

A Flexible, Low-Power, Programmable Unsupervised Neural Network Based on Microcontrollers for Medical Applications

Rafał Długosz^{1,2}, Tomasz Talaśka², Paweł Przedwojski², Paweł Dmochowski³

¹Institute of Microtechnology,
Swiss Federal Institute of Technology, Lausanne,
Neuchâtel, Switzerland
rafal.dlugosz@epfl.ch

²Faculty of Telecommunication and Electrical Engineering
University of Technology and Life Sciences (UTP)
Bydgoszcz, Poland
talaska@utp.edu.pl, pawel.przedwojski@wp.pl

³Victoria University of Wellington
School of Engineering and Computer Science
Wellington, New Zealand
pawel.dmochowski@vuw.ac.nz

Abstract:

We present an implementation and laboratory tests of a winner takes all (WTA) artificial neural network (NN) on two microcontrollers (μ C) with the ARM Cortex M3 and the AVR cores. The prospective application of this device is in wireless body sensor network (WBSN) in an on-line analysis of electrocardiograph (ECG) and electromyograph (EMG) biomedical signals. The proposed device will be used as a base station in the WBSN, acquiring and analysing the signals from the sensors placed on the human body. The proposed system is equipped with an analog-to-digital converter (ADC), and allows for multi-channel acquisition of analog signals, preprocessing (filtering) and further analysis.

Keywords:

WTA network, microcontrollers, wireless sensor networks, medical applications

1. INTRODUCTION

Artificial Neural Networks (NN) are commonly used in tasks requiring processing, classification and recognition of "difficult" signals, such as, for example, heuristic data and non-stationary signals. They find applications in medical health care, telecommunication, and various other electrical engineering areas. In literature one can find various implementation techniques of various NNs, both the software- and the hardware-based [1, 2, 3].

Taking into account such criteria as energy consumption and calculation capacity, *fully-custom* designed networks are the most efficient solutions. Such networks enable parallel data processing, and thus are faster than their software-based counterparts. The circuits, which are designed on transistor level, enable a good matching of the internal structure to a given task, and therefore usually consume much less energy [4, 5]. On the other hand, they require relatively complex and time consuming design process, which is expensive, especially in case of realization of short series.

In this paper we present an implementation of the Winner Takes All (WTA) NN using an alternative approach that is based on microcontrollers (μ Cs). Such

networks are significantly less efficient than those realized in the *full-custom* style, but are up to ten times more efficient compared to similar NNs realized on PC [5]. Using the Eagle 5.6 environment, we developed a prototype testing board with two μ Cs, namely the 8-bits AVR and the 32-bits ARM CortexM3. The device is described in details in Section 3.

The motivation behind the realization of this device was manifold. The main purpose was to realize a flexible, programmable and convenient to use system for the application in various medical healthcare areas. The device is to be used as a base station in the Wireless Body Sensor Network (WBSN) for an on-line monitoring of patients. For this purpose, future versions will be equipped with the wireless I/O module. The WBSNs become more and more popular in recent years and there promises to be more interest in the future [6].

An interesting aspect is using the NNs implemented in WBSNs in the analysis of various biomedical signals, such as electrocardiograph (ECG) and electromyograph (EMG) signals. In [7] it was demonstrated that NNs are a very efficient tool in the analysis of such signals. Three different learning rules were investigated, specifically the self organizing map (SOM), the back-propagation

(BP) NN and the learning vector quantization (LVQ) NN. The unsupervised SOM algorithm was shown to be more efficient than, for example, the BP NN, although the learning process takes more time in this case.

The other reason behind realization of this device was to build a system that could be used as a hardware model of the analog and mixed (analog-digital) WTA NNs realized by the authors as Application Specific Integrated Circuits (ASIC) in the CMOS technology [5, 8]. As mentioned above, the overall design process of such networks is expensive and time consuming. The authors' experience shows that modeling of such chips using only Matlab or C++ is not always sufficient, especially in case when analog signals are to be processed and analysed. For this reason the proposed device is equipped with the ADC and the DACs that enable multi-channel processing of analog signals in the real time. Transistor-level designed NNs can be tested using the Spice environment, but simulations of even relatively short times require hours on a typical PC, which is the bottleneck in optimization of such chips. The proposed platform enables fast reprogramming and thus optimization of even large WTA NN in a short time.

It is worth mentioning that *fully-custom* designed NNs enable realization of the WBSN of different types. In the first approach, in which the NN is implemented in the base station, the sensors can be very simple. Their role is reduced in this case to data collection, and some preprocessing, such as data compression. In an alternative approach the ultra-low power unsupervised NNs can be used directly in particular sensors. In this case data exchange with the base station is reduced to the necessary minimum. This approach enables significant increasing the battery life time, as the communication usually consumes more than 90% of total energy consumed in such systems.

In the literature one can find the attempts to realize various NNs using μ Cs and microprocessors (μ Ps) [9, 10]. An example realization by use of the Single Instruction Multiple Data (SIMD) processor has been described in [9], where different methods of detection of the winning neuron were studied. In general, the SIMD μ Ps are suitable for large and fast networks with even hundreds neurons. The main reason for this is relatively large power dissipation, as well as large cost of a single device. Another realization of a multi-layer network using the PIC18F45J10 μ C has been described in [10]. In this case the maximum of 256 weights (connections) can be realized, which in practice means about 50 neurons.

NNs realized on μ Cs find the application in different areas, mostly in control and in diagnostics. For example, a device described in [11] has been used as an intelligent wireless electronic nose node (WENN) used in classification and quantification of binary gas mixtures NH₃ and H₂S. A NN described in [12] has been used to control temperature of a furnace.

A very important aspect in hardware realized NNs is the complexity of the learning algorithm. It has the influence on the power dissipation and achievable data rate.

Microcontrollers are rather suitable for simple arithmetic operations. For example, the unsupervised trained WTA NN requires only simple operations, such as multiplications, summations and subtractions. For the comparison, the NN described in [10] requires *tanh* activation functions, relatively more difficult in hardware realization.

2. BASICS OF THE WTA NETWORKS

In this section we outline the fundamentals of the WTA learning algorithm. WTA NN [13] belongs to the group of networks trained without the supervision, making them relatively fast, which is important in applications such as telecommunications [14, 15]. The training process relies on presenting the NN with learning patterns, X , in order to make the neurons' weight vectors, W , resemble presented data. For each new pattern the network calculates the distance between the X and the W vectors in all neurons. Different measures of the similarity between both vectors can be found in the literature. One of them is the Euclidean distance (L2) defined as:

$$d(X, W_i) = \sqrt{\sum_{l=1}^n (x_l - w_{il})^2} \quad (1)$$

Another popular measure is the Manhattan (L1) distance, defined as:

$$d(X, W_i) = \sum_{l=1}^n |x_l - w_{il}| \quad (2)$$

In the L1 measure the squaring and rooting operations have been omitted, which allows for simplification of the learning algorithm. Both these measures have been implemented in the proposed device for the comparison. The adaptation of the winning neuron in the t^{th} iteration is performed in accordance with the formula:

$$W_i(t+1) = W_i(t) + \eta \cdot (X(t) - W_i(t)) \quad (3)$$

where η is the learning rate. Other neurons in the network that lose the competition remain unchanged.

A significant problem encountered in the WTA networks are the, so-called, dead neurons i.e. the neurons that take part in the competition but never win and their weights remain unchanged. One of the reasons for this problem are badly selected initial values of the weights [16]. Such neurons reduce the number of classes that can be discriminated, thus increasing the mapping (quantization) error of the network. For this reason reducing the number of dead neurons is an important objective. One of the very efficient methods in this task is using the conscience mechanism [5, 17]. Its role is to increase the likelihood of winning the competition for all neurons in the NN. For this reason, the conscience mechanism has been implemented in the proposed device. This mechanism can be described by:

$$d_{\text{cons}}(X, W) = d_{\text{L1/L2 norm}}(X, W) + L_{\text{count}} \cdot K \quad (4)$$

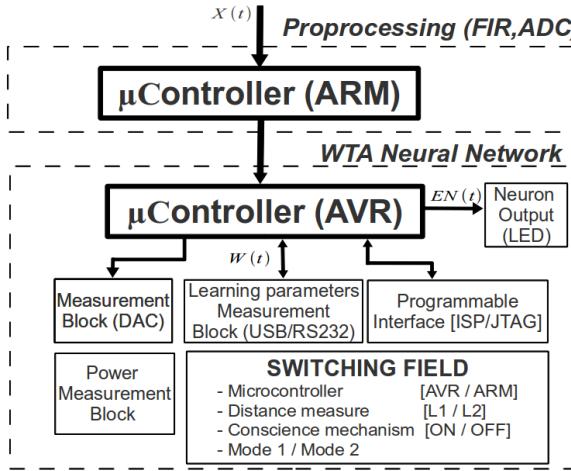


Figure 1. The proposed testing board with the programmable WTA NN based on two microcontrollers

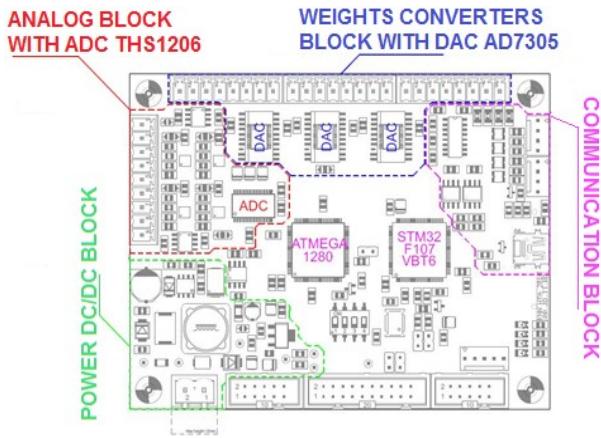


Figure 2. The proposed testing board of the programmable WTA neural network based on microcontrollers

The real distance $d_{L1/L2}(X, W)$ between the W and the X vectors is increased by adding a signal that is proportional to the number of the wins for a given neuron. Finally, in detection of the winning neurons a modified distance $d_{\text{cons}}(X, W)$ is being used. The L_{count} parameter is the number of the wins of a given neuron. The K coefficient is the gain factor that allows for controlling and optimizing the learning process by adjusting the strength of the conscience mechanism.

3. THE PROPOSED DEVICE

An overview of the realized system with two μ Cs is shown in Fig. 1, while the layout and the photograph of the testing board shown in Fig. 2. It is composed of several important blocks. One of them is the interface block, mentioned in Section 1, composed of the ADC and the DAC circuits that enable multi-channel data processing. A single 4-channel THS1206 ADC and three 4-channel AD7305 DACs have been used. This 12-bits ADC converts the analog input learning signals X , while the DACs enable a direct observation of selected neuron weights, w , on the oscilloscope. A direct observation is possible, for example, for 3 inputs and 4 outputs corresponding to 12 neuron weights. The higher number of the weights can be implemented, but in this case only selected weights can be observed directly, while the others can be viewed on PC. The interface block enables sending all weights, as digital signals to PC for a more detailed analysis. Digital signals are acquired throughout the USB and RS232 serial ports. The serial ports allow also for acquiring the learning signals X (in digital form), the calculated distances, d , between the X and the W vectors, the numbers of the wins of particular neurons and the quantization error for detailed analysis of the network performance. The numbers of the wins enable creation of statistics, which is very useful in many applications.

The overall device has been realized in a manner to enable on-line measuring of the power dissipation, separately for the ADC, the DAC blocks and for both μ Cs.

The core blocks are the μ Cs that are programmed by the use of the ISP/JTAG interfaces. The μ Cs can operate in different modes. In the first mode particular μ Cs work separately, performing the learning algorithm of the NN. In this case only one μ C is being used, while the second one is turned off to reduce the power dissipation. This mode enables a direct comparison of both devices in terms of the attainable data rate and power dissipation. In this mode the active μ C receives data from the ADC or directly the external digital signals. In the second case the ADC is turned-off to save energy.

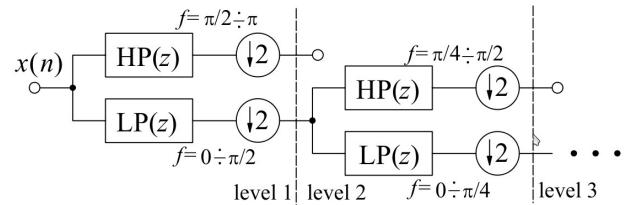


Figure 3. Cascaded discrete wavelet transform (DWT) realized using the QMF filter bank with the FIR filters

In the second mode, which is currently tested and optimized, both μ Cs are used at the same time, connected in series. The first μ C (ARM) is in this case used as a signal preprocessing / conditioning block, while the second one (AVR) performs the learning algorithm. Using the ARM μ C as a first block is necessary, as data preprocessing is usually more

complex than the subsequent WTA learning algorithm, requiring more computing resources. The used Cortex μ C is a 32-bits device, which is more convenient in data preprocessing (mostly filtering) than the 8-bits AVR μ C.

Table 1. An example realization of the wavelet Daubechies (Db10) transfer functions with reduced precision

LP theor.	LP round	LP binary	HP theor.	HP round	HP binary	HPx round
-0.0076	-1	1000001	-0.0189	-2	1000010	-2
0.0010	0	0000000	0.1331	15	0001111	17
0.0026	3	0000011	-0.3728	-41	1101001	-48
-0.0208	-2	1000010	0.4868	53	0110111	62
-0.0505	-6	1000110	-0.1988	-22	1010110	-25
0.0658	7	0000111	-0.1767	-19	1010011	-23
0.0901	10	0001010	0.1386	15	0001111	18
-0.1386	-15	1001111	0.0901	10	0001010	12
-0.1767	-19	1010011	-0.0658	-7	1000111	-8
0.1988	22	0010110	-0.0505	-6	1000110	-6
0.4868	53	0110111	0.0208	2	0000010	3
0.3728	41	0101001	0.0026	3	0000011	3
0.1331	15	0001111	0.001	0	0000000	0
0.0189	2	0000010	-0.0076	-1	1000001	-1

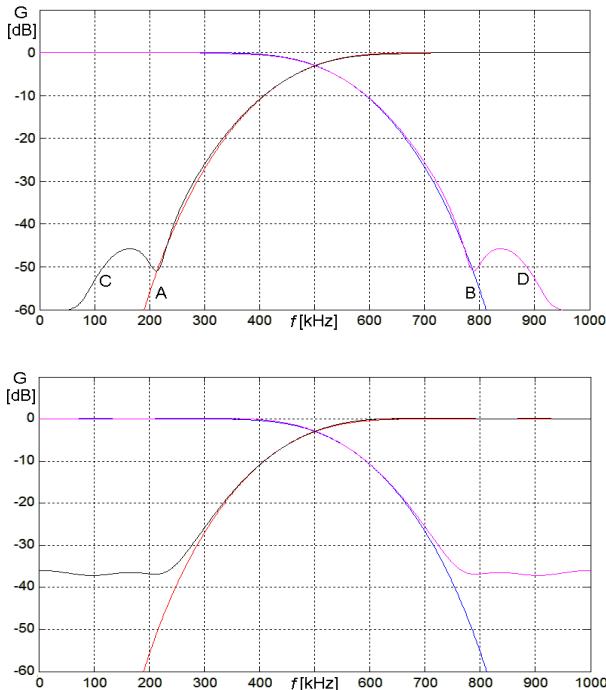


Figure 4. Frequency responses of the LP (B, D) and the HP (A, C) Daubechies (Db10) filters used in the DWT. The upper plot is for the optimal rounding, while the lower for not optimal rounding (HPx) of the filter coefficients.

Both of the above modes make the realized device a powerful autonomous system, suitable for WBSN. The second mode is important in case of the analysis of the ECG/EMG biomedical signals. Such signals must be first de-noised, and then some characteristic points must be extracted from the complexes [18].

Data preprocessing is based on the Finite Impulse Response (FIR) filtering. The Infinite Impulse Response (IIR) filters were also considered, but the FIR filters offer the attractive property of linear phase response. Data extraction is performed by means of a multistage discrete wavelet transform (DWT), shown schematically

in Figure 3. DWT is a series of the filtering operations performed by use of the quadrature mirror filter (QMF) bank composed of the lowpass, $LP(z)$, and the highpass, $HP(z)$, halfband FIR filters. Each stage is followed by decimation by a factor of 2.

The problem which must be considered at this stage is data resolution. The analog input signals X are converted into 8 – 12 bit digital form. As mentioned above, the proposed device is also used as the hardware model of the system that will be realized as an ASIC. In such implementations each bit significantly increases the number of transistors (and consequently, the chip area), and the power dissipation. For this reason data resolution has been limited to 16 bits only, and thus the FIR filter coefficients are rounded to 6 bits + 1 bit representing the sign, as shown in Table 1 for an example of Daubechies wavelet. In this situation the rounding operation must be performed very carefully to avoid the loss of the dynamic range of the filter, as shown in Figure 4. The non-optimal HPx case in Table 1 is shown at the bottom diagram. In this case the rounding factor is even less restrictive, but the loss of the attenuation exceeds 10 dB.

Data preprocessing is required as it simplifies the input signals provided to the NN, thus making the analysis performed by this network feasible. In particular, it minimizes the number of required network inputs, as particular features are provided to separate inputs. In [19] the ECG signals were decomposed into only four features, specifically, the span of the QRS wave, the interval of the R-R segment, the voltage and the slope of the S-T segment. The results obtained in [19] are in the good agreement with the diagnosis made by the medical staff. The analysis in [19] was performed using the WTA network. The results for the WTA NN were compared with the results obtained in case of using of the BP network. This conclusion is important, as the WTA NN is much simpler in the hardware realization than the BP algorithm.

Summarizing, the realized device enables operation in different modes, with either the L1 or the L2 distance measures, with or without the conscience mechanism, with different numbers of neurons, with the analog or digital input data. The system is still being developed and optimized. One of the modules that will be added in the next version is the wireless I/O module, to enable the application of the device in the WBSN.

4. LABORATORY TESTS

One of the important tests was to determine the maximum achievable data rate. This parameter depends on the number of neurons in the network, since all calculations inside the μ C are preformed serially. This is a disadvantage in comparison with the NNs realized as ASIC, in which a fully parallel data processing can be easily implemented. The input data rate can be estimated by use of the following equation:

$$f_{\text{data}}(n) = f_{\max} / (N_O \cdot n) \quad (6)$$

In (6) f_{\max} is the maximum clock frequency of a given μ C, equal to 16 MHz and 72 MHz for the AVR and the ARM μ Cs, respectively. The N_0 parameter is the number of the clock cycles per a single input pattern X , in the NN with n neurons. The N_0 parameter differs for the AVR and the ARM μ C. The ARM μ C is more efficient, so although f_{\max} is in this case only 5 times larger, the NN operates more than 7 times faster. The data rate depends on the distance measure, and is approximately two times larger for the L1 case, as no squaring / rooting operations are required in this case. This is shown in Figure 5, as a function of the number of neurons.

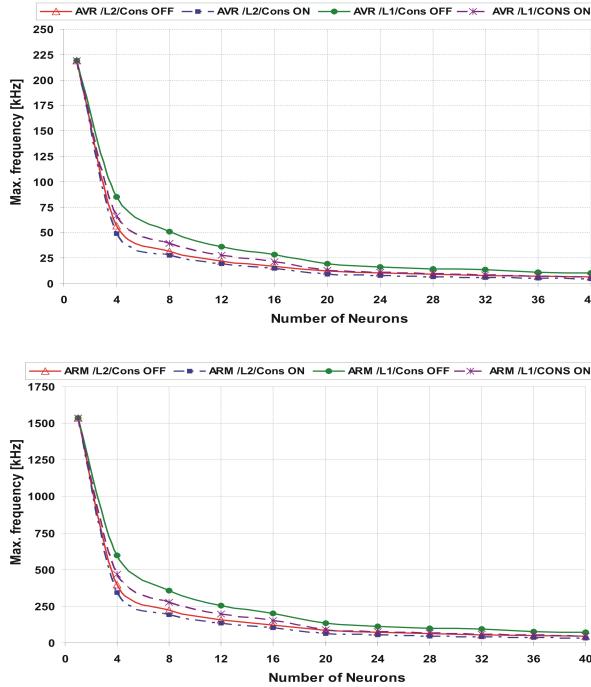


Figure 5. Achievable data rate of the WTA NN as a function of the number of neurons for both μ Cs for (top) the AVR, and (bottom) the ARM microcontroller.

Fig. 6 presents selected measurement results of the NN with 3 inputs and 4 outputs realized on the ARM μ C sampled at 135 kHz. An example adaptation process has been presented for two selected neurons. The EN is the output signal of the WTA block that determines the winning neuron, i.e. the neuron the most resembling the input pattern X . Only this neuron can adapt its weights. As can be observed, when the EN signal becomes a logical '1', the weights of the corresponding neuron are modified. The following input signals were provided to the NN:

- x_1 – the sine signal with the frequency of 5 kHz;
- x_2 – the triangular signal with the frequency of 10 kHz
- x_3 – the rectangular signal with the frequency of 2 kHz.

Figure 7 illustrates the influence of the conscience mechanism on the learning quality of the network. In this case data acquired throughout the serial port for the NN with 10 neurons are shown. The conscience mechanism is able to activate the neurons that otherwise would remain inactive. Figure 7 illustrates the influence of dead neurons on the mapping properties. In case (a) all neurons are active, becoming representatives of particular data classes. Case

(b) is shown for the “too weak” conscience mechanism. In this case the number of dead neurons is smaller than in (a), but is not zero. In the worst case (c) only several neurons took part in the competition.

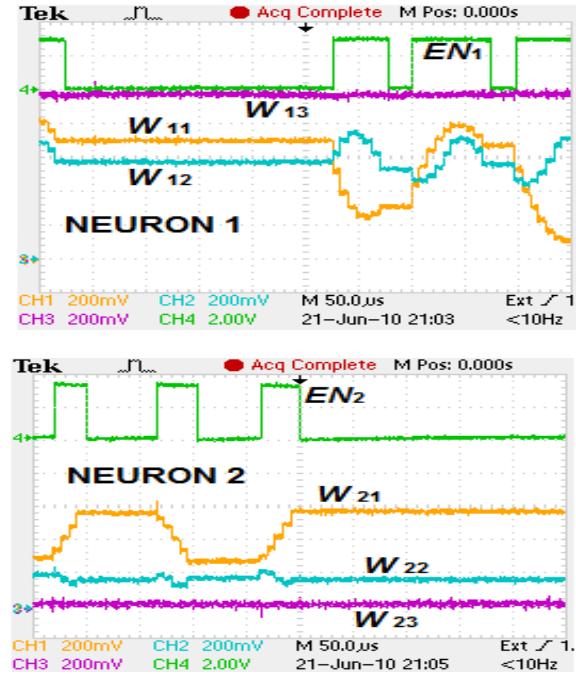


Figure 6. Example measurement results of the WTA NN with 3 inputs and 4 outputs (4 neurons). The diagrams present selected neuron weights, w , as analog signals.

An interesting aspect is the comparison of the NN realized on μ C with the analog network realized earlier by the authors in the CMOS 180nm technology. The realized device consumes an average power of 300 mW and 500 mW, for the AVR and the ARM μ C, respectively. For the comparison, the analog WTA network with 12 weights, sampled at 1 MHz, dissipated the power of 700 μ W, i.e. approximately 500 times less than in case of the realization on μ Cs. Taking into account that the sampling frequency is now two times smaller, the analog NN are 1000 times more efficient.

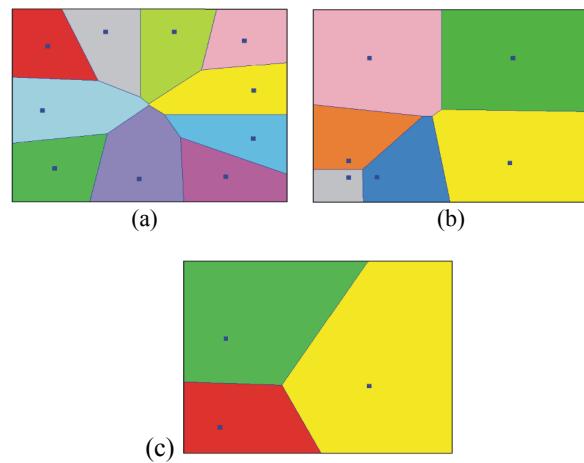


Figure 7. Voronoi's diagrams illustrating the influence of the conscience mechanism on the final placement of neurons. Depending on the strength of this mechanism the number of dead neurons varies in-between 0 and 70%. Te results are shown for the NN with 10 neurons and digital weights transferred to PC.

5. CONCLUSION AND FURTHER WORK

A new implementation of the unsupervised trained Winner Takes All neural network (WTA NN) on microcontrollers (μ C) with the AVR and the ARM cores has been presented. We realized a prototype testing board with both μ C operating either separately or cooperatively. In the second case the ARM μ C is used in data pre-processing / conditioning, which relies on the finite impulse response (FIR) filtering and extraction of useful information from the input signals. The output signals of the ARM μ C become training signals provided to the AVR μ C that in this case performs the WTA classification algorithm. The prospective application of this device is in the Wireless Body Sensor Network (WBSN) in an on-line analysis of the biomedical ECG and the EMG signals.

The measurement results show that in the comparison with the same NN realized on PC, the network realized on μ C is even 10 times more efficient taking into account such parameters as the achievable data rate and the power dissipation. On the other hand, such network is even 1000 times less efficient than the same network realized as analog chip. The advantage of the proposed realization is relatively low cost of a single device and small sizes.

In future work, the board will be equipped with the wireless I/O ports, enabling applications in the wireless medical diagnostics systems, serving as a base station acquiring biomedical signals from the sensors.

6. REFERENCES

- [1] Macq, D. *et al.*, "Analog Implementation of a Kohonen Map with On-Chip Learning", *IEEE Transactions on Neural Networks*, Vol. 4, No. 3, May 1993, pp. 456-461
- [2] M. Holler *et al.*, "An electrically trainable artificial neural network (ETANN) with 10240 'floating gate' synapses", *International Joint Conference on Neural Networks*, 1989, pp. 191-196
- [3] J. Choi *et al.*, "A programmable analog VLSI neural network processor for communication receivers", *IEEE Transactions on Neural Networks*, Vol. 4, No. 3, 1993 r.
- [4] Dlugosz R., Talaska T., "Low power current-mode binary-tree asynchronous Min/Max circuit", *Microelectronics Journal*, Vol. 41, no 1, Jan. 2010, pp. 64-73
- [5] R. Dlugosz *et al.*, "Realization of the Conscience Mechanism in CMOS Implementation of Winner-Takes-All Self-Organizing Neural Networks", *IEEE Transactions on Neural Networks*, Vol. 21, No. 6, 2010, pp. 961-971
- [6] C.C. Enz *et al.*, "WiseNET: An Ultra low-Power Wireless Sensor Network Solution", *Computer*, August 2004, Vol. 37, No. 8
- [7] Lin He *et al.*, "Recognition of ECG Patterns Using Artificial Neural Network", *International Conference on Intelligent Systems Design and Applications* (ISDA), Vol. 2, 2006, pp. 477-481.
- [8] Dlugosz R., et al., "Programmable triangular neighborhood functions of Kohonen Self-Organizing Maps realized in CMOS technology", *European Symposium On Artificial Neural Networks* (ESANN), Bruges, Belgium 2010, pp.529-534
- [9] B. Mailachalama and T. Srikanthan, "Area-time issues in the VLSI implementation of self organizing map neural networks", *Microprocessors and Microsystems*, Vol. 26, no. 9-10, 2002, pp. 399-406.
- [10] N.J. Cotton *et al.*, "A Neural Network Implementation on an Inexpensive Eight Bit Microcontroller", *International Conference on Intelligent Engineering Systems*, USA, 2008, pp. 109-114.
- [11] Young Wung Kim *et al.*, "An Intelligent Wireless Electronic Nose Node for Monitoring Gas Mixtures Using Neuro-Fuzzy Networks Implemented on a Microcontroller", *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, Italy, 2007, pp.100-104.
- [12] H. M. Mousa *et al.*, "On-Line Neurocontroller Based on Microcontrollers", *IEEE International Conference on Industrial Technology*, Hong-Kong, 2005, pp. 1252-1256.
- [13] T. Kohonen, *Self-Organizing Maps*, Springer Verlag, Berlin, 2001.
- [14] C. Amerijckx *et al.*, "Image Compression by Self-Organized Kohonen Map," *IEEE Transactions on Neural Networks*, Vol. 9, No. 3, 1998, pp. 503-507.
- [15] C. Chang *et al.*, "New Adaptive Color Quantization Method Based on Self-Organizing Maps," *IEEE Transactions on Neural Networks*, Vol. 16, No. 1, 2005, pp. 237-249.
- [16] T. Talaska and R. Dlugosz, "Initialization mechanism in Kohonen neural network implemented in CMOS technology", *European Symposium on Artificial Neural Networks* (ESANN), Bruges, Belgium, 2008, pp. 337-342.
- [17] D. DeSieno, "Adding a conscience to competitive learning", *IEEE Conference on Neural Network*, Vol. 1, 1988, pp. 117-124.
- [18] Paul S Addison P. S., "Wavelet transforms and the ECG: a review", *Physiological Measurement*, Vol. 26. No. 5, 2005
- [19] Chen Tian-hua *et al.*, "The Sorting Method of ECG Signals Based on Neural Network", *International Conference on Bioinformatics and Biomedical Engineering*, 2008, pp. 543 - 546.