ZBIGNIEW ŚWIERCZYŃSKI, TOMASZ GUSZKOWSKI

Wroclaw University of Technology
Faculty of Electronics
Chair of Electronic and Photonic Metrology
Poland, e-mail: zbigniew.swierczynski@pwr.wroc.pl, tomasz.guszkowski@pwr.wroc.pl;

APPLICATION OF SOM IN CLASSIFICATION OF EGG SIGNALS

The report presents problems associated with computer aided gastric diagnosis. The subject of the study are electrogastrographic (EGG) signals (non-invasively measured electrical signals generated by the human stomach). The signals were digitally recorded and then parametrized, with linear autoregressive models (AR). The data and parametrization method used in the study were the same as used by the authors in the previous study; therefore here they are only shortly described. The sets of numbers, obtained by these means, were treated as information vectors, and classified with the Self Organizing Map (SOM) classifier. The structure and parameters of the algorithm used for classification of the parametrized EGG data are described. The final efficiency of the whole system (SOM classifier with the parametrization method applied), reaching 80%, is promising. It is similar to the results of other classifiers. The ways to improve the effectiveness are also outlined.

Keywords: EGG, electrogastrographic signals, gastric electrical activity, SOM, Self Organizing Map, parametrization, classification, stomach

1.INTRODUCTION

Nowadays gastric diseases are very common. They are mainly caused by unhealthy diet and stressful life. Unfortunately, most of the gastric system (GS) diagnosing methods used in medicine are invasive (e.g. gastroscopy). This sometimes restrains the patient from visiting a physician when the disease is in its early stage. On the other hand, nobody declines ECG measurements which are simple, cheap and non-invasive. The situation could change, if similar methods for diagnosis of the GS system were introduced. GS, like the other systems/organs of the human body, generates electrical signals, which can be measured through sensors connected to the surface of the body. These signals are known as electrogastrographic signals (EGG). To put it strictly, the EGG signals are non-invasive recordings of gastric electrical activity (GEA), obtained with cutaneous electrodes [1, 2, 3]. GEA is related to all the processes in the GS generating electrical signals, but usually is associated only with the stomach. A useful EGG signal has a low frequency (up to 0.25 Hz) and amplitude [1, 2, 4, 5]. However, the signals cannot be used directly for medical diagnosis, because valuable information is difficult to extract from the signal shape. Additional difficulties arise from the fact that useful signals are masked by noise and artifacts generated by various organs of the human body. Because of the problems with the extraction of valuable information, special methods of analysis are required. In spite of the extensive investigation [5, 6, 7], an adequate method to classify EGG signals generated in different stomach or GS diseases has not yet been developed. For this reason, EGG applications are not common.

EGG signals are stochastic in their nature; therefore, to analyze them, non-random parameters and/or characteristics are required. If the signals are to be classified, parameters connected (at least partially) with the features of the generating object (GS, in general, or stomach, in particular) should be chosen. On the other hand, the features of interest in the task, i.e. those that can be useful in classification, are only the ones that contain information about the differences. If all such features can be collected and proper values assigned to them, then the information vectors could be created and used as the input data for a classifier. To

obtain satisfying results, the classifier should be matched and tuned to input data. Unfortunately the procedure described above is not simple in real tasks. Although diagnoses of the GS have been carried out for a long time, they are mainly based on invasive GEA [1, 2, 4]. Unfortunately, the EGG signals that are non-invasively measured are not identical to GEA, and only some similarities in the signal shapes can be observed [2, 4]. Another problem is the relation between the features of the stomach that are assessed during medical diagnosis (e.g. changes caused by an ulcer) and the features or parameters of the signals. This relation has not yet been defined, though there is one parameter of EGG signals (the frequency of the main signal component), which is accepted and used by almost all the researchers working in the field [1, 8]. Although it contains very useful information, used by some investigators to reveal different states of the stomach, and it is (or can be) related to some diseases [2, 8], the information is not sufficient to make a reliable medical diagnosis. Thus to form an information vector some general techniques that convert the signal into a set of parameters are required, e.g. time series modeling. To summarize: it is possible to find such a classification system, but due to lack of information, many attempts and investigations must be undertaken, with the degree of its applicability determined by the classification results.

A study was conducted to check the effectiveness of the Self Organizing Map (SOM) classifier [9, 10] applied in gastric data recognition. The basics of the SOM are presented in Section 3. The data passed on to the classifier have the form of a set of numbers (model parameters). Section 2 briefly describes the parametrization process, i.e. the way of transformation of the recorded EGG signal into the set of the model parameters. The combination of chosen methods (parametrization and pattern recognition) provides a good classification tool, which is proven by the results presented in Section 4. The overall performance of the system is promising and can be further improved, which is pointed out in future plans (Section 5).

2. MATERIALS AND PREPROCESSING

The measurement system used for acquisition of EGG signals was presented in [3, 4, 7], and information about the signals could be found in [4, 5]. The methods used for parametrization in the study were exactly the same as in [5, 6, 7], but to make the report complete, they are briefly repeated here. The same applies to the materials.

In the study, signals from our signal database were used. It consists of signals obtained from volunteers – normal subjects and patients (diabetic, with duodenal/gastric ulcer, after partial gastric resection and others). The measurements, most often, were saved as 3-channel recordings, with the duration from 5 to 60 minutes and the sampling frequency 20 Hz (cut-off frequency of anti-aliasing filter: 0.25 Hz). The selected signals had the best signal-to-noise ratio and were performed in the preprandial (empty stomach) state.

To represent the recorded EGG signals by sets of numbers as required for the next step, they were subjected to the parametrization process. The process was performed by means of linear autoregressive models (AR). Shortly, the idea of modeling can be explained starting with Youle's idea [11, 12] or with the Wold decomposition [12]. In both cases, the stochastic process, called white noise [11, 12], is converted into another process with a linear filter (Fig. 1).

Fig. 1. The idea of modeling.

In this situation, the output process $x(n)$ (with zero mean value) can be described by the equation:

$$x(n) = \sum_{k=0}^{\infty} h(k)w(n-k)\,, \tag{1}$$

where $w(n)$ is the white noise (with zero mean value and variance σ_w^2) and $h(k)$ is the impulse response of the system. Process $x(n)$ is a linear combination of realizations of the white noise. The process can be also treated as a weighted sum of the past values and white noise [11]:

$$x(n) = -\sum_{k=1}^{\infty} a(k)x(n-k) + w(n)\,. \tag{2}$$

When in the Eq. (2) only the first p coefficients $a(k)$ are not zero, which can be written as

$$x(n) = -\sum_{k=1}^{p} a(k)x(n-k) + w(n)\,, \tag{3}$$

then the output process is called an *autoregressive process* of order p and the system is called a linear autoregressive model (AR). To find the model parameters, the second order statistics of the process [11, 12] are employed, so the parameters of the system-model describe the output process in the meaning of second order statistics. Hence, when the stochastic process is autoregressive, and all the p parameters of its model are known, then its second order statistics are known. In some situations, especially when Gaussian processes are investigated, and the phase characteristics are not the subject of the investigator's interest, information included in model parameters ($a(k)$) and the white noise variance (σ_w^2) fully describes the process [11]. In this study, the results for the set of the model parameters, with and without the white noise variance, were checked.

As described in [4, 13], for EGG signals a model order of 18 (p = 18) is optimal. The sets of p numbers are model parameters or sets of autoregressive coefficients. They were obtained through conversion of 5 minute signals from single channels. The signals came from 40 recordings (3 channels per each recording) from 20 volunteers (10 normal subjects and 10 patients, 15 men and 5 women, age 24-40 years). 100 signals were selected: (30 recordings) × (3 channels) + (10 recordings) × (1 channel); the rejected signals had unsatisfactory signal-to-noise ratio. In this way, there were 100 sets of numbers for the next step. It should be emphasized here that the signals used in the study did not consist of carefully selected groups of subjects (volunteers and/or illnesses). Some of the patients suffered from more than one disease and could also have other than gastric health problems. The other differences basically concerned the course of the disease, treatment, diet, etc.

The basic aim for the authors was to answer the question: despite such a great diversity, can the methods based on SOM be useful in classification of EGG signals in a chosen parametrized form and can the results be generalized?

3. SOM THEORY AND APPLICATION

Self Organizing Maps (SOMs) are mostly used for data clustering and visualization since they map multiparameter objects onto shapes (mostly planes) that are easier to interpret by humans [9]. To be more precise, the simple neurons (units) with linear transfer function and no bias are organized on the plane with a rectangular or hexagonal lattice (Fig. 2). Each neuron has its defined topological neighbours. Usually, unsupervised learning algorithms are used. While learning the response of the best matching unit (BMU), which mostly resembles a given sample is amplified by directed modification of its weights. Also the answer of BMU's neighbours is intensified which leads to "attraction" of similar objects on the map which, in consequence, builds up clusters containing akin elements. Due to its simplicity (one-layer only, linear neuron transfer function, no bias) learning SOMs is fast and efficient. A general introduction to Self Organizing Maps is presented in [9].

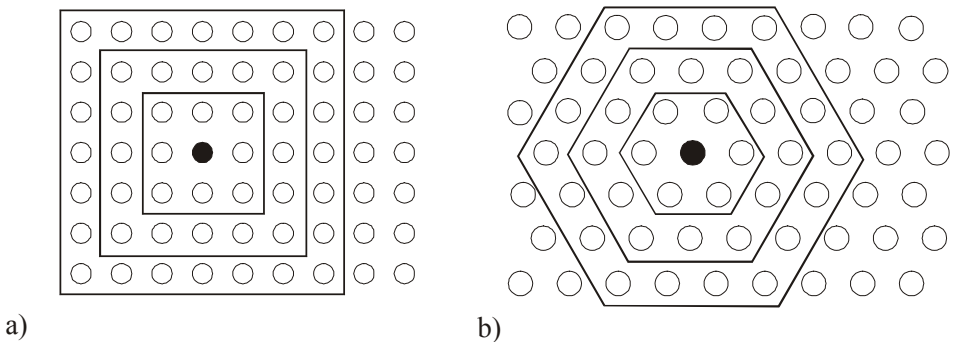

Fig. 2. A sample organization of SOM. Neighbourhood up to the third degree is shown.
a) Rectangular grid; b) Hexagonal grid.

In order to construct the classifier, a so-called Supervised SOM has been used. The tool is implemented in MATLAB as *somtoolbox2* [10]. Our map was created on a 30 units wide and 20 units high hexagonal plane. The other parameters of the network were set to the default values of the toolbox. Since the work was preliminary, only two classes (normal subject and patients) were considered. Then the training data was constructed by adding 1-of-2-coded class vector to the original, preprocessed data. After learning, the class of each map unit was determined by taking the maximum over the added columns, and applying a label, accordingly.

4. RESULTS

The database used in the study consisted of 100 EGG signals, 50 coming from normal subjects and 50 from patients. Each signal was processed as described in Section 2 and converted into the set of parameters. They were used to form information vectors of two types. The first type, of length $N = p$-1, was formed with parameters of model of order p. Due to the fact that in the AR model the first parameter always equals 1 [11], it was skipped. The second type, of length $N' = p$ - 1+1, was formed by extending the first type with the white noise variance of the model.

To evaluate the quality of the classifier the "put-aside-one-sample" technique was employed, where for N_A samples (the number of all the samples) exactly N_A classifier maps were built, with one sample from the training data successively excluded for evaluation of classifier's quality. The excluded data set was then submitted to the classifier of the number

equal to the number of the excluded sample. This procedure was looped over N_A classifiers. Then the recognition rate over the built maps was calculated as

$$R = \frac{N_G}{N_A} 100\%, \tag{4}$$

where N_G is the number of correctly classified samples. Table 1 presents the results for several vector length values (*N)*. The apostrophe " ' " indicates that the vector includes the white noise variance.

Table 1. Recognition (*R*) vs. number of parameters (*N* - vector's length). Apostrophe indicates the vector including white noise variance.

N	14	17	18'
R	80%	79%	78%

The presented results show that the increasing model order (*p*) above 15 failed to improve the effectiveness. It is not surprising, because in many cases, the model order of 15 is the correct (adequate) AR model of an EGG signal [4]. As some characteristics point (e.g. Akaike Information Criterion), further increase of the order does not significantly improve the errors [4, 11, 12]. However the results for the smaller model order (not presented here) were diverse. It was because such a model was more sensitive to the individual signal, although it was not general [4]. Moreover, adding up the white noise variance failed to increase the effectiveness (in most cases) which contradicted the initial expectations. Thus using the white noise variance in information vectors remains questionable.

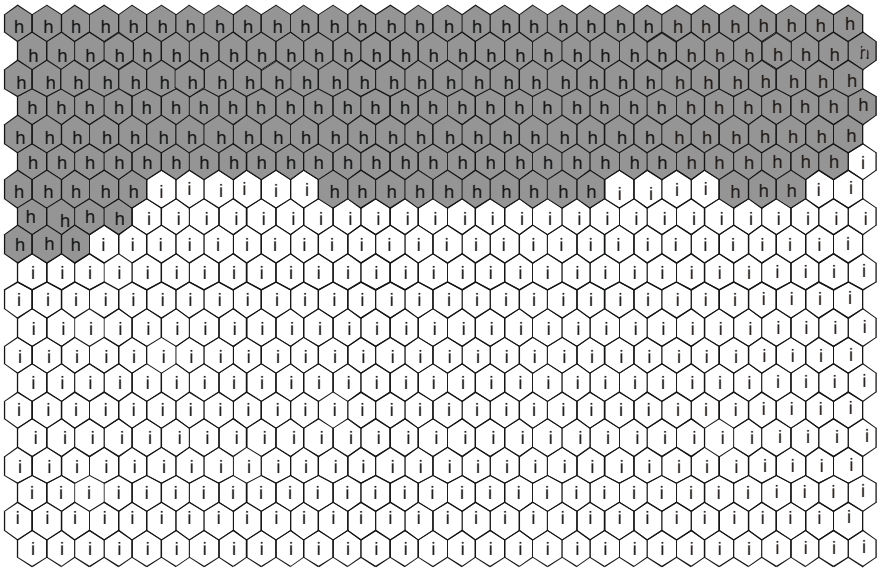

Fig. 3. Sample classifier map for a vector length equal 14. "h" mark for normal subjects, "i" mark for patient.

A new insight into the results was obtained when using a classifier map. A sample classifier map for $N = 14$ is presented in Fig. 3. It is easy to observe (on the map) the positions of data coming from normal objects and from patients and, what is very important, to identify quickly which one is "suspiciously" classified. An analysis of individual cells (data sets) through additional information about the measurement or the disease entity could be very useful. It could help to understand, for instance, why the specific set is falsely classified.

5. CONCLUSIONS

It is obvious that the requirements for the medical diagnosis system are far from being fulfilled by the effectiveness obtained in the study, even with the highest score of 80%. But the results of the early stages are promising, and they show that the EGG signal contains valuable information, and further studies should be undertaken. In this system, we can outline many problems concerning the following two aspects:

1) parameters of the classifying method;
2) adequate data representation and the sufficiency of the information they contain for the classification task.

The mentioned problems have various influence on the final result which requires many experiments to find a proper setup, producing the expected effectiveness. Taking this into consideration, there are elements that can be modified in the proposed system to obtain better results. As the above results show, the tuning of the classifying method can significantly enhance the overall performance. So, to better match the structure of the parametrized gastric signals, fine tuning of the map parameters, such as the map size, lattice, neighbourhood definition, learning rate, learning algorithm and input data normalization, should be considered. Moreover, on the basis of the experience, the element that requires a lot of work is preparation of an information vector of the EGG signals.

The result presented here is very similar to the results obtained with other classifiers [5, 6, 7]. This fact indicates that the SOM-based classifier is able to extract almost the same amount of information. In addition, although a higher effectiveness was expected, the achieved result can suggest that EGG data does not contain enough information for diagnostic purposes or, in the chosen configuration, there is too much useless or distorted information. To change this, further studies concerning the parameter vector should be undertaken. For instance, it could be completed with information concerning the disease, e.g. the onset, course, treatment and its duration, concurrent diseases and other features that could affect functioning of the stomach and carry relevant diagnostic information. Because the SOM method allows using different types of data in one vector, it seems to be a promising tool.

In our research the results presented in the report pertain only to the preprandial state. In the future, a study of the signals in the postprandial (after food stimulation) state is planned.

ACKNOWLEDGMENT

The authors express their gratitude to T. P. Sebzda, MD for his support in EGG data acquisition.

REFERENCES

1. Chen J., McCallum R.W.: *Electrogastrography: measurement, analysis and prospective applications*, Med. Biol. Eng. & Comput. 1991, 29:339-50.
2. Mintchev M.P., Kingma Y.J., Bowes K.L.: *Accuracy of cutaneous recordings of gastric electrical activity*, Gastroenterology, 1993, 104:1273-80.
3. Świerczyński Z., Hańczycowa H., Sebzda T., Leszczyszyn J., Ponikowski P., Głowacki M.: *Acquisition and analysis of electrogastrographic signals*, Acta Bio- Optica et Informatica Medica, 1, 1997, vol. 3, pp. 45-50. (in Polish)
4. Świerczyński Z.: *Parametrization of biomedical signals generated by the human stomach*, PhD Thesis, Institute of Telecommunication and Acoustics, Faculty of Electronics, Wroclaw University of Technology, Wrocław 2002. (in Polish)
5. Świerczyński Z., Zagańczyk A.: *Application of Neural Network and Genetic Algorithms in Computer Aided Gastric Diagnostic System*, BBE, 2005, vol. 25, no 1, pp. 49-58.

6. Świerczyński Z., Zagańczyk A.: *Application of NN and GA in Computer Aided EGG Diagnostic System*, Proc. of the Biocybernetyka i inżynieria biomedyczna. XIII Krajowa Konferencja Naukowa, 2003, p. 231-236. (in Polish)
7. Świerczyński Z., Mazur J.: *Application of SVM in Computer Aided Gastric Diagnostic System*, BBE, 2004, vol. 24, no 4, pp. 19-30.
8. Świerczyński Z.: *Application of harmonics estimation methods in extraction of frequency of the main component of EGG signals*, Proc. of the Kongres Metrologii, 2004, pp. 547-550. (in Polish)
9. Kohonen T.: *Self-Organizing Map*, Springer-Verlag, Berlin 1995.
10. Vesanto J., Himberg J., Alhoniemi E., Parhankangas J.: *SOM Toolbox for Matlab 5*, Libella Oy, Espoo 2000.
11. Box G. E. P., Jenkins G.M.: *Time series analysis. Prediction and control*, PWN, Warszawa 1983.
12. Kay S.M.: *Modern Spectral Estimation: Theory and Application*, Engelwood Cliffs, Prentice Hall, New Jersey 1988.
13. Muciek A., Świerczyński Z.: *Parametric models of biomedical signals*, Proc. of the Third Int. Symposium on Methods and Models in Automation and Robotics, 1996, vol. 2, pp. 683-687.

ZASTOSOWANIE SOM DO KLASYFIKACJI SYGNAŁÓW EGG

Streszczenie

Praca przedstawia problemy związane z komputerowo wspomaganym diagnozowaniem układu pokarmowego. Obiektem badań są tutaj sygnały elektrogastrograficzne – EGG (nieinwazyjnie mierzone sygnały elektryczne generowane przez żołądek człowieka). Sygnały te zostały zarejestrowane cyfrowo a następnie poddane parametryzacji przy pomocy liniowego modelu autoregresyjnego AR. Dane oraz metoda parametryzacji użyta w przedstawionych badaniach zostały opisane w poprzednich pracach autorów, więc tutaj ujęte są jedynie w zarysie. Zestawy liczb otrzymane w wyniku parametryzacji potraktowane zostały jako wektor parametrów i sklasyfikowane przy pomocy klasyfikatora opartego na samoorganizujących się mapach (SOM). W pracy przedstawiono strukturę i parametry użytego algorytmu. Ostateczna skuteczność całego systemu (tj. klasyfikatora SOM oraz zastosowanej metody parametryzacji) wyniosła 80 %, co jest wynikiem obiecującym i bardzo podobnym do tych jakie osiągnięto przy zastosowaniu innych metod klasyfikacji. Praca przedstawia również zarys metod poprawy efektywności opisanego systemu.