of the developed technology for the automatic formation of a digital voice portrait of a person is multilingualism - ensuring the functioning of personalized speech processing in most of the existing natural languages exclusively at the level of program interface procedures with minimal time and effort on the part of the user.

In the conditions of universal development, digitalization and implementation of innovative information technologies in all sectors of the economy, the development of intelligent human-machine systems based on speech technologies is one of the urgent tasks of our time. In this area, the creation of new technologies for the automatic formation of a digital voice image of a person and their introduction in intelligent speech systems are the main tasks of both speech recognition and speaker identification.

When forming a digital image of a person, the parameters of speech recognition algorithms are considered, taking into account the individual speech characteristics of a particular person. The developed algorithms and software tools, as well as the multilingual interface made it possible to analyze the features of speech signals with the possibility of presenting them in a unified format for further use in the system.

The practical implementation of the proposed technology is capable of generating a new market for specialized software products and services used in intelligent robotic systems, computer and digital devices for technical equipment and digitalization of various technological processes.

Speaker identification is an important research domain within the recent years. Speaker identification system specifies a human being according to his/her speech pattern. There are two research themes in the area thereof. It is features extraction from the voice signal and their comparison. Speaker identification basic system is shown on the Figure 1.

The system consists of two different stages. Recording or learning is the first stage, or recording is the first stage and the second stage is testing. At the learning stage every speaker, wishing to be recorded in the system, shall submit own speech patterns for preparing the reference model of all speakers having been recorded. At the stage of testing, an input voice signal of the speaker, pretending to identity, is used for unique feature extraction from it and compared to the stored features in order to obtain the identification result.

Amirgaliyev Yedilkhan are with Institute of Information and computing technologies of the Science Committee of RK MES and Al-Farabi Kazakh National University. (amir_ed@mail.ru), Yedilkhan Didar are with Institute of Information and computing technologies of the Science Committee of RK MES, Astana IT University (yedilkhan@gmail.com)

Waldemar Wojcik are with Lublin Technical University (waldemar.wojcik@pollub.pl). Musabayev Timur and Zhazira Amirgaliyeva are with Institute of Information and computing technologies of the Science Committee of RK MES(tmusab@yandex.ru),.

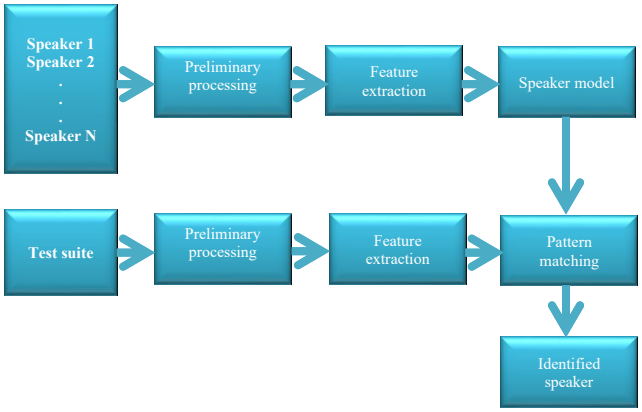

Fig. 1. Basic system of speaker identification

When an input voice signal precisely belongs to one of the speakers in the reference speech model of all registered speakers, then it is a speaker identification at the closed set [1]. If there is the probability of a speaker's voice signal nonconformity to any of the speakers in the reference speech model system, then it is the identification on the open set.

As well, a speaker identification system is subdivided into two types: text-dependent and text-independent. The system is text-dependent [2], if speaker identification depends on the text content in the spoken phrase and it is text-independent [3] if it does not depend on the text.

Features extraction various methods, used by many authors, include the basic techniques, based on spectral average values [4], pitch of tone, coding with linear prediction (LPC) [5], cepstral indices on the basis of the linear prediction (LPCC) [6], mel-cepstral coefficients (MFCC) [6] , etc.[7]

Speech recognition technology is widespread in different business areas:

- solutions "Smart house": voice interface of managing the system «Smart house»;
- household appliances and robots: electronic robots voice interface; voice control over household appliances, etc.;
- desktops and notebooks: voice input into computer games and applications;
- automobiles: voice control in passenger compartment -for instance, over navigation system [8].

In order to attain natural and individualized interaction between a robot and its users-people (HRI), it is important, that the first properly identifies its people-colleagues. Unique voice features are often used for reaching the goal thereof [11].

It is expected, that the speech recognition will be implemented by interactive robots, such as humanlike robots or robots-pets [12]. For example, the system, used by the robot Maggi and described in one of scientific works [9], has demanded, that the user shall be registered in the system. That phase consisted of questions, put forward by a robot to a user. Together with her name, age, language, a user selects a key phrase to adapt a robot to his/her voice. It might be a numeric password or something else. For that purpose, there has been used off-site package Loquendo ASR-Speaker Verification [10]. The main shortage was the fact, that a user for proper identification shall pronounce the same sentence he/she has used at registration stage [11]. At that, the people usually do not inform a robot such identification information during interaction. A robot might get a limited feedback information, whether the speaking identificatory has been true or not, observing or hearing speaker's emotional or behavioral reactions when an identified user calls a robot by name. That simple feedback cycle might be fulfilled through common interaction between people and robots. For example, that structure might be applied to robots-pets, which can distinguish the voices of their families' members. Every time, when people talk to the robots, a robot can upgrade its identification abilities [12].

Therefore, there has been taken a decision to develop speaker's voice identification system using computer-aided learning technologies based on neural network. Currently there is a big interest in developing and applying computer-aided learning algorithms to different science and engineering areas, including speech technologies. One of the examples is creation of the voice intellectual helper Alisa, having been developed by Yandex company and implemented by means of computer-aided learning technologies, including neural networks [13].

As the method of obtaining the features vector for training the neural network for solving the speaker voice identification task there has been selected the method based on mathematical statistical analysis of the fundamental voice tone. Such approach allows qualitatively reduce the vector dimensionality for training in the neural network, it does not demand much time for differentiating the unique speaker's features and it is described in some scientific works in the area of speech technologies and, in particular, for speaker's voice identification [14-16].

Into the voice identification system makeup there have been included the following information-bearing features, based on the fundamental voice tone statistics: upper quartile; lower quartile; interquartile range; median; arithmetic mean; minimal value; maximum value; entropy; excess; bias (asymmetry measure); dispersion; standard deviation; geometric average; harmonic mean.

In the course of the research herein there is being solved the task of selecting from the total multitude of accessible features such a set, which would maintain the highest accuracy of speaker identification according to the voice.

Likewise, analysis mathematical statistical techniques enter a composition of many software mathematical libraries for various development media. One of such libraries, containing all above enumerated methods is «Math.Net Numerics». It is directed to submission the methods and algorithms for numeric computations in science, engineering and in everyday usage. It includes some special functions, linear algebra, probabilistic models, random numbers, interpolation, integration, regression, optimization tasks, etc. It should be noted, that the library thereof is cross-platform and it operates in different operating systems and development, such as media Microsoft Visual Studio and Mono Develop [17, 18].

Further we will consider computer-aided learning libraries with a possibility to create the neural network for the speaker voice identification task solution.

II. METHOD OF RESEARCH

Selection of computer-aided learning algorithms and corresponding software libraries. Historically, an identification problem has appeared from classification task, conditioned with the necessity to separate some objects, which own similar

features, from other objects. The main process of every computer-aided learning method is a training selection. The bigger the selection, the more reliable the outcomes to be obtained, and as well it is possible to use more complex algorithms models [19].

To solve a speaker identification problem by means of computer-aided learning there has been used statistics of the fundamental voice tone. Statistics quantity is relatively small and has a finite sequence, consequently, a learning vector will have small dimensionality. Upon complication of perceptron structure there occurs learning duration increase in connection with the quantity of necessary hidden layers elements and weights correction of the relations between the elements. Under such conditions it is expedient to use a neural network architecture in the form of three-layered perceptron with one hidden layer. Such perceptron concept was offered in 1957 by an American scientist Rosenblatt F., and it was one of the earliest neural networks model [20]. Logic scheme of three-layered perceptron is shown on the Figure 2.

To solve a speaker identification task using the Person's Digital Voice Portrait (PDVP) there has been used a three-layer perceptron with one hidden layer. Let's denote a number of layers and neurons in a layer: N1 – neurons in the inlet layer, N2 – neurons in the hidden layer, N3 – outlet neurons, N - total number of layers in the network, including an inlet one, X – vector of network inlet signals, Y - vector of output signals.

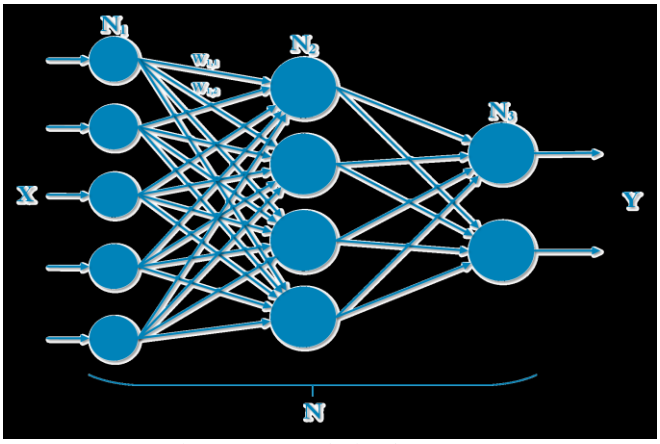

Fig. 2. Perceptron logic scheme

Algorithm of direct feedforward is described with the formulae:

$$S_{jl} = \sum_{i=1}^{N_{l-1}} w_{ijl} x_{ijl} \quad S_l = \begin{bmatrix} S_{1l} \\ \vdots \\ S_{N_l l} \end{bmatrix} \tag{1}$$

$$Q_{jl} = F_j(S_l) = \frac{e^{S_{jl}}}{\sum_{i=1}^{N_l} e^{S_{il}}}$$

$$x_{ij(l+1)} = Q_{il}$$

where an index i always will denote the input number, j — neuron number in the layer, l— layer number.

x_{ijl} - i-input signal of j-neuron in the layer l;

w_{ijl} – weighting factor of i-neuron input number j in the layer l;

S_{jl} - intermediate value of j-neuron in the layer l after the weighted sum of the given neuron input signals;

S_l - vector, consisting of all neuron signals in the layer l;

Q_{jl} – output signal of j-neuron in the layer l;

N_l – neurons quantity in the layer l;

Fj – j-value of SOFTMAX activation function.

For perceptron training there has been used an optimization method L-BFGS with a random first weights approximation. To compute the perceptron gradient error function there has been used the method of backpropagation.

Let's review computer-aided learning software libraries. We will give a concise description of advantages and disadvantages of enumerated in the survey software libraries. The main criteria will be the information about how many and which platforms (OS) are supported with a library, as well, programming languages support, whether the library is issued under the software open license (GPL, LGP, etc.) or under the closed license (proprietary), availability or absence of the documentation, supported functional and neuron networks types can be used through the library.

Review of computer-aided learning libraries. Proceeding from the conducted review of computer-aided learning libraries there has been compiled a comparison Table 1, describing the characteristics of the libraries having been surveyed. By reference to the statement of the task of speaker voice identification system development it has been decided to use features vectors for training the neural network based on the techniques of mathematical statistical analysis of the fundamental voice tone data, that is, to get the learning vector settled dimensionality, limited with the quantity of applied statistical methods. Therefore, the review has considered the computer-aided learning libraries, using neural networks architecture, based on multilayer perceptron with direct links. The most appropriate library for solving the given task, meeting the concurrent requirements to the quantity of supported programing languages and platforms, additional mathematical languages, built-in multithreading support, possibility of constructing the neural networks groups, availability of documentation and examples of using, has been selected an open library ALGLIB.

Let's set the computer-aided learning libraries comparison analysis:

1. Library ALGLIB:

Developer - ALGLIB Project;

License - PL and private license;

Platform - Windows, *nix similar;

Programing languages - C++, C#, Delphi, VB.NET, Python;

Functionality – support of additional mathematical functions, apart from neural networks (possibility of constructing the neural network assembly);

support of multithreading on payment basis; there is a free version under GPL license; substantial and detailed documentation with usage examples;

Multiprocessing or multithreading support - multithreading support in library's commercial version.

2.Library FANN:

Developer- Steffen Nissen –FANN library founder;

License – LGPL;

Platform - Windows, Linux;

Programing languages - Support about 15 programing languages; Functionality - flexible library with parameters big set for neural network learning; there is a graphic image mode of the program operation; concentrated only at the work and adjustment of neural networks, therefore there are no additional mathematical instruments; no built-in multithreading support;

Multiprocessing or multithreading support – No.

3. Library OpenNN:

Developer - Artificial Intelligence Techniques, Ltd;

License – LGPL;

Platform - Windows, Linux и MacOS;

Programing languages - C++. Functionality - available processor's parallelization by means of acceleration OpenMP and GPU with CUDA, only one programming language;

Multiprocessing or support of multithreading - Yes (OpenMP and CUDA).

Information system of speaker's voice identification. Identification procedure in different information system is a procedure in the result of executing it, there is detected user's unique identificatory in the information system, identifying it unambiguously [96]. Identification procedure should not be confused with authentication procedure. Result of speaker authentication procedure is assigning the user's identificatory in the system, in order to give a possibility to a user to go through the identification procedure. An identificatory in the developed system is a unique number (key) being stored in the Table of speakers' personal data. As the Table attributes there has been used the following set of speaker's personal data: ID; speaker surname; speaker's first name; speaker's date of birth; speaker's age; speaker's photo.

In the developed system of speaker's voice identification PDVP is used as a set of the fundamental voice tone data in the form of files of PDVP system format (Person's Digital Voice Portrait) [21]. For that purpose the speaker voice identification system addresses the system of forming the PDVP by means of the program operation terminal regime and sends the path to the speaker recorded voice file, receiving in reply the files in PDVP format, containing the filtered with CIS-filter, voiced areas of the fundamental voice tone. In the developed system of a speaker voice identification there are three main interfaces: speaker authentication interface, speaker voice identification interface, as well, the interface of identification model performance research under various fundamental voice tone statistics composition at the set of speakers voice patterns.

The technology of forming a digital voice portrait of a person was applied in the process of training the speaker's voice identification system. Figure 3 shows the structural diagram of the voice recognition system of the speaker.

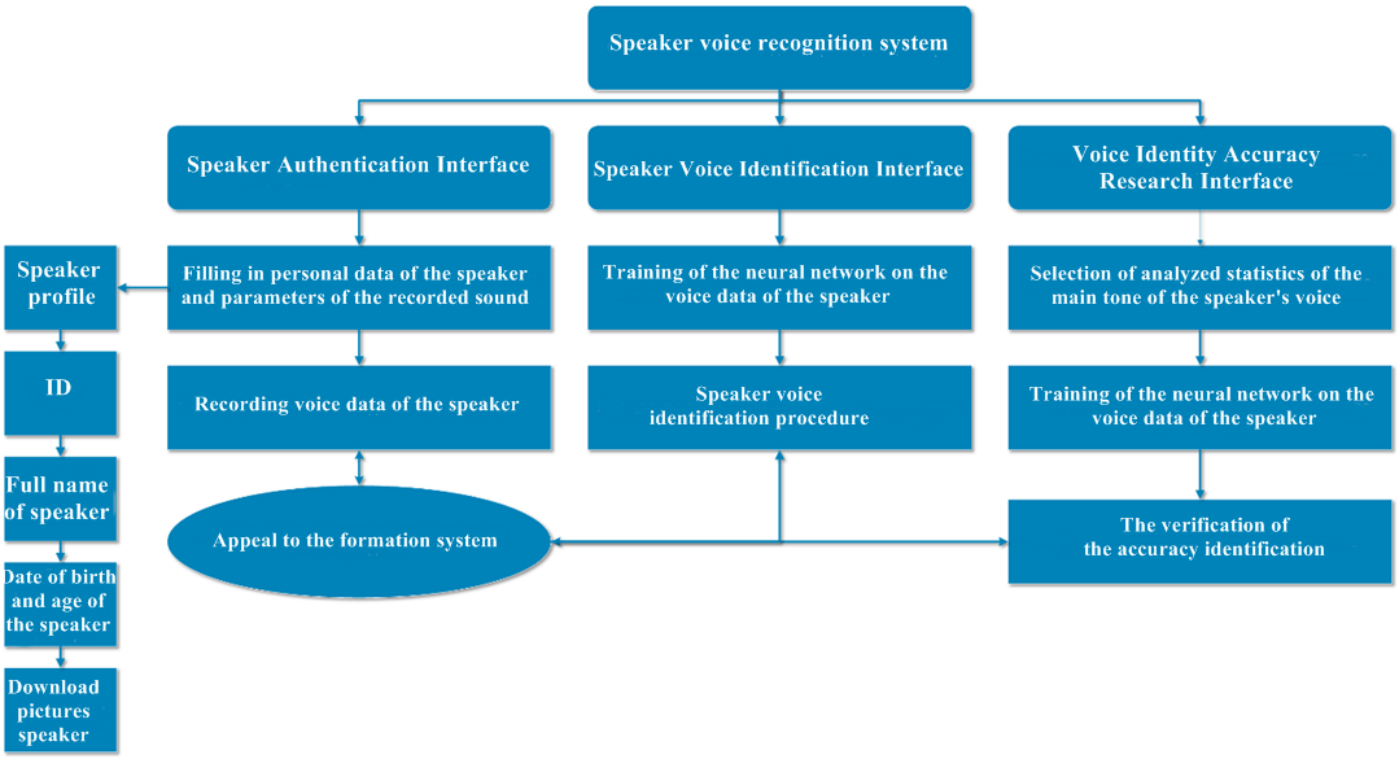

Fig. 3. Structural diagram of voice recognition system speaker

The result of the procedure are files containing statistics of the basic tone of the voice, which can be used as a neural network training vector. In the developed system, it is assumed that each speaker can dictate several sound patterns for which basic tone statistics will be calculated. This is necessary to build a learning matrix of a neural network. This matrix will be used to train the neural network during the speaker identification procedure.

Interface of identification model performance assessment at various fundamental voice tone statistics composition. When the model has been constructed, it is needed to assess the magnitude of its error at the test (or learning) set. If we talk about the classification problem, we can use two measures of error. The first and most widely used is the classification error (quantity or percentage of improperly classified cases). The second, not less known error measure is cross - entropy. In

ALGLIB package there is used an average cross-entropy per test set element, designed in bits (logarithm according to base 2). Using the average cross-entropy (instead of total cross-entropy) allows obtaining comparable scores for various test sets. Those error measures are widely known, and they do not need to be discussed [22].

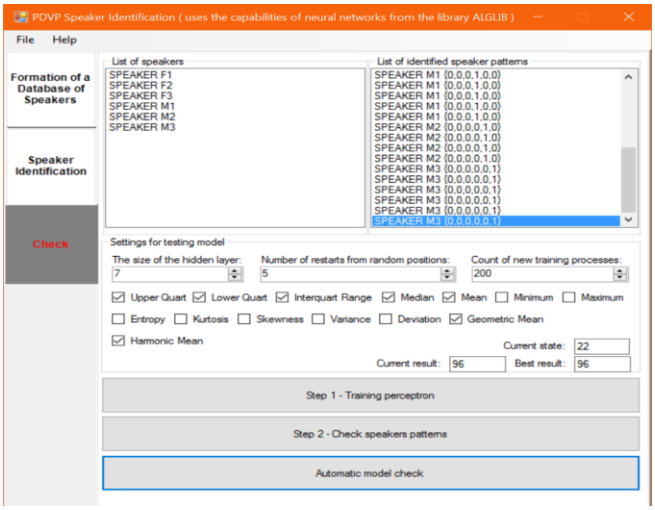

Fig. 4. Interface of identification model performance assessment

For percentage of improperly classified cases there has been developed a special interface of neural model identification performance assessment at different fundamental voice tone statistics composition of various speakers (Figure 4).

In the domain «List of speakers» there is a list of speakers, the voices of which have been used for investigating (Prefix F means, that it is a female speaker, and M – a male speaker). In the domain «List of identified speaker patterns» there are identified speakers voice patterns. Herein, we can calculate the percentage of improperly classified cases, if the speaker's voice pattern has been identified improperly.

In the domain «Settings for testing model» there assigned the parameters of neural network learning from the library ALGLIB, such as the quantity of hidden layers in the field «The size of hidden layer», in the field «Number of restarts from random positions», a number of randomized restarts from different random positions for one process of neural network learning. Number of independent processes of neuron network learning is assigned in the field «Count of new training processes». As well, in the given domain there can be fulfilled the selection of necessary fundamental voice tone statistics for research implementation. Current number of the model checking processes is displayed in the field «Current state», the percentage of correctly classified cases for the current checking process is shown in the field «Current result», the highest percentage of properly classified cases at all model checking processes is given in the domain «Best result».

For the model training process there has been used the set of voice sentences, recorded by the Center for Speech Technologies Research attached to Edinburgh university [23]. The set has been read by six different speakers, amongst them three female and three male speakers. RIFF voice files format, sampling rate 16 kHz, capacity 16 bits. Speech patterns amount for training– 60. Tested voice patterns – 30.

As a research object there used the fundamental voice tone statistics, enumerated below.

Used fundamental tone statistics: 1-upper quartile, 2-lower quartile, 3- interquartile range, 4-median, 5- arithmetic mean, 6-minimal value, 7-maximum value, 8-entropy, 9-excess, 10- bias (asymmetry measure), 11-dispersion, 12-standard deviation, 13-geometrical mean, 14-harmonic mean.

Training has been rendered at three-layers perceptron with one hidden layer. Hidden layer size constituted 100% off the input training vector dimensionality. Quantity of neural network randomized restarts from different random positions - 5. Number of independent checking processes - 200.

In the result of the conducted research on identification model performance assessment at various fundamental voice tone statistics makeup there have been obtained the following outcomes, which are demonstrated in the Table I and Figure 5.

TABLE I
OUTCOMES ON IMPROPERLY CLASSIFIED CASES AT VARIOUS STATISTICS MAKEUP OF THE SPEAKERS MAIN VOICE TONE IN PDVP STRUCTURE

Numbers of used statistics	Amount of hidden layer elements	Amount of randomized restarts of the neuron network from different random positions	Number of independent checking processes	Amount of improperly classified cases in %
Total	14	5	200	24
1, 2, 3, 4, 5, 13, 14	7	5	200	4
1, 2, 3, 4, 5, 6, 7, 13, 14	9	5	200	14
6, 7, 8, 9, 10, 11, 12	7	5	200	37

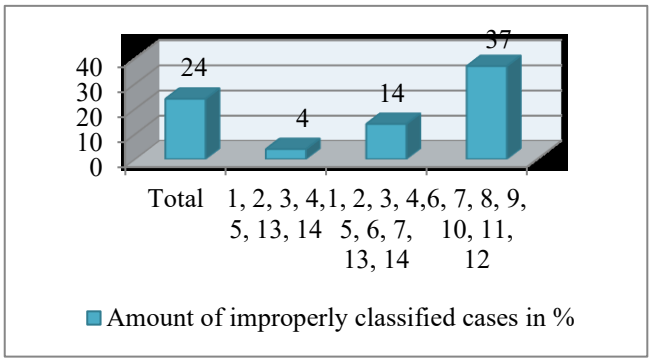

Fig. 5. Scheme of results of improperly classified cases at various statistics makeup of the speaker main voice tone the PDVP has been formulated for

In compliance with the obtained outcomes, we can make a conclusion, that the speaker fundamental voice tone statistics group, based on computing the average values (upper quartile, lower quartile, interquartile range, median, arithmetic mean, geometrical mean, harmonic mean) presents the lowest error upon the speaker identification, equal to four percent. Apparently, it can be related to a little adulteration of different voices in the voice files set, recorded by the Center for Speech Technologies Research attached to Edinburgh university. For instance, for the statistics, based on computing the average values, presence of outside male voice upon recording the

female one does not change much an identification result, comparing to the statistics, returning F0-circuit minimum or maximum values.

CONCLUSION

The article herein contains the information system description of a speaker automatic identification according to the voice to apply to robotic-verbal systems. Previously, there has been conducted computer-aided learning libraries and algorithms surveys and selected the most appropriate according to necessary criteria ALGLIB. There has been carried out the research of identification model operation performance assessment at various fundamental voice tone statistics. As an accuracy criterion there has been used the percentage of speaker identification improperly classified cases. According to the obtained outcomes, we can make a conclusion, that speaker fundamental tone statistics group, based on computing the average values (upper quartile, lower quartile, interquartile range, median, arithmetic mean, geometrical mean, harmonic mean) provides acceptable results on speaker identification accuracy. Classification available error of 4% is probably conditioned with availability of background noise and speech constituent in the analyzed voice signal. It has been detected, that for the statistics, based on computing the average values the presence of outside male voice upon analysis of a female one does not change much the identification result, comparing to the usage of statistics, returning the F0-circuit minimum or maximum values. Combination of information-bearing parameters from extended PDVP makeup, having provided the highest accuracy of speaker identification has been included into the final optimized PDVP composition.

The work has been executed under support of MES RK grant #AP05132648 «Creation of verbal-interactive robots based on advanced speech and mobile technologies», being fulfilled at the Institute of Information and Computer technologies SR MES RK, contract # 211 dated 19.03.2018.

REFERENCES

[1] J. P. Campell and Jr., Speaker Recognition: A Tutorial, Proceeding of IEEE, vol. 85, pp. 1437–1462, (1997).

[2] Osman Buyuk and Lavent M. Arslan, HMM-based Text-dependent Speaker Recognition with Handset-channel Recognition, IEEE ICSPCA, pp. 383–386, (2010).

[3] D. A. Reynolds and R. C. Rose, Robust Text-independent Speaker Identification using Gaussian Mixture Speaker Models, IEEE Transaction on SAP, vol. 3, no. 1, pp. 72–83, (1995).

[4] R. E. Wohiford, E. H. Jr. Wrench and B. P. Landell, A Comparison of Four Techniques for Automatic Speaker Recognition, Proceedings of IEEE ICASSP, vol. 5, pp. 908–911, (1980).

[5] B. Atal, Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification, The Journal of the Acoustical Society America, vol. 55, pp. 1304–1312, (1974).

[6] Sangeeta Biswas', Shamim Ahmadt and Md Khademul Islam Molladt, Speaker Identification using Cepstral Based Features and Discrete Hidden Markov Model, Proceedings of IEEE ICICT, pp. 303–306, (2007).

[7] Latha, Robust Speaker Identification Incorporating High Frequency Features, Procedia Computer Science, vol. 89, 2016, pp. 804-811.

[8] https://ru.wikipedia.org/wiki/ Speech recognition.

[9] F. Alonso-Martin, J. F. Gorostiza, M. Malfaz, and M. Salichs. Multimodal Fusion as Communicative Acts during Human-RobotInteraction. Cybernetics and Systems, 44(8):681–703, 2013.

[10] E. Dalmasso, F. Castaldo, P. Laface, D. Colibro, and C. Vair. Loquendo - Speaker recognition evaluation system. In Acoustics, Speech and Signal Processing, ICASSP 2009. IEEE

[11] F. Alonso Martin, A. Ramey, M. A. Salichs. Speaker identification using three signal voice domains during human-robot interaction. HRI'14. 2014.

[12] Y. Kida, H. Yamamoto, C. Miyajima, K. Tokuda, T. Kitamura. Minimum Classification Error Interactive Training for Speaker Identification. Proceedings. (ICASSP '05). 2005.

[13] Alisa (voice helper) // https://ru.wikipedia.org/wiki/Alisa: 24.11.2017

[14] Kovalj S.L., Labutin P. V., Malaya Ye. V., Proshina Ye. A. Speakers identification based on the main voice tone statistic comparison // Informatization and information security of law-enforcement agencies: proceedings of the XV International scientific conference — M.: Russia Ministry of the Interior Academy of management, 2006. –p.p. 324–327.

[15] Bulgakova Ye.V., Sholokhov A.V., Tomashenko N.A. Speakers identification method based on phonemes length statistics comparison // Scientific-technical vestnik of information technologies, mechanics and optics. –2015. – No 1. – p.p. 70–77.

[16] Lukiyanov D. I., Mikhailova A. S. Human being automatic identification according to the voice using an algorithm based on Gaussian mixtures model// Vestnik of RSRTU. – 2017. – No 61. – p.p. 19-24.

[17] Math.NET Numerics // https://numerics.mathdotnet.com/: 28.07.2017.

[18] Statistics – Math.NET Numerics Documentation. Extension methods to return basic statistics on set of data // https://numerics.mathdotnet.com/api/ MathNet.Numerics.Statistics/Statistics. htm:24.11.2017.

[19] Vetrov D.P., Kropotov D.A. Bayesian method of computer-aided learning. – Study guide – M., 2007. – 132 p.

[20] Glushkov V.M., Amosov N.M., Artyemenko I.A. Cybernetics encyclopedia. Volume 2. – K.: Main office of Ukrainian soviet encyclopedia, 1974. – 624 p.

[21] Mussabayev R.R., Amirgaliyev Ye. N., Tairova A.T., Mussabayev T.R., Koibagarov K. Ch. The technology for the automatic formation of the personal digital voice pattern // 10th IEEE International Conference on Application of Information and Communication Technologies (AICT). – Azerbaijan, Baku, 2016. – P. 422-426

[22] General concepts. Library of algorithms ALGLIB // http://alglib.sources.ru/dataanalysis/generalprinciples.php: 18.08.2017.

[23] Full set of sentence recordings for downloading. The Centre for Speech Technology Research // http://www.cstr.ed.ac.uk/projects/eustace/download.html: 25.08.2017.