A Connectionist Framework For Feature Based Speech Recognition System Using Artificial Neural Networks

Nalini Vasudevan * Anushruthi Rai † Arjun Jain ‡

Abstract

Keywords: formants, fourier transform, dft, dwt, lpc, template matching, dynamic warping, artificial neural networks

In this paper we study the various methods employed to recognize discrete speech.We design a recognition system which is capable of recognizing spoken language. The software takes spoken language and translates it into written text, or follow the spoken instructions to perform other functions. Here, we propose an unexampled method to recognize speech.We provide a basic connectionist framework to analyze speech wave.The spoken words are digitized(turned into sequence of numbers) and matched against pretrained samples in order to identify the words. The system is trained, requiring samples of actual words that will be spoken by the user of the system. The sample words are digitized, stored in the computer to match against future words. We propose a novel combination of extracting the characterestics of the audio signal using linear predictive coding and a computational approach of using artificial neural networks in indentifying the correct sample. The analog audio is converted into digital signals. This requires analog-to-digital conversion. Linear Predictive Coding is a correlation measure, a measure of similarity between two signals, and is used in the analysis of speech in our implementation. As speech recognition involves the ability to match a voice pattern against a provided or acquired vocabulary, a neural net is constructed to achieve maximum accuracy We show that this method gives salutary results using experimental observations. Then we provide conditions under which the system gives optimum results.

1 Introduction

Speech Recognition is the field of computer science that deals with designing computer systems that can recognize spoken words. They generally require an extended training session during which the computer system becomes accustomed to a particular voice and accent. Such systems are said to be speaker dependent. Many systems also require that the speaker speak slowly and distinctly and separate each word with a short pause. These systems are called discrete speech systems. Recently, great strides have been made in continuous speech systems – voice recognition systems that allow you to speak naturally. There are now several continuous-speech sys-

*R.V. College of Engineering, Computer Science & Engineering, Bangalore, India 560059 Tel: +91 94481 07482 Email: naliniv@gmail.com

†R.V. College of Engineering, Computer Science & Engineering, Bangalore, India 560059 Tel: +91 98451 07687 Email: anu_shruthi@yahoo.com

‡R.V. College of Engineering, Computer Science & Engineering, Bangalore, India 560059 Tel: +91 99451 24241 Email: arjunjain@gmail.com

tems available for personal computers. Because of their limitations and high cost, voice recognition systems have traditionally been used only in a few specialized situations. For example, such systems are useful in instances when the user is unable to use a keyboard to enter data because his or her hands are occupied or disabled. Instead of typing commands, the user can simply speak into a headset. Increasingly, however, as the cost decreases and performance improves, speech recognition systems are entering the mainstream and are being used as an alternative to keyboards. It appears that most computer users can create and edit documents more quickly with a conventional keyboard, despite the fact that most people are able to speak considerably faster than they can type. Using both keyboard and speech recognition simultaneously, however, can in some cases be more efficient than using any one of these inputs alone. Additionally, heavy use of the speech organs results in vocal loading. Also, the typical office environment with a high amplitude of background speechs are among the most adverse environment for current speech recognition technologies. For use with computers, analog audio must be converted into digital signals. This requires analog-to-digital conversion. For a computer to decipher the signal, it must have a digital database, or vocabulary, of words or syllables, and a speedy means of comparing this data with signals. The speech patterns are stored on the hard drive and loaded into memory when the program is run. A comparator checks these stored patterns against the output of the A/D converter.
Speech recognition is composed of two parts:

1. Feature Extraction

2. Pattern Classification

In [7] and [8] we studied the speech analysis with **Fourier transforms**. A modification of Fourier transform is **Discrete wavelet transform** as explained in [9]. However the accuracy level is very low due to inadequate feature extraction. In [10] and [11] we studied the **Template Matching** for pattern recognition and **Dynamic Warping** method which resulted in low exactitude. In this paper we extend the analysis by using linear predictive coding along with Artificial Neural networks for efficient speech recognition.

2 Feature Extraction

Feature extraction involves information retrieval from the audio signal. The fundamentals of speech analysis and information retrieval are discussed in [1], [2], [3], [4] and [5].

Fourier Transform:
In this paper we start the analysis using Fourier transforms. Since frequency is one of the important pieces of information necessary to accurately recognize sound, it is necessary to have a transformation that allows one to break a signal into its frequency components. The Fourier transform of a signal is the representation of the frequency and amplitude of that signal. Since the differential of a wave signal is not continuous, we get phantom frequencies. Common, everyday signals, such as the signals from speech, are rarely stationary. They will almost always have frequency components that exist for only a short period of time. Therefore, the Fourier transform is rendered an invalid when faced with the task of speech recognition.

Discrete Fourier Transform:
To overcome the above deficiency, discrete Fourier transform is used. The Discrete

Fourier Transform is symmetric, so the first half of the data is really all that is interesting. Short time fourier transform was used. A band pass filter is used to remove unwanted frequencies. Fourier transforms and its application in speech analysis is enumerated in [6], [7] and [8].

Discrete Wavelet Transform:
[9] uses wavelet transform for the analysis of sound patterns. DWT provides a compact representation that shows the energy distribution of the speech signal in time and frequency. The Wavelet Transform was more efficient than Short Time Fourier Transform (STFT) because STFT provided uniform time resolution for all frequencies whereas the DWT provided time resolution proportional to the frequency.

2.1 Linear Predictive Coding

In this paper we propose to use LPC, which is a modification of DFT. LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The method employed is a difference equation, which expresses each sample of the signal as a linear combination of previous samples. Such an equation is called a *linear predictor*, which is why this is called Linear Predictive Coding.

The basic assumption behind LPC is the correlation between the n-th sample and the p previous samples of the target signal. Namely, the n-th signal sample is represented as a linear combination of the previous p samples, plus a residual representing the prediction error:

$$x(n) = -a_1x(n-1) - a_2x(n-2) - ... - a_px(n-p) + e(n)$$

The equation is an autoregressive formulation of the target signal.

The coefficients of the difference equation (the prediction coefficients) characterize the formants, so the LPC system needs to estimate these coefficients. Minimizing the mean-square error between the predicted signal and the actual signal does the estimate.

It is more accurate[1] than DFT.

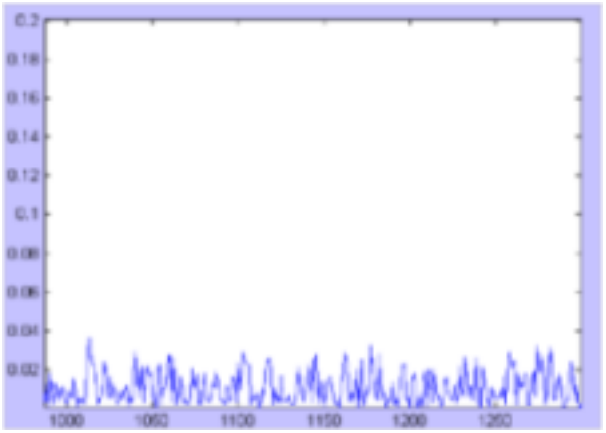

Figure 1: LPC of Letter 'A'

3 Pattern Classification Methods

Once the feature has been extracted , the task is to match the right pattern.

Template Matching:
Template matching is being used widely in recognition systems ([10] and [11]). Template Matching is one of the simplest methods to measure similarity. Initial samples are taken as reference (training sets). The test sample is compared with each of the training sets and the one with the best match is the one with the least Euclidean distance.

[1]For continuous speech at 56kbps, it is benchmarked to be 94% more accurate

Euclidean distance is given by:

$$E = \sqrt{\sum (t_j - o_j)^2} \quad (1)$$

t_i is the i^{th} lpc value of the training sample.
o_i is the i^{th} lpc value of the test sample.

Dynamic Time Warping:
Dynamic Warping Method because uses both frequency and time domain characteristics where as the previous method uses only frequency domain characteristics. In this method speech is divided into frames of 30 ms at every 15 ms intervals (allowing overlap). The lpc features of each frame are extracted.

A frame of the test sample is compared with the corresponding frame in the training sample by applying Euclidean formula:
x_i- i^{th} lpc value of the x^{th} frame of the test sample.
y_i- i^{th} lpc value of the y^{th} frame of the training sample.
The dynamic equation is given by:

$c(x,y) = min(c(x-1,y), c(x,y-1), c(x-1,y-1)) + ed(x,y)$

Where $c(x,y)$ measures the dissimilarity between the test sample (up to frame x) and training sample (Upto frame y). The test sample is compared with all the trained samples, and one with the least $c(x,y)$ gives the best match.
Dynamic programming methods and its application in word recognition are discussed in [13] and [12] respectively.

3.1 Artificial Neural Networks

In this paper we extend the analysis to use the concept of Neural networks([14],[15],[16])
A neural network is composed of a number of interconnected units (artificial neurons).Each unit has an input/output(I/O) characteristics and implements a local computation or function. The output of any unit is determined by the I/O characteristics, its interconnection to other units and (possibly) the external inputs. The applications of Neural Networks are enumerated in [18] and [19]

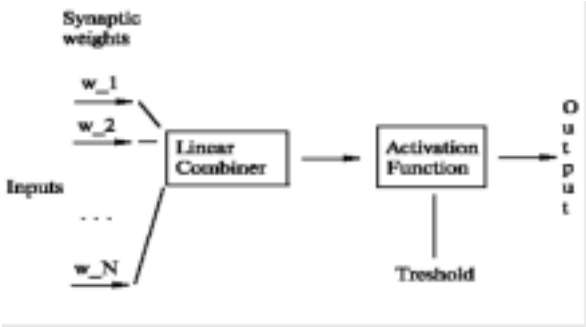

Figure 2: An 'Artificial Neuron'

A single neuron by itself is not a very useful pattern recognition tool. The real power of neural networks comes when we combine neurons into the multilayer structures, called neural networks.

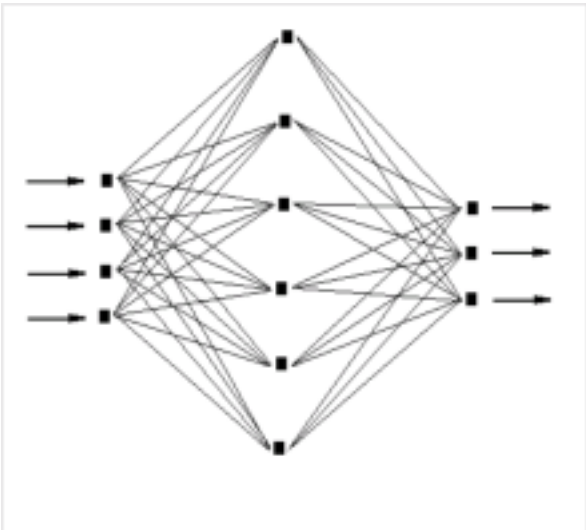

Figure 3: A 'Simple Neural Network'

The Neuron has:
Set of nodes that connect it to inputs, output, or other neurons, also called synapses. A Linear Combiner, which is a function that takes

all inputs and produces a single value. A simple way of doing it is by adding together the Input multiplied by the Synaptic Weight.

An Activation Function. It will take ANY input from minus infinity to plus infinity and squeeze it into the -1 to 1 or into 0 to 1 interval.

Finally, the threshold defines the *INTERNAL ACTIVITY* of a neuron, when there is no input.In general, for the neuron to fire, the sum should be greater than threshold. For simplicity, threshold can be replaced with an *EXTRA* input, with weight that can change during the learning process and the input fixed and always equal (-1). The first layer is known as the input layer, the middle layer is known as hidden layer and the last layer is the O/P layer.

3.2 Back Propagation

Neural networks are employed for machine learning [17]. Back propagation is one of the algorithms used for self-learning and recognition. The primary objective of this session is to explain how to use the back propagation training functions in the to train feed forward neural networks to solve speaker dependent speech recognition problems. There are generally four steps in the training process:

1. Assemble the training data
2. Create the network object
3. Train the network
4. Simulate the network response to new inputs

3.2.1 Feed forward Dynamics

When a BackProp network is cycled, the activations of the input units are propagated forward to the output layer through the connecting weights.

$$net_j = \sum w_j a_i \tag{2}$$

where a_i is the input activation from unit i and w_{ji} is the weight connecting unit i to unit j. However, instead of calculating a binary output, the net input is added to the unit's bias and the resulting value is passed through a sigmoid function:

$$F(net_j) = \frac{1}{1 + e^{-net_j + j}} \tag{3}$$

The sigmoid function is sometimes called a "squashing" function because it maps its inputs onto a fixed range.

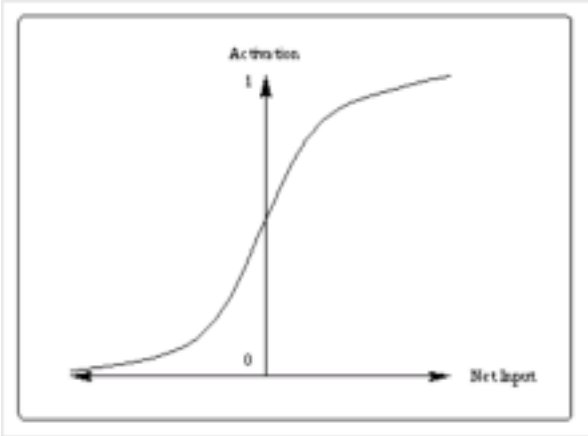

Figure 4: Sigmoid Activation Function

3.2.2 Gradient Descent

Gradient descent is an hill-descending algorithm that approaches a minimum of a function by taking steps proportional to the gradient (or the approximate gradient) at the current point.(Figure 5)

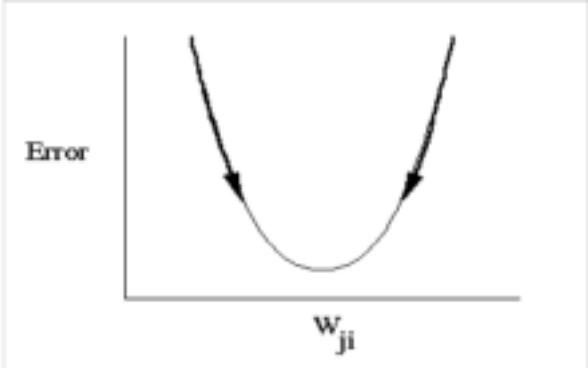

Figure 5: Gradient Descent

3.2.3 Input & Output of the Neural Network

The lpc values of each of the training sets is fed as input to the neural network. Each output neuron represents a voice command or a word.The target o/p is made 1 at the corresponding neuron. There are 15 inputs(lpc values) and n outputs(corresponding to the n words to be recognized).The training is iterated and the weights are adjusted for each training sample.

3.2.4 The Backpropagation Training Algorithm

The objective of the Backpropogation training algorithm is to minimize the error by adjusting the weights.

Initialization: Initial weights w_i set to small random values, learning rate $\eta = 0.1$
Repeat

1. For each training example (x, y)

   (a) Calculate the outputs using the sigmoid function:

   $$o_j = \sigma(s_j) = 1/(1 + e^{-s_j}),$$

   $$s_j = \Sigma_{i=0}^{d} w_{ij} o_j$$

   $$o_k = \sigma(s_k) = 1/(1 + e^{-s_k}),$$

   $$s_k = \Sigma_{i=0}^{d} w_{ik} o_k$$

   (b) Compute the benefit β_k at the nodes k in the output layer:

   $$\beta_k = o_k(1 - ok)[yk - ok]$$

   (c) Compute the changes for weights $j \rightarrow k$ on connections to nodes in the output layer:

   $$\Delta w_{jk} = \eta \beta_k o_j$$

   $$\Delta w_{0k} = \eta \beta_k o_j$$

   (d) Compute the benefit β_j for the hidden nodes j with the formula:

   $$\beta_j = o_j(1 - o_j)[\Sigma_k \beta_k w_{jk}]$$

   (e) Compute the changes for the weights $i \rightarrow j$ on connections to nodes in the hidden layer:

   $$\Delta w_{ij} = \eta \beta_j o_i$$

   $$\Delta w_{0j} = \eta \beta_j$$

2. Update the weights by the computed changes:

   $$w = w + \Delta w$$

   until termination condition is satisfied.

3.2.5 Feeding Test Data

The test sample is fed to the Neural Network.Using the trained weights the O/P is calculated at each neuron of the output layer. The word corresponding to the neuron that gives the maximum O/P is the match required.

4 Testing, Results and Comparision

The advantages of the LPC is its optimal time-frequency resolution and dynamics properties, as well as the continuous time processing. This indeed gives gives a more powerful in compressing the spectral information into a few filter coefficients. This is one important reason why we have used LPC for audio coding. As the number of training samples increase for a particular word or syllable, the accuracy of pattern matching increases. In previous methods used, such as euclidean distance and dynamic warping method, the accuracy is measurable variant of the number of training samples and there exists a quadratic or cubic dependency on the number of training samples. This dependency is quite low compared to our neural network approach. The accuracy increases almost exponentially using neural nets. Neural networks are an efficient, pervasive, and powerful means of computation. The creation of neural networks was inspired by the study of the human brain. Indeed, many aspects of neural networks attempt to emulate biological function, but neural networks do not accurately model biology. Neural networks are pattern classifiers. They do not store "knowledge" in a memory bank. The information is distributed throughout the network and is stored in the form of weighted connections. The most valuable characteristics of neural networks are adaptability and tolerance to noisy data. Thus, they are well suited for applications that involve classification of input (e.g., digital image, natural language, and speech processing). Neural networks are not appropriate for problems that require precise, unary answers, such as solving mathematical problems. With the combination of lpc and ANN, we have achieved an accuracy of 82%.

The accuracy is measured as :

accuracy = (no. of patterns recognized accurately) / (no. of patterns fed)

A performance factor as mentioned before can be defined as:

p = (accuracy) / (no. of training samples per syllable)

We see that p is 1.2-1.4 times greater than the conventional methods.

5 Conclusions and Future Work

In this paper we have studied artificial neural networks as a framework for recognizing words.We first considered the case where template matching was used. For this case, we show that the results are not accurate because it does not capture the time domain characteristics. Further,the next method- the dynamic method was deficient since it was difficult to choose the right window size as it focuses more on time domain characteristics. The neural network approach is quite general and can be extended to continuous speech to obtain high levels of pattern classifications and recognition. We are hoping to extend this idea from discrete words to continuous sentences and achieve speaker independency.

References

[1] Jonathan Foote, An Overview of Audio Information Retrieval, ACM Multimedia Systems, Vol.7, 1999, pp. 2-10.

[2] Rabiner and B.H. Juang, Fundamentals of speech recognition, Prentice Hall, Upper Saddle River, New Jersey 07458, 1993.

[3] G. Tzanetakis, P.Cook , A framework for audio analysis, Organised sound,Vol.4(3), 2000

[4] E. Wold et al., Content-based classification, search and retrieval of audio data, IEEE Multimedia Magazine, Vol. 3, No. 2, 1996

[5] M.Hunt, M.Lenning and P.Mermelstein. Experiments in syllable-based recognition of continuous speech, Proc. Inter.Conference on Acoustics, Speech and Signal Processing (ICASS), 1980

[6] B. Allen and L.R. Rabiner, A unified approach to short-time Fourier analysis and synthesis, Proc. IEEE, Vol. 65, No. 11, pp. 1558-1564, 1977

[7] M.R. Portnoff , Short-time Fourier analysis of sampled speech IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-29, pp. 364- 373, 1981.

[8] J.S. Lim et al., Signal estimation from modified short- time Fourier transforms, IEEE Trans. Acoust., Speech and Signal Pro-cessing, Vol. ASSP-32, pp. 236-243, 1984.

[9] R.Kronland-Martinet, J.Morlet and A.Grossman, Analysis of sound patterns through wavelet transformation, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 1(2)

[10] Marcus E. Hennecke, K. Venkatesh Prasad, and David G. Stork, Using deformable templates to infer visual speech dynamics, 28th Annual Asilomar Conference on Signals, Systems, and Computers volume 1, pages 578582, Pacific Grove, CA, November 1994 IEEE, IEEE Computer Society Press.

[11] Alan L. Yuille, David S. Cohen, and Peter W. Hallinan. Facial feature extraction by deformable templates Technical Report 88-2, Harvard Robotics Laboratory, 1988.

[12] H. Sakoe and S. Chiba, Dynamic programming op- timization for spoken word recognition, Proceed- ings of ICASSP-78, vol. 26, no. 1, pp. 43-49, 1997.

[13] Bellman and S. Dreyfus, Applied dynamic pro- gramming, Princeton, NJ: Princeton University Press, 1962.

[14] Anderson, James A. An Introduction to Neural Networks (1st ed.), 1995. MIT Press

[15] Fausett, Laurene V. "Fundamentals of Neural Networks : Architectures, Algorithms, and Applications", Englewood Cliffs, NJ: Prentice-Hall, 1994.

[16] Golden, Richard M. Mathematical Methods for Neural Networknalysis and Design (1st ed.), MIT Press, 1996.

[17] Mitchell, Tom M. "Artificial Neural Networks. In Machine Learning", pp. 81-127. New York: McGraw Hill Companies, Inc.1997.

[18] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zachs, and S. Levy, Inferring articulation and recognising gestures from acous- tics with a neural network trained on x-ray microbeam data, J. Acoust. Soc. Am., 92(2):688?700, August 1992.

[19] Widrow, Bernard, David E. Rumelhart, and Michael A. Lehr. 1994," Neural Networks: Applications in Industry, Business and Science",Communications of the ACM 37 (3): 93-105, 1994