

Neuro-heuristic voice recognition

Dawid Połap Institute of Mathematics Silesian University of Technology Kaszubska 23, 44-100 Gliwice, Poland Email: Dawid.Polap@gmail.com

Abstract—Protection of private data, signing electronic documents are selective use of identity verification. In this work, the problem of voice verification has been discussed. The method of verifying the voice based on the methods of artificial intelligence was presented. Numerous tests were performed to demonstrate the effectiveness of the presented solution - research results are shown and discussed in terms of advantages and disadvantages.

I. INTRODUCTION

THE digitization of today's world has started some time ago. At almost every step we associate with the technology in one form or another. Each one of us carries a cell phone or laptop and therefore has continuous access to the Internet. On the Internet, people handle all the daily issues buy goods, pay bills or store images and a variety of data. These are the times in which it is hard to remain anonymous what requires continuous improvements in data transmission and data protection.

One of the most advanced trends in computing today is artificial intelligence, which finds its use in almost any application. One of the first approaches of this trend were neural networks, which are mainly used as classifiers, i.e. classification of signatures [1], surface structures [2] and multimedia applications [3]. Modern methods of mathematics and artificial intelligence are increasingly finding application in medicine and other aspects. In [4] and [5], the authors presented an interesting approach to the processing of EEG signals. A very important achievement is also fuzzy logic, which is often combined with neural networks in order to not only increase precision, but allow the classification of other measures such as linguistics [6] and [7]. Similarly we can use voice recognition in AAL environments to help users [8] and improve data acquisition [9]. These aspects of artificial intelligence methods are also widely used in verification of people. On a daily basis, we meet with the problem of verification in banks, stores (e.g. when signing checks), companies or in different institutions for which a signature, voice, or even features of the iris are a confirmation of our identity. Voice verification is one of the least expensive in implementation of solutions of this type – just a microphone and expert application are needed. Different approaches are analyzed in order to maximize the precision of the voice classification. In [10], the authors presented the statistical approach, again in [11] used hidden Markov models for the same purpose. In this work, I would like to introduce an innovative way of extracting features of the voice sample using heuristic and neural classifier.

II. AUDIO SIGNAL PROCESSING

For the purposes of analysis of digital sound samples, sound file must be represented by a numerical value, or a function that enables analysis. One of the best known methods is the Discrete Fourier Transform (DFT) presented in [12]. Transform is called transformation of vector that represents the numerical values of the signal $\vec{x} = [x_0, x_1, \dots, x_{N-1}]$ into $\vec{z} = [z_0, z_1, \dots, z_{N-1}]$, where the value of $z_i \in \mathbb{C}$. DFT is performed according to the following formula

$$z_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n \exp(\frac{-2kni\pi}{N}),$$
 (1)

where k is harmonic number, n is number of signal samples and N is total number of samples.

One way to display the audio signal is to use a spectrogram [13] which is a graph of amplitude spectrum signal over time. The construction of the spectrogram operates on the principle of dividing the signal (obtained by the use of short-time fast Fourier transform with Hamming function) on parts for which the amplitude of harmonic components are calculated. During the analysis of this graph, we assume that

- the power at a given frequency is based on the color the value is higher if the color is warmer;
- the frequency increases with the height of the point on the chart.



Fig. 1: Graphical representation of sound samples with the sentence "My name is Han Solo": (a) the DFT of the signal (b) spectrogram of the recorded sound with noise (c) spectrogram of the clear signal.

Recorded sound is exposed to register all types of noise in the case of voice verification in a large company, people standing in a queue already make noise as well as any other voices. Noise is clearly visible on spectrograms interiors as a plurality of light color. In order to correct analysis of sound, noise should be removed to leave only the desired recording. Noise reduction can be achieved by a multi-band spectral subtraction described in [14]. The algorithm assumes that speech spectrum will be divided into N bands and the estimate of the clean speech spectrum can be obtained by performing the following formula for each band i for k value

$$\hat{S}_{i}(k)|^{2} = |Y_{i}(k)|^{2} - \alpha_{i}\delta_{i}|\hat{D}_{i}(k)|^{2} \qquad l_{i} \le k \le h_{i}, \quad (2)$$

where S means the magnitude spectra of the clean speech and \hat{D} of the noise, Y is the magnitude spectra of the incoming signal, l_i is the beginning frequency of the *i*th band and similarly h_i is the ending frequency, δ_i is a tweaking parameter which is set by empirical way. The parameter α_i can be calculated by

$$\alpha_i = \begin{cases} 5 & \psi_i < -5 \\ 4 - 0.15 * \psi_i & -5 \le \psi_i \le 20 \\ 1 & \psi_i > 20 \end{cases}$$
(3)

where

$$\psi_i = 10 \log_{10} \left[\left(\sum_{k=l_i}^{h_i} |Y_i(k)|^2 \right) / \left(\sum_{k=l_i}^{h_i} |\hat{D}_i(k)|^2 \right) \right].$$
(4)

The samples of this methods to process audio file are shown in Fig. 1. Another way to present the audio signal is a periodogram, which for the first time was shown by Arthur Schuster [15]. It is obtained by the modulus squared of the DFT and presents the Power Spectral Density (PSD) estimate, as shown in Fig. 3.

III. PREPROCESSING FOR NEURAL NETWORK

Creating samples is an important part for the pattern recognition problem. Each sample must not only represent input data but save as much information about the sample using the fewest number of values. For that purpose, a model of the general pattern based method on a large number of input samples is proposed. All samples representing the specific person will be created using general pattern.

A. Preparation of the aggregate sample

In the first step of processing sound samples, spectrograms are created. Each spectrogram is subjected to noise removal. On the basis of all samples, the aggregated sample is prepared. This process is about creating one sample spectrogram, which retain repeating features in all processed ones. At the beginning of the method, the $w \times h$ array is created, where w is the width and h the height of all samples - each point (x, y) on spectrogram is corresponding to one cell in the array. Initially, each cell is set to 0. For each point (x, y) color pixel on each bitmap is verified. In the case where the pixel is not white, the value on this position (x, y) is increased by 1 in the array. Color $[R_{old}, G_{old}, B_{old}]$ is selected and the new sample is created. For each pixel, the color is calculated using a value of m from an array and the following formula

$$C_{new} = C_{old} * (1 - m\varsigma), \tag{5}$$

where C is one of the color components (R, G or B) and ς means the color shade and it is calculated as

$$\varsigma = \frac{2}{\max_{\substack{0 \le i < w \land 0 \le j \le h}} m_{ij}}.$$
(6)

In the next section of this paper, created spectrogram will be called a *general spectrogram*. The process of creating a *general spectrogram* is shown in Fig. 2.



Fig. 2: Visualization of creating a *general spectrogram* based on input spectrograms with attached legend of colors - the darker shade of red, the point often occurs on the input samples.

B. Heuristic detection of key-points

Using the so-created *general spectrogram*, heuristic algorithm will be used to find the key points of the image. The coordinates of key points will allow to create the mask of the features by which we will be able in a quick and effective way to retrieve specific values from the spectrogram to perform training using verification vectors.

As a heuristic algorithm, Flower Pollination Algorithm (FPA) [16] was selected, which is a mathematical model of a natural phenomenon pollination of flowers in the spring. The original algorithm assumes some basic rules aimed at simplifying some dependence

- Global pollination (e.g.: biotic phenomenon) is represented as Levy flights,
- Local pollination is interpreted as abiotic and selfpollination,
- Pollen is carried by the wind what is modeled by a random factor $p \in [0, 1]$.

FPA is used to find the important points in the image. Population of flowers is placed in random places on *general spectrogram*. Then, the simulations of global and local pollination are performed in order to find the best points. Point selection is done according to the fitness function, which is defined as

$$F(x_i) = \begin{cases} \min_{\substack{0 \le j < N \\ \sqrt{B^2(x_i) - B^2(x_{best})}}} & \text{if } x_{best} = 0\\ \sqrt{B^2(x_i) - B^2(x_{best})} & \text{for others} \end{cases}$$
, (7)

where x_i is a point representing a pollen, x_{best} is a point with the best fitness value in actual iteration, and the function $B(\cdot)$ is the brightness of a pixel that takes values of [0, 1], N is the size of population.

In each iteration, two operations are performed – global and local pollination. Global pollination moves pollen x_i over the spectrogram image according to

$$x_i^{t+1} = x_i^t + L(x_i^t - x_{neighboring}), \tag{8}$$

where t is the number of iteration, $x_{neighboring}$ is the nearest point to x_i and $L(\cdot)$ is a function of Levy flight understood as

$$L(x,\kappa,\mu) = \sqrt{\frac{\kappa}{2\pi}} \frac{e^{-\kappa/(2(x-\mu))}}{(x-\mu)^{3/2}},$$
(9)

where κ and μ are specified parameters.

The second operation is a local pollination which takes place over the neighborhood pixels for a given point x_i and it is defined as

$$x_{i}^{t+1} = x_{i}^{t} + \epsilon (x_{j}^{t} - x_{k}^{t}),$$
(10)

where x_j and x_k are neighboring pollens.

The algorithm returns the best adapted solution with its motion path. From a mathematical point of view, the trajectory is a closed consisting of two-dimensional coordinates of key points found on the spectrogram. Created set will allow to get the appropriate features based on the spectrogram in a very short time. Moreover, for every person, FPA will find a different trajectory, and therefore the possibility of fraud during the verification respectively decreases.



Fig. 3: Periodogram of the sentence "Han Solo" with the selected maximum value of PSD.

C. A general vector for identification

In order to analyze every sound, a file should be saved as a vector of numbers representing as much as possible features. In addition, the vector should contain the minimum number of values, because then a learning process occurs much more efficient. The proposed model in the general formula for a vector in this work is the combination of the features extracted from the spectrogram and the periodogram using the method described in Sec. III.

The proposed model of the sample can be defined as follows

$$[B(x_0), B(x_1), \dots, B(x_n), f_{max}, id],$$
(11)

where $B(x_i)$ is brightness value of pixel at the key-point x_i obtained from FPA, f_{max} is the maximum value of the PSD determined from periodogram (see Fig. 3) and *id* is the designation of the owner of the sample.

IV. NEURAL NETWORK

The beginning of the history of neural network models is dated to the first half of the twentieth century [17]. Until today, new and modified models of the network, or even learning algorithms of this type of networks are created. A neural network is understood as a complex object composed of many layers. The simplest models contain only two - input and output layer. In this type of layers, learning vector is inserted into input layer and the calculation result is forwarded to the next layer. In more complex structures the hidden layers are added between input and output. Their number depends on the size of the problem, for which a structure has been created. Each layer is made of neurons through which the layers are connected - each neuron of one layer is connected to each neuron in the next one. Neurons use activation function, whereby the value is calculated. In [18], different activation functions were proposed, but the most common is a unipolar sigmoid function. Moreover, each connection between two neurons is burdened with a certain weight in the range of [0,1] and this contributed to a number of learning methods for improving received results.

One method of learning neural network is the backpropagation algorithm for the first time mentioned in [19]. To this day, it is one of the most commonly used training algorithms for neural networks. It is a type of supervised learning which is based on minimizing the error function of the output layer up to the achievements of the first hidden layer. The error calculated in this way is used to modify the weights on the connections between neurons. The formula for calculating an error for the neuron k is defined as follows

$$\delta_k = \begin{cases} out_k(1 - out_k)(ex_k - out_k) & \text{the output layer} \\ out_k(1 - out_k) \sum_{o \in out} w_{ok} \delta_k & \text{the hidden layer} \end{cases},$$
(12)

where *out* means the output values from neuron k, ex is the expected value at the output of the neuron. The calculated error values are used in updating the weights by

$$w_i = w_i + \delta w_i. \tag{13}$$

V. EXPERIMENTS

Using the method of pre-processing sound samples described in Sec. II and III are processed. In experimental research 100 different sound files belonging to four people were used. During the processing, the following parameters were used

• FPA – 100 flowers, 15 iterations, $\kappa = 0.4$, $\mu = 0.35$, After creating samples, obtained vectors have been placed in a database and used for training the network until the error value was smaller than 0.1 with the proportion of mixed samples – 80% to train and 20% to verify. Subsequently, using all samples for examined network an accurate measurements were

TABLE I: Average accuracy for the identification

Id	Correctly classified	Incorrectly classified	Accuracy
1	14	11	56%
2	20	5	80%
3	23	2	92%
4	18	7	72%

The results allow to calculate the effectiveness of the verification network, which reaches 75% with 25 samples of voice for one person. The results in the table indicate that for a person no. 1 efficacy haughty only 56% accuracy, again for a person no. 3 it was 92%. Such a large discrepancy between the results can be an effect of a bad sound recording or different voice tones. Another problem that might be a big obstacle in good verification is possible jitters or hoarseness. This can be accomplished by recording the samples for a few days—not only by one day.

VI. FINAL REMARKS

In times where almost everything revolves in the digital world, any method of data security become an important area of IT. Newer and more complex algorithms can increase the security not only of information in computer networks, but many companies, where the entrance is authorized.

The innovative idea of the features extraction from the processed audio samples is presented in this work. Experiments based on artificial intelligence methods have been performed and discussed. The proposed method of voice verification due to the 75% effectiveness is a good alternative to existing methods. Moreover, the use of heuristics to find a sequence of points can efficiently assist in verification of possibility of fraud - each sequence of points is different because of the basic assumption of randomness for heuristics. Therefore proposed solution seems to be right development to increase security in man-machine interactions.

VII. ACKNOWLEDGMENT

Author acknowledge contribution to this project of Operational Programme: "Knowledge, Education, Development" financed by the European Social Fund under grant application POWR.03.03.00-00-P001/15.

REFERENCES

 D. Połap and M. Woźniak, "Flexible neural network architecture for handwritten signatures recognition," *International Journal of Electronics* and *Telecommunications*, vol. 62, no. 2, pp. 197–202, 2016, DOI: 10.1515/eletel-2016-0027.

- [2] G. Capizzi, G. Lo Sciuto, C. Napoli, E. Tramontana, and M. Woźniak, "Automatic classification of the fruit defects based on co-occurrence matrix and neural networks," in *Proceedings of the Federated Conference on Computer Science and Information Systems - FedCSIS*'2015. 13-16 September, Lodz, Poland: IEEE, 2015, pp. 861–867, DOI: 10.15439/2015F258.
- [3] M. Knop, T. Kapuscinski, W. K. Mleczko, and R. A. Angryk, "Neural video compression based on RBM scene change detection algorithm," *Lecture Notes in Artificial Intelligence - ICAISC'2016*, vol. 9693, pp. 660–669, 2016, DOI: 10.1007/978-3-319-39384-1_58.
- [4] R. Damaševičius, M. Vasiljevas, I. Martišius, V. Jusas, D. Birvinskas, A. Venčkauskas, and M. Woźniak, "BoostEMD: an extension of EMD method and its application for denoising of EMG signals," *Elektronika IR Elektrotechnika*, vol. 21, no. 6, pp. 57–61, 2015, DOI: 10.5755/j01.eee.21.6.13763.
- [5] I. Martisius, D. Birvinskas, R. Damasevicius, and V. Jusas, "EEG dataset reduction and classification using wave atom transform," in *Proceedings* of the 23rd International Conference on Artificial Neural Networks and Machine Learning; ICANN 2013 - Volume 8131. New York, NY, USA: Springer-Verlag New York, Inc., 2013, pp. 208–215, DOI: 10.1007/978-3-642-40728-4_26.
- [6] K. Cpałka, O. Rebrova, R. Nowicki, and L. Rutkowski, "On design of flexible neuro-fuzzy systems for nonlinear modelling," *International Journal of General Systems*, vol. 42, no. 6, pp. 706–720, 2013.
- [7] C. Napoli, G. Pappalardo, E. Tramontana, R. K. Nowicki, J. T. Starczewski, and M. Woźniak, "Toward automatic work groups classification based on probabilistic neural network approach," *Lecture Notes in Artificial Intelligence - ICAISC'2015*, vol. 9119, pp. 79–89, 2015, DOI: 10.1007/978-3-319-19324-3_8.
- [8] R. Damaševičius, M. Vasiljevas, J. Salkevicius, and M. Woźniak, "Human activity recognition in aal environments using random projections," *Computational and Mathematical Methods in Medicine*, vol. 2016, pp. 4073 584:1–4073 584:17, 2016, DOI: 10.1155/2016/4073584.
- [9] M. Woźniak, D. Połap, R. K. Nowicki, C. Napoli, G. Pappalardo, and E. Tramontana, "Novel approach toward medical signals classifier," in *IEEE IJCNN 2015 - 2015 IEEE International Joint Conference on Neural Networks, Proceedings.* 12-17 July, Killarney, Ireland: IEEE, 2015, pp. 1924–1930, DOI: 10.1109/IJCNN.2015.7280556.
- [10] Y. Zhang, S. Sankaranarayanan, and F. Somenzi, "Statistically sound verification and optimization for complex systems," in *Automated Tech*nology for Verification and Analysis. Springer, 2014, pp. 411–427.
- [11] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, *IEEE*, vol. 29, no. 6, pp. 82–97, 2012.
- [12] S. Winograd, "On computing the discrete fourier transform," *Mathematics of computation*, vol. 32, no. 141, pp. 175–199, 1978.
- [13] J. L. Flanagan, Speech analysis synthesis and perception. Springer Science & Business Media, 2013, vol. 3.
- [14] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE international conference on acoustics speech and signal processing*, vol. 4. Citeseer, 2002, pp. 4164–4164.
- [15] A. Schuster, "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena," *Terrestrial Magnetism*, vol. 3, no. 1, pp. 13–41, 1898.
- [16] X.-S. Yang, "Flower pollination algorithm for global optimization," in Unconventional computation and natural computation. Springer, 2012, pp. 240–249.
- [17] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.
- [18] P. Sibi, S. A. Jones, and P. Siddarth, "Analysis of different activation functions using back propagation neural networks," *Journal of Theoretical and Applied Information Technology*, vol. 47, no. 3, pp. 1264–1268, 2013.
- [19] P. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," 1974.

calculated in Tab. I.