Boltzmann Machine Neural Network for Arabic Speech Recognition

H.R.Mohamed
Department of Computer Science , College of Education for Girls, University of Kufa

Abstract

Boltzmann machine neural network has been used to recognize the Arabic speech. Fast Fourier transformation algorithm has been used to extract spectral features from an a coustic signal .

The spectral feature size is reduced by series of operations in order to make it salable as input for a neural network which is used as a recognizer by Boltzmann Machine Neural network which has been used as a recognizer for phonemes . A training set consist of a number of Arabic phoneme repesentations, is used to train the neural network.

The neural network recognized Arabic. After Boltzmann Machine Neural network training the system with few selected Arabic phonemes, the results came out to be very encouraging .

Introduction

Neural network was applied to speech recognition in the mid-1980 s.Neural network has been proposed by speech recognizers. Their function is to provide a statistical model capable of associating a vector of speech features with the probability of any of a given number of phoneme (1).

Neural networks have here the function of statiscal machines . The syllable in Arabic speech is based on the constructive components that are contained in its structure. Each syllable has a main part that stands out and has prominence . This part is referred to as the (nucleus) of the syllable (2).

The remaining componects are referred to as (margined factors). The features used by speech recognition systems may be as simple as the energy or zero-crossing rate, and other related features of the waveform during each speech frame.

A more elegant and robust method for feature extraction is based on the source / filter model of the vocal tract .

The most important parametric representation of speech is the short time spectral envelope. Spectral analysis methods are therefore generally considered as the core of the signal processing in speech recognition system (3).

There are many methods to process singnals, some of them in time-domain and the others in frequency domain. The Fourier transom method (FFT) is asset of highly efficient algorithms for evaluating signals . Any signal x(t) , expressed as a function of the time – in the so called time domain , can be instead expressed in the frequency domain x(w) , in terms of its frequency spectrum (2).

The continuous –time Fourier integral provides the means for obtaining the frequency-domain representation of signal from its time representation and vice-versa. The digital version , becuase it is a sampled system , only requires that the basis function to operate at a number of discrete frequencies it is there fore known as the Discrete Fourier Transform or DFT(4).

Training and learning are not the same for speech recognition by neural network training it is the procedure by which the network learns, learning is the end result of that procedure .Trainig is a procedure external to the network , learing is an internal process or activity . Training is done by example speech by neural network (4).

Boltzmann machine : The Boltzmann machine solved the complex prolem for Hopfield net. Hopfield net suffer from a tendency to stabilize to a local rather than a global minimum of the energy function . This problem is larely solved by Boltzmann machine in which the neurons change state in a statistical rather than a deterministic fashion. There is a close analogy between these methods and the way in which a metal is annealed ,hence the methods are of ten called simulated annealing(1).

Metal is annealed by heating it to a temperature above its melting point , and then letting it cool gradually (4).

At given temperature s , the probability distributionof system energies is determined by the Boltzmann probability of system energies determined by the Boltzmann probability factor (1).

P=Exp(-Eik T) ……………………..[1]

Where

E=system energy

K=bltzmznn 's Constant
T=temperature

Boltzmann machines are essentially an extension of simple stochastic associative networks to include hidden units vnis not involved in the pattern vector (1)

Both voiced and unvoiced sounds are subdivided into several subcategories :-

1. Fricative sounds are produced as the result of air friction within the oral cavity.
2. Plosive sounds (stop sounds) are called so since they sound like small explosions coming from the vocal cavity .
3. Nasal sounds are head when you block your oral tract by closing your mouth .And for any air into your nasal cavity . They are voiced sounds since they utilize the vocal cords.

4.Affricate sounds are similar to the fricative sounds since they are slso produced by air turbulence.

5. Glide (liquid) sound is a voiced sunod similar to a vowal sound. The difference is that you change the shape of your vocal cavity while pronouncing the sound .

We must find useful in phonology to define the phoneme . It is the smallest unit in speech where substitution of one unit for another makes a distinction of meaning (3).

There is a wide variation in the acoustic properties representing a particular phoneme . In some cases these differences are merely the result of the influence of the neighboring sounds on the positions of the tongue and other articulators.

This effect is known as coarticulation. The difference might be a feature that has developed for the language over a period of time(2).

The coarticulated sounds are called Allophones . Each phoneme has several allophones associated with it .

Design of system

The aim of which is to transcribe spoken utterances into phonemes and eventually to transform then into recognition words by Boltzmann machine network (written text).

The speech singnal input to the processing unit is recorded in a room environment and based on speaker dependent. All recordings were collected in a passive manner in a digital data base to consist of Arabic syllable each of 4 different utterances .

Wave shaper sound of creative lab is used to convert speech signal into digital form, every sample represented as 8 bit integer . Each fram is defined by 128 points Hamming window which covers 16 ms, it is transformed into its corresponding spectra by using FFT algorithm . The result is the 64 power spectra components and selected from FFT output .

Each component is written as :

$AK(m) = \sum X_n(m).EXP(-j2\pi k_n / N)$................[2]

$1 \leq K \leq N$

The output of the channel j at time m is compared with the mean of its output a time m-1 and m+1 and modify its value by

$$Aj(m) = \begin{Bmatrix} Aj(m)................ifAj(m) \geq Cj \\ Cj.......................otherwise \end{Bmatrix}......................[3]$$

By this step we reduced such variation to the minimum in order to ehance the recognition process accuracy .

Window used to reduce effects to using a finitelength recode .

They are two input signals x(n) , y(n) which are defined for the interval $0 \leq n \leq N-1$

Out side this interval it is assumed that x(n) and y(n) are 0.

We wish to compute the correlation function .

$Rxy(m) = 1/N \sum (x(n) - x)(y(n+m) - Y)$........................[4]

$0 \leq m \leq L-1$

Where x and Y are the estimated means of x(n) and y(n)

$x = 1/N \sum x(n)$..[5]

$\dot{Y} = 1/N \sum y(n)$ ....[6]

The number of samples of the singal is large compared to the number of correlation lags of interest , an efficient procedure for implementing Eq[4] is to use FFT method to accumulate partial correlation sums and then inverse transform the result to give the desired correlation function . If the size of the FFT is M points , then M/2+1 values of the correlation function are obtained thus FFT size must be at lest twice the desired number of correlation values L.

The features collected contained 64×M (M frames) points . A neural network must have at least one connection to each one of its input value.

The flow chart of the system is shown in figure(2).

Performance

The speech signal is broken in to 128 different frequency contributions . These may appear at relative amplitudes of 1 to 1.5 . The Boltzmann machine has 128 multi –valued input units, 40 hidden units and 8 output units. Each syllable in Arabic speech corresponds to an 8-bit code . Each hidden unit was connected to all input, output and other hidden units . There are 7000 links in all. After sufficient learing , the machine recognized about 90% of patterns accurately .

Most of the learning was complete within about 50 epochs.

132 training cases were used .

The algorithm for segmentation is :

1. Begin
2. Input the Arabic speech
3. For each apeech frame m

We sum all the magnitudes of amplitude in time domain

$$Mag \quad t(m) = \sum Xr(m)| \ldots\ldots\ldots [7]$$

n=1,2,3……….,N

4.Calculate the sum of magnitudes of frequencies in frequency domain

$$mag_f(m) = \sum \sqrt{Ar_k^2}(m) + Ai_k^2(m) \ldots\ldots\ldots\ldots [8]$$

Where $1 \le k \le N$

$A_K^R(M)$ is the real part of A_k(M) and Ai_k(m) is its imaginary part .

5.Computed the ratio of mag_t(m) to mag_f (m)to mag_f(m)

$$R(m) = \frac{mag_t(m)}{mag_f(m)} \ldots\ldots\ldots\ldots\ldots\ldots [9]$$

6. Calculate the SUM of the indexes of the four bands with the highest power

$FR(m) = \sum NR_j(m) 1 \le j \le 4$

The Upper frequency of each band is determined $\mathrm{Frq}_k = \frac{4f_s}{N} K$

F_s: is the sampling rate
K: band sequence number

Conclusions

1. The vowels are characterized by their low frequencies with high energies, where as most consonants are characterized by high freqencies
2. Concerning the limitation for the application of Boltzmann network there is no certainty about how big the learn factor E should be selected in the algorithm and there are no estimates for the number of states to be determined .
3. Boltzmann network weights are symmetric but allows for Hidden units. Weights adjusted through stochastic update rule based on simulated annealing .
4. After training the neural netwok , we notice that most of the output layer cells, each one responds to only one phoneme and few of them were responding to more than one phoneme.
5. Boltzmann machine provides a learning algorithm which adapts all the connections weights in the network given only the probability distribution over the visible units .
6. All units in a Boltzmann machine compute an "energy gap"

$\Delta Ei = E_{-1} + E_{+1} = \sum W_j \delta_j$

7. We are taking the difference of two noisy random variables in order to estimate the correlation , the error only decreases as√Nsamples.

References

1. Hinton,G.E. and Sejnowski, T.J. Learning and relearning in Boltzmann machine , in parallel distributed processing , Cambrideg, MA:MIT Press, 1:282-317.
2. Mohsin,A.H. (2000). Hybrid scheme for Arabic speech recognition , Ph.D thesis of computer science, BasrahUniversity.
3. Rabinar ,L.R. and Juang,B. (1993). Fundamentals of speech recognition, Prentice –Hall.

4. Fauseelt,L. (1994). Fundamentals of neutal networks, Prentice Hall International , Inc .

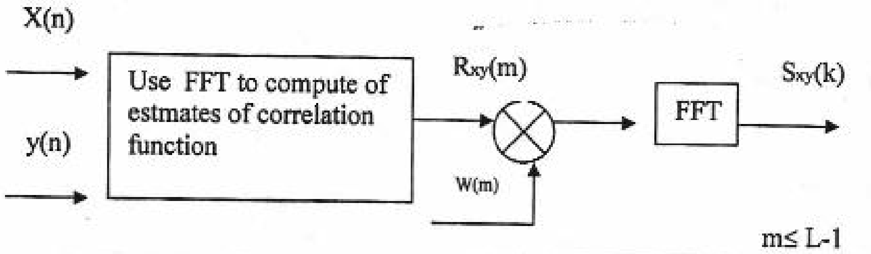

Fig.(1) Block diagram of the correlation method of spectrum analysis

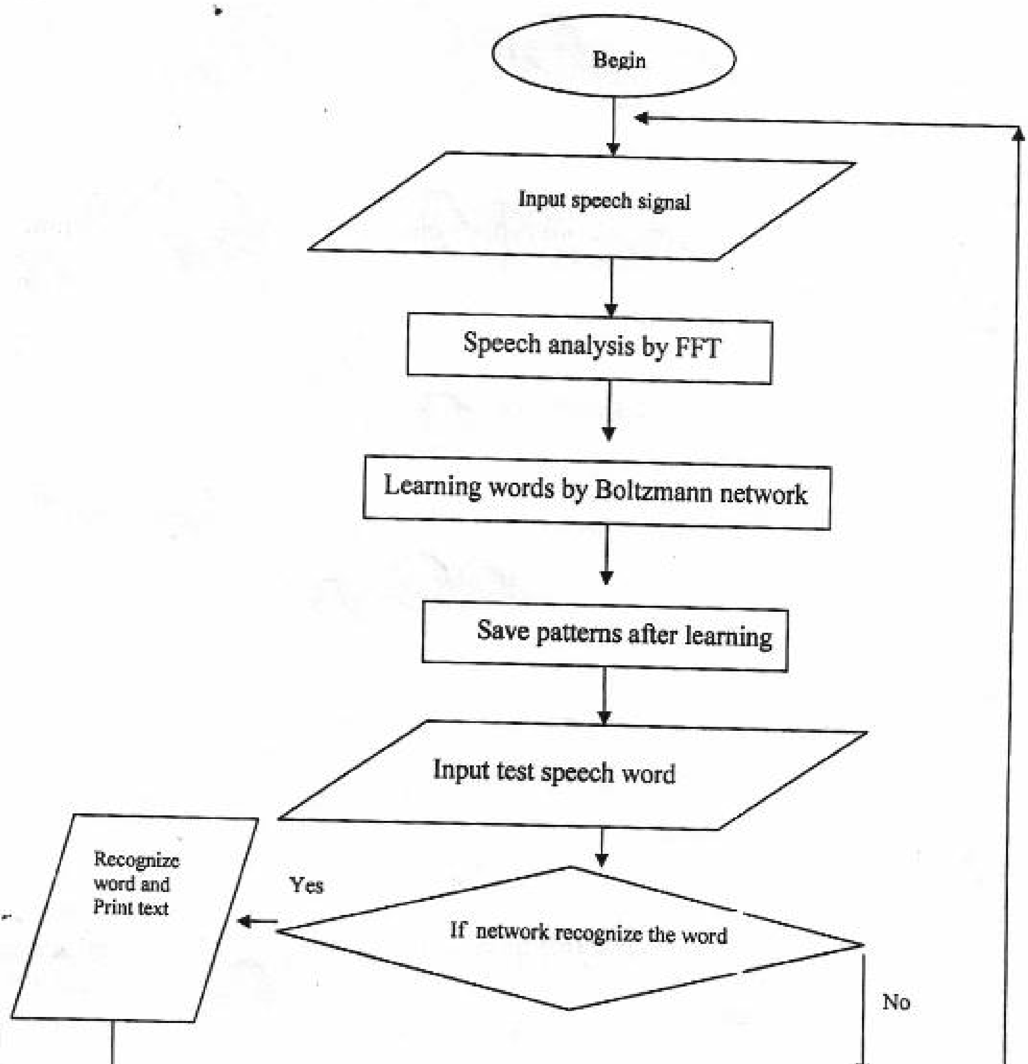
Begin
Input speech signal
Speech analysis by FFT
Learning words by Boltzmann network
Save patterns after learning
Input test speech word
Recognize word and Print text
Yes
If network recognize the word
No
No
Yes
Stop
New input

Fig.(2) Flow chart of the system

الشبكة العصبية المسماة ماكنة Boltzmann لتمييز الكلام العربي

هند رستم محمد

قسم الحاسبات ، كلية التربية للبنات ،جامعة الكوفة

الخلاصة

استخدمت خوارزمية فورير السريعة (FFT) لاستخلاص طاقة الترددات التي تحتويها الاشارة الصوتية تم تقليص حجم البيانات بعد ذلك بسلسلة من العمليات بغية الوصول الى مجموعة مناسبة للشبكة العصبية التي استخدمت كمميز للأصوات العربية .

استخدم في هذا البحث شبكة ماكنة Boltzmann كمميز للأصوات حيث دربت مجموعة تدريب تحتوي على عدد من تمثيلات الأصوات العربية التي تم الحصول عليها من مرحلة التمثيل. استطاعت هزة الشبكة من خلال التدريب ان الأصوات العربية حيث كانت النتائج مشجعة جدا" بعد تدريب الشبكة العصبية ماكنة Boltzmann.