ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
December 2014,
Vol. 7, No. (3):
Pgs. 331-336

Artificial Intelligence Technique for Speech Recognition based on Neural Networks

MOHAMMAD AL- RABA BAH[1], ABDUSAMAD AL-MARGHILANI[1] and M. A EYAD[2]

[1]Department of Computer Science, Northern Border University, Arar, KSA.
[2]Jizan University, Jazan, KSA

(Received: November 01, 2014; Accepted: December 05, 2014)

ABSTRACT

Creation of natural human sources to communicate with the computer is currently one of the greatest challenges of modern science. The speech input facility is the most user-friendly way, adopted by development of speech recognition based on sophisticated technologies. Scientists began the selection of informative signs, describing the voice signal, afterwards the task of classification of speech signals as a set of informative signs. The development of methods of signal processing in the absence of sufficient models lead to questions about the processes of generation signals using artificial neural networks, As a result; when building a signal processing system, the structure of the network should be selected; according to parameters of the signals and training network using an algorithm to maximize the use of the information contained in the data of the experiment. This article proposes the Application of wavelet transform for reduction of the value of artificial neural networks for speech recognition tasks this method a present Study a new modification of neural networks a neural network with inverse Wavelet Decomposition of the signal. For example, the speech recognition task, the analysis of the proposed method. The effectiveness of the method is proved by the results of computer simulations.

Key word: Artificial intelligence, speech recognition, neural network, Wavelet transforms.

INTRODUCTION

In this area, significant progress has been achieved, but the main problem of modern speech recognition is to achieve the robustness of the process. Unfortunately, programs that could show equivalent human quality speech recognition under any conditions, not yet created [2]. The immaturity of existing technologies is associated primarily with the issues of recognition of noisy and continuous speech.

Known methods are various advantages such as good account of temporal structure of speech signal (shear strength), resistance to the variance signal resistance to noise, low resource consumption, size of the dictionary. but the problem is that for high-quality speech recognition to match these benefits in the same method of recognition

In speech recognition takes into account the following:

The temporary nature of the signal.

Speech Variation due to: local distortion of scale, interaction (spirit) sounds intonation, the human condition.

Speech signal Variability due to: conditions of entry (distance from the recording device, its features, etc.), score of ambient sound (noise).

Continuous Speech Recognition and Speaker independent recognition.

Speech technology

Speech recognition is the process of extracting text transcriptions or some form of meaning from speech input.

Speech analytics can be considered as the part of the voice processing, which converts human speech into digital forms suitable for storage or transmission computers.

Speech synthesis function is essentially reverse speech analysis-they convert speech data from digital form to one that is similar to the original entry and is suitable for playback.

Speech analysis processes can also be called digital speech coding (or encoding) and

The high variability due to local scale as shown in [3]. processing of time signals requires devices with memory. This issue [4] calls the problem of temporary structures,

The problem of temporary distortions It was that speech comparison samples of the same class can be used only if the timescale conversions of one of them. In other words, say the same sound with different durations, and Moreover, the various parts of the sounds may have different duration as part of a class. This effect allows you to talk about "local distortions of scale along the time axis.

You need to combine the advantages of different methods in one that leads to the idea of applying specialized neural networks. In fact, the artificial neural network technology, no limited in theory, perspectives and opportunities, most flexible and most intelligent. But the need to take into account the specifics of the speech signal the easiest to implement, through the use of a priori information in neural network structure, which requires specialization. In this work, offer specialized architecture neural networks with Wavelet Decomposition vector, or target the neural network with inverse Wavelet Decomposition.

Related Works

Reduce the value of artificial neural networks

Neural network speech recognition scheme implies a number equal to the number of classes of recognition. Each entry gives a value to indicate the probability of belonging to a given class, or a measure of closeness of this fragment to this speech resolves to sound. For simplicity we confine ourselves to describing a class of pattern recognition. Our reasoning without losses can be transferred to a more general case.

Usually the voice signal is broken into small pieces-frames (segments), each frame is subjected to pre-treatment, for example, by using the window Fourier transform. This is done to reduce space and increase of attributive stripped classes [5].

As a result, each frame is characterized by the set of coefficients, called acoustic

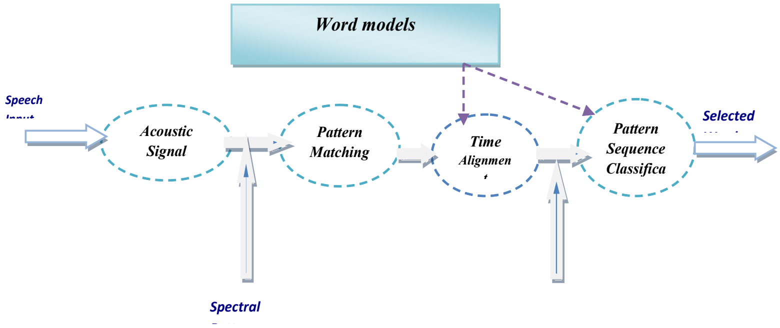

Fig. 1: Diagram of the processing of speech signals *planning*

characteristic vector. Denote the length of the frame as Δt, and the length of the characteristic vectors as N_{ϕ}, and the acoustic characteristic vector at time $t + n\Delta$ As $x(n$.

B this case is expected to assess the probability that the speech portion of the class at the time $n_0 t$ You must consider the voice signal in the final period $(t - n\Delta t; t + n\Delta$, where Δt - the length of the frame in time, à $n \in N$. This cut is usually called a window. We have a neural network solution to display input parameters x(n0-n), x(n0-n+1), … , x(n0), … , x(n0+n-1), x(n0+n) the output value of the y(n0) [6].

Improving recognition quality in standard approach is associated usually with manipulations on the input signal, or selecting the conversion and improvement of preprocessing.

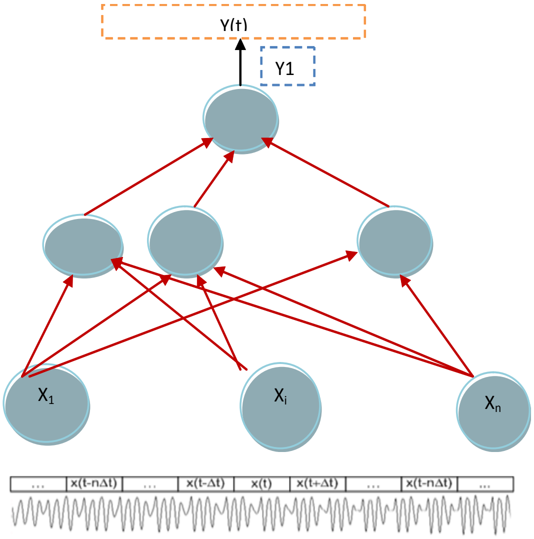

Fig. 1. A standard approach to recognition

It does not take into account that the output value (see Figure 1) is a function that depends on time, with a small velocity changes. To use the properties of the output signal in order to improve the quality of the learning process the authors developed technique to reduce the field values.

g(f(x)) Function f(x), with its values E(f) one-to-one displayed on many values (E) approximating the neural network, When this $(E(g(f)) \cap E \cap I) \neq \varnothing$ (2.1)

and ; $(E(g(f)) \setminus E) = \varnothing$(1)

I – many surplus values

An excess of neural network for the region, its value or exception of surplus values will be referred to as the one-to-one conversion of h (f (x)) of the function f (x), at which many excess values are partially or entirely outside the field values $(h(I) \setminus E) = \varnothing$ (2).

An example of the exception of surplus values INCE may be scaling, or multiplication by a factor of standard neural network has a limited output, that is, each component of the output vector is within a certain range, usually either (-1.1), either (0; 1) [7]. For simplicity, we will assume that this range (0; 1). This means that the value is a hypercube. If you know that the value of a certain output neural network according to the problem conditions does not exceed a certain value k, It is advisable to increase the components of target vectors on k. so, we cut the excess values $\left(\frac{1}{k};1\right)$ scaling is a linear transformation-scale components of the target vector with shift.

that the neural network has a significant impact view output in the training set of the neural

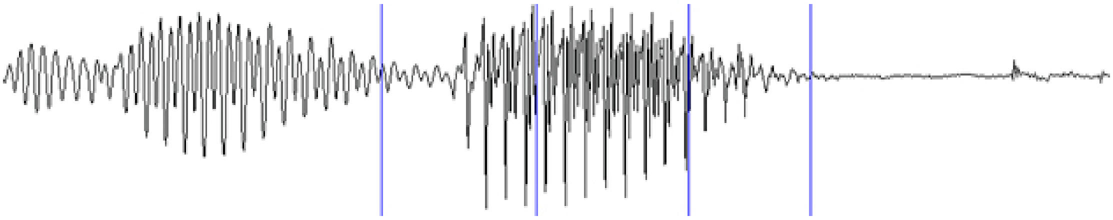

Fig. 2: Example of interaction of sounds-write the words

network. This should reduce the redundancy in the description of the objective function that entails a reduction in the values area. Therefore, the selection of this view is an important task of designing a neural network. To view target vectors is affectedly three factors: the chosen objective function, selection of presentation, reducing the redundancy in the description of the objective function, and method of reducing the field values.

Note that the speech recognition task, you can reduce the value of thousands of up to three hundred thousand times.

Methods

For the application of the method of reducing the value of artificial neural network for recognition of phonemic tasks we need to choose a target function and analyze its properties.

One of the problems with speech recognition is that you cannot select a stand-alone speech sound. The form of sound is very dependent on the sound environment of a Signal that goes after him, and the sound that goes before him [8].

It is known that waveform is a smooth transition from one sound to another.

The clear boundaries between the sounds correctly, better to talk about the intermingling signals the sounds and the background of zero (Figure 2).

As a model of the phoneme in question proposes a model of phonemic zone is close to pure sound-area of phonemic another zone.

As the target function to measure the similarity of pure sound, arbitrarily drawn or highlighted in the speech processing. Let us introduce the similarity measure P (t, Ù)of the speech signal at time t the pure sound of Ù. Under sound refers to a signal, separately, without subsequent and earlier phonemes [9].

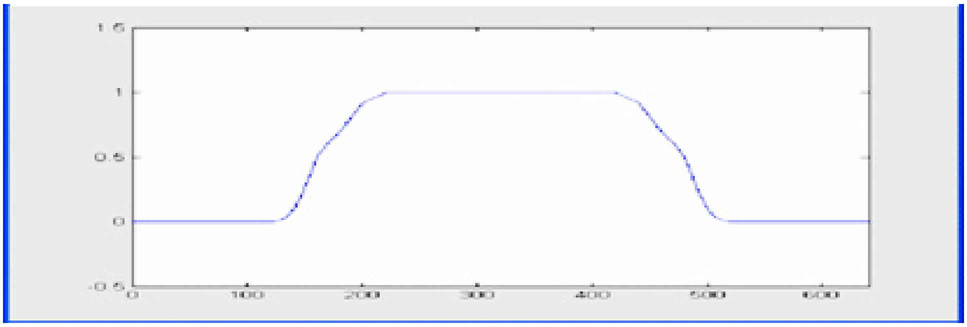

Fig. 3: An example of the similarity function.

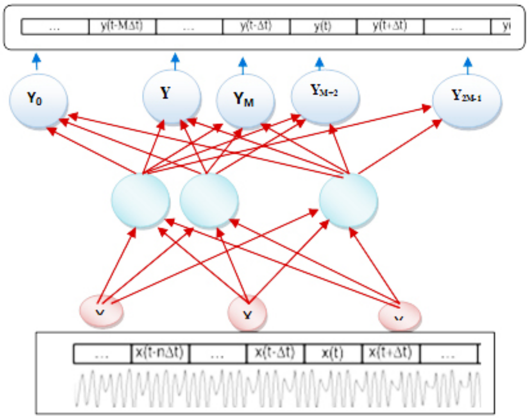

Fig. 4: The present configuration of the neural network for recognition of phonemic tasks.

Properties of functions $P(t,\Omega)$:
$P(t,\Omega)=0$ and $P(t,\Omega)<\varepsilon_0$
If the beep sound is not ε, or is not a speech anyway, where e on 0 threshold is close to zero;
$P(t, \Omega) = 1$ if the beep is sound;
$P(t, \Omega)$ E (0.1) in the zone of phonemic interface, and P is the phonemic seamlessly interface to sound and gradually decreases in the phonemic joint after.
An example of a function of similarity can be seen in Figure 3.
in t1,t2,…,tk,… - readouts of time, ti+1=ti+Δt for

any iεN. then

$$\sup\{P(t_{i+1})-P(t_i)\} << \sup\{P(t_i)\} \quad ..(3)$$

Since the function, so $\sup\{P(t_i)\}=1$ and 3.1 can be rewritten as $\sup\{P(t_{i+1})-P(t_i)\}<<1$

..(4)

Extend the standard configuration of a neural network (described in the previous section) so that the neural network was looking for a vector with length values at once M - y(n0), y(n0+1), … , y(n0+M), moments in time $n_0\Delta t$, (n_0+1) … , $(n_0+M$. The number of input vectors, respectively, to increase the M - x(n0-n), x(n0-n+1), … , x(n0), … , x(n0+n-1), x(n0+n), … , x(n0+n+M-1), x(n0+n+M).

In accordance with the method of reducing the value of the select transformation that eliminates the redundancy in the description of the output signal. As in our case, it is known that [10]

$P(t_{2i+1})-P(t_{2i})<<1$, i=1,2,3,…, It makes sense to replace values $P(t_{2i+1}), P(t_{2i})$ on $o(t_{2i+1}), o(t_{2i})$,

Moreover , $o(t_{2i+1})=\dfrac{P(t_{2i+1})+P(t_{2i})}{2}$

and $o(t_i)=\dfrac{P(t_{2i+1})-P(t_i)}{2}$ ..(5)

Convert easily to formulas $P(t_{2i+1})=o(t_{2i+1})+o(t_{2i})$

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

$$o(t_{2i}) \sup o(t_{2i})=\frac{1}{k}$$

$\frac{1}{k}< 1$. Thus, the values in the interval $\left[\frac{1}{k};1\right)$ redundant

$$k_i=\frac{2}{\sup\{P(t_{i+1})-P(t_i)\}} \quad ...(7)$$

Then we can scale every second output by a factor of.

Compare two versions of a neural network. The first is with the uncalled exits, the second with scaled outputs. First, in the case of uncalled input values, the neural network is broader, and in order to achieve the same accuracy as the scaled inputs required to go through several iterations until accuracy is achieved $\sup_i\left\{\frac{2}{k_i}\right\}$

The resulting accuracy is less than in the first case. Show it.

Let be the error of the neural network outputs in this iteration. In this case, the error (sum of squares) of the first neural network will

$$E_1=2e^2 \quad ...(8)$$

$$E_2=e^2+\left(\frac{e}{k_i}\right)^2=\left(1+\frac{1}{k_i}\right)e^2 \quad ..(9)$$

The results of practical experiments

The speech was chosen for the experiments of 37 words. Verification of neural network with inverse Wavelet Decomposition was carried out at the isolated sound recognition task ”a”. The task was complicated by the fact that the length of the sound in the varied fourfold from the minimum.

Purpose of the experiment was to compare two identical neural networks with identical architecture, trained by the same algorithm, one of which it studied with the wavelet, and another without it [11]

Neural network estimation was carried out as follows. There have been several runs of the system (here are the results for 20 runs). In each run was chosen as the test case with the worst result and calculated mathematical expectation errors on all examples of control sample. Based on the results of twenty runs was chosen the best result according to both criteria.

During these experiments yielded the following results (see table 1):

The criterion here is “No. 1” is the best mathematical expectation errors on the

Table. 1: Results of the experiments.

Method	Criterion No1,%	Criterion No2,%	Dispersion
Perceptron	11,57	17,83	0,03028
Perceptron with Wavelet module	0,00330	0,00744	0,00016

control sample results 20 training neural networks, Accordingly, the criterion of ”No. 2" is the best mathematical expectation the supermom errors on the control sample results 20 training neural networks.

Based on the results of the experiments we can talk about very notable results. In this case, as you can see on the dispersion, this result is very resistant.

CONCLUSION

Use model of speech recognition based on artificial neural networks. Training the neural network approach is being developed using genetic algorithm. This approach will be implemented in the system identification numbers. Coming to the realization of the system of recognition of voice commands It is also planned to develop system of automatic recognition of speech keywords that are associated with the processing of telephone calls or area security.

Reducing the likelihood of lower learning in local minimum, so you can talk about the following advantages of wavelet transforms target values:

1. Convergence acceleration due to the conversion of the gradient.
2. Improvement of the result accuracy.
3. Reduction in the number of iterations by greater initial localization solutions.
4. Decrease the probability of falling into a local minim mum.

REFERENCE

1. Tebelskis, J. Speech Recognition using Neural Networks: PhD thesis … Doctor of Philosophy in Computer Science/ Joe Tebelskis; School of Computer Science, Carnegie Mellon University.– Pittsburgh, Pennsylvania, 1995.– 179 c.
2. Jain, L.C., Martin, N.M. Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Applications/ Lakhmi C. Jain, N.M. Martin.– CRC Press, CRC Press LLC, 1998.– 297c.
3. Handbook of neural network signal processing/ Edited by Yu Hen Hu, Jenq-Neng Hwang.– Boca Raton; London; New York, Washington D.C.: CRC press, 2001.– 384c.
4. Principe, J.C. Artificial Neural Networks/ Jose C. Principe// The Electrical Engineering Handbook/Ed. Richard C. Dorf.– Boca Raton: CRC Press LLC, 2000.– 2719c.
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. Arisoy, E., Chen, S. F., Ramabhadran, B., and Sethy, A. (2013). Converting neural network language models into back-off language models for efficient decoding in automatic speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 8242–8246. IEEE.
8. Dahl, G., Yu, D., Li, D., and Acero, A. Large vocabulary continuous speech recognition with context-dependent dbn-hmms. In ICASSP (2011).
9. Graves, A., Jaitly, N., and Mohamed, A. Hybrid speech recognition with deep bidirectional LSTM. In ASRU (2013).
10. Grezl, Karafiat, and Cernocky. Neural network topologies and bottleneck features. Speech Recognition (2007).
11. T. A. Al Smadi ,An Improved Real-Time Speech In Case Of Isolated Word Recognition, *Int. Journal of Engineering Research and Application,* **3** (5), 01-05 (2013).