

Limited receptive area neural classifier for recognition of swallowing sounds using short-time Fourier transform

Oleksandr Makeyev, *Associate Member, IEEE*, Edward Sazonov, *Member, IEEE*, Stephanie Schuckers, *Member, IEEE*, Ed Melanson, and Michael Neuman, *Member, IEEE*

Abstract— In this paper we propose a sound recognition technique based on the limited receptive area (LIRA) neural classifier and short-time Fourier transform (STFT). LIRA neural classifier was developed as a multipurpose image recognition system. Previous tests of LIRA demonstrated good results in different image recognition tasks including: handwritten digit recognition, face recognition, metal surface texture recognition, and micro work piece shape recognition. We propose a sound recognition technique where spectrograms of sound instances serve as inputs of the LIRA neural classifier. The methodology was tested in recognition of swallowing sounds. Swallowing sound recognition may be employed in systems for automated swallowing assessment and diagnosis of swallowing disorders. The experimental results suggest high efficiency and reliability of the proposed approach.

I. INTRODUCTION

Many signal processing methods have been developed for analysis of non-stationary signals, such as short-time Fourier transform. The difficulty with these techniques in pattern recognition applications is that the size of the feature space is multiplied from a one-dimensional signal to a two-dimensional “image”. More research is needed to develop techniques to extract features for pattern recognition applications. In this paper, we propose to apply LIRA-based image recognition technique to the “images” of the time-frequency decomposition of a sound instance.

Limited receptive area (LIRA) neural classifier was developed as a multipurpose image recognition system based on the paradigm of Random Local Descriptors (RLDs) [1], [2]. RLDs play the role of features to be extracted from an image. The advantage of this type of features is a possibility to create sufficiently general description of the image.

The LIRA neural classifier was tested in different image recognition tasks including: handwritten digit image

recognition [3], micro device assembly [4], mechanically treated metal surface texture recognition [5], face recognition [6], and micro work piece shape recognition [1]. The best result for handwritten digit recognition on the MNIST database [7] was the error rate of 0.55% [3] and for face recognition on the ORL database [6] it was the error rate of 0.1% [6]. The results obtained in micro device assembly, metal texture and micro work piece shape recognition are also promising [1].

In this paper we propose a sound recognition technique combining the LIRA neural classifier and short-time Fourier transform (STFT). The suggested sound recognition methodology is tested in the task of recognition of swallowing sounds. This is an important task in bioengineering because swallowing sound recognition may be employed in systems for automated swallowing assessment and diagnosis of abnormally high rate of swallowing (aerophagia) [8], which is the primary mode of ingesting excessive amounts of air, and swallowing dysfunction (dysphagia) [9]-[12], that may lead to aspiration, choking, and even death, and represents a major problem in rehabilitation of stroke and head injury patients.

In current clinical practice videofluoroscopic swallow study (VFSS) is the gold standard for diagnosis of swallowing disorders. However, VFSS is not portable, time-consuming, and results in some radiation exposure. Therefore, various non-invasive methods are proposed for swallowing assessment based on evaluation of swallowing sounds, recorded by microphones and/or accelerometers and analyzed by digital signal processing techniques [9]-[12]. Swallowing sounds are caused by a bolus passing through pharynx. It is possible to use swallowing sounds to determine pharyngeal phase of the swallow and characteristics of the bolus [9].

Several techniques are proposed for automated detection of swallowing and breath sounds. In [10] an algorithm based on multilayer feed forward neural network is proposed for decomposition of tracheal sounds into swallowing and respiratory sound segments. The algorithm is able to detect 91.7% of swallows correctly for healthy subjects. In [11] a wavelet transform based filter with iterative sequences of multiresolution decomposition and reconstruction is proposed to identify swallowing sounds from breath sounds in healthy and dysphagic subjects. This algorithm is able to

This work was supported in part by National Institutes of Health grant 5R21HL083052-02.

O. Makeyev, E. Sazonov, and S. Schuckers are with the Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY 13699 USA (e-mail: mckehev@cias.clarkson.edu, esazonov@cias.clarkson.edu, sschuckers@cias.clarkson.edu).

E. Melanson is with the Center for Human Nutrition, University of Colorado Health Sciences Center, Denver, CO 80262 USA (e-mail: ed.melanson@uchsc.edu).

M. Neuman is with the Department of Biomedical Engineering, Michigan Technological University, Houghton, MI 49931 USA (e-mail: mneuman@mtu.edu).

detect 93% of the swallowing sound segments correctly.

However, sound artifacts such as talking, throat clearing, and head movement that may be confused with swallowing and breath sounds decrease the efficiency of the recognition [12]. In [12] two sets of hybrid fuzzy logic committee neural networks (FCN) are proposed for recognition of dysphagic swallow from artifacts and normal swallow from artifacts correspondingly based on bandpass filtered acceleration signals obtained by an ultra miniature accelerometer attached to the skin in the midline of the throat at the level of thyroid cartilage. Evaluation results revealed that FCN correctly identified 31 out of 33 dysphagic swallows, 24 out of 24 normal swallows, and 44 out of 45 artifacts. An ability to recognize swallow signal and eliminate artifacts with high accuracy is very important for development of home/teletherapy biofeedback systems [13].

In this paper we demonstrate the recognition of swallowing sounds using STFT in combination with the LIRA neural classifier.

I. METHODOLOGY

A. Data collection

Commercially available miniature throat microphone (IASUS NT, IASUS Concepts Ltd.) located in the area caudal to the laryngeal prominence was used during the data collection process. Throat microphones convert vibration signals from the surface of the skin rather than pick up waves of sound pressure, thus reducing the ambient noise. Throat microphones also pick up such artifacts as head movements and talking that should not be confused with swallowing sounds.

Twenty sound instances were recorded for each of three classes of sounds (swallow, talking, head movement) for a healthy subject without any history of swallowing disorder, eating or nutrition problems, or lower respiratory tract infection. An approval for this study was obtained from Institutional Review Board and the subject was asked to sign an informed consent form. To record the swallowing sound the subject was asked to consume clear water in boluses of arbitrary size. The subject was asked to turn his head to the left side and back to record the sounds that correspond to head movements. To record talking the subject was asked to say the word “Hello”.

Sound signals for each class were amplified and recorded with a sampling rate of 44100 Hz.

B. Data preprocessing

Swallowing, head movement, and talking sounds were extracted from recordings in segments of 65536 samples (approximately 1.5 s) each using the following empiric algorithm: time limits of each sound instance under recognition were found using the signal-to-noise ratio thresholding under monitoring of the signal in the time and frequency domains, center of mass was calculated for each

sound instance and used to center the corresponding extraction segment.

The power spectrum of each segment was calculated with a window of 512 samples extracted using a Hanning window algorithm and processed by STFT with 50% window overlap. Due to limited signal bandwidth the higher frequencies do not contain significant energy of the original time domain signal and can be eliminated from the spectrogram. Truncating the spectrogram from 512x256 pixels to 256x256 pixels preserved most of the signal energy and eliminated insignificant harmonics. Examples of the spectrograms for swallowing sounds, talking, and head movements are presented in Fig. 1, columns *a*, *b*, and *c*.

A fourth class of outlier sounds was introduced to demonstrate the ability of the neural classifier to reject sounds with weak intra-class similarity and no similarity with other three classes. These were obtained with the same processing techniques as above and include random segments of music recordings. Examples of the spectrograms for the outlier class are presented in Fig. 1, column *d*.

These 80 grayscale spectrogram images, 20 for each of 4 classes, compose the image database that was used to test our system.

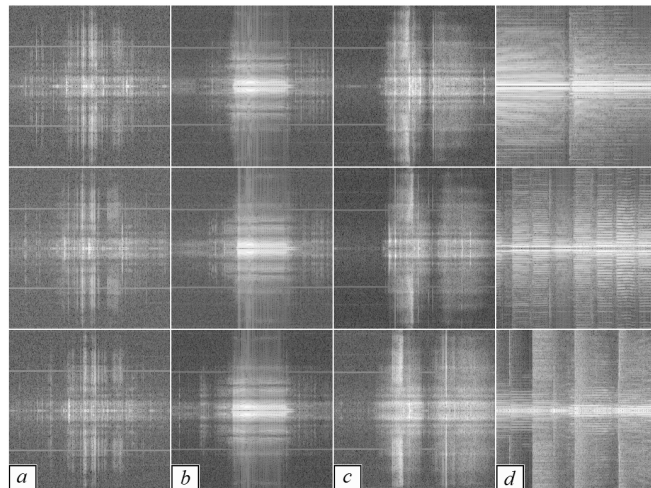


Fig. 1. Examples of spectrograms of (columns): *a*) swallowing sounds *b*) talking, *c*) head movements, *d*) outlier sounds.

C. LIRA neural classifier

LIRA neural classifier was developed on the basis of the Rosenblatt perceptron [14]. The three-layer Rosenblatt perceptron consists of the sensor *S*-layer, associative *A*-layer and the reaction *R*-layer. The first *S*-layer corresponds to the input image. The second *A*-layer corresponds to the feature extraction subsystem. The third *R*-layer represents the system’s output. Each neuron of this layer corresponds to one of the output classes.

The associative layer *A* is connected to the sensor layer *S* with randomly selected, non-trainable connections. The set of these connections can be considered as a feature extractor.

The *A*-layer consists of 2-state neurons; their outputs can be equal either to 1 (active state) or to 0 (non-active state). Each neuron of the *A*-layer is connected to all the neurons of the *R*-layer. The weights of these connections are modified during the perceptron training.

To adapt the LIRA neural classifier for grayscale image recognition we have added an additional 2-state neuron layer between the *S*-layer and the *A*-layer. We term it the *I*-layer (intermediate layer). The structure of the LIRA neural classifier is presented in Fig. 2.

Coding procedure

Each input image defines unique activations of the *A*-layer neurons. The binary vector that corresponds to the associative neuron activations is termed the image binary code $A = (a_1, \dots, a_N)$, where N is the number of the *A*-layer neurons. The procedure that transforms an input image into corresponding binary vector A is termed the image coding.

We connect each *A*-layer neuron to *S*-layer neurons randomly selected from a randomly generated window of height h and width w that is located in the *S*-layer (Fig. 2).

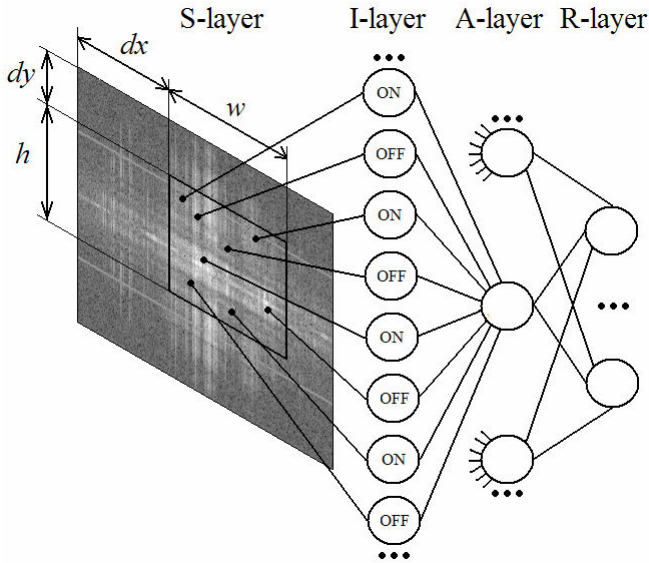


Fig. 2. Structure of the LIRA neural classifier.

The distances dx and dy are random numbers selected from the ranges: dx from $[0, W_S - w]$ and dy from $[0, H_S - h]$, where W_S and H_S stand for width and height of the *S*-layer. We create the associative neuron masks that represent the positions of connections of each *A*-layer neuron with neurons of the window $h \cdot w$. The procedure of random selection of connections is used to design the mask of *A*-layer neurons. This procedure starts with the selection of the upper left corner of the window $h \cdot w$ in which all connections of the associative neuron are located.

The following formulas are used:

$$dx_i = \text{random}_i(W_S - w),$$

$$dy_i = \text{random}_i(H_S - h),$$

where i is the position of a neuron in associative layer *A*, $\text{random}_i(z)$ is a random number that is uniformly distributed in the range $[0, z]$. After that position of each connection within the window $h \cdot w$ is defined by the pair of numbers:

$$x_{ij} = \text{random}_{ij}(w),$$

$$y_{ij} = \text{random}_{ij}(h),$$

where j is the number of the connection with the *S*-layer.

Absolute coordinates of a connection to the *S*-layer are defined as:

$$X_{ij} = x_{ij} + dx_i,$$

$$Y_{ij} = y_{ij} + dy_i.$$

The input of each *I*-layer neuron is connected to one neuron of the *S*-layer and the output is connected to the input of one neuron of the *A*-layer. All the *I*-layer neurons connected to one *A*-layer neuron form the group of this *A*-layer neuron. There are two types of *I*-layer neurons: ON-neurons and OFF-neurons. The outputs of ON- and OFF-neurons are computed according to the formula:

$$ON_i = \begin{cases} 1, & b_i > \theta_i \\ 0, & b_i \leq \theta_i \end{cases},$$

$$OFF_j = \begin{cases} 1, & b_j < \theta_j \\ 0, & b_j \geq \theta_j \end{cases},$$

where ON_i and OFF_j are the outputs of ON-neuron i and OFF-neuron j correspondingly, θ_i and θ_j are the thresholds of ON-neuron i and OFF-neuron j correspondingly, and b_i and b_j are the values of brightness of the image pixels that correspond to ON-neuron i and OFF-neuron j correspondingly. Thresholds θ_i and θ_j are selected randomly from the range $[0, b_{max}]$, where b_{max} is maximal brightness of the image pixels. For example, in Fig. 2 the group of eight *I*-layer neurons, four ON-neurons and four OFF-neurons, corresponds to one *A*-layer neuron. The i -th neuron of the *A*-layer is active ($a_i = 1$) only if outputs of all the neurons of its *I*-layer group are equal to 1 and is non-active ($a_i = 0$) in the opposite case.

After execution of the coding procedure every image has an associated image binary code that is to be used during the training and recognition procedures.

Training procedure

Before starting the training procedure the weights of all the connections between neurons of the *A*-layer and the *R*-layer are set equal to 0. Then training is performed using the

following algorithm:

Step 1. Calculation of excitation.

An image binary code A is presented to the LIRA neural classifier. R -layer neuron excitations E_i are computed according to the formula:

$$E_i = \sum_{j=1}^N a_j \cdot w_{ji},$$

where E_i is the excitation of the i -th neuron of the R -layer, a_j is the output signal (0 or 1) of the j -th neuron of the A -layer, w_{ji} is the weight of the connection between the j -th neuron of the A -layer and the i -th neuron of the R -layer.

Step 2. Excitation adjustment.

Excitation adjustment is performed after calculation of the neuron excitations of the R -layer. The excitation E_c of the R -layer neuron that corresponds to the correct class c is recalculated according to the formula:

$$E_c^* = E_c \cdot (1 - T_E),$$

where $0 \leq T_E \leq 1$ determines the reserve of excitation the neuron that corresponds to the correct class must have. In our experiments the value T_E varied from 0.1 to 0.5.

Next, the neuron with the largest excitation is selected. This winner neuron represents the recognized class j .

Step 3. Adjustment of weights.

If the winning neuron corresponds to the correct class c ($j = c$) then no modification of weights is needed. If $j \neq c$ then following modification of weights is performed:

$$\begin{aligned} w_{ic}(t+1) &= w_{ic}(t) + a_i, \\ w_{ij}(t+1) &= w_{ij}(t) - a_i, \end{aligned}$$

where $w_{ij}(t)$ and $w_{ij}(t+1)$ are the weights of the connection between the i -th neuron of the A -layer and the j -th neuron of the R -layer before and after modification, a_i is the output signal (0 or 1) of the i -th neuron of the A -layer.

The training process is carried out iteratively. In each training cycle all the images of the training set are presented to the neural classifier.

Performance of a LIRA neural classifier can be improved with implementation of distortions of input images during training and recognition [1]. The decision whether to use distortions and selection of a particular type of distortions depends on specificity of the image database and available computational resources. Since the images in our database are not centered perfectly, in our experiments we used different combinations of horizontal, vertical and bias image translations.

Recognition procedure

Image distortions have been used both for training and recognition. There is an essential difference between implementation of distortions for training and recognition. In the training process each distortion of the initial image is considered as an independent new image. In the recognition process it is necessary to introduce a decision-making rule in order to be able to make a decision about a class of the image under recognition based on both the image itself and all of its distortions. Decision-making rule that we have used consists in calculation of the R -layer neuron excitations for all the distortions sequentially:

$$E_i = \sum_{k=0}^d \sum_{j=1}^N a_{kj} \cdot w_{ji},$$

where E_i is the excitation of the i -th neuron of the R -layer, a_{kj} is the output signal (0 or 1) of the j -th neuron of the A -layer for the k -th distortion of the initial image, w_{ji} is the weight of the connection between the j -th neuron of the A -layer and the i -th neuron of the R -layer, d is the number of applied distortions (case $k = 0$ corresponds to the initial image).

The neuron with the largest excitation (winner neuron) represents the recognized class.

II. RESULTS

Testing was performed on the database of grayscale spectrogram images described in section II. The database contained 80 spectrograms, 20 for each of four classes that correspond to instances of swallowing sounds, talking, head movements and outlier sounds. The spectrograms in the database were randomly divided into the training and validation sets. The number of images in training set varied from two to ten. In our experiments we used holdout cross-validation, i.e. the validation set for each class was chosen randomly from the database and the rest of the database was used for training. In each experiment we performed 50 runs of the holdout cross-validation to obtain statistically reliable results. A new mask of connections between the S -layer and the A -layer and a new division into the training and validation sets were created for each run. The following statistics were calculated for the obtained sample of 50 error numbers: mean, standard deviation and 95% confidence interval for the mean [15]. Mean recognition rate was calculated based on the mean number of errors for one run and the total number of images in the validation set.

Table 1 presents the recognition rates obtained for various numbers of images in training and validation sets for the following set of parameters: window $h \cdot w$ width $w = 10$, height $h = 10$; reserve of excitation $T_E = 0.3$; the number of training cycles is 30; the number of ON-neurons in the I -layer neuron group that corresponds to one A -layer neuron is 3, the number of OFF-neurons is 5; 8 distortions for training including ± 1 pixel horizontal, vertical and bias image

translations and 4 distortions for recognition including ± 1 pixel horizontal and vertical image translations; the total number of associative neurons $N = 512,000$. It can be seen that even in case of using only 2 images for training and 18 for testing the LIRA neural classifier achieves a mean recognition rate of 98.75%.

Table 1. Dependency of the recognition rate on the number of images in training set

T/V *	Mean number of errors for one run / Number of images in the validation set for one run	Standard deviation	Confidence interval (95%)	Mean recognition rate (%)
2/18	0.9 / 72	1.07	[0.59, 1.21]	98.75
4/16	0.24 / 64	0.48	[0.1, 0.38]	99.63
6/14	0.16 / 56	0.51	[0.01, 0.31]	99.71
8/12	0.04 / 48	0.2	[0, 0.1]	99.92
10/10	0 / 40	-	-	100

* T is the size of the training set for each class, V is the size of the validation set for each class.

III. DISCUSSION

A distinctive feature of the LIRA neural classifier is the random assignment of connections between the S -layer and the I -layer that assures a good description of an image without prior knowledge about the image content. Large total number of random features (associative neurons) ensures that there will be a sufficient number of significant features. Features that do not provide useful information for separation of classes will not obtain significant weights during training.

In our experiments we set the parameter values to maximize efficiency of the LIRA neural classifier. The amount of time needed for one run of classifier coding, training and recognition with the set of parameters presented in Section 4 is approximately 2 min (90 sec for coding, 30 sec for training and 1 sec for recognition) on a computer equipped with AMD Athlon 64 X2 4400+ Dual Core processor and 2.00 GB of RAM. Such computational time is justified by the increase in the recognition rate and allows practical utilization of the classifier with the highest recognition rate. In particular, achieved recognition speed allows real-time recognition of swallowing instances in sound streams.

Results obtained in the testing suggest high efficiency and reliability of the proposed method, though tests on a larger database would be needed for a conclusive proof. An important advantage of the proposed method is utilization of a double-redundant approach to identification of significant features. First, STFT provides a redundant description of a sound instance, therefore increasing chances for random selection of a significant feature. Second, massively redundant links between the sensor and associate layers ensure multiplicity of extracted features. Such an approach

presents a novel deviation from relatively small sets of empirically-selected statistics (such as number of zero crossings, average power, average frequency, etc) that are traditionally used as features in sound recognition.

IV. CONCLUSION

In this paper we propose a sound recognition technique based on the limited receptive area (LIRA) neural classifier and short-time Fourier transform (STFT). The proposed technique works by applying a LIRA-based image recognition system to the spectrograms of sound instances.

The suggested methodology is tested in recognition of four classes of sounds that correspond to swallowing sounds, talking, head movements and outlier sounds. The database of 80 sound instances is used during the testing. Experimental results suggest high efficiency and reliability of the proposed method with the recognition rate of 100% obtained for database divided in half into training and validation sets.

The proposed method may be employed in systems for automated swallowing assessment and diagnosis of swallowing disorders and has potential for application to other sound recognition tasks.

ACKNOWLEDGMENT

The authors gratefully acknowledge E. Kussul and T. Baidyk, National Autonomous University of Mexico (UNAM), for the constructive discussions and helpful comments.

REFERENCES

- [1] E. Kussul, T. Baidyk, D. Wunsch, O. Makeyev, A. Martin, "Permutation coding technique for image recognition systems," *IEEE Trans Neural Networks*, vol. 17/6, pp. 1566-1579, 2006.
- [2] E. Kussul, T. Baidyk, D. Wunsch, O. Makeyev, A. Martin, "Image recognition systems based on random local descriptors," in *Proc. International Joint Conference on Neural Networks IJCNN'2006*, Vancouver, Canada, 2006, pp. 4722-4727.
- [3] E. Kussul, T. Baidyk, "Improved method of handwritten digit recognition tested on MNIST database," *Image and Vision Computing*, vol. 22, pp. 971-981, 2004.
- [4] T. Baidyk, E. Kussul, O. Makeyev, A. Caballero, L. Ruiz, G. Carrera, G. Velasco, "Flat image recognition in the process of microdevice assembly," *Pattern Recogn Lett.*, vol. 25/1, pp. 107-118, 2004.
- [5] O. Makeyev, T. Baidyk, A. Martin, "Limited receptive area neural classifier for texture recognition of metal surfaces," in *Proc. IFIP WCC AI2006*, Santiago de Chile, Chile, 2006, pp. 10.
- [6] E. Kussul, T. Baidyk, M. Kussul, "Neural network system for face recognition," in *Proc. IEEE International Symposium on Circuits and Systems, ISCAS*, Vancouver, Canada, 2004, vol. V, pp. V-768-V-771.
- [7] Y. LeCun, MNIST OCR data. Available: <http://yann.lecun.com/exdb/mnist/index.html>
- [8] A. K. Limdi, M. J. McCutcheon, E. Taub, W. E. Whitehead, E. W. Cook, "Design of a microcontroller-based device for deglutition detection and biofeedback," in *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Seattle, USA, 1989, vol. 5, pp. 1393-1394.
- [9] T. Nakamura, Y. Yamamoto, H. Tsugawa, "Measurement system for swallowing based on impedance pharyngography and swallowing sound," in *Proc. 17th IEEE Instrumentation and Measurement Technology Conference*, Baltimore, Maryland, USA, 2000, vol. 1, pp. 191-194.

- [10] M. Aboofazeli, Z. Moussavi, "Automated classification of swallowing and breath sounds," in *Proc. 26th Annual International Conference of the Engineering in Medicine and Biology Society*, San Francisco, California, USA, 2004, vol. 2, pp. 3816-3819.
- [11] M. Aboofazeli, Z. Moussavi, "Automated Extraction of Swallowing Sounds Using a Wavelet-Based Filter," in *Proc. 28th Annual International Conference of the Engineering in Medicine and Biology Society*, New York, New York, USA, 2006, pp. 5607-5610.
- [12] A. Das, N. P. Reddy, J. Narayanan, "Hybrid fuzzy-neural committee networks for recognition of swallow acceleration signals", *Computer Methods and Programs in Biomedicine*, vol. 64, pp. 87-99, 2000.
- [13] N. P. Reddy, V. Gupta, A. Das, R. N. Unnikrishnan, G. Song, D. L. Simcox, H. P. Reddy, S. K. Sukthankar, E. P. Canilang, "Computerized biofeedback system for treating dysphagic patients for traditional and teletherapy applications," in *Proc. International Conference on Information Technology Application in Biomedicine ITAB'98*, Piscatway, New Jersey, USA, 1998, 100-104.
- [14] F. Rosenblatt, "Principles of neurodynamics". Spartan books, New York, 1962.
- [15] D. J. Sheskin, "Handbook of parametric and nonparametric statistical procedures". Chapman & Hall / CRC Press, Boca Raton, 2003.