Manuscript received December 19, 2016; revised April, 2017. DOI: 10.1515/eletel-2017-0021

Voice recognition through the use of Gabor transform and heuristic algorithm

Marcin Woźniak and Dawid Połap

***Abstract*—Increasingly popular use of verification methods based on specific characteristics of people like eyeball, fingerprint or voice makes inventing more accurate and irrefutable methods of that urgent. In this work we present voice verification based on Gabor transformation. Proposed approach involves creation of spectrogram, which serves as a habitat for the population in selected heuristic algorithm. The use of heuristic allows for feature extraction to enable identity verification using classical neural network. The results of the research are presented and discussed to show efficiency of the proposed methodology.**

***Keywords*—Neural networks, voice recognition, Gabor transformation, heuristic algorithm, swarm**

I. Introduction

EVER-INCREASING expansion of large companies and corporations is not only the reason for reducing unemployment but above all, motivation to create new software operated by these companies. One example is a system to verify identity of employees who come to work and confirm their presence. In a situation where a person enter the system, other information can be linked, for example, authorization to enter specific rooms or record start and end of work.

The past has shown that signing and use of cards are not an appropriate methods to verify identity of employees due to large number of forgeries. Recent scientific studies show that verification of identity may be based primarily on signal processing. It is area in which we distinguish analysis of fingerprints, signatures or voice samples. From the perspective of computer analysis of a fingerprint or a signature we can say that object is examined similarly as two-dimensional image. For this purpose the image must be processed in order to minimize the amount of information on the input file. One of the most known algorithms of image processing is edge detection which helps to find only a specific area in image. In [1], authors presented the idea of creation image descriptors based on that processing with a crawler algorithm. Obtained results have shown that this combination may be used in further research in image recognition. Again in [2], they presented the idea of feature extraction and indexation for large database system. Of course, these operations are used in further works, eg. fingerprint recognition. These type of identification may be based on comparing specific constituents, or even areas of images what was shown in [3]. It is similar with handwriting recognition, where similar techniques can be used. Moreover, in case of comparing signatures we may also use dynamic characteristics as pressure of a pen or speed of signing. One such study has been shown in [4] where authors presented a new algorithm for such identification.

Another group of identification is voice recognition where input file is a sound sample. In [5], an analysis of sounds to find voice disorders is presented. The idea of using voice recognition in biometric systems is shown in [6]. Interesting approach to analysis of sound is shown in [7], where the system identifies singer based on power of sound. Another approach is to use acoustic-spectrographic action in these field what was introduced in [8].

In this paper, we want to show processing of voice sample using Gabor transform and its interpretation in graphical form – spectrograms, which will allow for feature extraction and recognition based on heuristic algorithm and neural network.

Authors acknowledge contribution to this project of the "Diamond Grant 2016" No. 0080/DIA/2016/45 from the Polish Ministry of Science and Higher Education.

Marcin Woźniak and Dawid Połap are with Institute of Mathematics, Silesian University of Technology, Poland, (e-mail: Marcin.Wozniak@polsl.pl, Dawid.Polap@gmail.com)

II. Gabor Transformation

Signal analysis is impossible without use of a transformation that will unify specific signals and allow them to be compared. The most famous is Fourier transform. In 1946, Hungarian physicist and Nobel Prize winner, Dennis Gabor noted that Fourier transform has a drawback. More specifically, the transformation is not suitable for analysis of small fragments of selected time signal. This observation led him to propose a definition of time window and windowing operation, which indirectly helped to analyze only fragments of signals.

Definition A time window is called function $w(n)$ which satisfy the following conditions

- $w(n)$ is non-zero function in a finite time interval,
- $\max w(n)$ is achievable in the middle of time interval,
- graph of $w(n)$ is symmetrical with respect to $\max w(n)$.

There are many different window functions, an example may be a Hanning window determined as follows

$$w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N}\right), \tag{1}$$

where $n = 0, 1, \ldots, N$ and N is maximum division of signal.

Definition Windowing is called sampling of signal s with the use of time window $w(n)$, which can be represented as

$$g(n) = s(n)w(n), \qquad -\infty < n < \infty. \tag{2}$$

Windowing operation allowed us to minimize imperfections of Fourier transform. The use of window makes spectral characteristics become more smooth, and thus result in reduced blur of signal spectrum.

In order to use windowing operation for transformation, was introduced continuous transformation (called a Gabor transformation).

Definition A continuous Gabor transform of the signal s is called transformation defined as

$$G_s(x,f) = \int_{-\infty}^{\infty} s(\tau)w(\tau - x)\exp(-if\tau)d\tau. \quad (3)$$

For practical purposes (mainly computing), we introduce a discrete form of the transform.

Definition A discrete Gabor transform of signal s is called transformation defined as

$$G(k) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} C_{nm}g_{nm}(k), \quad (4)$$

where

$$g_{nm}(k) = s(k - mN)\exp\left(\frac{2j\pi nk}{N}\right), \quad (5)$$

and the signal s is divided into M time intervals of length N (so $l = M * N$), and C_{nm} means a specific values of signal s.

Field of time-frequency analysis deals with distribution of amplitudes in time, their density as well as characterization of spectrum in terms of width or harmonic components. In this area, signal must be subjected to transformation in order to extract characteristics of the sample, which may be presented in a visual form through spectrograms. Spectrogram illustrates graph of signal amplitude spectrum given on two axes represented by time and frequency. Moreover, amplitude of a given point is illustrated by giving it a value of a specific color of so-called intensity of the point. Analysis of this type graphic image is based on the intensity, which formed structures represented by darkest color (where the amplitude is highest).

III. Heuristic Algorithm

Faster growth of interest in search for alternative solutions to optimization problems allowed for large growth of heuristic algorithms that relay primarily on randomness.

A. Dragonfly Algorithm

The ever-increasing popularity of such solutions suggest newer methods as well as numerous practical applications. In [9], author presented mathematical model of dragonflies life in terms of static and dynamic movement. It covers primarily a way of avoiding enemies and acquiring food.

In order to model a natural phenomena of life, several assumptions are made. Each individual in population will be interpreted as a point $X = (x, y)$ in two-dimensional space, which has velocity V (the initial value is a random number). Each individual can only see a certain area what is described as radius r_0. In addition, number of individuals in a population is constant and described as N wherein initial position is selected at random. To describe movement of these creatures, several factors must be defined. Individuals do not move together but live in herd, so spread of dragonflies in space will be signed as S_i. In addition, we are introducing an alignment factor A_i, consistency factor C_i, attractiveness rate of given flight direction toward food as F_i and coefficient of dispersion enemies in space as E_i. All these parameters are calculated as

$$\begin{cases} S_i = -\sum_{j=1}^{N} X_i - X_j \\ A_i = \frac{1}{N}\sum_{j=1}^{N} V_j \\ C_i = \frac{1}{N}\sum_{j=1}^{n} X_j - X_i \\ F_i = X^+ - X_i \\ E_i = X^- + X_i \end{cases}, \quad (6)$$

where X_j is a neighbor for dragonflies X_i, as X^+ is understood as food source, and X^- the enemy.

Movement of dragonflies in the space will be modeled as redeployment in accordance with

$$X_i^{t+1} = X_i^t + \Delta X_i^{t+1}, \quad (7)$$

where t is the number of current iteration and ΔX_i^{t+1} is correlation of all coefficients defined as follows

$$\Delta X_i^{t+1} = (sS_i + aAi + cC_i + fF_i + eE_i) + \Delta X_i^t, \quad (8)$$

where parameters $\{s, a, c, f, e\} \in \langle 0, 1\rangle$ are weights for specific coefficients. In first iteration, each dragonfly has $\Delta X_i^{t=0} = 0.5$.

In order to introduce greater randomness of movement in population there is a modification of equation (7) defined as

$$X_i^{t+1} = X_t + L(d)X_t, \quad (9)$$

where $L(x)$ is a function of Lvy's flight described as

$$L(x) = 0.01\frac{r_1\gamma}{|r_2|^2}, \quad (10)$$

where r_1, r_2 are random values in the range of $\langle 0, 1\rangle$, β is a constant value and γ is

$$\gamma = \left(\frac{\Gamma(1+\beta)\sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right)\beta 2^{\frac{\beta-1}{2}}}\right)^{1/\beta}, \quad (11)$$

where $\Gamma(x) = (x-1)!$.

In each iteration of the algorithm, each individual is evaluated in terms of fitness function $f(X_i)$. Then, the best X^+ and the worst X^- are selected. The full algorithm modeling dragonflies life is depicted in Algorithm 1.

IV. Voice Verification

The whole process of voice verification contains few steps. Creation of a database containing samples of the voice is a combination of three elements – creation of voice samples, processing and saving them. After that, the classifier should learn to recognize the person on basis of voice samples in database. The last step is to classify in real time which means a combination of following operations – save a new voice sample, fast processing and classification. In these section, proposed method of verification is presented.

Define the following parameters of the algorithm N, t, r_0, s, c, f, e, w;
Define search space and fitness function $f(x)$;
Generate initial population at random;
$T = 0$;
while $T < t$ **do**
  Evaluate individuals according to $f(x)$;
  Find the best X^+ and the worst individual X^-;
  Update coefficients $\{s, c, f, e\}$ by increasing or decreasing them by a constant;
  Update all parameters using (6);
  if X_i *has a neighbor in the field of view* r_0 **then**
    Update position of dragonfly X_i by (7);
  else
    Update position of dragonfly by (9);
  end
  if $T == t - 1$ **then**
    Evaluate individuals according to $f(x)$;
  end
  T++;
end
Return the best dragonflies in the population;

Algorithm 1: Dragonfly Algorithm.

A. Data Extraction

Data extraction is to extract characteristics of a person. For this purpose, heuristic algorithm will move over spectrograms in order to locate areas which are characteristic for that person. Moreover, the use of heuristic algorithm allows for high randomness and unpredictability in verification process. Heuristics described in Sec. III will move over spectrogram created based on Gabor transform introduced in Sec. II. Each individual in population will be represented by pixel in the image. Evaluating given individual will occur in terms of darkness (of the pixel) described in the following way

$$f(x) = -\frac{\eta + \delta}{2}, \tag{12}$$

where

$$\begin{cases} \eta = \min(R(X), G(X), B(X)) \\ \delta = \max(R(X), G(X), B(X)) \end{cases}, \tag{13}$$

function $R(\cdot)$, $G(\cdot)$, $B(\cdot)$ is specific component of a given pixel X in the RGB model. So defined function allows to search for the darkest areas of the image. Heuristic algorithm returns the top k individuals from last iteration. These points indicate areas where adaptation is the best in practice, it means the maximum amplitude for that specific voice sample.

For each person, the algorithm is executed for several (assume m) spectrograms, and the best points are selected and combined into a vector consisting $n = k * m$ values in the following way

$$[X_1, X_2, \ldots, X_n] = [(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]. \tag{14}$$

Defined vector allows for modeling learning vector that will be created for each spectrogram for a specific person defined as follows

$$[f(x_1, y_1), f(x_2, y_2), \ldots, f(x_n, y_n)], id_1, \ldots, id_m], \tag{15}$$

where $f(\cdot)$ is a function of darkness shown in (12) and a sequence value of $\{id_1, \ldots, id_m\}$ is identifier of specific person where m is length of this identifier. This value depends on number of persons in database, for example, if the value will be three, so this length is sufficient for eight people (it is a combination of binary number $\{0, 1\}$).

B. Neural Classifier

Problem of classification is very important due to different automated systems, which should properly divide knowledge and data – even these that deviate from accepted norm. Neural Networks (NN) are mathematical structures that model mechanism for information flow in human brain [10], [11]. NN are one of the most common classifier.

1) Artificial Neural Network: The network model is composed of layers. First layer is referred to as input, last one as output and middle ones as hidden. Input layer is responsible for acceptance of training vector, output for returning results of network (classification results), and hidden layers for data flow. Each layer is made up of smaller units called neurons.

Neuron is theoretical unit which takes output values of other neurons o (from previous layer) and weights w (neurons between layers connected to each other, and each combination is burdened with a weight). Neuron has task to calculate the output value Θ according to the following formula

$$\Theta\left([o_1, \ldots, o_m], [w_1, \ldots, w_m]\right) = \Phi\left(\sum_{i=1}^{m} o_i \cdot w_i\right), \tag{16}$$

where $\Phi(\cdot)$ is activation function defined as

$$\Phi(x) = \left(1 + e^{-\alpha x}\right)^{-1}, \tag{17}$$

and $\alpha \in \langle 0, 1 \rangle$.

2) The Back Propagation Algorithm: For the correct classification of objects using neural network, the structure must be trained to recognize it. For this purpose, back-propagation algorithm was proposed in [12], which is one of the most popular methods of this type. Network error is calculated as square of deviation described as

$$d([w_1, \ldots, w_m]) = \frac{1}{2}\left[\Phi\left(\sum_{i=1}^{m} o_i w_i\right) - t\right]^2. \tag{18}$$

where t is expected result. For so constructed error function, the algorithm works by modifying weights using gradient of that function defined as

$$\frac{\partial d(w_1, \ldots, w_k)}{\partial w_j} = (\Theta - t)\Phi'(s)o_j. \tag{19}$$

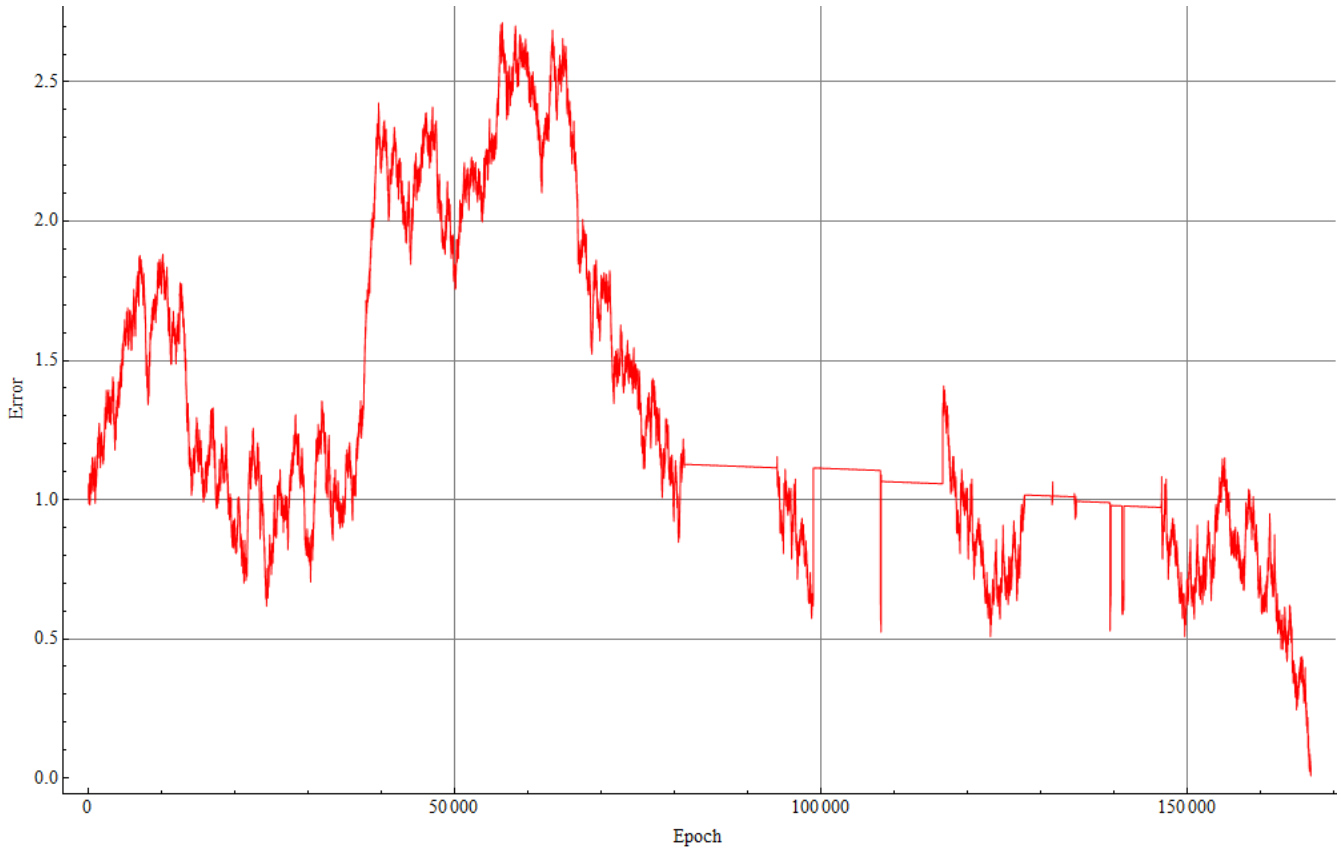

Fig. 1: An example of a graph showing an error of NN's training.

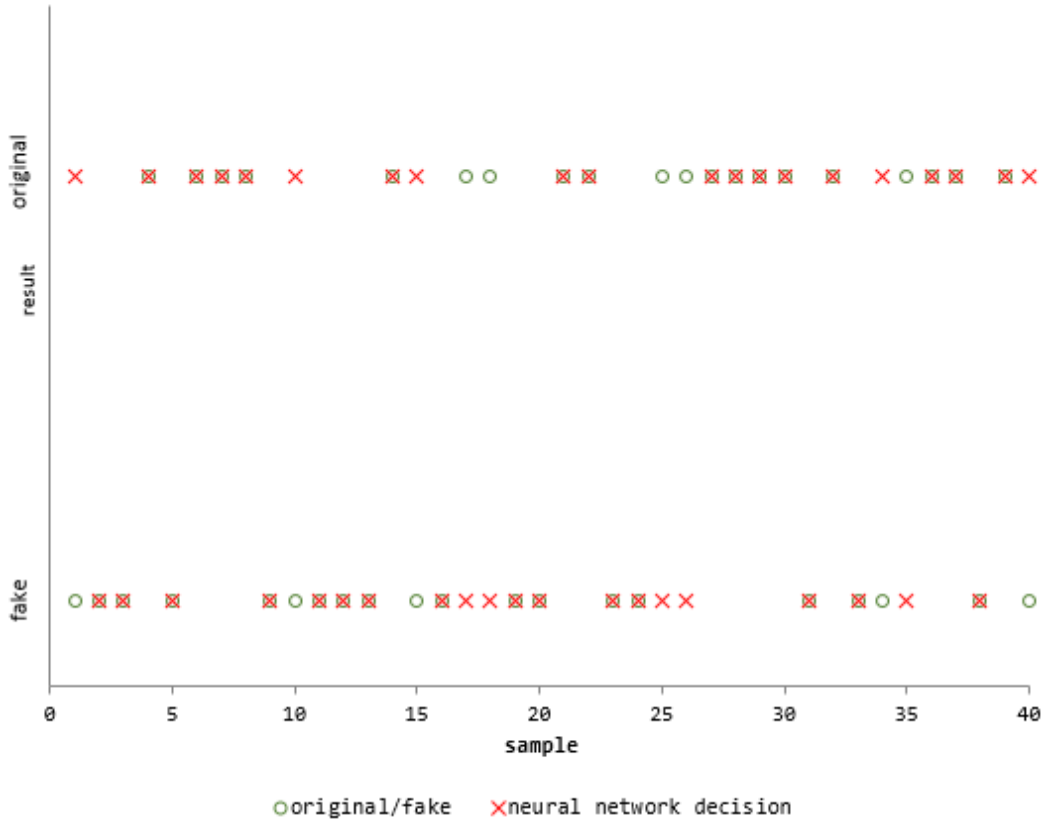

Fig. 2: Sample verification process over probe containing 40 randomly selected voice samples from verification set.

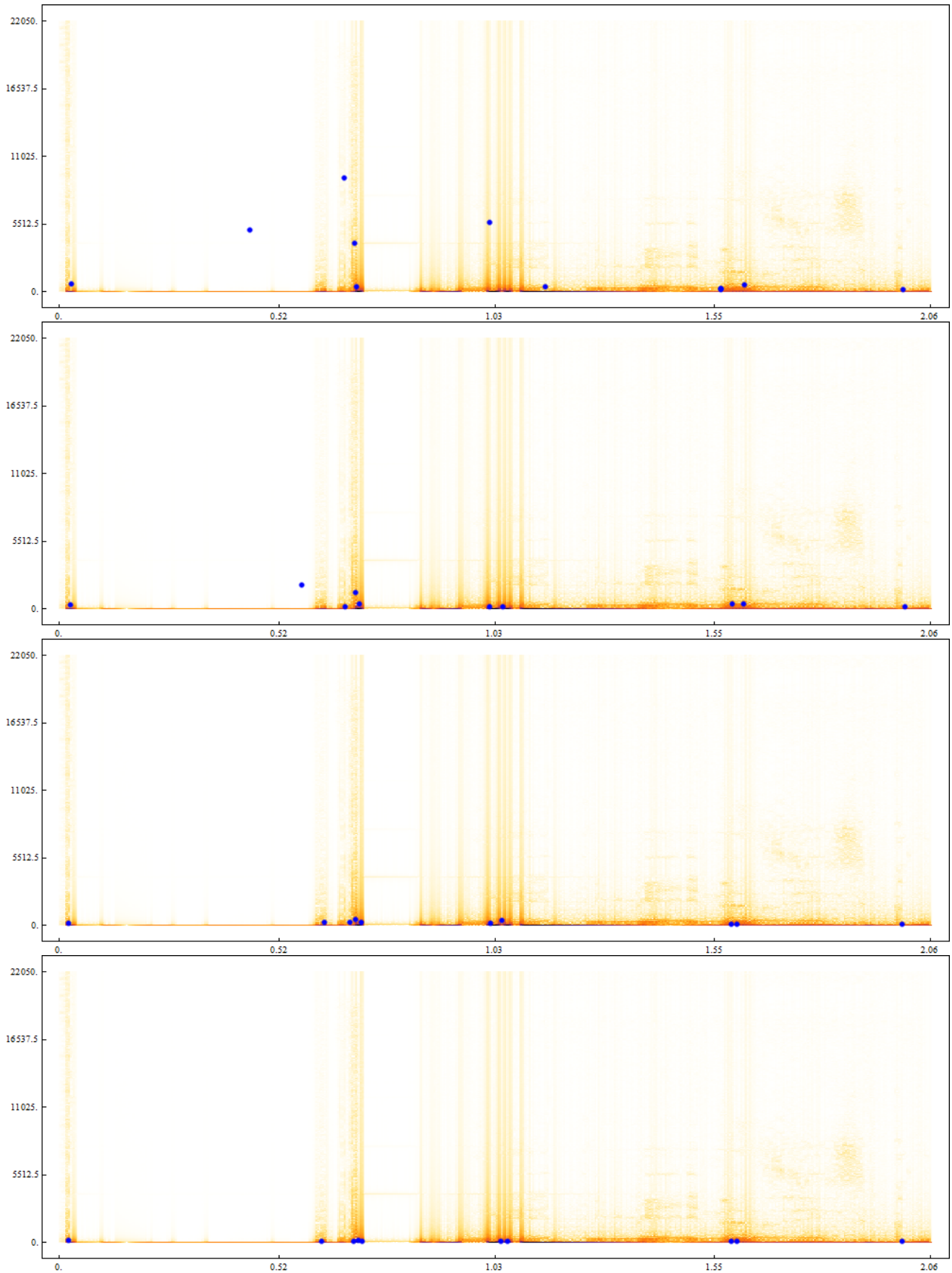

Fig. 3: Movement of the population on the spectrogram in 25, 50, 75 and 100 iterations.

V. EXPERIMENTS

For proposed method, the base of 200 samples of voice with the name "James T. Kirk" and 150 with the name "Hikaro Sulu" was created. Then, heuristic algorithm found suitable series of values composed of 10 individuals. During the tests, 100 iterations have been used in both cases. Sample movement of individuals is shown in Figure 3. Then added sequentially identifiers 0 and 1. For each sample vector has been created, then neural network was trained for so formed samples. Training was stopped at the level of 0.01 what can be seen in Figure 1. During training, all samples were divided $70 : 30$ (training/verification). Trained neural network was tested with verification samples. Obtained result was at 75% efficiency. In Fig. 2 we can see the test of the network for random samples.

VI. CONCLUSIONS

Proposed solution is not only attractive in terms of spectrogram analysis, but also by using heuristic which for one person may create an infinite number of various representative vectors. Performed tests show that this is a good solution from theoretical point of view. Correct classification rate of 75% is a good result, but maybe still not sufficient for practical use. An additional drawback is time of training - in case of practical implementation of the solution in companies, there is a risk of potential frequent re-learning, for example, in a situation of dismissal or employment of one employee. Of course, these defects can be eliminated by using other tools. Moreover, classification result should be increased, which may be possible using a vector with more numerical values.

REFERENCES

[1] R. Grycuk, M. Gabryel, M. Scherer, and S. Voloshynovskiy, "Image descriptor based on edge detection and crawler algorithm," in *International Conference on Artificial Intelligence and Soft Computing*. Springer, 2016, pp. 647–659.

[2] R. Grycuk, M. Gabryel, R. Nowicki, and R. Scherer, "Content-based image retrieval optimization by differential evolution," in *Evolutionary Computation (CEC), 2016 IEEE Congress on*. IEEE, 2016, pp. 86–93.

[3] J. Kim, K. Oh, A. B.-J. Teoh, and K.-A. Toh, "Finger-knuckle-print for identity verification based on difference images," in *Industrial Electronics and Applications (ICIEA), 2016 IEEE 11th Conference on*. IEEE, 2016, pp. 1073–1077.

[4] K. Cpałka, M. Zalasiński, and L. Rutkowski, "A new algorithm for identity verification based on the analysis of a handwritten dynamic signature," *Applied soft computing*, vol. 43, pp. 47–56, 2016.

[5] S. N. Awan, N. Roy, D. Zhang, and S. M. Cohen, "Validation of the cepstral spectral index of dysphonia (csid) as a screening tool for voice disorders: development of clinical cutoff scores," *Journal of Voice*, vol. 30, no. 2, pp. 130–144, 2016.

[6] M. Pal and G. Saha, "On robustness of speech based biometric systems against voice conversion attack," *Applied Soft Computing*, vol. 30, pp. 214–228, 2015.

[7] M. Usha, Y. Geetha, and Y. Darshan, "Objective identification of prepubertal female singers and non-singers by singing power ratio using matlab," *Journal of Voice*, 2016.

[8] E. Krasnova, E. Bulgakova, and V. Shchemelinin, "Performance evaluation of acoustic-spectrographic voice identification method in native and non-native speech," *Performance Evaluation*, vol. 5656, p. 10004820, 2016.

[9] S. Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," *Neural Computing and Applications*, vol. 27, no. 4, pp. 1053–1073, 2016.

[10] P. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," 1974.

[11] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[12] D. Williams and G. Hinton, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.