Rough Set Flow Graphs and Ant Based Clustering in Classification of Disturbed Periodic Biosignals

Krzysztof Pancerz[1], Arkadiusz Lewicki[1], Ryszard Tadeusiewicz[2], and Jan Warchoł[3]

[1] University of Information Technology and Management
Sucharskiego Str. 2, 35-225 Rzeszów, Poland
{kpancerz,alewicki}@wsiz.rzeszow.pl
[2] AGH University of Science and Technology
Mickiewicza Av. 30, 30-059 Krakow, Poland
rtad@agh.edu.pl
[3] Medical University of Lublin
Jaczewskiego Str. 4, 20-090 Lublin, Poland
jan.warchol@umlub.pl

Abstract. In the paper, we are interested in classification of disturbed periodic biosignals. An ant based clustering algorithm is used to group episodes into which examined biosignals are divided. Disturbances in periodicity of such signals cause some difficulties in formation of coherent clusters of similar episodes. A quality of a clustering process result can be used as an indicator of disturbances. A local function in the applied clustering algorithm is calculated on the basis of temporal rough set flow graphs representing an information flow distribution for episodes.

Keywords: rough set flow graphs, ant based clustering, biosignals

1 Introduction

An increasing interest is now evident in classification and clustering for time series and signals using a variety of methodologies (cf. [5]). One of the frequently examined problems concerns analysis of biosignals (cf. [6], [12]). In our research, we consider a problem of classification of voice signals in order to detect some disturbances indicating the possible presence of laryngeal pathologies.

Periodicity is one of the main features of some biosignals (e.g., voice, ECG). However, in case of some diseases, this periodicity can be disturbed. Especially, it is expressed by different shapes in selected time windows corresponding to periods of biosignals. The main idea of the proposed approach is based on recognition of temporal patterns and their replications in a selected fragment of the signal being examined. In case of some disturbances, temporal patterns cannot be found. This problem has been considered by us in examination of a voice signal for a non-invasive diagnosis of selected larynx diseases. We have used different approaches to solve this problem, for example, recurrent neural networks

[13], [14], [15], mining unique episodes in temporal information systems [10], ant based clustering with similarity measures [9].

In this paper, an ant based clustering algorithm working on the basis of temporal rough set flow graphs is proposed. Rough set flow graphs [11] introduced by Z. Pawlak are a useful tool for the knowledge representation. We use them as a tool for representing the knowledge of transitions between consecutive samples of examined signals. A quality of a clustering process result can be used as an indicator of disturbances in periodicity of biosignals. If episodes included in time windows corresponding to periods of the examined signal are similar (there is a lack of non-natural disturbances), then they should be grouped into very cohesive clusters. If significant replication disturbances in time appear, then time windows are grouped in several clusters or they are scattered on the grid without any distinct groups. To provide such a classification ability, we use a special algorithm for clustering a set of well categorized objects, originally proposed by us in [8]. A set of well categorized objects is characterized by a high similarity of objects within classes and a relatively high dissimilarity of objects between different classes. The algorithm is based on versions of ant based clustering algorithms proposed earlier by Deneubourg, Lumer and Faieta as well as Handl et al. Temporal rough set flow graphs representing an information flow distribution for episodes are used in the clustering algorithm for calculation of the local function.

2 Episode Information Systems and Temporal Rough Set Flow Graphs

In this section, we recall the basic concepts concerning information systems and rough set flow graphs which are crucial to understand the approach proposed in the paper as well as we introduce a notion of episode information systems and show how to create rough set flow graphs for such systems.

An information system is a pair $S = (U, A)$, where U is a set of objects, A is a set of attributes, i.e., $a : U \rightarrow V_a$ for $a \in A$, where V_a is called a value set of a. Any information system can be represented as a data table, whose columns are labeled with attributes, rows are labeled with objects, and entries of the table are attribute values.

In our approach, we will use information systems to represent time series data. Biosignals recorded in electronic devices have a digital form and they can be treated as time series, i.e., sequences of signal samples. Each consecutive sample represents a value of the signal at a given time instant. Therefore, we assume that a set of attributes in an information system is ordered in time, i.e., $A = \{a_t : t = 1, 2, \ldots, n\}$, where a_t is the attribute determining values of the signal at time instant t. Each object in such a system is said to be an episode and the whole system will be called an episode information system and denoted by S_e. An example of the episode information system $S_e = (E, A)$, where E is a nonempty finite set of episodes and A is a nonempty finite set of attributes

ordered in time, is shown in Figure 1. This system includes five episodes of real voice signals and each episode consists of five consecutive samples.

Table 1. An example of the episode information system S_e

E/A	a_1	a_2	a_3	a_4	a_5
e_1	-9362.00	-10168.00	-11222.00	-11749.00	-12400.00
e_2	-9734.00	-10261.00	-11346.00	-12090.00	-12524.00
e_3	-10261.00	-11160.00	-11656.00	-12524.00	-13268.00
e_4	-9672.00	-10664.00	-11191.00	-11718.00	-12400.00
e_5	-8990.00	-9517.00	-10013.00	-10664.00	-10912.00

Rough set flow graphs have been defined by Z. Pawlak [11] as a tool for reasoning from data. A flow graph is a directed, acyclic, finite graph $G = (N, B, \sigma)$, where N is a set of nodes, $B \subseteq N \times N$ is a set of directed branches and $\sigma : B \to [0, 1]$ is a flow function.

An input of a node $x \in N$ is the set $I(x) = \{y \in N : (y, x) \in B\}$, whereas an output of a node $x \in N$ is the set $O(x) = \{y \in N : (x, y) \in B\}$. $\sigma(x, y)$ is called a strength of a branch $(x, y) \in B$. We define the input and the output of the graph G as $I(G) = \{x \in N : I(x) = \emptyset\}$ and $O(G) = \{x \in N : O(x) = \emptyset\}$, respectively. $I(G)$ and $O(G)$ consist of external nodes of G. The remaining nodes of G are its internal nodes.

With each node $x \in N$ we associate its inflow $\delta_+(x)$ and outflow $\delta_-(x)$ defined by:

- $\delta_+(x) = \sum_{y \in I(x)} \sigma(y, x)$,
- $\delta_-(x) = \sum_{y \in O(x)} \sigma(x, y)$.

For each node $x \in N$, its throughflow $\delta(x)$ is defined as follows:

$$\delta(x) = \begin{cases} \delta_-(x) & \text{if } x \in I(G), \\ \delta_+(x) & \text{if } x \in O(G), \\ \delta_-(x) = \delta_+(x) & \text{otherwise.} \end{cases} \tag{1}$$

With each branch $(x, y) \in B$ we may also associate its certainty $cer(x, y) = \frac{\sigma(x,y)}{\delta(x)}$, where $\delta(x) \neq 0$.

A directed path $[x \dots y]$ between nodes x and y in G, where $x \neq y$, is a sequence of nodes $x_1, x_2, \dots, x_n$ such that $x_1 = x$, $x_n = y$ and $(x_i, x_{i+1}) \in B$, where $1 \leq i \leq n - 1$. For each path $[x_1 \dots x_n]$, we define its certainty as $cer[x_1 \dots x_n] = \prod_{i=1}^{n-1} cer(x_i, x_{i+1})$. In our approach, we will use different operators (not only product) for calculating a certainty of a given path.

To represent an information flow distribution in an episode information system we can use rough set flow graphs. A very similar problem has been considered

Algorithm 1: Algorithm for creating a temporal rough set flow graph corresponding to an episode information system

Input : An episode information system $S_e = (E, A)$, where $A = \{a_t : t = 1, 2, \ldots, n\}$.

Output: A temporal rough set flow graph $G = (N, B, \sigma, cer)$ corresponding to S_e.

$N \longleftarrow \emptyset$;
$B \longleftarrow \emptyset$;
for *each* $i = 1, 2, \ldots, n$ **do**
  $N_i \longleftarrow \emptyset$;
  for *each attribute value* v *of* a_i **do**
    Create a node n_v representing v and add it to N_i;
  end
  $N \longleftarrow N \cup N_i$;
end
for *each* $i = 1, 2, \ldots, n-1$ **do**
  for *each node* $n_v \in N_i$ **do**
    for *each node* $n_w \in N_{i+1}$ **do**
      Create a branch $b = (n_v, n_w)$;
      $\sigma(b) \longleftarrow \frac{card(E(a_i,v) \cap E(a_{i+1},w))}{card(E)}$;
      $cer(b) \longleftarrow \frac{card(E(a_i,v) \cap E(a_{i+1},w))}{card(E(a_i,v))}$;
      $B \longleftarrow B \cup \{b\}$;
    end
  end
end

in [4]. The algorithm presented in this section is modeled on that presented in [4]. Let $S_e = (E, A)$, where $A = \{a_t : t = 1, 2, \ldots, n\}$, be an episode information system. A rough set flow graph G corresponding to S_e consists of n layers. Nodes in the i-th layer of G represent values determined by the attribute a_i, where $i = 1, 2, \ldots, n$. Since the attribute set A is ordered in time, we can call the graph G as a temporal rough set flow graph. It represents temporal flow distribution in an episode information system. In order to construct a temporal rough set flow graph G corresponding to an episode information system S_e we may perform Algorithm 1. A graph constructed using this algorithm is supplemented with a certainty function which assigns a number called certainty from the interval $[0, 1]$ to each branch in G. Hence, we have $G = (N, B, \sigma, cer)$, where N is a set of nodes, B is a set of directed branches, $\sigma : B \to [0, 1]$ is a flow function, and $cer : B \to [0, 1]$ is a certainty function. Certainty factors of branches will be used in our approach presented here.

Let $S_e = (E, A)$, where $A = \{a_t : t = 1, 2, \ldots, n\}$ be an episode information system. By $E(a_i, v)$ we denote the set of all episodes in E for which the attribute a_i has the value v.

3 Ant Based Clustering Algorithm

Ant based clustering of time series was considered by us in [7]. Next, the approach presented there has been modified in our last work [8]. Here, we briefly remind this approach. It is based on algorithms proposed earlier by Deneubourg [1], Lumer and Faieta [3] as well as Handl et al. [2] with some our modifications. A set of steps is formally listed in Algorithm 2. In this algorithm:

- E is a set of episodes being clustered (the size of E is n),
- N is a number of iterations performed for the clustering process,
- $Ants$ is a set of ants used in the clustering process,
- $p_{pick}(e)$ and $p_{drop}(e)$ are probabilities of the picking and dropping operations made for a given episode e, respectively (see Formulas 3 and 4).

Let P be a set of all places of the square grid G on which objects are scattered. Each place p of G is described by two coordinates, i and j, written as $p(i, j)$. The neighborhood $\pi(p)$ of $p(i, j)$, where a given object is foreseen to be dropped or whence it is foreseen to be picked up, is defined as a square surrounding $p(i, j)$, i.e.:

$$\pi(p) = \{p'(i', j') \in P : abs(i' - i) <= r \text{ and } abs(j' - j) <= r\},$$

where r is a radius of perception of ants and abs denotes the absolute value.

For the neighborhood $\pi(p)$, a local episode information system $S_e(p)$ is created. $S_e(p)$ includes all episodes placed in $\pi(p)$. Next, on the basis of $S_e(p)$, a temporal rough set flow graph $TRSFG(p)$ is built using Algorithm 1. $TRSFG(p)$ serves as a base for calculating a local function value.

Let $S_e(p) = (E, A)$, where $A = \{a_t : t = 1, 2, \ldots, n\}$, be a local episode information system and e^* be the episode designated to pick up from or to drop at the place p. If e^* is described by the following attribute values: $a_1(e^*) = v_1$, $a_2(e^*) = v_2$, ..., $a_n(e^*) = v_n$, then a local function $f_{loc}(e^*)$ is calculated as:

$$f_{loc}(e^*) = \overset{n-1}{\underset{i=1}{Op}}\, cer(n^i_{v_i}, n^{i+1}_{v_{i+1}}), \tag{2}$$

where:

- Op is an aggregation operator, for example, median, arithmetic average, average weighted with different ways, etc.,
- $n^i_{v_i}$ is the node in the i-th layer of $TRSFG(p)$ corresponding to value v_i,
- $n^{i+1}_{v_{i+1}}$ is the node in the $(i+1)$-th layer of $TRSFG(p)$ corresponding to value v_{i+1}.

It is easy to see that a local function is calculated as the aggregation of certainty factors of branches belonging to the path in $TRSFG(p)$ determined by attribute values of the episode e^*.

To avoid forming smaller clusters of very similar episodes, the threshold density $minDens$ is used. The density $dens(e)$ of the neighborhood $\pi(p)$ for the episode e designated to pick up from or to drop at the place p is calculated as a

ratio of a number of all episodes placed in the neighborhood $\pi(p)$ to a number of all places in the neighborhood $\pi(p)$.

Picking and dropping decisions for the episode e can be formally expressed by the following threshold formulas:

$$p_{pick}(e) = \begin{cases} 1 & \text{if } f_{loc}(e) <= \vartheta_{sim}^{pick}, \\ \frac{1}{(1-\vartheta_{sim}^{pick})^2}(f_{loc}(e)-1)^2 & \text{otherwise} \end{cases} \tag{3}$$

and

$$p_{drop}(e) = \begin{cases} 1 & \text{if } f_{loc}(e) >= \vartheta_{sim}^{drop}, \\ \frac{1}{(\vartheta_{sim}^{drop})^2} f_{loc}(e)^2 & \text{otherwise.} \end{cases} \tag{4}$$

Thresholds ϑ_{sim}^{pick} and ϑ_{sim}^{drop} for picking and dropping operations, respectively, are fixed. These values are selected experimentally for given sets of signals.

4 Classification Procedure

It was mentioned in the introduction that we are interested in periodical biosignals. The classification procedure proposed in this paper is as follows. We divide a given fragment of a periodic biosignal into time windows. Each time window represents one period of the signal. A signal included in one time window is called an episode. The set of episodes is used in the clustering process.

In the experiments, voice signals collected by J. Warchoł [16] were examined. Examples of two voice signals, divided into episodes are shown in Figures 1 and 2, respectively. The first signal does not include disturbances of periodicity whereas the second one includes them. For comparison, we also present a pure sine signal (see 3) divided into episodes.

Before clustering, we apply some pre-processing procedures:

1. Values of signal samples are normalized to the interval $[-1.0, 1.0]$.
2. Each episode is transformed into the so-called delta representation, i.e., values of samples have been replaced with differences between values of current samples and values of previous samples. After transformation, each episode is a sequence consisting of three values: -1 (denoting decreasing), 0 (denoting a lack of change), 1 (denoting increasing) (cf. [7]). This transformation enables us to obtain a multistage decision transition system with discrete values.

Experiments for the ant based clustering have been performed with the following parameters:

- *grid size*: 100×100,
- *no. of ants*: 5,
- *radius of perception*: 2,
- *no. of iterations*: 10000,
- *aggregation operator*: arithmetic average.

Algorithm 2: Algorithm for Ant Based Clustering of Episodes

```
for each episode e_i ∈ E do
    Place e_i randomly on a grid G;
    Set e_i as dropped;
end
for each ant a_j ∈ Ants do
    Place a_j randomly on a grid place occupied by one of episodes from E;
    Set a_j as unladen;
    workTime(a_j) ← 0;
end
for k ← 1 to N do
    for each ant a_j ∈ Ants do
        if a_j is unladen then
            if place of a_j is occupied by episode e then
                Draw a random real number r ∈ [0, 1];
                if dens(e) < minDens or r ≤ p_pick(e) then
                    set e as picked;
                    set a_j as carrying the episode;
                else
                    move a_j randomly to another place occupied by one of
                    episodes from E;
                end
            else
                move a_j randomly to another place occupied by one of episodes
                from E;
            end
        else
            Draw a random real number r ∈ [0, 1];
            if r ≤ p_drop(e) then
                move e carried by a_j randomly to a new place on a grid;
                set e as dropped;
                set a_j as unladen;
                workTime(a_j) ← 0;
            else
                workTime(a_j) ← workTime(a_j) + 1;
            end
        end
        if workTime(a_j) > maxWorkTime then
            set e carried by a_j as dropped;
            set a_j as unladen;
            move a_j randomly to another place occupied by one of episodes
            from E;
            workTime(a_j) ← 0;
        end
    end
    increase minDens;
end
```

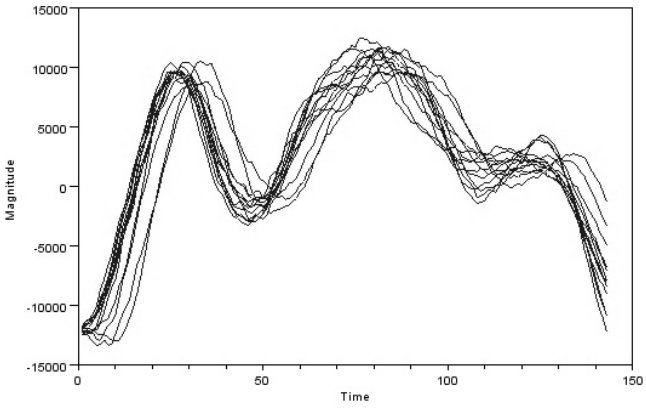

Fig. 1. Selected episodes of the signal without disturbances of periodicity

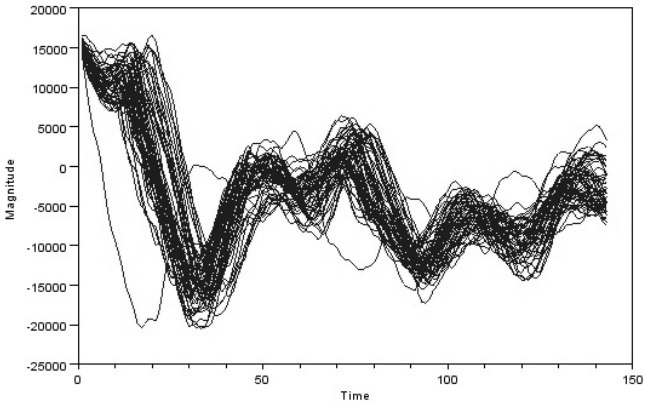

Fig. 2. Selected episodes of the signal with disturbances of periodicity

After a clustering process, we expect to have two distinguishable situations. If episodes are similar (there are no disturbances), then they are grouped into very cohesive clusters. If significant replication disturbances appear, then episodes are grouped in several dispersed clusters or they are scattered on the grid. A result of clustering is an indicator used to classify the examined biosignal.

Exemplary results of ant based clustering, for the signals without and with disturbances of periodicity, are shown in Figures 4 and 5, respectively. For com-

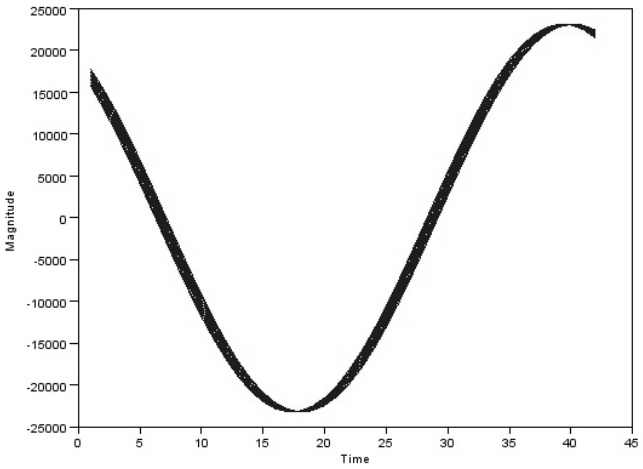

Fig. 3. Selected episodes of the pure sin signal

parison, we also present a result of ant based clustering for a pure sine signal (see Figure 6).

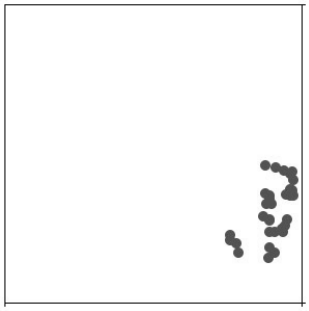

Fig. 4. The result of clustering the signal without disturbances of periodicity

5 Conclusions

In the paper, the first attempt to application of ant based clustering based on temporal rough set flow graphs for classification of disturbed periodic biosignals

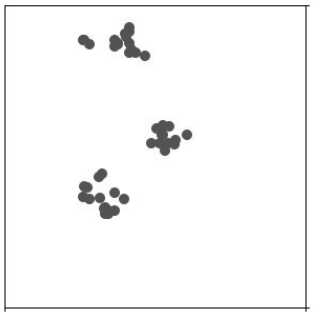

Fig. 5. The result of clustering the signal with disturbances of periodicity

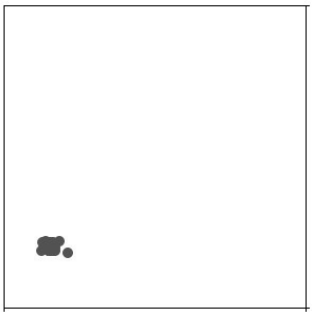

Fig. 6. The result of clustering the pure sin signal

has been presented. At the beginning, we were interested in simple classification, i.e., periodicity of the examined signal is disturbed or no. In the future, we plan to make more detailed analysis of results of a clustering process. This should give us additional information about the scale and character of disturbances.

Acknowledgments

This paper has been partially supported by the grant No. N N519 654540 from the National Science Centre in Poland.

References

1. Deneubourg, J., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., Chrétien, L.: The dynamics of collective sorting: Robot-like ants and ant-like robots. In: Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 1. pp. 356–365. MIT Press, Cambridge, MA (1991)

2. Handl, J., Knowles, J., Dorigo, M.: Ant-based clustering and topographic mapping. Artificial Life 12(1), 35–62 (2006)
3. Lumer, E., Faieta, B.: Diversity and adaptation in populations of clustering ants. In: Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3. pp. 501–508. MIT Press, Cambridge, MA (1994)
4. Matusiewicz, Z., Pancerz, K.: Rough set flow graphs and max - * fuzzy relation equations in state prediction problems. In: Chan, C.C., Grzymala-Busse, J.W., Ziarko, W. (eds.) Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science, vol. 5306, pp. 359–368. Springer-Verlag, Berlin Heidelberg (2008)
5. Mitsa, T.: Temporal Data Mining. CRC Press (2010)
6. Nait-Ali, A. (ed.): Advanced Biosignal Processing. Springer-Verlag, Berlin Heidelberg (2009)
7. Pancerz, K., Lewicki, A., Tadeusiewicz, R.: Ant based clustering of time series discrete data - a rough set approach. In: Panigrahi, B.K., et al. (eds.) Swarm, Evolutionary, and Memetic Computing, Lecture Notes in Computer Science, vol. 7076, pp. 645–653. Springer-Verlag, Berlin Heidelberg (2011)
8. Pancerz, K., Lewicki, A., Tadeusiewicz, R.: Ant based clustering of two-class sets with well categorized objects. In: Greco, S., Bouchon-Meunier, B., Coletti, G., Fedrizzi, M., Matarazzo, B., Yager, R. (eds.) Advances in Computational Intelligence, Communications in Computer and Information Science, vol. 299, pp. 241–250. Springer-Verlag, Berlin Heidelberg (2012)
9. Pancerz, K., Lewicki, A., Tadeusiewicz, R., Szkoła, J.: Classification of speech signals through ant based clustering of time series. In: Computational Collective Intelligence Technologies and Applications. Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin Heidelberg (2012), to appear
10. Pancerz, K., Paja, W., Wrzesień, M., Warchoł, J.: Classification of voice signals through mining unique episodes in temporal information systems: A rough set approach. In: Proceedings of the 21th international Workshop on Concurrency, Specification and Programming (CS&P 2012). Berlin, Germany (2012), to appear
11. Pawlak, Z.: Flow graphs and data mining. In: Peters, J., Skowron, A. (eds.) Transactions on Rough Sets III, pp. 1–36. Springer-Verlag, Berlin Heidelberg (2005)
12. Semmlow, J.: Biosignal and Medical Image Processing. CRC Press (2009)
13. Szkoła, J., Pancerz, K., Warchoł, J.: Computer-based clinical decision support for laryngopathies using recurrent neural networks. In: Hassanien, A., et al. (eds.) Proc. of the ISDA'2010. pp. 627–632. Cairo, Egypt (2010)
14. Szkoła, J., Pancerz, K., Warchoł, J.: Computer diagnosis of laryngopathies based on temporal pattern recognition in speech signal. Bio-Algorithms and Med-Systems 6(12), 75–80 (2010)
15. Szkoła, J., Pancerz, K., Warchoł, J.: Recurrent neural networks in computer-based clinical decision support for laryngopathies: An experimental study. Computational Intelligence and Neuroscience 2011 (2011), article ID 289398
16. Warchoł, J.: Speech Examination with Correct and Pathological Phonation Using the SVAN 912AE Analyser (in Polish). Ph.D. thesis, Medical University of Lublin (2006)