# Speech Recognition from PSD using Neural Network

Amin Ashouri Saheli*, Gholam Ali Abdali**, Amir Abolfazl suratgar***

***Abstract:*** **in this paper we present a system for speech recognition using neural network from obtained data of power spectral density peaks. For this work, a small size vocabulary containing the word "yes" and "no" is chosen. Spectrum features is extracted from estimated power spectrum by Autoregressive parametric method in each frame of speech signal. And it is given to feed forward Back propagation neural network with gradient descent with adaptive learning rate training algorithm. Network is trained for classification to two classes.**

***Keywords***: ***Autoregressive, PSD, Back propagation, burg.***

## *1. Introduction*:

Goal of this work is to obtain a system based of neural network for recognition of speech with using latent information of data's PSD. Speech recognition is done using feed forward network over features that obtained from estimated PSD by autoregressive method (Burg) in each frames of speech signal. Autoregressive method is parametric approach for estimation of Spectrum of data that can achieve with small data set. This method unlike other parametric methods utilizes linear equation. One property of AR method is smoothed spectrum in comparison with FFT.

The steps involved in the process of speech recognition are as follows:

- Sampling and digitizing the speech signal
- Computing PSD in each frame of sequence
- Spectrum features extraction (amplitude and location of peaks in each frame)
- Exertion achieved features to neural network for classification

* M.sc student of electronics engineering, Azad Islamic University, Arak, Iran, ashouri85@gmail.com

** M.sc student of electronics engineering, Azad Islamic University, Arak, Iran, G.A.abdali@gmail.com

***Assistant Professor, Faculty of Engineering, Arak University,a-suratgar@aut.ac.ir

At first, speech signals are sampled with 44100 hertz sample rate and then it is down sampled to 8000 hertz (for reduced stored samples and ease of processing). Then for each frame of data, spectrum estimation is performed with using of autoregressive method. Burg method for spectrum estimation is chosen. Reason of this choice is robustness of this method in proportion of other methods of AR. In next step data that is obtained from PSD, is fed to neural network.

Rest of the paper is organized as follows. Section 2 shows the Overall architecture of process, continued by Section 3 which deals with feature extraction of estimated PSD. In section 4, we introduce the neural network based system and numerical experiments are given in section 5.

## *2. System Architecture*

The overall architecture of this speech recognition process is shown in figure 1:

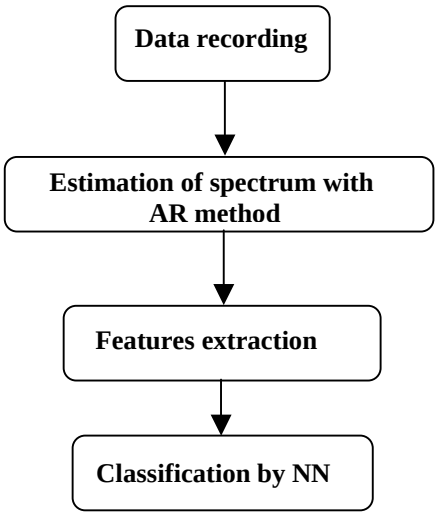


Figure (1) : Architecture of process

### *3. Features Extraction*

After sampling, its rate is decreased from 44100 Hz to 8000 Hz. Reason of this work is ease the process and reducing number of stored samples. Because in the speech signal most of information are occurred in frequency band in size 4 KHz, chosen of 8000 KHz sample rate will be appropriated. And each sample of it is stored with 16 bits. Obtained signal is divided to 30 ms frames.

Reason of this is that speech signal is naught stationary, and because speech signal in frame along 10 to 30 ms is considered locally stationary therefore signal is divided to 30 ms frames. Process done in each frame by burg method for compute of PSD based obtained coefficients from AR system order 14.

Whereas in speech signal, vertexes of spectrum are most important than valleys and since obtained vertexes of spectrum by AR method are estimated better than valleys, and athwart periodogram method that obtained spectrum of it, is very ragged and has very swings. Therefore Autoregressive methods and here burg method present smooth spectrum from signal. This method is chosen for this study.

After PSD computing in each frame of sequence, amplitude of vertexes unitage of dB, and their location is stored for sake features of each frame of sequence. For error reducing, between vertexes that happened in close vicinity, vertex that is more than others are chosen and other vertexes in this vicinity is relinquished. Also than vertexes that their amplitude was lesser than -20 dB, is relinquished.

This work is done in each frame for yield a features dataset. In figure 2, a paradigm of achieving of spectrum features in a frame is shown. Figure2, shows spectrum of a frame of "yes" that estimated by burg method. And figure3 shows extracted features of this frame of word.

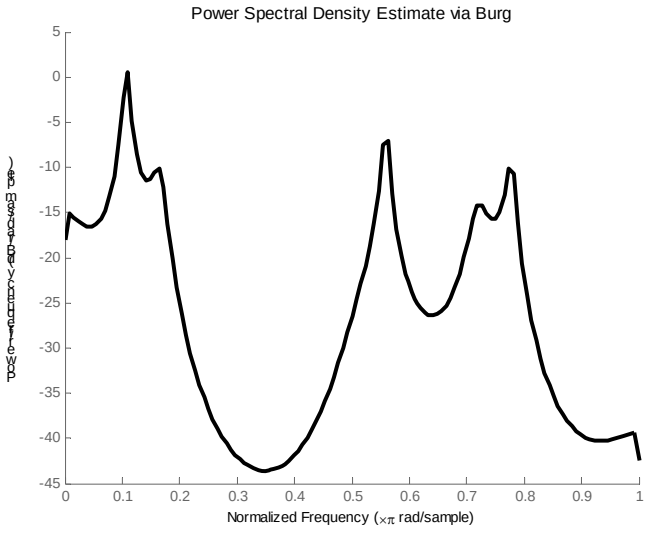


**Figure (2):** PSD estimated via burg of a frame of speech signal

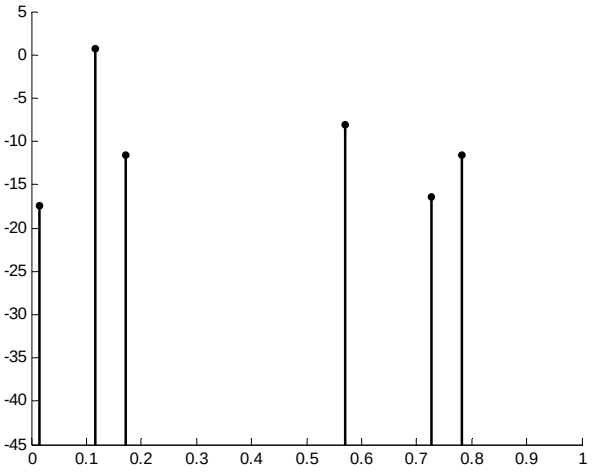

**Figure (3):** feature of this frame

For each frame of signal, six intervals for lying the vertexes is considered. If vertex is lied inside of interval, amplitude and location of that vertex is stored. Providing lack of vertex in this interval, zero is stored in dataset and next interval is surveyed.

### *4. Network Structure*

For word recognition, a feed forward Back propagation neural network is utilized. It is trained by gradient descent with adaptive learning rate training algorithm. This network involves 194 input and 2 hidden layers and 2 output neurons. Input layer with 194 input, generates 8 elements that are exerted to next layer and second layer generates 12 elements and feeds these elements to final layer (output layer).

Two neurons form output of network. For training this network two paradigm of each word is used.

First and second layer have tan sigmoid function. And for output layer, since its output must be between zero and one, its function is chosen log sigmoid.

### *5. Result and Discussion*

The network was trained on a training set of subject one's spoken "yes" and "no". Thus supervised learning is used to the network. We utilized just 2 samples of each word for training of back propagation network that were pronounced by one person. The maximum epochs were 3000, and the goal was set to 0.000001. After training, for test of this network, forty samples of each word, which were pronounced by a person, were tested. Results of network for all of 80 tests (equal for each word, 40 tests for "yes" and 40 tests for "no") were true.

In this work, a simple method for speech recognition with using of latent information of data's PSD with concentration on vertexes is presented. The system provided satisfactory results. Through the encouraging

success of the current system is achieved based on a limited vocabulary, the system can be expanded to a larger vocabulary by expanding the number of outputs used in architecture.

In conclusion the network performs optimally in speaker dependent and limited vocabulary contexts. The ultimate goal of this study is to establish an accurate speech recognition system, which is capable of dealing with distinct speech inputs, such as distorted speech or stressed speech rather than just normal speech. Further work toward these goals is currently being pursued.

## References

[1] S.M. Kay.1987, ***Modern spectral estimation: theory and application***, Englewood Cliffs, New york 07632.

[2] D Polur,P. Zhou,R. Yang,J. Adnani,F. Hobson,R. (2001) ***" isolated speech recognition using artificial neural networks"*** proceeding of the 23rd annual EMBS international conference, October 25-28 ,Istanbul, turkey.

[3] Maaly,I. Obaid,M. (2006) ***"Speech Recognition using Artificial Neural Networks"***, IEEE.

[4] Marple,S.L. (1989) ***"a tutorial overview of modern spectral estimation "*** , IEEE.

[5] Harma,A. (2001) " ***Frequency-warped autoregressive modeling and filtering"*** Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing.

[6] Halavati,R. Shouraki,B,S. Tajik,H. Cholakian,A. Razaghpour,M. (2006) ***"a Novel Approach to Very Fast and Noise Robust,Isolated Word Speech Recognition"*** The 18th International Conference on Pattern Recognition.