CARPATHIAN J. MATH.
21 (2005), No. 1 - 2, 127 - 136

Experimental Results on Isolated Speech Recognition System

GAVRIL TODEREAN, MARIETA GÂTA and ALEXANDRU CĂRUNTU

ABSTRACT. In this paper we described the development of Romanian ANN speech recognition system, which uses artificial neural networks. The network, which is composed by a three layer (a Multilayer Perceptron), is trained by conventional Back-propagation algorithm. The ANN speech recognition system based on Mel Frequency Cepstral Coefficients was developed using Matlab toolkit. The system was tested on the 50 isolated word Romanian speech corpus and it was training on a 50 distinct isolated word recognition task. The word recognition accuracy obtained was about 76%.

1. NEURAL NETWORKS AND THE APPLICATION OF THE NEURAL NETWORKS TO SPEECH RECOGNITION

Figure 1 shows a conceptual block diagram of speech understanding system [9] based on a model of speech perception in human beings. The acoustic input signal is analyzed by an “ear model” that provides spectral information about the signal and stores it in a sensory information store. Other sensory information (from vision or touch) is available in the sensory information store and is used to provide several “feature-level” descriptions of the speech. Long-term (static) and short-term (dynamic) memory is available to the various feature detectors. After several stages of refined feature detection, the final output of the system is an interpretation of the information in the acoustic input.

The system from Figure 1 is meant to model the human speech understanding system. The auditory analysis is based on our understanding of the acoustic processing in the ear. The different features analysis represents processing at different levels in the neural pathways of the brain. The short and long-term memory provides external control of the neural processes in ways that are not well understood. The form of the model is that of a feed forward connectionist network.

1.1. **Neural Networks.**

A neural network (or connectionist model, neural net, parallel distributed processing model) is basically a compact interconnection of simple, non-linear, computational elements of the type shown in Figure 2. It is assumed that are N inputs, labelled $x_1, x_2, \ldots, x_N$, which are summed with weights $w_1, w_2, \ldots, w_N$, thresholded and then nonlinearly compressed to give the output y, defined as

$$y = f\left(\Sigma w_i x_i - \Phi\right), \quad i = 1, N \tag{1.1}$$

where Φ is an internal threshold or offset and f is a nonlinearity of one of the types given below:

Received: 14.06.2005; In revised form: 07.10.2005
2000 *Mathematics Subject Classification:* 68N15, 68T10, 68T50, 92B20.
Key words and phrases: *Speech recognition, artificial neural networks.*

a) hard limiter

(1.2) $f(x) = \{+1 \text{ for } x \leq 0;\ -1 \text{ for } x < 0\}$

b) sigmoid functions

(1.3) $f(x) = \tanh(\beta x),\ \beta > 0 \text{ or } f(x) = \dfrac{1}{(1 + e^{-\beta x})}\ ,\ \beta > 0$

The sigmoid nonlinearities are used more often because they are continuous and differentiable.

The biological basis of the neuron network is a model by McCullough and Pitts of neurons in the human nervous system. This model presents all the properties of the neural element including excitation potential threshold for neuron firing and non-linear amplification, which compresses strong input signals.

1.2. **Neural Networks Topologies.** Artificial neural networks model different phenomena. There are several results in the design of artificial neural networks where we define an artificial neural network as an arbitrary connection of simple computational elements of the type shown in Figure 2. One key result is network topology that is how the simple computational elements are interconnected. There are three standard topologies:

- single / multilayer perceptrons
- Hopfield or recurrent networks
- Kohonen or self-organizing networks

In the single / multilayer perceptron the output of one or more simple computational elements at one layer will form the input to a new set of simple computational elements of the next layer. Figure 3 shows a three-layer preceptron. The single-layer preceptron has N inputs connected to M outputs in the output layer. The three-layer preceptron has two hidden layers between the input and output layers. What distinguishes the layers of the multilayer preceptron is the nonlinearity at each layer that enables the mapping between the input and output variables to possess certain particular classification / discrimination proprieties. For example, it can be proven that a single-layer perceptron can separate static patterns into classes with class boundaries characterized by hyperplanes in the $(x_1, x_2, \ldots, x_n)$ space [9]. Similarly, a multilayer perceptron, with at least one hidden layer, can realize an arbitrary set of decision regions in the $(x_1, x_2, \ldots, x_N)$ space. Thus, for example, if the inputs to a multilayer perceptron are the first two speech resonances (F_1 and F_2) the network can implement a set of decisions regions that partition the (F_1 and F_2) space into the 10 steady state vowels (Figure 4) [8].

The Hopfield network is a recurrent network in which the input to each computational element includes both inputs as well as outputs. Thus with the input and output indexed by time, $x_i(t)$ and $y_i(t)$, and the weight connecting the i^{th} node and the j^{th} node denoted by w_{ij}, the basic equation for the i^{th} recurrent computational element is

(1.4) $y_i(t) = f\left[x_i(t) + \Sigma w_{ij} y_i(t-1) - \Phi\right]$

The most important property of the Hopfield network is that when w_{ij}=w_{ji} and when the recurrent computation (1.4) is performed asynchronously, for an arbitrary constant input, the network will eventually settle to a fixed point where $yi(t) = yi(t-1)$ for all i. These fixed relaxation points represent stable configurations of the network and can be used in applications that have a fixed set of patterns to be matched in the form of a content addressable or associative memory. A simple interpretation of the Hopfield network shows that the recurrent network has a stable set of attractors and repellers, each forming a fixed point in the input space. Every input vector, x, is either "attracted" to one of the fixed points or "repelled" from another of the fixed points. The power of this type of network is its ability to correctly classify "noisy" versions of the patterns that form the stable fixed points.

The third type of neural network topology is the Kohonen, self-organizing feature map, which is a clustering procedure for providing a codebook of stable patterns in the input space that characterize an arbitrary input vector, by a small number of representative clusters.

1.3. **Network Characteristics.** Four model characteristics must be specified to implement an arbitrary neural network:

1. Number and type of inputs - This choice is similar to those involved in the choice of features for any pattern-classification system.

2. Connectivity of the network - This characteristic involves the size of the network (the number of hidden layers and the number of nodes in each layer between input and output). There is no good rule of solicitation as to how large or small such hidden layers must be. Praxis and intuition says that if the hidden layers are large, then it will be difficult to train the network (there will be too many parameters to estimate) and if the hidden layers are too small, the network may not be able to precisely classify all the desired input patterns. It's obvious that practical systems must balance these two competing effects.

3. Choice of offset - This choice of threshold, Φ, for each computational element must be made as part of the training procedure, which chooses values for the interconnection weights w$_{ij}$ and the offset Φ.

4. Choice of nonlinearity - Praxis indicates that the exact choice of the nonlinearity, f, is not very important in terms of the network performance. It must be continuous and differentiable for the training algorithm to be applicable.

1.4. **Training of Neural Network Parameters.** To complete specify a neural network, values for the weighting coefficients ($w_{ij}(t)$, which connects i^{th} input node with j^{th} output node) and the offset threshold (Φ_j, offset to a particular computational element) for each computational element must be determined, based on a labelled set of training data. By a labelled training set of data, we mean an association between a set of Q input vectors $x_1, x_2, \ldots, x_Q$ and a set of Q output vectors $y_1, y_2, \ldots, y_Q$ where $x_1 \Rightarrow y_1, x_2 \Rightarrow y_2, \ldots, x_Q \Rightarrow y_Q$. For multilayer perceptions a simple, iterative, convergent procedure exists for choosing a set of parameters whose value asymptotically approaches a stationary point with a certain optimality property (e.g. Local minimum of the mean squared error). This procedure, called back propagation learning, is a simple, stochastic gradient technique. For a simple, single-layer network, the training algorithm can be realized through a few convergence steps.

The perceptron convergence procedure is a slow, methodical procedure for estimating the coefficients of a system (a neural network or a classifier) based on a mean squared error criterion. The algorithm is simple and is guaranteed to converge, in probability, under a restricted set of conditions. The algorithm's speed of convergence is not sufficiently fast in many cases. In these cases it can be apply alternative procedures for estimating neural network coefficients.

1.5. **Advantages of Neural Networks.** The reasons for which neural networks are given serious consideration for many problems (including speech recognition) are:

1. They can quickly implement a great number of parallel computations. A neural net is a highly parallel structure of simple, identical, computational elements and therefore they are applied in massively parallel computation (analog or digital).

2. They posses a great amount of robustness or errors tolerance. The information embedded in the neural network is spread to every computational element within the network, this structure is natural among the least sensitive of networks to noise or imperfections within the structure.

3. The connection weights of the network need to be enforced to be fixed. They can be adapted in real time to improve performance. This is the basis of the notion of adaptive learning, which is present in neural network structure.

4. A sufficiently large neural network can approximate any nonlinearity or non-linear dynamical system by reason of the nonlinearity within each computational element. Therefore neural networks provide an easy way of implementing non-linear transformations between arbitrary inputs and outputs. Often they are more efficient than alternative physical implementations of the nonlinearity.

1.6. **Neural Network Structures for Speech Recognition.** Conventional artificial neural networks are structured to deal with static patterns. Because of the dynamic nature of the speech, some modifications to the simple structures are required, except for simple problems.

The simplest neural network structure that incorporates speech pattern dynamics is the time delay neural network (TDNN) computation element [11]. This structure radiates the input to each computational element to include N speech frames (spectral vectors that cover a duration of $N\Delta$ seconds, where Δ is the time separation between adjacent speech spectra). By expanding the input to N frames (N is on the order on 15) various types of acoustic-phonetic detectors become practical through the TDNN [9].

1.7. **Implementation of an Artificial Neural Network for Speech Recognition.** Presentation of the algorithm (of the used network):

The paper presents an application that implements a type of the artificial neural network: MLP (Multi Layer Perceptron). This network is based on Backpropagation learn, which is the adjustment of the weights according to the global error of the network. This adjustment is done in backward order of processing. This type of networking has a number of standard steps, which must be kept. These steps are:

Initializations of the work with aleatory weights lower than one.

Presentation of an example of training. This step is consisted by two components: presentation of an example to the entrance of the network and at the same time initialization of the desired output.

Forward step is the step in which it is calculated, for each layer, independently, the outputs for each neuron, depending on the synaptic and concordantly weights.

Backward step is the step in which are calculated the new weights and the new thresholds for each neuron separately, with the aim of computing the error and then the adjustment of these weights depending on the error.

The calculus of the global error of the neural network for the presented examples.

After the presentation of all examples it is passed to the adjustment of the weights.

Next step retakes again the steps 2-6 until the global error diminishes below a certain value of the threshold, set to the value 0.01.

1.8. **Structure of Our Speech Recognition System.** The aim of this research was to develop a baseline small vocabulary isolated word Romanians ANN based speech recognition system.

The ANN system has already been successfully applied in connected and continuous English digits recognition tasks. In case of recognition systems developed for Romanian language we are still at the beginning, but some results were obtained for isolated word recognition [1], [3] and it have been made some tentative for continuous speech recognition [4].

In the next sections we present some results obtained by our team for an isolated words recognition system developed by ourselves.

The research presented here is new and original because it is based on the study upon learning rate, number of epochs, training time, number of neurons in hidden layer applied to Romanian isolated word recognition for our condition: database with 100 records, closed room and neural network described bellow.

The platform for research and development of our recognition system is using a program made by us, which was developed in Matlab R12.

In our experiments we used a feed-forward neural network, the network was trained with Backpropagation algorithm. This MLP network has 150 neurons on the input layer, 100 neurons on the hidden layer and 10 on the output layer. If in the network we add another intermediate layer this will affect the calculus time.

Our database is formed by ten speakers, each speaker record ten different words, in this case the number from zero to nine, the type of recording for each file will be wav format. A half of the utterance (wav files) was used in the training step and the other half in the testing step of the recognition process. As a training function we use tansig in the hidden layer and logsig in the output layer.

The values of the threshold assure a natural generalization of the net. The constraint of very well learning of a set of examples it goes to the errors of recognition through particularization and locking of the certain and rigid set of training. Therefore the training must be done until is achieved a certain value of the error, named critical value, below which the network doesn't generalize, and over which the networks doesn't offer proper answers. Obviously, in the case in which network learns a certain logical function, as XOR, then the error must be as little as possible.

2. Experimental Results

In this paper we achieved an application for speech recognition. We conducted a series of experiments using this application for different words and different speakers. As a base for our result we use ten Romanian words: digits from zero to nine.

Figure 5 shows an example of "zero" word recorded by the program as wav file. Figure 6 presents spectral component of the same word "zero" uttered by the same speaker as in Figure 5.

Our contribution is the study upon learning rate, number of epochs, training time, number of neurons in hidden layer in speech recognition for our condition: database with 100 records, closed room and neural network described before.

Figure 8 shows recognition rate for ten Romanian word, number of epochs 5000, learning rate 0.1, and sum-squared error 0.01. Figure 7 presents an example of Sum-Squared Network Error for 5000 epochs for same experimental dates as in Figure 8.

We obtained several results in this paper, which are presented in Table 1 and Table 2:

Number of epochs (for learning rate = 0.1)	Recognition rate(%)	Training time (min)
1000	73%	1
2000	76%	2
4000	74%	3
5000	74%	3.5
7000	73%	3.5
9000	70%	4

Table 1

Learning rate (for learning rate = 0.1)	Recognition rate (%)	Training time (min)
1000	73%	1
2000	76%	2
4000	74%	3
5000	74%	3.5
7000	73%	3.5
9000	70%	4

Table 2

3. Conclusions and Discussions

First of all we observed that the rate of recognition is decreasing with number of epochs. Other observation is that if learning rate is increasing than recognition rate is decreasing. Regarding the training time we observe that this is growing with the number of epochs and also is growing with learning rate usually.

The best recognition rate obtained by us is 76%. We obtained the same result even if we decreased learning rate to 0.01 and number of epochs remains the same, 500.

Regarding the number of neurons in the hidden layer, in our experiments we obtained as the best value 100 (depending of obtained results and calculus time). In our tests this value was between the values 50 and 250.

Also we observed that if learning rate is high then the algorithm is unstable and if learning rate is small then the number of steps for convergence is high.

3.1. **Future Work.** We are at the start age of the development of our neural networks speech recognition system. In order to improve a number of factors (recognition rate, window size, training method, architecture of the network) which can influence the performance of the system we have to work more. For conclude the final decisions the experiments should be extended to a much bigger speech databases and a large number of speaker.

Here we present a number of experiments involving neural networks used in speech recognition. Our goal was to test our neural networks library that we have implemented, to study the influence of a few parameters -such as learning rate, number of epochs, training time, number of neurons in hidden layer- in speech recognition for our condition of work. Our future tests will assume continuous speech recognition for Romanian language.

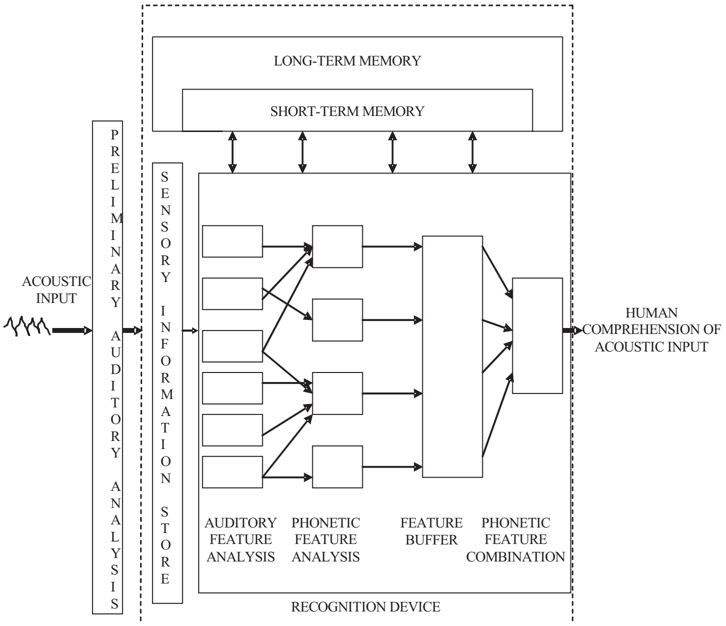
LONG-TERM MEMORY
SHORT-TERM MEMORY
PRELIMINARY AUDITORY ANALYSIS
SENSORY INFORMATION STORE
ACOUSTIC INPUT
HUMAN COMPREHENSION OF ACOUSTIC INPUT
AUDITORY FEATURE ANALYSIS
PHONETIC FEATURE ANALYSIS
FEATURE BUFFER
PHONETIC FEATURE COMBINATION
RECOGNITION DEVICE

Figure 1

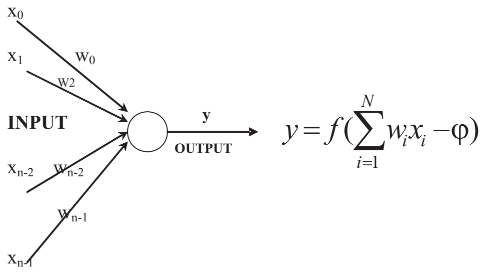

Figure 2

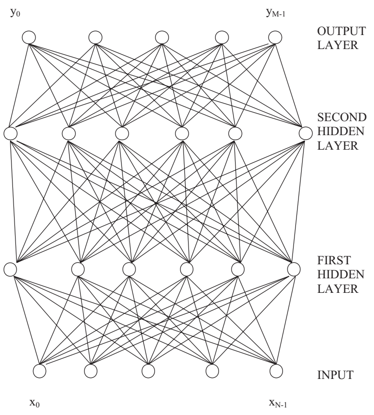

Figure 3

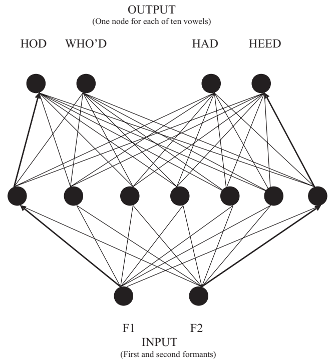
OUTPUT
(One node for each of ten vowels)
HOD
WHO'D
HAD
HEED
F1
F2
INPUT
(First and second formants)

Figure 4

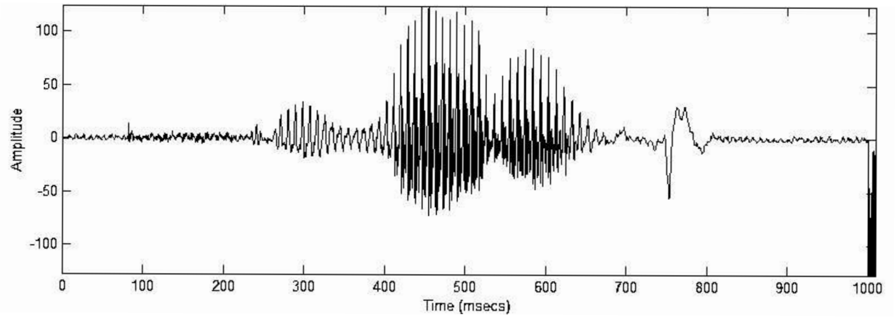

Figure 5

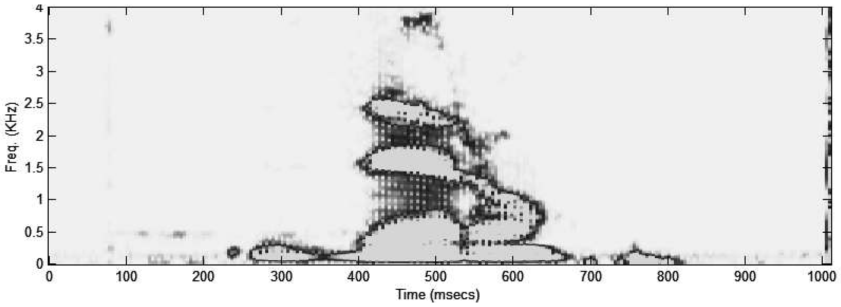

Figure 6

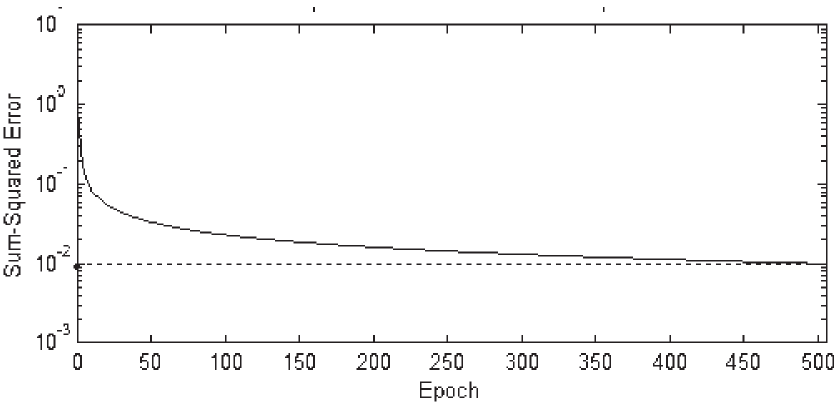

Figure 7

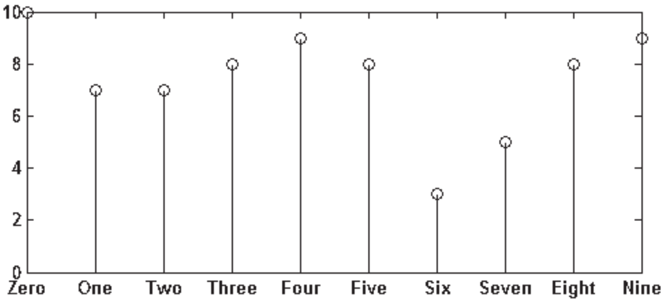

Figure 8

REFERENCES

[1] Căruntu, A., Toderean, G., *A Comparative Study of the Methods Used in Isolated Words Recognition*, 2^{nd} Conference on Speech Technology and Human Computer Dialogue, SpeD 2003, Bucharest, 155-159, April 10-11, 2003

[2] Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V., *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, 1996

[3] Giurgiu, M., *Contributions to the Isolated Word Recognition for Romanian Language*, (in Romanian) PhD Thesis, TUCN, 1996

[4] Giurgiu, M., *Experimental Informations Retrieval System Based on Romanian Continuous Speech Recognition*, 2^{nd} Conference on Speech Technology and Human Computer Dialogue, SpeD 2003, Bucharest, 161-166, April 10-11, 2003

[5] Green, P., Renals, S., *Speech Technology*, 2002

[6] Huang, X., Acero, A., Hon, H. W., Reddy, R., *Spoken Language Processing: A Guide to Theory, Algorithm & System Development*, 2001, Prentice Hall PTR

[7] Hagan, M. T., Demuth, H. B., and Beale, M. H., *Neural Network Design*, 1996, Boston, MA PWS Publishing

[8] Lippmann, R., *An Introduction to Computing with Neural Nets*, IEEE ASSP Mag., 4(2):4-22, April 1987

[9] Rabiner, R L., Juang, B.H., *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, USA, 1993

[10] Toderean, G., Costeiu, M., Giurgiu, M., *Artificial Neural Networks* (in Romanian), Microinformatica Cluj-Napoca, 1995

[11] Weibel, A., Hanazawa, T., Hinton, G., Shikano K., Lang, K.J., *Phoneme Recognition Using Time Delay Neural Networks*, IEEE Trans. Acoustics Speech Signal Processing, ASSP-37: 328-339, 1989

TECHNICAL UNIVERSITY OF CLUJ-NAPOCA
FACULTY OF ELECTRONICS AND TELECOMMUNICATIONS
DEPARTMENT OF COMMUNICATIONS
G. BARIŢIU 26-28, 400027 CLUJ-NAPOCA, ROMANIA
E-mail address: Gavril.Toderean@com.utcluj.ro

NORTH UNIVERSITY OF BAIA MARE
DEPARTMENT OF MATHEMATICS
AND COMPUTER SCIENCE
VICTORIEI 76, 430122 BAIA MARE, ROMANIA
E-mail address: marietag@ubm.ro

TECHNICAL UNIVERSITY OF CLUJ-NAPOCA
FACULTY OF ELECTRONICS AND TELECOMMUNICATIONS
DEPARTMENT OF COMMUNICATIONS
G. BARIŢIU 26-28, 400027 CLUJ-NAPOCA, ROMANIA
E-mail address: Alexandru.Caruntu@com.utcluj.ro