Voice Recognition by Neuro-Heuristic Method

Dawid Połap* and Marcin Woźniak

Abstract: The amount of digital data is increasing every day. At every step of our daily lives, we deal with technologies in which our data are stored (e.g., mobile phones and laptops), and this is one of the main reasons for the design of various types of encryption and user identity verification algorithms. These algorithms are meant not only to fulfill the desire of protecting data but also to address the possibility of granting access of specific digital data to selected individuals. This process brings with it the problem of identity verification. This paper discusses the problem of voice verification and presents a voice verification method based on artificial intelligence methods. Numerous tests are performed herein to demonstrate the effectiveness of the presented solution. The research results are shown and discussed in terms of the advantages and disadvantages of the solution.

Key words: neural network; heuristic algorithms; Discrete Fourier Transform (DFT); image processing

1 Introduction

Technology is involved in various aspects of computer and multimedia systems. Digital information can be processed in computer systems using advanced methods of computational intelligence. People nowadays are using smart phones, televisions, laptops, tablet computers, etc., which are all efficient multimedia processors. Moreover, people handle all daily transactions on the Internet, such as buying goods, paying bills, booking holiday trips, communicating, storing images, and accessing entertainment. In these times, we are almost unable to stay anonymous; therefore, continuous improvements in digital protection and access verification methods are necessary.

Today, one of the most advanced trends in computing is artificial intelligence, which can be efficiently implemented into security and communication in

multimedia systems. A survey of multimedia methods can be found in Ref. [1]. Neural networks are one of these methods that were widely applied as classifiers and co-work with other computational techniques in user verification systems. Security and safe communication are important trends in computer science. Various devoted systems have been developed to support safe communication[2], especially for mobile communication where developed architectures must support high requirements from users[3]. Each day, we face verification problems in banks, stores (e.g., when signing checks), companies, or in various public institutions where a signature, voice, or feature of the iris can confirm our identity. People have been using personal signatures to confirm their identities since the invention of writing. Correspondingly, the format of such signatures has also developed with time, from traditional wax stamps and feathers with inks in the medieval times to pens and now electronic devices. Based on similarity of the features of signatures, we can determine whether a particular signature belongs to a particular person. However, to perform verification, some initial operations must be introduced. In most cases, input data should be normalized for the classifier input. One of the latest solutions to this problem is to use Dynamic Time Warping (DTW)[4], various

● Dawid Połap and Marcin Woźniak are with Institute of Mathematics, Silesian University of Technology, Kaszubska 23, Poland. E-mail: Dawid.Polap@gmail.com; Marcin.Wozniak@polsl.pl.
∗ To whom correspondence should be addressed.

algorithms, and neural network classifiers[5].

Another way of verifying the identity of a person is to examine the features of his iris or fingerprint, which are almost impossible to forge because each human is born with particular shapes of fingerprints, colors of iris, etc., which comprise our natural features. However, this type of identity check is rarely used in practice because of high production costs and complicated implementation. Therefore, research on improvements of any idea and technology is important. A learning pairwise filter for iris verification was presented in Ref. [6]. The authors in Ref. [7] discussed an approach to a bimodal biometric verification system based on fingerprint and face recognition.

Voice verification is another method of solving the user identification problem. Compared with sophisticated methods and devices for iris verification, voice identity is a fairly cheap solution in terms of implementation. For this method, a microphone and a trained expert application are required to classify sounds. The core for this technology is an efficient voice classifier. Each sample sound must be processed to compare certain features that characterize a particular voice. To do this, we have developed devoted models based on intelligent technologies. Different approaches can influence efficiency and maximize the precision of classification. In Ref. [8], the authors presented a statistical approach to maximize classification, whereas those in Ref. [9] used hidden Markov models. The present study discusses an innovative and devoted combination of neural networks and a heuristic technique to extract the most important features of a voice sample to be processed by a developed classifier.

2 Audio Signal Processing

Before the voice sample analysis, we first transform the voice using a function that enables more efficient analyses. The Discrete Fourier Transform (DFT) proposed in Ref. [10] is one of the best transforms used to adapt a sample into numerical values suitable for various verification methods. DFT uses the sample transformation $x = [x_0, x_1, \ldots, x_{N-1}]$ into a vector $z = [z_0, z_1, \ldots, z_{N-1}]$ which represents numerical values of a signal, where the values are the complex and can be written as $z_i \in \mathbb{C}$. DFT uses the following formula:

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

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

Audio signals can be displayed using a spectrogram[11], which is a graph of the amplitude spectrum signal over time. The spectrogram is constructed using the principle of dividing the signal (obtained using short-time fast Fourier transform with Hamming function) into parts, for which the amplitudes of harmonic components are calculated. We make the following assumptions during the analysis:

• The color corresponds to the power at a given frequency: the warmer the color, the higher the value;

• The frequency is higher with higher points of the chart.

Sound is exposed to various noises during recording, and these noises will be reduced (i.e., voice identification in a company can be conducted in a crowd of workers waiting at the check desk while there is background noise that includes talking and sounds from office devices). For correct sound analyses, each noise should be removed to leave only the voice for verification. We chose to use the spectrogram because the voice is clearly visible on this chart in a plurality of light color. With this, noise reduction can be achieved by implementing the multi-band spectral subtraction proposed in Ref. [12]. The discussed algorithm assumes that the speech spectrum will be divided into N bands while clear sound from the speech spectrum will be obtained for each band i of k values using the following formula:

$$|\hat{S}_i(k)|^2 = |Y_i(k)|^2 - \alpha_i \delta_i |\hat{D}_i(k)|^2, \quad l_i \leqslant k \leqslant h_i \quad (2)$$

where $\hat{S}$ is the magnitude spectra of the clean voice, $\hat{D}$ is the magnitude spectra of the noise, Y is the magnitude spectra of the incoming signal, l_i is the initial frequency of the i-th band, h_i is the final frequency, and δ_i is a tweaking parameter mainly set in empirical manner. For this equation, the α_i parameter is calculated as follows:

$$\alpha_i = \begin{cases} 5, & \psi_i < 5; \\ 4 - 0.15\psi_i, & 5 \leqslant \psi_i \leqslant 20; \\ 1, & \psi_i > 20 \end{cases} \quad (3)$$

where

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

We made an initial study of the efficiency of the proposed method after dividing our model space into smaller subsets. The best results for such values for above equations were obtained in the tests. Hence, we present our ideas for the values received as the result of our research. Figure 1 depicts the samples processed using this method from the initial audio file.

Similarly, we implemented periodograms to present the input audio signals. This approach was developed by Arthur Schuster in 1898[13]. Each of the periodograms was obtained using the modulus squared of the DFT projected to present the Power Spectral Density (PSD) estimation. Figure 2 plots the sample result.

3 Preprocessing for Neural Network

Preprocessing samples is very important for the efficiency of pattern recognition systems. This process must be flexible enough to match various input samples. At the same time, it is necessary to preserve as much important information about the input as possible using the lowest possible number of values to improve the recognition speed. For this purpose, we propose a general model for pattern recognition, where the input samples describing the same input are generalized into one pattern that takes the information of all the most important features.

3.1 Aggregated sample composition

We calculated the spectrograms for each input sample at the beginning of processing. Afterwards, the noise from each spectrogram was removed. We then introduced the developed method to aggregate samples into a general pattern preserving all most important information (Fig. 3).

The array $w \times h$ was composed at the beginning of the proposed method, where w represents the width and h denotes the height of all samples. Each cell in this array exactly matched one point (x, y) of the aggregated spectrograms. Each cell value was initially set to 0. We then verified the color of each point (x, y) in the spectrogram. If the color is other than white, the value corresponding to this position in the aggregated array is increased by 1. We selected color $[R_{old}, G_{old}, B_{old}]$, for which a new sample was composed using the proposed model. For each pixel, a new color was calculated using the value of m from the aggregated array in the following formula:

$$\begin{cases} R_{new} = R_{old} \cdot (1 - m_\varsigma), \\ G_{new} = G_{old} \cdot (1 - m_\varsigma), \\ B_{new} = B_{old} \cdot (1 - m_\varsigma) \end{cases} \quad (5)$$

where ς is the color shade calculated as follows:

$$\varsigma = \frac{2}{\max\limits_{0 \leqslant i < w, \, 0 \leqslant j < h} m_{ij}} \quad (6)$$

The recomposed spectrogram was called the general spectrogram. Figure 3 depicts the schematic model of the general spectrogram composition.
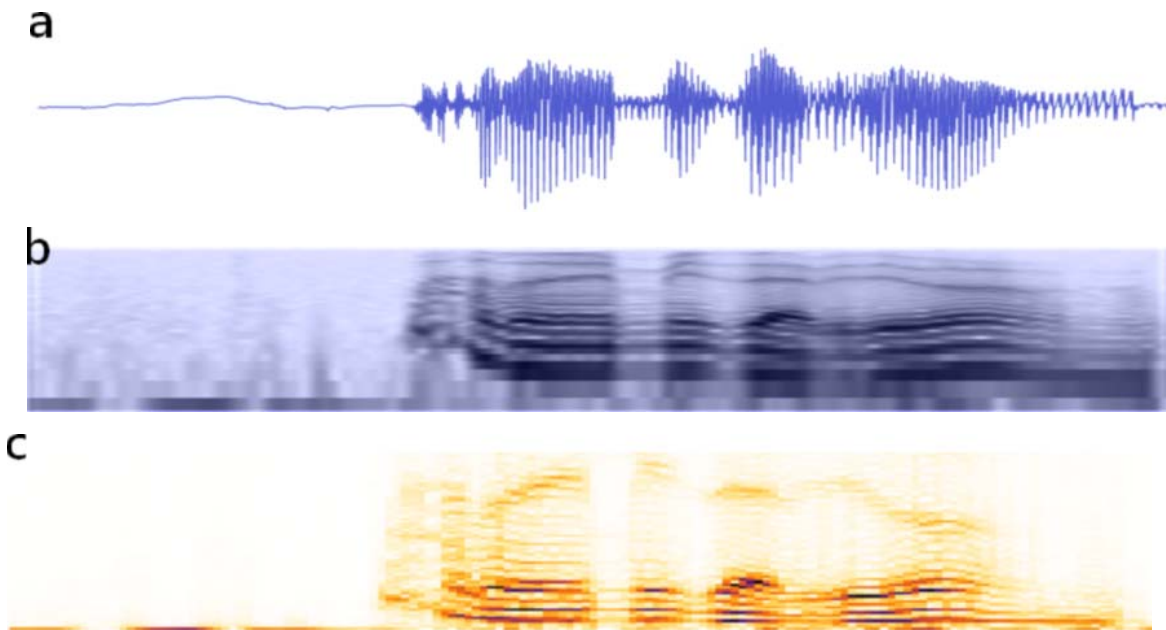

Fig. 1   The sentence "My name is Han Solo" is presented as a sound sample using (a) DFT of the signal, (b) spectrogram of the recorded sound with noise, and (c) spectrogram of a clear signal.
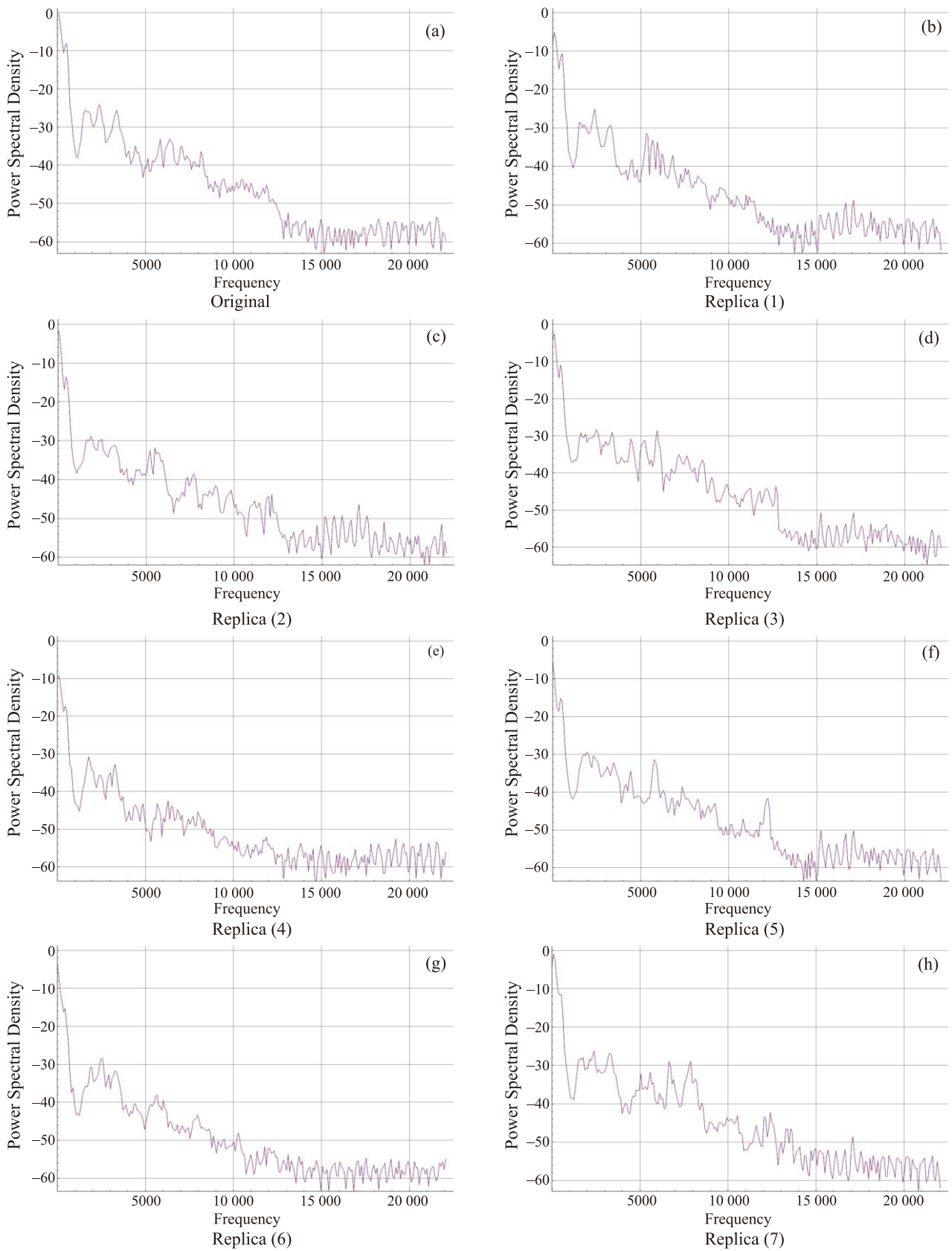
Fig. 2 **Periodograms with the selected maximum value of the PSD for the sample user verification sentence "James T. Kirk" presented for the original vice sample (a) and its various replicas used for verification (b–h).**
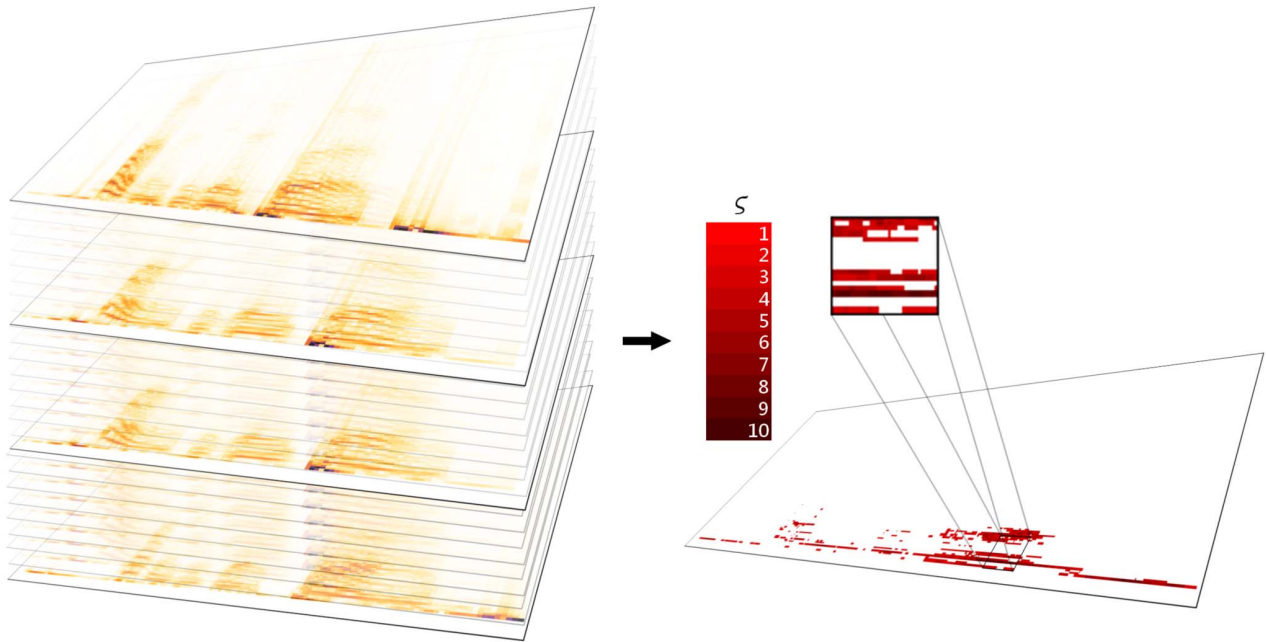
Fig. 3 **Visualization of the general spectrogram composition based on the information from the input spectrograms. The explanation of the proposed method is very simple: the darker the shade of red, the more important the information from all the inputs (the more often the information is present in all the inputs, the more important it is for the recognition; therefore, we use it to compose a general model).**

3.2 Heuristic detection of key-points

The applied heuristic method uses aggregated spectrograms to find the key points. These key points will be used to compose a special mask of features for the input, which will enable efficient retrieval of information to compose the training vectors for the verification system.

Accordingly, we implemented the Flower Pollination Algorithm (FPA) for these purposes[14]. The FPA simulates the natural phenomenon of flower pollination in spring. This algorithm is easy to implement and enables faster search in the space. The model of this algorithm assumes the following basic rules aimed to simplify the relations and dependences between operations:

• The global search (represented in a biotic pollination phenomenon) is simulated using Levy flights.

• The local search is simulated as abiotic and self-pollination.

• In nature, pollen is carried by the wind, which is not predictable; thus, in the algorithm, this is simulated using random factor $p \in [0, 1]$.

To find the most important points in the spectrogram image, the FPA first randomly locates the first population of flowers over the image of the aggregated spectrogram.

Subsequently, the global and local search simulations are performed. The key point selection is conducted using the following developed fitness function:

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

where x_i represents the points of the aggregated spectrogram, x_{best} is the best key point with highest fitness function value in current iteration, function $B(\cdot)$ represents the brightness of the pixel in values of $[0, 1]$, and N is the population size. Value χ is a value that reduces the brightness determined by the test method.

We simulated searching for the best key points in each iteration using the global and local searches. In the global search, we simulated pollen motion over spectrogram points x_i using the following formula:

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

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

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

where κ and μ are specified parameters.

After the local search, the operation was simulated in the surroundings of each selected global search point x_i according to the following formula:

$$x_i^{t+1} = x_i^t + \epsilon(x_j^t - x_k^t) \tag{10}$$

where x_j and x_k are the neighboring points.

The proposed algorithm returned the best adapted solution with its motion path. Figure 4 shows the sample trajectory returned by the proposed FPA. From a mathematical point of view, this trajectory was a closed set consisting of two-dimensional coordinates representing the key points of the analyzed aggregated spectrogram. From this set, we can use the information representing the most important features of the spectrogram and therefore input the signal in a very short time. Moreover, the proposed FPA will find a different trajectory for every input sample from the identified person, which will drastically improve the recognition efficiency and make the proposed method harder to get cheated on by frauds during verification.

3.3 General vector for identification

For the analysis of the input sounds, we converted these sounds into numerical vectors representing the input features. However, to speed up the verification process, these vectors should be composed only of the necessary information defined in the following form:

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

where $B(x_i)$ is the brightness value of the pixel found key points x_i returned by the proposed FPA, $f_{\max}$ is the maximum value of PSD determined from periodogram (Fig. 2), and id is the indexation of the voice sample (i.e., corresponding to the user ID). Therefore, the proposed definition for the general vector composition is a combination of the most important features from the analyzed spectrogram and periodogram using the method presented in Section 3. This makes the vector contain the most important information about the verified user and improves the system security because eventual fraud is difficult.

4 Neural Network

Various models and applications in decision support systems have been discussed since the beginning of research on neural networks in the first half of the 20th century[15]. We generally understand the neural network as a complex system in which neurons are connected between layers to forward the information signal. Applied architectures comprise various numbers of layers and neurons, where the input data is retrieved in all the following layers using the activation function. In more complex neural structures, hidden layers are added between the input and the output to improve the classification efficiency and to better adapt the network to the problem. Activation functions are mainly adapted to the solution space; however, the following unipolar sigmoid function is most often used:

$$y(x) = \frac{1}{1 + \exp(-\beta x)} \tag{12}$$

where β is the parameter adopted to increase the signal processing sensitivity. Each connection between neurons is associated with the weight in the range of [0, 1] scaled during the training process.

References [16, 17] presented a study of various activation functions, while Refs. [18, 19] discussed different approaches to training neural networks. One
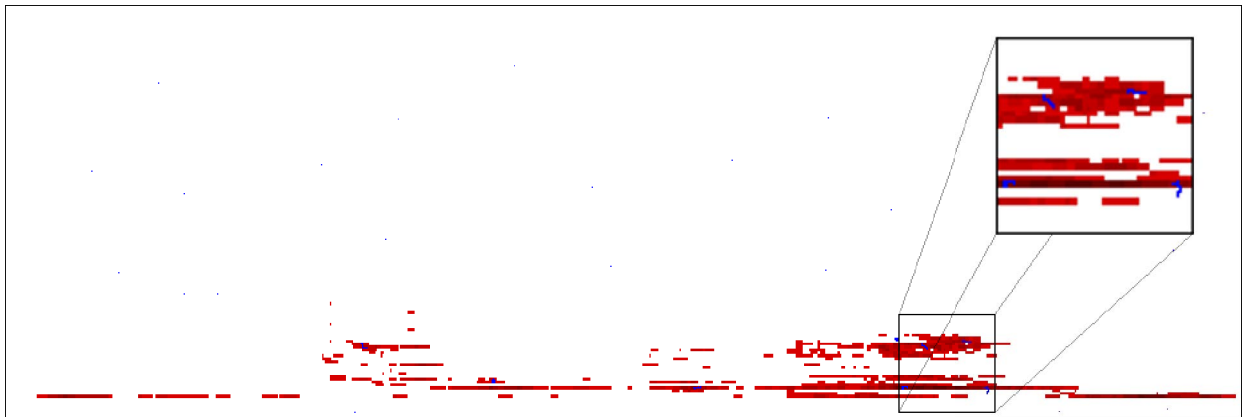

Fig. 4 Aggregated spectrogram covered with pollen trajectories while searching for the key points using proposed version of FPA.

of these methods is the back-propagation algorithm, which is the most commonly used algorithm because of ease of implementation and relatively high performance for various purposes[20]. Algorithm 1 is a type of supervised learning, where we minimize the error function on the output layer using the results from all the previous layers according to the following formula:

Algorithm 1   Feature extraction from spectrograms

1: Start,
2: Load samples into spectrograms,
3: **for** each spectrogram **do**
4:      Remove noise according to Eq. (2),
5: **end for**
6: Create an array of $w \times h$ filled with values 0,
7: **for** each spectrogram **do**
8:      **for** each point (x, y) **do**
9:          **if** color of pixel (x, y) is not white **then**
10:             Increase value of x,y cell by 1,
11:         **end if**
12:     **end for**
13: **end for**
14: Define color $[R_{old}, G_{old}, B_{old}]$,
15: Create empty bitmap B of size $w \times h$,
16: **for** each value in array **do**
17:     Create new color using Eq. (5),
18:     Set new color of pixel in bitmap B,
19: **end for**
20: Define all FPA coefficients and fitness function using Eq. (7),
21: Take aggregated spectrogram as search space for FPA,
22: Create randomly initial population,
23: Find the best solution x_{best} in actual population,
24: $t = 0$,
25: **while** $t < epoch$ **do**
26:     **if** $rand > p$ **then**
27:         Do global search using Eq. (8),
28:     **else**
29:         Randomly select two solutions from population,
30:         Do local search using Eq. (10),
31:     **end if**
32:     Evaluate solution,
33:     Save all trajectories,
34:     **if** $L(x_{new}) < L(x_{old})$ **then**
35:         Replace x_{old} with x_{new},
36:         Increase value of x_{new} by 0.08,
37:         Decrease value of x_{old} by 0.08,
38:     **end if**
39:     **for** each solution x_i **do**
40:         **if** $B(x_{best}) < B(x_i)$ **then**
41:             Replace x_{best} with x_i,
42:         **end if**
43:     **end for**
44:     $t + +$,
45: **end while**
46: Return the best trajectory,
47: Save coordinates of this trajectory as mask of features,
48: Stop.

$$\delta_k = \begin{cases} \text{out}_k(1 - \text{out}_k)(\text{ex}_k - \text{out}_k), & \text{output;} \\ \text{out}_k(1 - \text{out}_k)\sum_{o \in \text{out}} w_{ok}\delta_k, & \text{hidden} \end{cases} \quad (13)$$

where out is the output value of neuron k and ex is the expected value of the neuron. The error calculated in this manner is used to modify the weights of the connections between neurons using the following formula for each layer:

$$w_i = w_i + \delta_k w_i \quad (14)$$

5   Experimental Research

The proposed method of sound sample preprocessing described in Sections 2 and 3 (and proposed system is illustrated in Fig. 5) was used for 10 different users. Each one recorded 25 sound files. For processing, we used the proposed FPA algorithm with a population of 100 individuals, 15 iterations, and coefficients selected on the assumption of the fastest convergence to dark areas $\kappa = 0.4$ and $\mu = 0.35$. Composed aggregated vectors were used for training the neural network until the error value fell below 0.1 with a mixture of samples -80% to train and 20% to verify. Figure 6 illustrates the average error values during the training process. Table 1 presents the research results for the proposed method. The results were the averaged values of 10 tests performed for each set of 25 voice samples taken from each participating user.

From the results, we calculated the efficiency of the proposed verification to be approximately 75% with an average 25 samples per user. The results showed that the proposed method reached an approximately
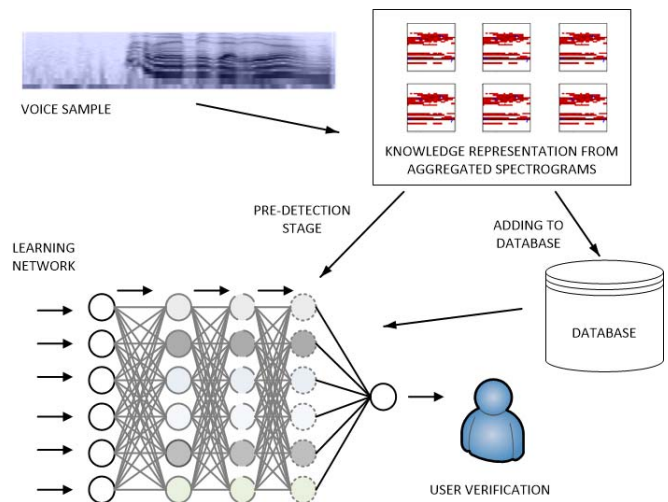

Fig. 5   Schematic information processing in the proposed system with the neural network and the heuristic method.
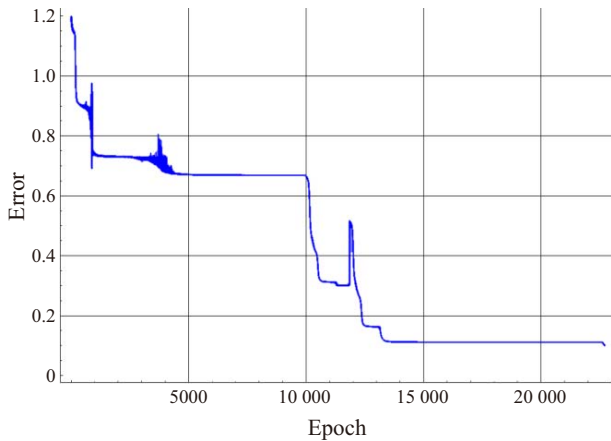
Fig. 6 Error change during the training process of the implemented neural network.

Table 1 Average accuracy for user verification.

Id	Correctly classified	Incorrectly classified	Accuracy (%)
1	14	11	56
2	20	5	80
3	23	2	92
4	18	7	72
5	21	4	84
6	20	5	80
7	19	6	76
8	22	3	88
9	18	7	72
10	15	10	60
AVG.	19	6	76
Std. deviation	287	287	11.47

56% accuracy for user 1 and an above 92% accuracy for user 3. This result is attributed to the quality of recording and the user name given for verification (i.e., users speaking loudly and clearly into the microphone increased precision, whereas quiet and indistinct voices may be recorded with some similarities to other users).

Similarly, possible jitters, hoarseness, and other factors that depend on the user in conjunction with the quality of the recording device may slightly influence the system. We plan to work on these features in future research. Nevertheless, the proposed method reached an accuracy above 75%, which shows a good potential for further research and development.

6 Final Remarks

Security has become increasingly important for digital systems with almost all data now being stored in a digital form. Accordingly, we need to develop new methods of access verification for these systems. The applications of new and devoted methods can increase security levels in computer networks and can improve information protection, user verification, authorization of access in various companies, etc.

This study discusses an innovative method of using a devoted combination of analyzing spectrograms of input audio samples by applying a heuristic algorithm with a neural network tool implemented for verification. The experimental research confirmed that the proposed method generally has an efficiency above 75%. This method can be an interesting alternative to the existing methods and can provide a good reason for further research on this solution.

The proposed application of the heuristic method in spectrogram sequence analyses provided good results. It can be recommended for verification of fraud possibility because for each analyzed sequence of points, we can achieve unique results because of the basic assumption of randomness of heuristics. In other words, the proposed solution can contribute to increasing the security in man-machine interactions.

References

[1] S. Lu, T. Mu, and S. Zhang, A survey on multiview video synthesis and editing, *Tsinghua Science and Technology*, vol. 21, no. 6, pp. 678–695, 2016.

[2] H. Zhang, J. Wen, J. Cui, and S. Zhang, Efficient conditional privacy-preserving and authentication scheme for secure service provision in vanet, *Tsinghua Science and Technology*, vol. 21, no. 6, pp. 620–629, 2016.

[3] D. Triboan, L. Chen, F. Chen, and Z. Wang, Towards a service-oriented architecture for a mobile assistive system with real-time environmental sensing, *Tsinghua Science and Technology*, vol. 21, no. 6, pp. 581–597, 2016.

[4] A. Fischer, M. Diaz, R. Plamondon, and M. A. Ferrer, Robust score normalization for dtw-based on-line signature verification, in *Proc. 13th Int. Conf. on Document Analysis and Recognition (ICDAR)*, Tunis, Tunisa, 2015, pp. 241–245.

[5] Y. Guerbai, Y. Chibani, and B. Hadjadji, The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters, *Pattern Recognition*, vol. 48, no. 1, pp. 103–113, 2015.

[6] N. Liu, M. Zhang, H. Li, Z. Sun, and T. Tan, Deepiris: Learning pairwise filter bank for heterogeneous iris

verification, *Pattern Recognition Letters*, vol. 82, no. 2, pp. 154–161, 2015.

[7] X. Wang, Y. Zhang, and H. Yang, A bimodal biometric verification system based on fingerprint and face, in *Proc. 5th Int. Conf. Electronics Information and Emergency Communication*, Beijing, China, 2015, pp. 325–327.

[8] Y. Zhang, S. Sankaranarayanan, and F. Somenzi, Statistically sound verification and optimization for complex systems, in *Automated Technology for Verification and Analysis*, Cham, Germany, 2014, pp. 411–427.

[9] 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*, vol. 29, no. 6, pp. 82–97, 2012.

[10] S. Winograd, On computing the discrete fourier transform, *Mathematics of Computation*, vol. 32, no. 141, pp. 175–199, 1978.

[11] J. L. Flanagan, *Speech Analysis Synthesis and Perception*. Berlin, Germany: Springer Science & Business Media, 2013.

[12] 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*, Orlando, FL, USA, 2002, p. 4164.

[13] 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.

[14] X. S. Yang, Flower pollination algorithm for global optimization, in *Unconventional Computation and Natural Computation*, Berlin, Germany, 2012, pp. 240–249.

[15] 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.

[16] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, Multilayer feedforward networks with a nonpolynomial activation function can approximate any function, *Neural Networks*, vol. 6, no. 6, pp. 861–867, 1993.

[17] 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.

[18] M. F. Møller, A scaled conjugate gradient algorithm for fast supervised learning, *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[19] B. Wirow, A. Greenblatt, Y. Kim, and D. Park, The no-prop algorithm: A new learning algorithm for multilayer neural networks, *Neural Networks*, vol. 37, pp. 182–188, 2013.

[20] P. Werbos, Beyond regression: New tools for prediction and analysis in the behavioral sciences, PhD dissertation, Havard University, Cambridge, MA, USA, 1974.

Marcin Woźniak received the MS degree from Silesian University of Technology in 2007 and PhD degree from Częstochowa University of Technology in 2012. He is an associate professor at Institute of Mathematics of the Silesian University of Technology in Gliwice, Poland. His main scientific interests are neural networks with their applications together with various aspects of applied computational intelligence. He is a scientific supervisor in editions of "the Diamond Grant" and "The Best of the Best" programs for highly talented students from the Polish Ministry of Science and Higher Education. He served as an editor for various special issues, and as an organizer or a session chair at various international conferences and symposiums.

Dawid Połap is with Institute of Mathematics at the Silesian University of Technology in Poland, where he received the diplomas in computer science and applied mathematics (with honors), in 2016 and 2017, respectively. In his career, he was awarded by the Ministry of Science and Higher Education with a "Diamond Grant" for the most talented students. He was participating in various scientific projects (both as a lead investigator and scientific investigator) from Polish and Italian universities. His main scientific interests are image processing and various aspects of computational intelligence.