Journal of Advanced Sciences and Engineering Technologies (2022) 5 (1):23-32

DOI: http://doi.org/10. 10.32441/jaset.05.01.03

Journal of Advanced Sciences and Engineering Technologies

https://isnra.net/ojs/index.php/jaset/index/

Artificial neural networks for voice activity detection Technology

***Takialddin Al Smadi*[1*], *Ahmed Handam*[2] ,*Mahmoud Ababneh*[1]**

1 College of Engineering, Jerash University, Jerash, Jordan.
2 Renewable Energy engineering Department, Faculty of Engineering, Amman Arab University Jordan.

Keywords:

voice biometrics, Automatic detection, activity detection, speech recognition

A R T I C L E I N F O

Article history:

Received	03 June. 2021
Accepted	28 Aug. 2021
Available online	17 Jan. 2022

Citation: Al Smade, T., Handam, A. , Ababneh M (2022). Artificial neural networks for voice activity detection Technology. ournal of Advanced Sciences and Engineering Technologies, 5(1), 1-15.**http://doi.org/10. 10.32441/jaset.05.01.03**

A B S T R A C T

Currently, the direction of voice biometrics is actively developing, which includes two related tasks of recognizing the speaker by voice: the verification task, which consists in determining the speaker's personality, and the identification task, which is responsible for checking the belonging of the phonogram to a particular speaker. An open question remains related to improving the quality of the verification identification algorithms in real conditions and reducing the probability of error. In this work study Voice activity detection algorithm is proposed, which is a modification of the algorithm based on pitch statistics; VAD is investigated as a component of a speaker recognition system by voice, and therefore the main purpose of its work is to improve the quality of the system as a whole. On the example of the proposed modification of the VAD algorithm and the energy-based VAD algorithm, the analysis of the influence of the choice on the quality of the speaker recognition system is carried out.

[*] Corresponding Author: E-mail: Takialddina@gmail.com, https://orcid.org/0000-0002-1322-9707, College of Engineering, Jerash University, Jerash, Jordan.

Introduction

Artificial neural networks are the simplest mathematical models of the brain. To understand the basic principles of building, you can consider them as a set network of individual structures neurons. Very roughly the structure of a biological neuron can be described as follows. The neuron has soma - a body, a tree of entrances - dendrites, an exit - an axon. [1].On the soma and on the dendrites, the endings of the axons of other neurons, called si-naps, are located. The synapses received by the signals Tend to either excite the neuron or slow down. When the total excitation reaches a certain threshold.

The neuron is excited and sends a signal to the other neurons along the axon. Each synapse has A unique synaptic force that, in proportion to its value, changes the input signal to the neuron. In accordance with the above Description, the mathematical model of the neuron is a summing threshold element, The direct distribution of signal is layer, starting with the input layer, use the calculated amount of the input signals for each neuron and the Function is generated by the Activation Response of the neuron, which is distributed in the next layers, into account the weight of the neural connection on the fig (1). As a result of this step we get a vector of output values of the neural network.

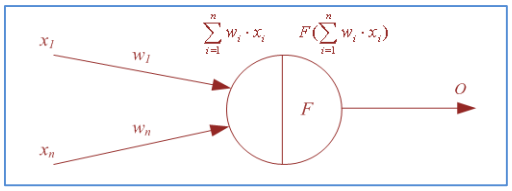


Fig 1 an artificial neuron the formula for triggering of the neuron:

$$O = F(< W^T, X >) =$$

$$F(\sum_{i=1}^{n} w_i \cdot x_i), = O = \begin{cases} 1, (< W^T, X >) \geq 0, \\ 0, ... \end{cases}$$

The neural networks (**NA)** learning occurs on some training sample, for each sample of which all current outputs are determined and compared with the desired values. If the difference is unacceptable, the weights change. The end of training is the situation when a common error on all samples is permissible. All algorithms for learning neural networks are a variety of learning algorithms based on the method of error correction, which is carried out in different ways. The idea of changing the (NA) weights is to find a general measure of the quality of the network, which is usually chosen as the network error function [2, 3] in order to find the right weights, The most common method of finding the minimum is the method of gradient descent. For the case of a function with one variable, weights change in the direction opposite to the derivative The Reverse Algorithm Error Distribution involves the calculation of the error, as the output layer and each neuron network, as well as correction weights of neurons in accordance with their current values. In the first step of the algorithm of the weight of all the ties are initialized with small random values (0 to 1). After the initialization of weights in the learning process of the neural network to perform the following steps:

- Direct distribution of signal;
- Error calculation of the neurons of the last layer;
- Inverse distribution of errors.

Recognition based on neural networks inter-disciplinary

Speech recognition is an interdisciplinary subfield of computational linguistics that develops methodologies and technologies that allow you to recognize and translate spoken language into text by computers. It is also known as automatic speech recognition and computer speech recognition" or simply "speech in the text. It includes knowledge and research in the field of linguistics, informatics and electrical engineering.

The input signal is divided into 20 frames, each of which contains 512 samples. For each Frame a gives 255 Spectral objects, on the input neurons in the neural network. On the basis of the input data and the output requirements in direct neural network.

Discrete wavelet transform (DWT) algorithm in speech recognition Definition of words

The word determination can be performed by comparing numeric forms signals or by comparing the spectrogram of the signals. The comparison process in both cases should compensate for the different lengths of the sequence and the non-linear nature of the sound.

The DWT algorithm manages to resolve these problems by finding the deformation corresponding to the optimal distance between two rows of different lengths there are 2 features of the algorithm:

1. Direct comparison of numerical waveforms. In this case, for each numerical sequence a new sequence is created, the dimensions of which are much smaller. A numerical sequence can have several thousand numeric values [4].

While a subsequence can have several hundred values, reducing the number of numerical values can be accomplished by removing them between corner points. This process of reducing the length of a numerical sequence must not change its representation. Undoubtedly, the process leads to a decrease accuracy of recognition. However, taking into account the increase in speed, accuracy, in fact, is increased by increasing the words in the dictionary.

Problem of speaker verification by voice

A system of automatic voice verification based on Gaussian mixture models (GMMs) and support vector machine (SVM) classification, described in detail in [8], is considered. Mel-frequency spectral coefficients (MFCC) were chosen as informative acoustic features. In this problem, it is necessary to carefully control the nature of the sound data supplied to the input of the training and test data, not allowing processing of non-speech fragments of the signal,

Since the system under consideration is very sensitive to this kind of mistakes. To assess the quality of the system, the standard for tasks is used. Pattern recognition criterion - resulting value equal probable pass/reject error of the entire verification system (equal error rate, EER).

ERR =FA = FR,

Where FR (false reject) is the probability of a false rejection (the probability of an error of the first kind), FA (false acceptance) - the probability of false identification (Probability of error of the second kind). Acoustic features are fed to the input of the verification algorithm, Phonograms selected on speech segments. Segmentation into sections speech / non-speech is the result of the algorithm, a typical example using VAD in the verification system.

***2. Representation of spectrogram signals and application of the DTW algorithm for comparison of two spectrograms*.**

The method consists in dividing the digital signal into some number of intervals that will overlap. For each pulse, Intervals of real numbers (sound frequencies), will calculate the Fast Fourier transform, and will be stored in the matrix of the sound spectrogram. Options will be the same for all computational operations: pulse lengths, Fourier transform lengths, overlap lengths for two consecutive pulses. The Fourier transform is symmetrically connected with the center, and the complex numbers on one side are related to the numbers on the other hand. In this regard, only the values from the first part of the symmetry can be saved, so the spectrogram will represent the matrix of complex numbers, the number of lines in such a matrix is equal to half the length of the Fourier transform, and the number of columns will be determined depending on the length of the sound. (DTW) will be applied to the matrix of real numbers as a result of conjugation of the spectrogram of values; such a matrix is called the energy matrix [5].

Developed a neural network has demonstrated the expected behavior associated with the study and generalization error. It was found that even if the error synthesis, decreases with increasing training sequence, errors starts oscillating regardless of the introduction of dynamic learning speed. In the network were prepared enough to meet the requirements for the generalization error, but, nevertheless, there is still a possibility to improve aggregate error show the fig 2

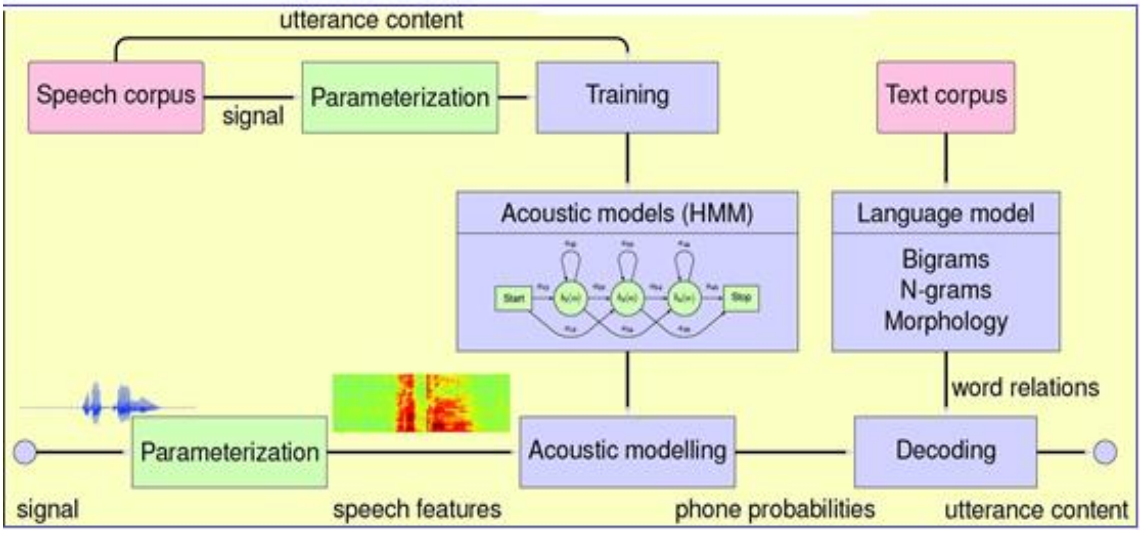


Figure 2.Block Scheme of speech Recognition

.

Signs of nonlinear dynamics

The maximum characteristic index of the emotional state of a person, to which corresponds certain geometry of the attractor phase portrait; person's emotional state, from "calmness" to "anger" deformation and subsequent shift of the speech signal spectrum. [6].

For a group of signs of non-linear dynamics, the speech signal is considered as a scalar quantity observed in the human vocal tract system. The process of speech formation can be considered nonlinear and analyzed by the methods of nonlinear dynamics. The problem of nonlinear dynamics consists in finding and studying in detail the basic mathematical models and real systems that come from the most typical suggestions about the properties of the individual elements making up the system and the laws of interaction between them. Currently, the methods of nonlinear dynamics are based on the fundamental mathematical theory, which is based on the Taken theorem, which provides a rigorous mathematical basis for the ideas of nonlinear auto regression and proves the possibility of reconstructing the phase portrait of an attractor from a time series or from one of its coordinates. An attractor is defined as a set of points or a subspace in the phase space to which the phase trajectory approaches after decay of transient processes. Estimates of signal characteristics from reconstructed speech trajectories are used in constructing nonlinear deterministic phase-space models of the observed time series. The revealed differences in the form of attractors can be used for diagnostic rules and signs allowing recognizing and correctly identifying various emotions in an emotionally colored speech signal [7].

Related Works

Extraction of vocalized areas of speech is based on the fact that the methods used by experts in the field of voice biometrics use vowels and nasalized consonants. The negative side is the loss of some consonants. On the other hand, explosive consonants and affricates have less identification significance, so it can be assumed that the loss of some part of insignificant speech material will be compensated by the qualitative removal of non-speech areas.

Vote assignment is used in case voting by experts in the field of voice biometrics, vowels, and nasal reveal cases. The negative reason is the loss of some cases. On the other hand, the detection of blastomeric and joints, which leads to a decrease in arterial hypertension, in connection with an injury that may occur, the loss of a significant part of the river substance not important to the quality of removal will be compensated by non-speech areas show in fig 3

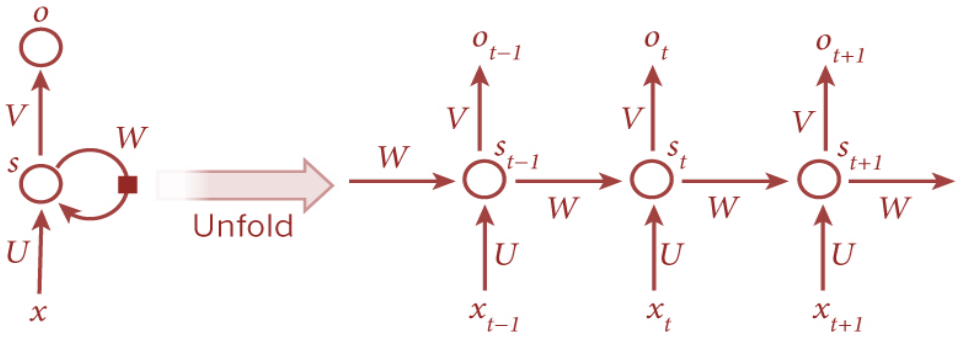


Figure 3 the process of the structure neural network

The network at time t accepts an input vector x_t, the latent condition in the previous step st 1 and calculates the output vector. After this the new state set is transferred to the next iteration of the process. Such networks allow processing a sequence of unknown length, given the connection between the present and the past. The main method of teaching such networks is the method return an error distribution in time. [8]It works as follows, all of the learning algorithms of neural networks are the varieties learning algorithm the method of error correction, which is carried out in different ways. The idea of changing weights to find common measures the quality of the network, as usually choose the function of network error. Then, in order to find the right weight loss, it is necessary to minimize the error function. The most common method of finding a minimum is the method of gradient descent. In the case of function with one variable, the weight of the change in the opposite direction of the derivative the fair formula.

$$\partial_e F(W) = \frac{\lim_{t\to 0}(F(W+et)-F(W))}{t} \Rightarrow$$

e = (0, 0...1...0) Define the private differential.

$$\partial_i F(W) = \frac{\lim_{t\to 0}(F(w_1, w_2,\ldots, w_{i+et},\ldots, w_n)-F(W))}{t} \Rightarrow$$

$$\partial F(W) = ((-\partial F(w_1), -\partial F(w_2), \ldots, -\partial F(w_i), \ldots, -\partial F(w_n))^T$$

To determine the generalized functions let's look at the tutorial sample $\{(x^k, K^k)\}$, k = 1, ..., K. The accumulated in all epoch's error

$$E = \sum_{k=1}^{K}(E^k) = \sum_{k=1}^{K}\left(\sum_{i=1}^{m} 1/2\|O_i - Y_i\|^2\right) \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

He formula for modification of the weights of the NA

$$W^{n+1} = W^n - h\cdot \partial E/\partial W$$

$$O_i = < W_i, X_i > \Rightarrow$$

$$\partial E/\partial W = -(Y_i - O_i)\cdot X, \Rightarrow$$

$$W^{n+1} = W^n - h\cdot (Y_i - O_i)\cdot X. \Rightarrow$$

$$W^{n+1} = W^n - h\cdot \delta \cdot X, \ldots\ldots\ldots\ldots\ldots (3)$$

Properties of functions (t,Ω):

$$p(t,\Omega) = 0. \,\&\, p(t,\Omega) < \varepsilon_0$$

P (t, Ω) ∈ (0.1)..... (4)

The exam similarity function chow in Figure 4.

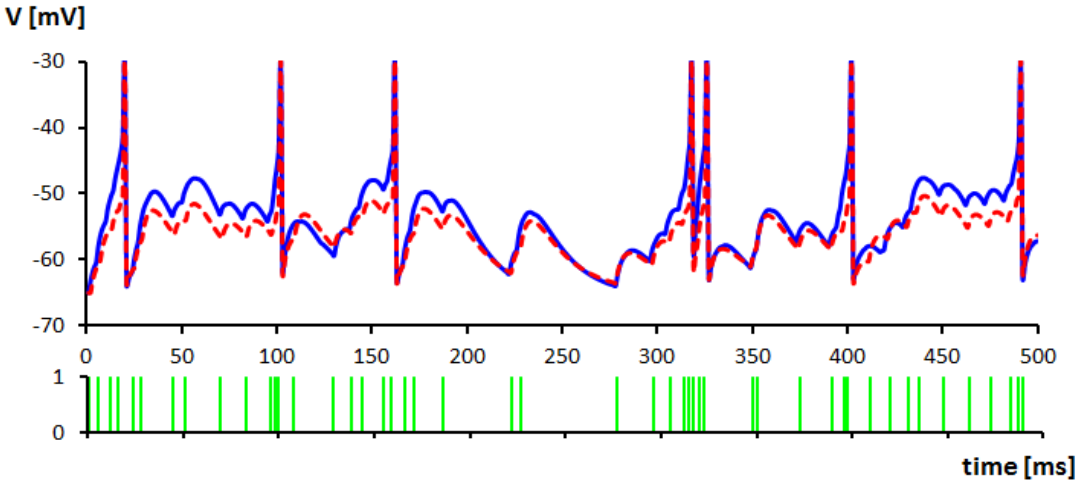


Figure 4 an the similarity function

$$\sup\{P(t_{i+1}) - P(t_i)\} << \sup\{P(t_i)\} \ldots\ldots\ldots(4)$$

Since the function $P(t_i) \in [0,1]$, so

$\sup\{P(t_i)\} = 1$ and 3.1can be rewritten as

$$\sup\{P(t_{i+1}) - P(t_i)\} << 1 \ldots\ldots\ldots\ldots\ldots(5)\cdot$$

Extend the standard configuration of a neural network was looking for a vector with length values at once

$$M - y(n_0), y(n_{0+1}), \Leftrightarrow y(n_0 + M)$$

$\Rightarrow n_0\Delta t$, $(n_0+1)\Delta t$, $(n_0+M)\Delta t$. The number of input vectors, respectively to increase the

$$\Rightarrow M - x(n_o - n), y(n_o - n + 1), x(n_o)x(n_o + n_{-1}), x(n_o + n), (x(n_o + n + m - 1), x(n_o + n + M$$

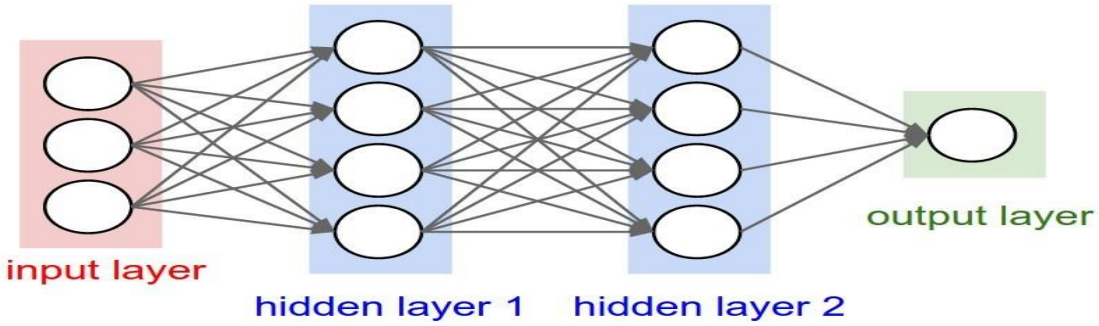


Figure 5 neural networks for recognition of phonemic tasks.

These layers are referred to as linear, as is the multiplication of the input vector by a matrix of weights w(k) for k-th layer. In practice, such a transformation is usually add to the confusion of bk , i.e. the output vector of k-th layer of u(k) is calculated through the previous layer

u(k) = φ(k)(A(k) • u(k−1) + bk), A(k) ∈ Rd1 ×d2 , bk ∈ Rd1

[10, 11 and 12].

$P(t_{2i+1}) - P(t_{2i}) << 1$, i=1 to 3,...,

- $P(t_{2i+1}), P(t_{2i})$on $o(t_{2i+1}), o(t_{2i})$,
  $o(t_{2i+1}) = \frac{P(t_{2i+1}) + P(t_{2i})}{2}$,
- and

$$o(t_{2i}) = \frac{P(t_{2i+1}) - P(t_{2i})}{2} \ldots\ldots(6)$$

- Convert easily to formulas $P(t_{2i+1}) = o(t_{2i+1}) + o(t_{2i})$
- $P(t_{2i}) = o(t_{2i+1}) - o(t_{2i})$.....(7)
- $o(t_{2i})$. $\sup o(t_{2i}) = \frac{1}{k}$
- $\frac{1}{k} << 1$. The redundant interval $\left[\frac{1}{k};1\right)$

$$k_i = \frac{2}{\sup\{P(t_{i+1}) - P(t_i)\}} \text{.....(8)}$$

The unclaimed goes second with large-scale activities. First, in the case of the unclaimed input values, the neural network is more to achieve the same accuracy, scaled to want to take a few iterations. $\sup_i \left\{\frac{2}{k_i}\right\}$.

Find the errors of the neural network results in this iteration, in this case, the errors of the first neural network.

$$E_1 = 2e^2 \text{.....(9)}$$

$$E_2 = e^2 + \left(\frac{e}{k_i}\right)^2 = \left(1 + \frac{1}{k_i}\right)e^2 \text{....(10)}$$

Algorithm based on energy signal estimation

An algorithm based on the estimation of the energy level, calculation of the ratio of the average signal power per speech signal and subsequent setting of the power threshold, according to take over the decision about the presence of speech in a special section of the signal . To remove possible interference that is not in the speech range frequencies (for example, mains interference at a frequency of 50-60 Hz), is used filter with an infinite impulse response that suppresses low signal frequency. Such a filter, calculated by the method, provides minimal signal distortion in amplitude and in group delay in the pass region, clearing the speech signal from network interference.

The model interprets a short sequence of words

Fixed the size of the Classic the recurrent neural networks take on the entrance of arbitrary length, but at the outlet is still emit the vector of fixed size. One possible solution is the architecture of the encoder-decoder the main idea of this method is the first of the vector of fixed size, describing the input sequence and then deploying already in the output sequence [13]. In more detail, the process usually occurs as follows:

- Login is passed through the neural network and cell state (instead of in the hidden state) after passing the whole sequence is considered to be a vector describing the entrance.
- This phase encodes the data in a vector; therefore, the neural network is called encoder.
- At the second stage (decoding) is the task of the vector understands the output sequence. For this the cell states the initial state of also technology in the following network, which consistently generates symbols from the output Alphabet until it generates a special terminal symbol. Often, this architecture is used in the tasks of machine translation, so here was a concept of the output alphabet, as in the classical generated at time t is input to the network in the next moment of time t 1. But in the subject in the work of the task, it is not necessary, the output alphabet (encoded many words) has a small volume, and the symbols of the received letters, which in turn is a lot of words, do not form between any relationships. Therefore, used described the architecture without Reference generated by the character further in the process of decoder. The procedure structure can be seen in Figure6

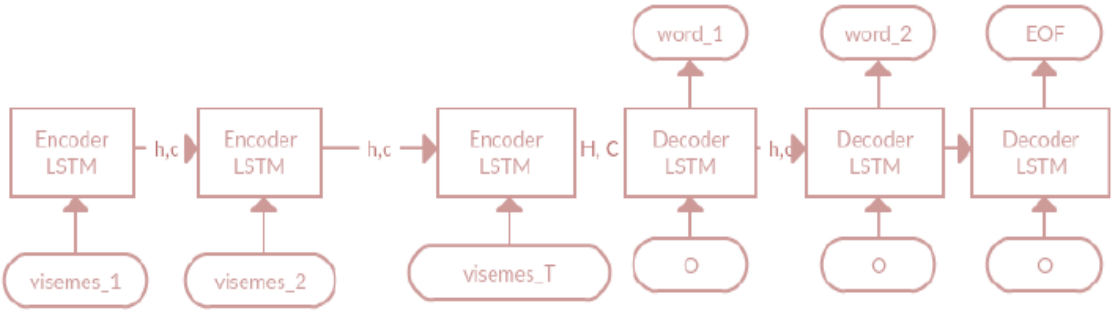


Figure 6 Model of the encoder-decoder LSTM

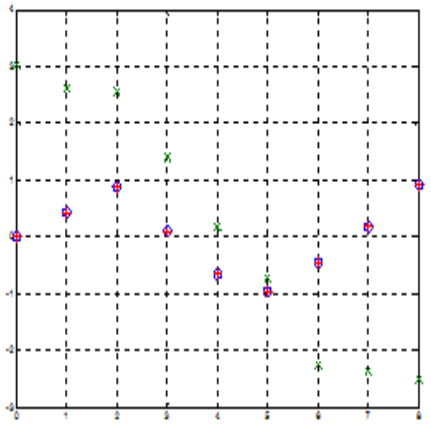
As you exit the expected sequence of marks of words, spoken on this video, too short words were as empty word, actually, this means that all the short words were in one large class.

THE RESULTS

The example uses a two-layer perceptron with two nonlinear neurons in the first layer and one in the second layer. The operation of the algorithm for back propagation of the error is broken down into the following steps: the assignment of the input and the desired output, the direct passage of the input signal to the output, the inverse propagation of the error, and the change in the weights. Variables permitting to trace the operation of the algorithm for reverse propagation of error

Table 1 The results of the phased implementation of the back propagation algorithm

Stage	Direct distribution of the input signal	Direct distribution of the input signal	Changing weights
$A_1(1), A_1(2)$	*Log sig* (W_1P+B_1) = = [0,321, 0,368]	Not running	Not running
A_2	*Pure line* (W_1P+B_1) = = 0,446	»	»
e	$t - A_2 = 1{,}261$	»	»
$N_1(1), N_1(2)$	Not running	$\frac{\partial \log sim(N_1)}{\partial N_1 \cdot W_2 \cdot N_2} =$ $= [0{,}049,\ 0{,}100]$	»
N_2	Not running	$-2 \cdot \frac{\partial purelin(N_2)}{\partial N_2 \cdot e} = -2{,}522$	»
$W_1(1)$ $W_1(2)$	»	Not running	$W_1 = W_1 - \text{lr} \cdot N_1 \times P =$ $= [-0{,}265, -0{,}420]$
$B_1(1), B_1(2)$	»	»	$B_1 = B_1 - \text{lr} \cdot N_1 =$ $= [-0{,}475, -0{,}140]$
B_2	»	»	$B_2 = B_2 - \text{lr} \cdot N_2 = 0{,}732$
$W_2(2)$	»	»	$W_2 = W_2 - \text{lr} \cdot N_2 \times N_1 =$ $= [0{,}171,\ 0{,}077]$



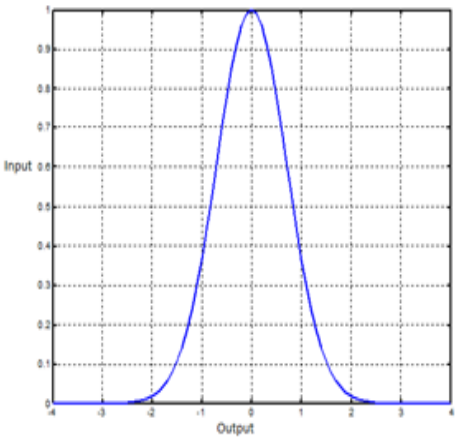


Fig 7 Result of approximation of vectors by a two-layer perceptron and Fig 8. Radial basis function

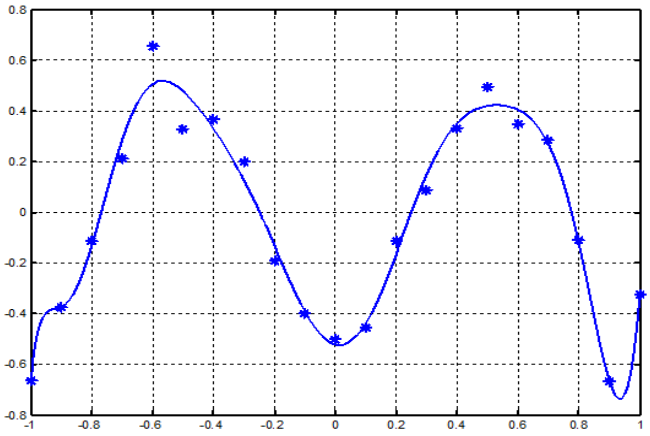


Fig 9 approximation by means of a radial basis neural network test

Conclusion

The article describes a mathematical model of a middle ear person with the help of psychoacoustic perception approach heights and obtained from him the image classification. Are the results of the experiments speech recognition based on neural networks, The advantages of this method, you can make it sufficient simplicity of implementation, as well as the very obvious analogy with the processes taking place in the real organ of hearing rights. Disadvantage is the level of

Errors in the recognition (13-23%) which is offered to reduce the use of contextual recognition offer Recognition of individual voice commands. automated Key words from the stream of speech, which are associated with the processing of telephone calls or sphere of security. The ability to learn and summarize the accumulated knowledge, the neural network has the features of artificial intelligence. A network trained on a limited set of data is able to generalize the information obtained and to show good results on data not used in the learning process.

A characteristic feature of the network is also the possibility of its implementation with the use of technology of a very large degree of integration. The difference in network elements is small, and their repeatability is enormous. This opens the prospect of creating a universal processor with a homogeneous structure, capable of processing a variety of information. Decisively different tasks. For example, in the task of determining the boundaries of speech sections, the proposed algorithm will not have such an advantage as in the task of recognizing a speaker by voice. The study of the behavior of the developed algorithm in various conditions is a further direction of research.

Reference

[1]. Ghaidan, K. A., Al Smadi, T. A., Aljumailly, T. A., & Al-Taweel, F. M. (2011). Development of a new approach for transmitting Digital message on a Frequency Limited Communication Channel transmission. Journal of Advanced Computer Science and Technology Research, 1, 52-62.

[2]. Smadi, T. A., Al Issa, H. A., Trad, E., & Smadi, K. A. A. (2015). Artificial Intelligence for Speech Recognition Based on Neural Networks. Journal of Signal and Information Processing, 06(02), 66–72. http://dx.doi:10.4236/jsip.2015.62006

[3]. Graves, Alex,Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks."Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013.

[4]. Abdel-Hamid, Ossama, et al. "Convolutional neural networks for speech recognition."IEEE/ACM Transactions on audio, speech, and language processing22.10 (2014): 1533-1545.

[6]. Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., and Penn, G. (2012).Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition.In ICASSP, pages 4277–4280. IEEE.

[7]. Dede, G., Sazlı, M.H.:Speech recognition with artificial neural networks. Digital Signal Processing 20, 763–768 (2010)

[8].Dahl, G., Yu, D., Li, D., and Acero, A. (2011). Large vocabulary continuous speech recognition with context-dependent dbn-hmms.In ICASSP.

[9]. Takialddin, A.S., Al Smadi, K. & AL-Smadi, O.O., 2017. High-Speed for Data Transmission in GSM Networks Based on Cognitive Radio. American Journal of Engineering and Applied Sciences, 10(1), pp.69–77. Available at: http://dx.doi.org/10.3844/ajeassp.2017.69

77.

[10]. Toscano, J.C., McMurray, B.:Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics.Cognitive Science 34, 434–464 (2010)

[11]. Al Smadi ,An Improved Real-Time Speech In Case Of Isolated Word Recognition, Int. Journal of Engineering Research and Application, Vol. 3, Issue 5, Sep-Oct 2013, pp.01-05

[12].Siniscalchi, Sabato Marco, et al. "Exploiting deep neural networks for detection-based speech recognition." Neurocomputing 106 (2013): 148-157.

[13]. A.Smadi, K. & Al Smadi, T., 2017. Automatic System Recognition of License Plates using Neural Networks. International Journal of Engineering and Manufacturing, 7(4), pp.26–35. Available at: http://dx.doi.org/10.5815/ijem.2017.04.03

[14]. Al-Wahshat, H., Al-Maitah, M. & Al-Smadi, T., 2017. Voice Quality for Internet Protocol Based on Neural Network Model. Journal of Signal and Information Processing, 08(04), pp.195–202. Available at:http://dx.doi.org/10.4236/jsip.2017.84013

[15]. Al-Wahshat, H., Al-Maitah, M. & Al-Smadi, T., 2017. Voice Quality for Internet Protocol Based on Neural Network Model. Journal of Signal and Information Processing, 08(04), pp.195–202. Available at:http://dx.doi.org/10.4236/jsip.2017.84013